Multimodal representation learning has been an area of machine learning that increasingly draws the attention of the research community. Combining information from different modalities, such as images, and text, allows more informative representations, as they provide complementary insights for the same instances. Several works focus on using both vision and language modalities, introducing tasks such as visual question answering (Agrawal et al., 2016), visual reasoning (He et al., 2021a), visual commonsense reasoning (Zellers et al., 2019), visual entailment (Xie et al., 2018), image captioning (Stefanini et al., 2021), image-text retrieval and inversely text-image retrieval (Dubey, 2021), referring expressions (Krishna et al., 2018), visual explanations (Hendricks et al., 2016) and grounding (Endo et al., 2017), visual-language navigation (Anderson et al., 2018), visual generation from text (Reed et al., 2016c), visual storytelling (Huang et al., 2016b) and its inverse task of story visualization (Li et al., 2019c), and visual dialog (El-Nouby et al., 2019).

Some of the first attempts that combine vision and language face several limitations due to the restricted capacity of sequential models for language, such as recurrent neural networks (RNNs), LSTMs (Hochreiter and Schmidhuber, 1997) and GRUs (Cho et al., 2014), which struggle to represent long textual sequences. The area of multimodal learning has faced significant advancements especially since the introduction of the Transformer (Vaswani et al., 2017). Several powerful transformer-based variants, such as BERT (Devlin et al., 2019) and GPT-3 (Brown et al., 2020), set the foundations for the surge of visiolinguistic (VL) transformers. The extension of single-modality pre-training requires the introduction of visual features, which allow to infer masked linguistic components and vice versa, enabling learning cross-modality relationships from aligned data. In the meantime, pre-training tasks applied independently on the visual or linguistic components permit intra-modality learning. Fine-tuning on task-specific VL datasets, addressing the tasks of text-image retrieval (Lin et al., 2014), visual question answering (Ren et al., 2015; Agrawal et al., 2016; Hudson and Manning, 2019; Krishna et al., 2016; Zhu et al., 2016; Gao et al., 2019), visual reasoning (Suhr et al., 2017), visual commonsense reasoning (Zellers et al., 2019) and others, follows the pre-training stage. Models such as LXMERT (Tan and Bansal, 2019), VisualBERT (Li et al., 2019a), ViLBERT (Lu et al., 2019, 2020), VL-BERT (Su et al., 2020), UNITER (Chen et al., 2020), OSCAR (Li et al., 2020c), ViLT (Kim et al., 2021), CLIP (Radford et al., 2021), SIMVLM (Wang et al., 2021c) and many others have demonstrated state-of-the-art results in multiple VL tasks.

As for recent transformer-based approaches, despite pre-training on large amounts of aligned VL data, usually from Conceptual Captions, (Sharma et al., 2018) COCO (Ren et al., 2015) and Visual Genome (Krishna et al., 2016) datasets, the learned concepts remain limited and
lack further explicit information regarding commonsense knowledge, abstract entities or real-world events. Relevant issues in the natural language processing field were addressed by leveraging knowledge graphs, thus resulting in knowledge-enhanced approaches for several tasks, such as Language Modeling (Logan et al., 2019; Lu et al., 2021), Natural Language Inference (NLI) (Chen et al., 2018b,a), Language Generation (Yu et al., 2021b), Dialog Generation (Cui et al., 2021a), Entity Disambiguation (Ji et al., 2020), multilinguality (Liu et al., 2021a), Contextualized Language Embeddings (Sun et al., 2020), and models such as KnowBert (Peters et al., 2019), E-BERT (Poerner et al., 2020), ERICA (Qin et al., 2021), ERNIE (Zhang et al., 2019b), ERNIE-NLI (Bauer et al., 2021), LUKE (Yamada et al., 2020) and others. Therefore, the incorporation of large-scale knowledge graphs and ontologies can be also critical to the quality of multimodal representations and the success of relevant models on the various downstream tasks.

While previous surveys (Baltrušaitis et al., 2017; Kafle et al., 2019; Guo et al., 2019; Mogadala et al., 2021; Zhang et al., 2020; Uppal et al., 2020; Cao et al., 2020; Du et al., 2022) provide analysis and taxonomies over models, tasks and datasets regarding multimodal representations, they do not analyze knowledge-enhanced approaches. In contrast to those works, we focus on the integration and importance of external knowledge to VL models. Even though the current trends focus on transformer based implementations, for the sake of completeness, other techniques that have contributed to the field of knowledge-enhanced VL (KVL) learning are also included. Overall, we target to bridge the gap between knowledge representation and multimodal deep learning: we provide a broad and comprehensive analysis of both fields, and consequently collect models that have served the various KVL tasks. Finally, we discuss current challenges and limitations of existing datasets and approaches, upon which we suggest potential future directions of this evolving field. To the best of our knowledge, there are no extended works covering the intersection of those two very fundamental fields of AI, which have demonstrated promising directions in many aspects of state-of-the-art research when combined.

The current survey consists of four main parts. The first part (Section 2) covers the preliminaries of multimodal deep learning, analyzing trends, methods, models and tasks which set the basis for knowledge-enhanced VL (KVL) models. An analysis regarding graph structures and their representation follows in Section 4. Section 5 is dedicated to a taxonomy of knowledge senses and types, as well as the presentation of popular knowledge bases. Section 6 provides a taxonomy and analysis as per KVL task where the usage of external knowledge has been attempted, accompanied with datasets and evaluation methods. Finally, some needs and possible future directions regarding the usage of knowledge in multimodal learning are identified and analyzed.

2 Background

Multimodal learning is a large and diverse field that involves a variety of data sources, architectural approaches and tasks. By focusing on VL tasks, which exploit text and image data, we can identify a variety of relevant applications. The nature of each task defines the chosen backbone architecture, upon which all consequent approaches are built. More specifically, VL tasks can be divided in discriminative tasks, where the goal is either to provide a matching between modalities or understanding the one modality based on the other, and generative tasks, which target image or text generation. Discriminative VL tasks present a long line of research initially based on recurrent neural networks (RNNs) for text representation, with most contemporary approaches favoring the Transformer (Vaswani et al., 2017) framework for its indisputable advantages. On the other hand, generative tasks demonstrate an interesting variability in architectural approaches: while language generation tasks conditioned on image are addressed by architectures based on RNNs or Transformers, image generation tasks conditioned on text are mainly tackled by Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), and more recently by Transformers (Esser et al., 2020) and Diffusion models (Dhariwal and Nichol, 2021).

Knowledge-enhanced VL models (KVL models) usually step on existing approaches for VL representation and then employ various strategies to integrate knowledge. One common first step between most VL approaches is the independent encoding of text $T$ and images $I$, followed by the interaction of these encodings in order to acquire a joint representation. The choice of text encoding (RNN or Transformer) heavily influences the overall architecture of a VL model. On the other hand, image encoding adapts to the needs defined by text encoding, and variability in chosen image encoders serves particular improvements, such as performance boosting or reduction of trainable parameters.

Transformer-based VL models consist of two main stages: pre-training on large amounts of aligned image-text data, such as images and their captions, and then fine-tuning on smaller task-specific datasets. Pre-training learns a generic joint VL representation from independent image and text encodings using certain pre-training tasks (or pre-training objectives) that enforce cross-modality in-
teraction. Fine-tuning is performed independently on each task-specific dataset, leveraging the learned representation from the pre-training stage.

There is a variety of ways to incorporate knowledge $K$, with most approaches favoring external knowledge sources in the form of widely used knowledge graphs (KGs). In this case, a representation based on graph neural networks (GNNs) is the most popular approach towards providing an appropriate encoding for $K$. However, other knowledge types can be integrated as well, either in the form of knowledge stored in neural network weights (implicit knowledge) or linguistic knowledge from the web, embedded with a text encoder. Enhancements to VL model performance can be also realized via self-acquired (internal) knowledge: without leveraging external knowledge sources, automatic extraction of specific characteristics from either modality boosts and guides learning, improving knowledge-free baselines. Knowledge $K$ can be fused either in early stages together with text $T \in \mathcal{W}$ and image $I \in \mathcal{V}$ instances resulting in a KVL representation, or alternatively in later stages and independently of the VL stream, refining and correcting the predictions of the VL model.

KVL models can either target one task at a time (single-task models), or multiple tasks simultaneously (multi-task models). Single-task models present a large variety of architectural implementations so far, while multi-task models are exclusively built upon multimodal transformers, as they heavily rely on the pre-training fine-tuning scheme. Discriminative tasks are tackled by both single- and multi-task models. On the other hand, generative tasks can be only handled by single-task models, as they are harder by nature. The evaluation of the overall model performance is realized as per task, based on appropriate task-specific evaluation metrics.

A broad overview of KVL models is provided in Figure 1. The joint representation module is shown as a black box, as most architectural variations analyzed in following sections are happening within this stage.

3 Multimodal representation learning

The core of multimodal deep learning revolves around the ways the various modalities are represented independently and interact with each other. Especially for text and images, several representation techniques have been developed throughout the years, due to the advancements of image classification (He et al., 2016; Simonyan and Zisserman, 2015) and object detection (Girshick, 2015; Ren et al., 2017) models for vision, as well as distributed language representations (Mikolov et al., 2013; Pennington et al., 2014), recurrent neural networks (Hochreiter and Schmidhuber, 1997; Schuster and Paliwal, 1997; Cho et al., 2014), and attention-based models (Vaswani et al., 2017; Devlin et al., 2019; Liu et al., 2019b; Brown et al., 2020) for text. Apart from vision and language, more modalities can potentially contribute in a joint representation, such as speech, music, graphs and others.

3.1 Text representation

Text representation offers several variations over implementations, imposing significant influences towards multimodal learning in total. The two major categories of language embeddings include recurrent architectures (RNN/LSTM/GRU) and Transformers. The architectural choices for language representation guide the taxonomy of models per KVL task, due to the diversity they impose in the resulting implementations.

Distributed word representations Traditional and widely used representations such as Word2Vec (Mikolov et al., 2013) have contributed in several components of the various VL tasks, often used as initializations for other methods. Doc2vec (Le and Mikolov, 2014) extends word2vec, achieving a vector representation of a group of words. GloVe (Pennington et al., 2014), a distributed word representation model, is trained on a global word-word co-occurrence frequency matrix and successfully captures both local and global statistics. Fast-text (Bojanowski et al., 2017) represents each word with a bag of character n-grams in order to capture the internal structure of words. This approach can effectively utilize the morphology of words, thus it is able to represent rare formations occurring in morphologically rich languages. Word semantics cannot be successfully captured by static word representation methods, so contextualization is needed especially in cases of polysemy, an issue tackled in CoVe (McCann et al., 2018). ELMo (Peters et al., 2018) is another deep word contextualization model which leverages character level information to form robust representations based on morphological information. Universal Language Model Fine-tuning (ULMFiT) (Howard and Ruder, 2018) achieves robust inductive transfer learning on a variety of downstream NLP tasks, where only fine-tuning is required.

Recurrent neural networks The basic idea behind recurrent neural architectures (RNNs) is the sequential processing of elements belonging to a finite sequence $(x_1, x_2, ..., x_T)$, one $x_t$ at a time, while retaining context information from the previous elements in the form of the previous node’s hidden state $h_{t-1}$. Both $x_t$ and $h_{t-1}$ contribute to the calculation of the current hidden state $h_t$,
which consequently participates in defining the current output \( y_t \). Feed forward neural networks implement each time step as a layer, with shared weights across layers. Backpropagation through time (BPTT) is utilized to train recurrent neural networks (Lipton et al., 2015). Other recurrent neural architectures such as Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) dominated the field of language embeddings in many approaches, with more refined variants such as bidirectional LSTM (BiLSTM) (Schuster and Paliwal, 1997), Gated Recurrent Unit (GRU) (Cho et al., 2014) and bidirectional GRUs (BiGRUs) following in later works.

This sequential nature of RNNs fits naturally to language processing, inspiring several relevant implementations in the NLP domain, as well as in multimodal tasks. To this end, earlier works in VL architectures heavily rely on distributed word embeddings together with sequential models for language representation. However, limitations tied with sequential processing such as vanishing gradients and inability of parallel processing directed research interest towards novel frameworks, such as the Transformer (Vaswani et al., 2017).

**Language transformers**  The introduction of attention mechanisms and especially the Transformer framework (Vaswani et al., 2017) opened a whole new world of possibilities for language embeddings and several downstream linguistic tasks. The Transformer model relies on encoder-decoder structure and utilizes an attention mechanism to construct dependencies between input and output data. The layers of the encoder and the decoder are stacked the one upon the other, containing sub-layers with Multi-Head self-attention and position-wise fully connected feed-forward layers. Residual connections followed by layer normalization are used between the sub-layers of the encoder. The decoder has an additional encoder-decoder multi-head attention sub-layer that helps focusing on the appropriate parts of the encoded input sequence. Moreover, the decoder’s self-attention modules are modified so that they force the prediction at a certain position to be based only on the known predictions of previous positions. Transformer architectures prove that there is no need for convolutions or recurrent units to achieve state-of-the-art performance in linguistic tasks. Currently, most state-of-the-art VL architectures utilize attention mechanisms within their implementation.

Transformer models for NLP tasks consist of a **pre-training** stage on a large corpus of data, followed by **fine-tuning** on certain downstream tasks. Language transformers can be divided in two major categories: **autoregressive (AR)** and **autoencoding (AE)**, depending on whether pre-training is performed in a unidirectional or a bidirectional way. AR language models attempt to estimate the probability distribution of a text corpus, while AE models learn to reconstruct manipulated inputs with the help of
BERT (Devlin et al., 2019) is a popular bidirectional transformer-based language representation model, able to handle a variety of natural language processing tasks by just fine-tuning one additional output layer. It uses masked language modeling (MLM), randomly hiding some input tokens in order to be inferred from the surrounding words. The pre-training stage is based on unlabeled data, which enable parameter learning. Those parameters are then fine-tuned with labelled data corresponding to some certain tasks. RoBERTa (Liu et al., 2019b) offers an optimized extension to BERT, suggesting that longer training, more data and larger batch size, as well as training on longer sequences, dynamically altering the masking patterns and removing the next sentence prediction loss are factors that contribute to advanced performance of the original BERT model. XLNet (Yang et al., 2019b) combines AR and AE language modelling in a single approach. It introduces the pre-training objective of permuted language modelling, which attempts to collect information from permutations of the factorization order of text tokens with respect to AR likelihood estimation, practically inducing bidirectional capabilities to the learning process.

Generative NLP models have demonstrated impressive results in recent literature. GPT-2 (Radford et al., 2019) is pre-trained on a very large dataset (40GB of text data from over 8 million documents) and utilizes 1.5 billion parameters in order to construct a powerful language model. GPT-3 (Brown et al., 2020) is an AR language model of 175 billion parameters which achieves zero-shot, one-shot and few-shot capabilities. It is able of synthesizing results in a human-like level, so that writing articles or code in some programming language, learning and reusing new words, unscramble words and other tasks can be realized indistinguishably to humans. T5 (Raffel et al., 2020) introduces a unified format where both inputs and outputs are text. Without changing the model architecture, the hyperparameters or the loss function, T5 generates text to address tasks such as question answering, text summarization, machine translation, as well as classification tasks. BART (Lewis et al., 2019) utilizes a bidirectional encoder, enabling attention from both left and right directions, and an autoregressive (unidirectional) decoder, which allows attending to past tokens only. It is able to handle generative tasks such as question-answering, text summarization, conditional text generation, mask filling, and also classification tasks.

As for text encodings for VL models, BERT (Devlin et al., 2019) has become a golden standard for several transformer-based approaches, while fewer implementations utilize variants such as RoBERTa (Liu et al., 2019b). GPT2 (Radford et al., 2019), T5 (Raffel et al., 2020) and BART (Lewis et al., 2019) have also served as language encoders.

3.2 Visual representation
There is much less diversity in the representation of the visual modality compared to text encoding. Most works rely on widespread convolutional architectures without significant variations, and only recently some works attempted encoders based on image transformers, which however do not enforce architectural modifications.

Convolutional Neural Networks Representation of images can involve object level or image level features, depending on the granularity of information that needs to be imbued in the representation. A global image representation can be achieved by employing widely used image classification models as feature extractors. Many works rely on CNN based classifiers such as VGG (Simonyan and Zisserman, 2015) and ResNet (He et al., 2016), while others prefer more fine-grained local representations supported by object detectors, such as Fast-RCNN (Girshick, 2015) and Faster-RCNN (Ren et al., 2017). The fixed pre-trained models for object feature extraction somehow limit the expressivity of VL transformers, while being slow. Solutions to this limitation include the usage of grid features as visual tokens (Huang et al., 2020b), or discretized grid features (Huang et al., 2021).

Image Transformers Recent advancements in Visual Transformers as an extension of the aforementioned language Transformers (Vaswani et al., 2017) have influenced the field of image representation. ViT (Dosovitskiy et al., 2021) suggests patch-like parsing of images for feature extraction, resulting in more powerful image representations. Swin Transformer (Liu et al., 2021c) is a more efficient implementation due to the usage of self-attention in local image patches contrary to global self-attention of other approaches, which results in quadratic computation complexity compared to image size. Swin Transformer achieves linear complexity by hierarchically merging larger and larger image patches across layers, with self-attention acting only within each patch. Similar to NLP Transformers scaling capabilities, Swin Transformer V2 of 3 billion trainable parameters serves as the largest dense vision model so far (Liu et al., 2022).

3.3 Sequential models for VL tasks
Even though Transformer-based approaches have dominated the field of multimodal learning, sequential models have offered a variety of interesting solutions in sev-
eral tasks, and still serve as the way to go in some of them. Most sequential-based techniques address the vision-language co-operation through encoding-decoding schemes which utilize a CNN for images $I$ and an RNN/LSTM/GRU structure for text $T$. There are different ways for visual and language embeddings to interact, depending on the downstream task: for example, a CNN-based image encoding can be fed as a conditioning to an RNN/LSTM/GRU decoder structure for tasks requiring text generation from image. Alternatively, input text can be embedded using an RNN/LSTM/GRU encoder, and then a feed-forward neural network can learn the correlations between text embeddings and CNN-based image embeddings. This variant can also serve text-image matching tasks. Sequential structures for VL learning remain rather popular in tasks that require language generation, especially in underexplored ones where more refined architectures have not been attempted yet.

3.4 Multimodal Transformers

Multimodal transformers have revolutionized the field of multimodal learning, with almost any state-of-the-art model built upon them. There are some certain steps followed from receiving the input data until the final result on multimodal tasks, as presented in Figure 2. Initially, given a multimodal dataset, for example a VL dataset $D$ comprised of image-text pairs $(I, T)$ which consist of visual features $v$ and textual features $w$ respectively, an appropriate encoding scheme, such as the one described in sections 3.1 for $T$ and 3.2 for $I$ should be decided. The input embedding contains a tokenized text representation, an image encoding and other special embeddings. All transformer-based architectures, either targeting vision, language or both consist of the pre-training and fine-tuning stages. For this reason, a multimodal encoding module is designed to receive the embedded input, enabling the interaction between modalities by jointly learning complementary information with the help of well-designed cross-modal pre-training tasks. Finally, a fine-tuning stage adapts the pre-trained model to the downstream task by training on a smaller labelled dataset.

3.4.1 Special input tokens and embeddings

Special tokens need to be appended to the input before entering the VL transformer model in order to discern between different modalities, as well as the start and sometimes the end of the sequence. An input embedding is formed by combining text representation, visual representation, special tokens and other embedding information to guide training. Despite different VL transformers following slightly different strategies regarding input representation, the main constituents are analyzed below. Figure 3 provides a general overview of the input tokens and embeddings.

The **input token** denoted as $[CLS]$ is a special classification token that defines the start of the input sequence. Linguistic information from text $T$ is appended after the $[CLS]$ token. Usually WordPiece (Wu et al., 2016c) tokenizer, a sub-word based tokenizer framework, transforms words into tokens. Afterwards, a numerical format of tokens is obtained by assigning a unique embedding per subword token so that it can be further processed. The **text embedding** will then be $w = w_1, w_2, ..., w_n \in \mathbb{R}^d$, where $n$ corresponds to the textual sequence length and $d$ is the embedding dimension. A **segment token** $[SEP]$ is appended after $w$ to separate different modalities. Following the separator $[SEP]$, a **visual token** $[IMG]$ indicates the visual modality $I$. There are different usages of $[IMG]$, either marking the start of visual features $v$, or multiple $[IMG]$ instances acting as masks that locate the existence of visual elements within the **token embedding**. Some architectures may even discard $[IMG]$ tokens. A flattened visual feature vector is extracted as a result of the chosen visual encoder, denoted as $v = v_1, v_2, ..., v_k \in \mathbb{R}^d$, where $k$ is the number of visual features and $d$ the embedding dimension. The **end token** $[END]$ marks the end of the sequence. Many VL transformers ommit $[END]$ token.

**Segment embeddings** attribute the source of each input element by assigning a unique label to each of them. For example, input textual and visual features would be assigned with a different **segment embedding** $s_v$ and $s_w$ respectively, as they come from different modalities. Therefore, sequences $s_v = s_{v1}, s_{v2}, ..., s_{vk}$ are going to be summed with the **text embeddings** and similarly, $s_v = s_{v1}, s_{v2}, ..., s_{vk}$ will be summed with **visual embeddings**. Alternatively, if the input contains two sentences (such as question and answer pairs), each of those will be aligned with different segment labels. **Position embeddings** $p_i$ denote the positioning of tokens in a sequence $\{w_1, ..., w_i, ..., w_n\}$, as originally used in BERT (Devlin et al., 2019).

The summation of **token embeddings**, **segment embeddings**, **position embeddings** and visual embeddings can act as starting point towards the **input representation**, followed by an appropriate transformer structure. The **input embedding** generally follows the $\{[CLS], w_1, w_2, ..., w_n, [SEP], v_1, v_2, ..., v_k, [END]\}$ format, where $w$ and $v$ are the final **text and visual embeddings** respectively, after the summation of $w$ and $v$ with the **segment** and/or **position embeddings**.
3.4.2 Vision and Language joint encoding

Encoded modalities need to be projected to the same vector space and interact in order to achieve a meaningful joint representation. A concise separation of cross-modality encoding is provided in Du et al. (2022), where two main categories of encoding schemes are identified: fusion encoder and dual encoder. Those two approaches can be even combined. Fusion encoder concerns an abundance of VL transformers, and can be further divided in single-stream encoding and double-stream encoding.

Double-stream fusion encoder utilizes two separate transformer modules to process images and text respectively. First VL approaches such as ViLBERT (Lu et al.,
2019) and LXMERT (Tan and Bansal, 2019) fall into this category, where they naturally extend BERT to also process images. Specifically, ViLBERT decomposes images in non-overlapping patches, similar to how tokens serve as inputs in the case of BERT. Text is tokenized and fed together with positional embeddings to its transformer stream. A co-attention module enables interaction and alignment between modalities given their intermediate representations. An extension of ViLBERT to 12 downstream tasks was presented in Lu et al. (2020). LXMERT (Tan and Bansal, 2019) refines the cross-modal part to achieve advanced performance in downstream tasks. More recent models employing a double-stream fusion encoder are ALBEF (Li et al., 2021), Visual Parsing (Xue et al., 2021) and WenLan (Huo et al., 2021). In general, double-stream encoders demand training of two transformer models (one for each stream), which is computationally inefficient. Therefore, succeeding approaches focus on single-stream encoders.

**Single-stream fusion encoder** concerns the majority of state-of-the-art VL models. A single transformer network, usually stepping on the BERT (Devlin et al., 2019) backbone is used to process images and text representations simultaneously, where alignments are discovered via self-attention. Segment tokens to separate modalities, together with position tokens to indicate aligned token pairs are added to input data before they are concatenated and fed into the encoder. VisualBERT (Li et al., 2019a), VL-BERT (Su et al., 2020), PixelBERT (Huang et al., 2020b), InterBERT (Lin et al., 2021), VLP (Zhou et al., 2019a), Unified-VLP (Zhou et al., 2019b), B2T2 (Alberti et al., 2019), UNITER (Chen et al., 2020), XGPT (Xia et al., 2020), ViLT (Kim et al., 2021), VL-T5 (Cho et al., 2021), SOHO (Huang et al., 2021), SimVLM (Wang et al., 2021c) belong to the single-stream encoder category. Some of the latest models even present zero-shot learning capabilities, enabling more out-of-the-box capabilities in VL tasks.

**Dual encoder** is favored by a small family of recent VL models which exploit contrastive learning to provide image-text representations. Separate encoders embed each modality independently, and their representations are projected onto the same vector space with the goal of learning similarity and dissimilarity properties. This is why contrastive learning naturally fits: paired image-text samples are trained to stay close together, while being apart from the rest. CLIP (Radford et al., 2021) implements this strategy utilizing more than 400 million image-text pairs for training, achieving even zero-shot capabilities in retrieving previously unseen matches. ALIGN (Jia et al., 2021) follows the same recipe using over one billion image-text pairs, demonstrating that large-scale uncurated data can compensate for the presence of noise. FLORENCE (Yuan et al., 2021) exploits both image-to-language contrastive loss and its reverse, language-to-image contrastive loss, actually forming a bi-directional contrastive loss applied on image-label-description triplets.

**Fusion and dual encoder** A couple of models leverage the benefits of both encoding approaches (Du et al., 2022). FLAVA (Singh et al., 2021) is a holistic approach that utilizes a dual encoder to integrate visual and text representations into a multimodal one, providing unimodal and multimodal reasoning capabilities in the same model. Additionally, VLMo (Wang et al., 2021a) offers a flexible format by providing a fusion encoder for classification task, and a dual encoder for retrieval tasks.

### 3.4.3 Self-supervised pre-training

**VL pre-training datasets** A VL model can be pre-trained on unimodal (text and/or visual data independently) and multimodal data (paired image-text data). Unimodal data can be unlabelled, noisy and abundant, such as sets of documents or images, while multimodal data are labelled and cleaner, so that each modality presents another point of view for the same instance. Large-scale datasets containing visual and linguistic information are necessary for pre-training. Pre-training aspires to instill a generic understanding of the visual world, natural language and their in between interactions. Widely used datasets can either be in-domain, meaning that their data distribution is very close to task-specific datasets used in the fine-tuning stage, or out-of-domain, containing less similar data to the downstream tasks, but usually being much larger in size. Most VL transformers leverage a corpus consisting of COCO (Lin et al., 2014), Visual Genome (Krishna et al., 2016), SBU (Ordonez et al., 2011) and Conceptual Captions (CC) (Sharma et al., 2018) for pre-training, with fewer models either excluding some of those, or adding more datasets to this corpus.

**COCO** (Lin et al., 2014) consists of around 106K images/533K captions in the train split, and 25K images/5K captions in the test split, with 5 sentences provided per caption from different annotators. Those sentences are designed to provide an overall (global) understanding of their corresponding images, which represent scenes of multiple objects and their in-between relationships. As most VL tasks are built atop COCO, it is considered to be
an \textit{in-domain} dataset. Because of the usage of COCO for downstream tasks, subsets of COCO used in pre-training and fine-tuning must be mutually exclusive to avoid any data leakage.

**Visual Genome (VG)** (Krishna et al., 2016) is another \textit{in-domain} dataset, from which several VL tasks emerge. It contains more than 100K images of complex scenes, providing multiple annotations regarding per image scene graphs, objects (3.8M instances), relationships (2.3M), attributes (2.8M), visual questions and answers (1.7M), region descriptions (5.4M) and region scene graphs (3.7M). Region descriptions are captions grounded to image regions, acting as dense captions per image. As VG images have a partial overlap with COCO images, any COCO image used in downstream tasks should be excluded from the VG during pre-training.

**SBU Captions** (Ordonez et al., 2011) consists of 990K images/990K captions in the train split, and 10K images/10K captions in the test split, with each image corresponding to one caption. It is an \textit{out-of-domain} dataset, larger in size than COCO and VG.

**Conceptual Captions (CC)** (Sharma et al., 2018) is the largest \textit{out-of-domain} dataset used for pre-training, consisting of more than 3 million images for training and 14K for validation with 1 caption each.

**Pre-training objectives** \textit{Uni-modal objectives} concerning either \textit{T} (language objectives) or \textit{I} (vision objectives) at a time, as well as \textit{cross-modal objectives} which take into account both modalities at once are used for pre-training. Such objectives teach the models to infer missing information by understanding their surroundings. Self-supervised learning is the most prevalent technique, enabling learning with the help of a corrupted part of the input or adversarially matched pairs in a contrastive fashion.

\textit{Language objectives} are designed to implicitly learn linguistic rules and patterns so that a model pre-trained on them acquires some ‘understanding’ of natural language. Some language modelling objectives are analyzed below.

**Masked Language Modeling (MLM)** is a pre-training objective introduced in BERT (Devlin et al., 2019). Tokens in the input sentence are replaced at random with a special [MASK] token. The model needs to uncover the actual token by learning the context from its surroundings, thus permitting contextualized representations. The default probability of masking out a token is 0.15. The MLM objective function is bidirectional, which means that a token can be predicted from either its right ones, or its left ones. \textit{Prefix Language Modeling (PrefixLM)} (Wang et al., 2021c) differs from standard MLM such that it enables bi-directional attention on the prefix sequence.

**Next Sequence Prediction (NSP)** refers to the objective that given a pair of sentences, tasks the model to predict whether they could serve as consecutive sentences in a corpus or not.

\textit{Vision objectives} apply similar ideas on the visual modality. Due to the more high-dimensional nature of images compared to text, the masking tasks are challenging in design.

**Masked Region Modeling (MRM)** applies zeros over image regions, asking the model to infer missing parts. Semantic and pixel level information can be retrieved, naturally leading to two sub-tasks:

- **Classification** of regions can be used to obtain a discrete signal corresponding to \textit{semantic} entities.
- **Regression** of region features provides a more continuous, \textit{pixel-level} understanding of missing parts.

**Random Pixel Sampling (RPS)** (Huang et al., 2020b) tackles overfitting similarly to the Dropout mechanism (Srivastava et al., 2014): in each iteration, a subset of pixels is chosen to be inserted in the transformer network. The corrupted image that the model receives each time enables more robust representations, as relying on \textit{semantics} instead of single \textit{pixels} is encouraged.

**Cross-modal objectives** jointly leverages information from both text and image modalities for learning. The design of \textit{cross-modal} pre-training tasks is more challenging compared to their \textit{unimodal} counterparts, as it is required to ensure that the model does not depend learning on a single modality exclusively (Chen et al., 2020; Wang et al., 2021c; Du et al., 2022).

**Unidirectional (seq2seq) Language Modeling (seq2seqLM)** is an autoregressive variant of MLM, which means that a masked token can only attend to previous tokens and not to future ones. Seq2seqLM attempts to directly maximize the likelihood of a text, image pair \((T, I) = x \in (W, V)\) from dataset \(D\):

\[
\mathcal{L}_{LM} = -\mathbb{E}_{(W,V) \in D} \log P_\theta(x)
\]

**Masked Language Modeling with Image (MLMI)** attempts to recover the corrupted text tokens by also consulting the image apart from the linguistic part exclusively. Inferring the masked token is achieved by minimizing the negative log-likelihood, where \(w_m\) are the masked tokens, \(w_m\) are the unmasked ones, and \((W, V) \in D\):

\[
\mathcal{L}_{MLM} = -\mathbb{E}_{(W,V) \in D} \log P_\theta(w_m|w_{\backslash m}, V)
\]
Whole Word Masking (WWM) (Kim et al., 2021) is an extension of MLMI which masks out entire words rather than subword tokens. Therefore, the model is encouraged to consult the visual part to infer the corrupted word instead of guessing the missing part from its unmasked constituents. This procedure enforces a more difficult task to the transformer, achieving better cross-modal alignment.

Masked Region Modelling with Language (MRCL) is the complementary task of MLMI. Instead of text tokens, image region features are masked out with a probability (by default 0.15). Masking does not include a special token for the visual modality; it is implemented by filling the corresponding image regions with zeros. The model is tasked to reconstruct visual regions with zeros. The model token for the visual modality; it is implemented by filling image region features are masked out with a probability instead of guessing the missing part from its unmasked constituents. This procedure enforces a more difficult task than subword tokens. Therefore, the model is encouraged to consult the visual part to infer the corrupted word instead of guessing the missing part from its unmasked constituents.

An extension of MRC can consider the overall distribution of the object detector label predictions instead of exclusively focusing on the top class. In that case, the objective function would attempt to minimize the distance between the object detector distribution and the transformer’s distribution of predictions, which is actually equivalent of minimizing the KL divergence between those two distributions. The objective function for Masked Region Classification with KL-Divergence (MRC-KL) for a distribution of object detector labels \(\tilde{c}(v_i)\) instead of top object detector label \(c(v_i)\) can be written as:

\[
L_{\text{MRC-KL}} = -E_{(W, V) \in D} \sum_{i=1}^{M} D_{KL}((FC(h_{vi}), \tilde{c}(v_i))
\]

Image-Text Matching (ITM) enables learning visiolinguistic matches in a global level. It can be viewed as a multimodal extension of next sentence prediction (NSP), where the model needs to recognize whether a given pair of text and image is in fact matched or not, as both positive and negative pairs are sampled. The alignment probability between text and image is provided by a score function \(s_\theta\), and the binary cross-entropy (CE) needs to be optimized:

\[
L_{\text{VLM}} = -E_{(W, V) \in D}[y \log s_\theta(W, V) + (1 - y) \log(1 - s_\theta(W, V))]
\]

Word-Region Alignment (WRA) is a more fine grained version of ITM, where words have to be grounded to image regions.

Contrastive objectives act upon data pairs projected on the same semantic space, so that the model learns a representation based on their similarity.

Cross-Modal Contrastive Learning (CMCL) learns to place close together matching image-text pairs, while pushing apart any mismatched ones. The contrastive loss for the i-th and j-th pairs sampled from \(D\), where \(v_i\) refers to the image of the i-th pair and \(w_j\) refers to the text of the j-th pair, \(s_\theta(v_i, w_j) = v_i^Tw_j\) is a scoring function which is maximized when matching pairs occur, and \(\sigma\) serves as a learnable temperature parameter:

\[
L_{\text{CMCL}} = -\frac{1}{n} \sum_{i=1}^{M} \log \frac{\exp(v_i^Tw_i/\sigma)}{\sum_{j=1}^{M} \exp(v_i^Tw_j/\sigma)}
\]

3.4.4 Task-specific fine-tuning

Current VL tasks are usually created by extending existing tasks in NLP or vision domains. They may either be discriminative or generative, and usually there are variants addressing both problems (Mogadala et al., 2021).
Visual Question Answering (VQA) Given an image \( I \in V \) and a natural language question \( q \in W \), a VQA model is tasked to predict the correct answer \( a \in W \). VQA is an extension of NLP question answering (QA) to include the visual modality. It can be viewed as a classification task, where the predicted answer can be selected among a set of candidate answers. Alternatively, free-form answers can be generated, forming a generative VQA task. Widely used datasets for VQA, containing complex scenes of objects and relationships, are the original VQA (Agrawal et al., 2016), VQAv2 (Goyal et al., 2016), GQA (Hudson and Manning, 2019), Visual7W (Zhu et al., 2016), Visual Genome QA (Krishna et al., 2016) and COCO QA (Ren et al., 2015). Additionally, datasets providing explanations for answer selections, such as VQA-E (Li et al., 2018) have recently emerged.

Visual Entailment (VE) extends the default task of textual entailment by answering whether an image \( I \in V \) acting as the premise semantically entails the given textual hypothesis \( h \in W \). The hypothesis \( h \) can either entail, contradict or remain neutral with respect to the premise, providing an answer \( a \) to the visual premise \( I \). SNLI-VE (Xie et al., 2019) is a dataset used for VE. Moreover, eSNLI-VE (Do et al., 2020) corrects label errors present in SNLI-VE due to its automatic assembling while providing human-written explanations in natural language for the corrected SNLI-VE corpus.

Visual Referring Expressions (VRE) or Visual Grounding extends NLP referring expressions, attempting to ground a textual phrase \( s \in W \) to an image object or region for each image \( I \in V \). Datasets used for VRE are CLEVR-Ref+ (Liu et al., 2019a), RefCOCO, RefCOCO+, RefCOCO-g (Yu et al., 2016) and GuessWhat (de Vries et al., 2016).

Visual Dialog (VD) is the analogue of chatbots, aiming to maintain a meaningful conversation by responding to consecutive textual inputs \( q \in W \). VD is tasked to create such a dialog upon a given image \( I \in V \). VisDial (Das et al., 2016) is a dataset for VD that was proposed together with the introduction of the task. Other datasets for multimodal dialog are GuessWhat (de Vries et al., 2016) and CLEVR-Dial (Kottur et al., 2019).

Image retrieval from text or Text-Image Retrieval (TIR) attempts to return the most suitable image \( I \in V \) within a database according to a natural language description \( c \in W \). There is also the inverse task of text retrieval from image or Image-Text Retrieval (ITR) that searches the optimal \( c \in W \) given an image \( I \in V \). Cross-modal retrieval, referring to retrieving any modality from the other one is an extension of the NLP task of document retrieval. Common datasets used in TIR/ITR are COCO (Lin et al., 2014) and Flickr8k/30k (Young et al., 2014) which contain images paired with 5 captions each.

Image Captioning (IC) is a generative task, extending natural language generation (NLG) to describe images: given an image \( I \in V \), provide a sentence \( c \in W \) that describes it. IC can be viewed as a generative counterpart of text retrieval from image (ITR). Dense Captioning is a fine-grained analogue of IC that requires generation of descriptions for image regions instead of providing a global visual caption. Conceptual Captions (CC) (Sharma et al., 2018), SBU (Ordonez et al., 2011), COCO (Lin et al., 2014) and Flickr8k/30k (Young et al., 2014) are widely used datasets for image captioning.

Visual Storytelling (VIST) is the extension of image captioning to a sequence of \( N \) related images \( I_1, I_2, ..., I_N \in V \). Generated captions \( c_1, c_2, ..., c_N \in W \) should be consistent with each other throughout the sequence, forming a textual ‘story’. Datasets related to VIST are Visual Storytelling Dataset (VIST) (Huang et al., 2016a) which models social language regarding visual concepts, and New York City Storytelling (NYCStorytelling) (Park and Kim, 2015), which contains narratives from blogs.

Multimodal Machine Translation (MMT) is assisted from the visual modality to translate between two languages, as an extension of the machine translation task. Multi30K-MMT (Elliott et al., 2016) contains multilingual descriptions of images in English, German, French, and Czech languages.

Visual Reasoning (VR) extends visual perception tasks, such as object detection and classification, semantic segmentation and others. VR needs to predict meaningful relationships between image entities, which is similar to creating a scene graph. Compositional reasoning refers to the task where attributes need to be combined so that the identity of the whole can be inferred. Popular datasets for VR are Compositional Language and Elemental Visual Reasoning (CLEVR) (Johnson et al., 2017), Relational and Analogical Visual rEasoNing (RAVEN) (Zhang et al., 2019a), Natural Language Visual Reasoning (NLVR) (Suhr et al., 2017) and Natural Language Visual
Visual Commonsense Reasoning (VCR) attempts to understand an image $I \in V$ by incorporating commonsense knowledge relationships to explain the answer $a \in \mathcal{W}$ derived for each question $q \in \mathcal{W}$. It can be also viewed as an extension of the VQA task, where instead of merely providing an answer $a$ to a given visual question $q$, a rationale $r \in \mathcal{W}$ justifying the choice of the answer needs to be returned as well. The answer $a$ and rationale $r$ can be also generated, thus yielding the Visual Commonsense Generation (VCG) task. Widely used datasets are VCR (Zellers et al., 2019) and Visual COMmonsense rEasoning in Time (Visual COMET) (Park et al., 2020).

Vision-and-Language Navigation (VLN) can be the equivalent of either visual navigation or linguistic navigation in the multimodal domain. Some datasets for VLN are Cooperative Vision-and-Dialog Navigation (CVDN) (Thomason et al., 2019), Action Learning From Realistic Environments and Directives (ALFRED) (Shridhar et al., 2020) and others.

3.5 Image generation

There is a lot of architectural variation when the modality to be generated is the visual one, greatly diverging from architectures employed for discriminative visual tasks. Image generation can be performed by leveraging Generative Adversarial Networks (GANs), Transformers, Diffusion models or a combination of them.

3.5.1 Conditional image generation tasks

Conditional Image Generation (cIG) addresses the synthesis of an image $I$ guided by textual information $s \in \mathcal{W}$, extending uncoditional image generation. Traditionally, text to image generation is performed by conditional GANs (cGANs) (Mirza and Osindero, 2014), which attempt to generate realistic images semantically corresponding to a text description. Image generation from text can be considered as the generative counterpart of image retrieval from text (TIR), and can be also regarded as the inverse task of image captioning (IC). Text to image GANs have been benchmarked on a plethora of datasets, either containing simpler images with simple conditioning such as ImageNet (Deng et al., 2009), Oxford Flowers (Nilsback and Zisserman, 2008), FFHQ (Karras et al., 2018) and CIFAR (Krizhevsky, 2009), longer conditioning such as captions in natural language, as in CUB (He and Peng, 2020; Reed et al., 2016b), or even more complex scenes accompanied with captions as conditioning, such as COCO (Lin et al., 2014).

Story Visualization (SV) refers to the synthesis of a visual sequence $I_1, I_2, ..., I_N$ based on an input story $c_1, c_2, ..., c_N \in \mathcal{W}$, the inverse task of visual storytelling (VIST). Once again, GAN architectures are leveraged to produce the sequence of images, which should not only be realistic, but also maintain the serial progression from frame to frame while remaining relevant to their textual description. Datasets accompanying SV research are CLEVR-SV (Li et al., 2019c), Pororo-SV (Kim et al., 2017), Flinstones-SV (fli) and DiDeMo-SV (Hendricks et al., 2017).

3.5.2 Generative VL architectures

Text-to-image synthesis Adversarial text to image synthesis has demonstrated a long line of impressive results by converting textual inputs such as captions, interactive dialogs, sequential story-like captions or structured formats of textual inputs (such as scene graphs and layouts) to plausible images. Some of the first ventures (Reed et al., 2016a,d) achieve synthesizing images, even though their quality is fairly low. To resolve this limitation, subsequent implementations synthesize images in stages, increasing image resolution step by step. StackGAN (Zhang et al., 2017) and StackGAN++ (Zhang et al., 2018a) exploit stacked generators and discriminators, dedicated to coarser or finer resolutions. AttnGAN focuses on individual words rather than whole sentences to synthesize finer details of the image (Xu et al., 2018). This idea is extended by SEGAN (Tan et al., 2019), which attends on relevant keywords from the sentence rather than all existing words. The most significant parts of the sentences with respect to the image contribute in synthesizing images with reduced fuzziness in details, as reported in DM-GAN (Zhu et al., 2019).

Sequential text-to-image synthesis StoryGAN (Li et al., 2019c) was the model that introduced the Story Visualization (SV) task, which concerns synthesizing a sequence of related images, maintaining consistency across story frames. The main StoryGAN architecture consists of a generator and two discriminators (image and story discriminators) guided by an RNN structure responsible of encoding the textual story. Improvements on the basic model were performed in (Zeng et al., 2019; Li et al.,...
More recent models substituted the RNN encoding scheme with transformers (Maharana et al., 2021; Maharana and Bansal, 2021; Tsakas et al., 2023), following the same trend as in other multimodal tasks.

Text-to-image generative VL Transformers Surprisingly, VL transformers in their original form are not able of generating realistic images, despite their impressive capabilities in other visual tasks. This issue is attributed to the fact that regression based training objectives, such as the one used in LXMERT (Tan and Bansal, 2019), are not able to handle feature generation in high dimensional spaces. Getting VL transformers one step further, X-LXMERT (Cho et al., 2020) as an extension of LXMERT, generates caption-conditioned high fidelity images consistent to their descriptions. X-UNITER follows the same extension logic as X-LXMERT, based on UNITER architecture (Chen et al., 2020). The image generation capabilities of X-UNITER are comparable to X-LXMERT, showcasing the general applicability of this approach.

DALL-E (Ramesh et al., 2021) uses a 12 billion parameter version of GPT-3 to generate fine-grained and highly diverse images based on corresponding textual descriptions. It demonstrates a large range of conditional synthesis capabilities, such as controlling certain visual attributes, as well as object positioning in an accurate way, capturing and visualizing 3D scenes, performing several natural effects like reflections, inferring missing details from descriptions, combining unrelated concepts in one image, performing zero-shot reasoning in the visual domain, incorporating external spatial and temporal knowledge and others. Once again, a connection between model scale and advanced performance is observed in terms of zero-shot generation and regarding the range of generalization capabilities.

Text-to-image diffusion models are setting new baselines for state-of-the-art conditional image generation (Dhariwal and Nichol, 2021). They follow a process of gradually adding Gaussian noise on an image and then learning to reconstruct it. During the last year, the field of diffusion-based image synthesis has received a variety of interesting works. Stable Diffusion (Rombach et al., 2021) enables previously computationally demanding image synthesis even within limited resources scenarios by applying the diffusion training process in the latent space of autoencoders rather than pixel-level operations in the image space. DALL-E2 (Ramesh et al., 2022) extends the high-quality result of its predecessor (Ramesh et al., 2021) by using learned text-conditioned image embeddings obtained from CLIP (Radford et al., 2021) as conditionings to a diffusion model acting as the decoder. Synthesized images are photorealistic and faithful to their conditioning, while zero-shot language-guided manipulation on a source image is also possible. The concurrent work of Imagen (Saharia et al., 2022) exploits large pre-trained language models, such as T5 (Raffel et al., 2020) for language encoding and proceeds with image synthesis based on the diffusion process. DreamBooth (Ruiz et al., 2022) steps upon Imagen to contextualize image synthesis, given variable context described in text. Therefore, different variations of visual subjects can be obtained, maintaining high synthesis quality.

Combined GAN/Transformer text-to-image models Hybrid architectures utilize existing GAN generators together with powerful VL transformers such as CLIP (Radford et al., 2021). Given an input text prompt, CLIP is responsible of guiding image synthesis in the latent space, based on the text-image similarities it has learned. Fuse-Dream (Liu et al., 2021b) follows this paradigm by optimizing the latent space of pre-trained GANs for efficient navigation. Similar hybrid text to image implementations are Big Sleep\(^1\) and VQGAN+CLIP\(^2\).

3.6 Evaluation metrics for VL models

3.6.1 Classification metrics

**Discriminative** VL tasks that provide an answer chosen among pre-defined candidates can be evaluated via classification metrics. Such tasks include visual question answering (VQA), visual referring expressions (VRE), visual reasoning (VR), visual entailment (VE).

**Accuracy@k** measures the proportion of correct answers over all answers, where an answer is considered to be correct if it belongs within the top-k answers. It can serve as a generic measure of quality for discriminative VL tasks, irrespectively of the modality that is predicted as the answer: for example, in the case of VQA, accuracy refers to the linguistic modality, as a textual answer needs to be selected. When bounding boxes have to be predicted, as in the case of VRE, **Intersection over Union (IoU)** provides a measure of success regarding the overlap between ground truth and predicted bounding boxes. Higher values are better, indicating a larger percentage of overlap.

3.6.2 Ranking metrics

**Ranking metrics** (Manning et al., 2008) provide further insights regarding the success of retrieving the right an-

\(^1\)BigSleep notebook
\(^2\)VQGAN+CLIP notebook
swear by providing information related to the position the ground truth answer was ranked. Tasks commonly using ranking metrics are image captioning (IC), visual storytelling (VIST), visual dialog (VD), machine translation (MT). **Recall@k (R@k)** measures the proportion of total ground truth instances that were found in the top-k rank, without taking into account their ordering. Higher Recall@k scores are better. **Precision@k (P@k)** is the percentage of ground truth answers in the top-k over all retrieved top-k items, without taking into account their ordering. Higher Precision@k scores are better. When ordering is to be considered, **Mean Reciprocal Rank@k (MRR@k)** acts as a useful performance measure. As reciprocal rank of an answer is considered the inverse of the rank position of the -first- right answer. By averaging over all instances, MRR is derived, with higher values indicating better performance. For N instances, if \( rank_i \) denotes the position of instance \( i \), MRR can be written as:

\[
MRR = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{rank_i}
\]

Another order-sensitive metric is **Discounted cumulative gain (DCG)**, which measures the gain an answer offers based on its ranking by taking into account a graded relevance scale. The graded relevance value is supposed to reduce logarithmically with respect to the position, thus penalizing highly relevant answers that appear lower in the rank. For \( rel_i \) the graded relevance at position \( i \) and \( p \) a particular rank position, DCG score for \( p \) is defined as:

\[
DCG_p = \sum_{i=1}^{p} \frac{rel_i}{\log_2(i+1)}
\]

**Normalized Discounted Cumulative Gain (NDCG)** is a normalized rank quality score that represents the ratio between the DCG score of the rank returned by an algorithm divided by the ideal order DCG (iDCG).

**Median Rank (MedRank)** is the median position of ranked ground truth answers when all answers have been considered. In a similar fashion, **Mean Rank** denotes the average position of ranked ground truth answers when all answers have been considered. Lower median and mean rank values are better.

### 3.6.3 Similarity metrics

**Wu-Palmer similarity (WUPS)** (Wu and Palmer, 1994) reports the degree of similarity between two words based on their least common subsumer on WordNet (Fellbaum, 1998) taxonomy. Various thresholds can define the agreement between two candidates, with the most typical being 0.0 and 0.9.

### 3.6.4 Language generation metrics

Language generation tasks such as IC, VIST and generative versions of VQA and VCR (VCG) evaluate produced outputs using the following language generation metrics.

One of the oldest metrics is **Bilingual Evaluation Understudy (BLEU)** (Papineni et al., 2002), originally developed to evaluate machine translation. It compares how much a machine generated linguistic output matches text written by a human in a precision-oriented way. Therefore, it is supposed to demonstrate high agreement with human perception regarding the quality of generated text, with higher BLEU score indicating higher perceived quality and maximum BLEU score being equal to 1 or 100%. BLEU can not only measure individual word matches, but also n-gram (grouped words) matches, providing BLEU-N scores for different N (most usually N = 1 for unigrams, 2 for bigrams, 3 for trigrams, 4 for quadrigrams). Nevertheless, it cannot assess linguistic diversity of generated text and sometimes it cannot appropriately perform in practical settings. BLEU is rather reliable for long sentences, but not very helpful in short/monolectic ones.

**Recall Oriented Understudy for Gisting Evaluation (ROUGE)** (Lin, 2004) is designed to evaluate the quality of summarization compared to a human-made summary. Similarly to BLEU, n-grams of varying N are utilized to calculate ROUGE-N scores. ROUGE-N offers not only recall, but also precision and F1 evaluation. Specifically, **ROUGE-N Precision** measures how many overlapping n-grams were found over the total generated n-grams:

\[
ROUGE-N_{\text{precision}} = \frac{\text{common n-grams in generated and reference}}{\text{n-grams in generated}}
\]

**ROUGE-N Recall** assesses how many overlapping n-grams were found compared to the human-made reference:

\[
ROUGE-N_{\text{recall}} = \frac{\text{common n-grams in generated and reference}}{\text{n-grams in reference}}
\]

Finally, **ROUGE-N F1 score** takes into account both ROUGE-N precision and recall:

\[
ROUGE-N_{F1} = \frac{2 \times (ROUGE-N_{\text{precision}}) \times (ROUGE-N_{\text{recall}})}{(ROUGE-N_{\text{precision}}) + (ROUGE-N_{\text{recall}})}
\]

Additionally, there is **ROUGE-L** score, which measures the longest common subsequence (LCS) between generated and ground truth text, with a longer LCS implying higher similarity. ROUGE-L can operate in sentence-level or summary-level and demonstrates precision, recall and F1-score variants. Specifically, **ROUGE-L Precision**...
measures the LCS length compared to generated n-grams count:

$$\text{ROUGE-L}_{\text{precision}} = \frac{\text{LCS (common n-grams)}}{\text{count of n-grams in generated}}$$

Moreover, ROUGE-L Recall measures the LCS length compared to ground truth n-grams count:

$$\text{ROUGE-L}_{\text{recall}} = \frac{\text{LCS (common n-grams)}}{\text{count of n-grams in reference}}$$

**ROUGE-L F1 score** takes into account both ROUGE-L precision and recall:

$$\text{ROUGE-L}_{F1} = \frac{2 \times (\text{ROUGE-L}_{\text{precision}}) \times (\text{ROUGE-L}_{\text{recall}})}{(\text{ROUGE-L}_{\text{precision}}) + (\text{ROUGE-L}_{\text{recall}})}$$

There are also some other rarely used variants of ROUGE score. ROUGE-W searches for the Weighted Longest Common Sub-sequence. ROUGE-S (Skip-Bigram Co-Occurrences Statistics) measures overlap of non-consecutive n-grams between generated and reference sequences. Similar to BLEU, ROUGE scores are more effective in long sentences rather than short ones.

**Metric for Evaluation of Translation with Explicit Ordering (METEOR)** (Banerjee and Lavie, 2005) tackles the need for exact word matching as in BLEU score, and instead enforces semantic matching, taking into account possible synonyms and paraphrases of words in the reference text. Semantics matches are made possible due to the usage of WordNet (Fellbaum, 1998). Unigram alignment between ground truth and generated sequences contribute to the METEOR score. Multiple possible alignments between the two sequences are resolved by selecting the alignment with the fewest crosses, when ground truth and generated unigrams are matched. METEOR presents high agreement with human perception regarding the similarity of generated text sequences.

**Consensus-based Image Description Evaluation (CIDEr)** (Vedantam et al., 2014) is a metric inspired from human agreement when ground truth and generated sentences are compared. Such similarity can be expressed via TF-IDF score of n-grams across reference sentences, instructing frequently occurring n-grams in the whole dataset to have lower weight in the final score, as it is more possible to be less informative (IDF term), while at the same time increasing the weight of frequent n-grams within a reference sentence (TF term). Similar to METEOR, CIDEr exploits semantic matching by comparing stemmed versions of words in sentences. A cosine similarity score between generated and ground truth vectors computed from TF-IDF n-gram weights provides the final CIDEr score.

**Semantic Propositional Image Captioning Evaluation (SPICE)** (Anderson et al., 2016) is an automated evaluation metric for language generation operating over scene graphs, which are synthesized from ground truth captions and generated captions via dependency parsing. WordNet is utilized for disambiguation and synonym detection. Even though it resolves shortcomings of previous methods related to n-gram overlaps, the scene graph construction stage is possible to induce errors early in the evaluation process.

**BLEURT** (Sellam et al., 2020) is a BERT-based learned evaluation metric that addresses the shortcomings of traditional BLEU and ROUGE metrics in order to better correlate the resulting scores with human perception. It combines the merits of learning pure linguistic associations, as well as human assessments over language metrics. Therefore, pre-training on synthetic sentence pairs, which are designed to capture semantic, syntactic and lexical information assists in identifying possible dissimilarities between real and generated text. Afterwards, fine-tuning adapts automatically captured disagreements to actual human ratings.

### 3.6.5 Image generation metrics

**Inception Score (IS)** (Salimans et al., 2016) evaluates the quality and diversity of generated images by utilizing a pre-trained image classifier, such as Inception V3 (Szegedy et al., 2015). The classifier is tasked to provide a probability of whether an image is generated or real. IS can have a lowest value of 1.0 and a highest value equal to the number of classes the pre-trained classifier has seen, which in case of Inception V3 is 1000 (ImageNet (Deng et al., 2009) classes). Higher IS implies better quality and more diversity, therefore higher IS is better.

**Fréchet Inception Distance (FID)** (Heusel et al., 2017) compares the distribution of generated images with the distribution of the real ones by taking into account the mean and the standard deviation of the distributions. A pre-trained Inception-V3 (Szegedy et al., 2015) model is utilized to extract summary statistics for real and generated images. Lower FID scores are better, indicating higher similarity between generated and real images.

**Learned Perceptual Image Patch Similarity LPIPS** (Zhang et al., 2018b) is a metric designed to reflect the perceptual similarity between real and generated images using deep features of image classifiers. Lower LPIPS values indicate more similar images.

**R-precision** is another widely used metric to evaluate synthesis quality. Despite primarily being a retrieval metric, it can serve conditional generative models: generated images are used as queries to ground truth descriptions,
with higher scores indicating better quality of generated images.

4 Graphs

A graph structure $G = (V, E)$ consists of a set of nodes $V$ interconnected with weighted or unweighted edges from a set $E$. Edges also can be either directed or undirected depending on the constrains of relationships they express. Considering two nodes $v_i \in V, v_j \in V$, their in between edge is denoted as $(v_i, v_j) \in E$. Nodes and edges can additionally contain features. There are some distinct subcategories of graphs mentioned below, often present in real world data representation scenarios (Hamilton).

Different types of relationships raise the need for distinct edge representations, which is satisfied via Multipartite graphs. The edge notation needs to be altered to contain the edge type $\tau$, so that the multi-relational edge notation being $(v_i, \tau, v_j) \in E$. Heterogeneous graphs extend multi-relational graphs by introducing varying node types, so that nodes can contain labels forming non-overlapping node sets. Therefore, the node set $V$ can be expressed as a union of node labels $V = \mathcal{V}_1 \cup \mathcal{V}_2 \cup \ldots \cup \mathcal{V}_n$ for $n$ distinct and disjoint node categories. Those categories most often impose constrains on the edge types as well in order to remain meaningful. Multipartite graphs are heterogeneous graphs which exclusively contain edges that connect nodes of different types.

4.1 Knowledge Graphs

A Knowledge Graph (KG) $\mathcal{G}$ is a structured representation of facts $\mathcal{F}$, which consist of entities $\mathcal{E}$, relationships $\mathcal{R}$ and semantic descriptions. Entities can describe either abstract concepts or actual objects, relationships form the meaningful connections between entities, and semantic descriptions incorporate types and properties of those objects and relationships. KGs are directed and heterogeneous structures that describe human knowledge in the form of triplets head - relationship - tail, often denoted as $(h, r, t) \in \mathcal{F}$, or equivalently subject - predicate - object, denoted as $(s, p, o) \in \mathcal{F}$. (Ji et al., 2021)

While existing edges express known facts, there are two scenarios for missing edges (Ji et al., 2021): Open World Assumption (OWA) assumes that unobserved facts are either missing or false; Closed World Assumption (CWA) assumes that all unobserved facts are false.

4.2 Graph representation

Graph representation has been a field of increasing interest, as it affects numerous applications in artificial intelligence. Similar to language representations, KG embeddings are low-dimensional mappings of the graph entities and relationships. These vector representations capture the semantic information contained in KGs, which can consequently be used for a variety of downstream tasks (Ji et al., 2021; Hamilton et al., 2018).

Node embeddings constitute a family of graph embedding algorithms aiming to represent the nodes’ position and context in a vector space. Popular methods regarding node embeddings are based on random walks, presenting widely used implementations such as node2vec (Grover and Leskovec, 2016), DeepWalk (Perozzi et al., 2014), Large-scale Information Network Embedding (LINE) (Tang et al., 2015), Graph2vec (Narayanan et al., 2017) and others. However, standalone node embeddings are inadequate for multi-relational graph representations, which require an overall representation of nodes, edges and attributes. To this end, shallow translational models exploit geometric capabilities of the vector space to achieve multi-relational graph representations (Hamilton).

Graph Neural Networks (GNNs) refer to a broad framework that achieves graph structure and feature representations using deep learning. The basic idea behind GNNs is neural message passing, meaning the exchange and aggregation of neighboring node information through their connections, resulting in updating the node embedding itself. The process repeats by expanding in more distant neighborhoods (hops) in every step, integrating more information regarding graph structure and neighboring features (Hamilton).

More specifically, each node $v \in V$ of a graph $G$ is initialized with a random embedding $h_0$ and the following hidden embedding states $h_v$ are computed via a local transition function $f$ such that:

$$h_v = f(x_v, x_{co[v]}, h_{N_v}, x_{N_v})$$

where $x_v$ are the features of the node $v$, $x_{co[v]}$ are the edge features connected to $v$, $h_{N_v}$ are the embedding states of the neighbors of $v$ and $x_{N_v}$ are their features. The output is given from a local output function $g$:

$$o_v = g(h_v, x_v)$$

Graph Convolutional Networks (GCNs) are a variant of CNNs operating on graphs. However, different than images, graphs contain unordered nodes with varying numbers of connections. Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{X})$, the GCN receives as an input the feature matrix $\mathcal{X}$ containing node features, and an adjacency matrix $A$ representing the graph structure. The convolution operation is generalized so that the representation of a current node is obtained by aggregating its own features, as well as the features of its neighbors. A non-linear transformation is then applied
GCNs can effectively be applied on relational data and serve a variety of relevant downstream tasks. Relational Graph Convolutional Networks (R-GCNs) comprise a GCN variant for relation-specific encodings for KGs, i.e. encodings depending on edge directivity and type. Therefore, different weights will be assigned to different relationships (Schlichtkrull et al., 2017).

**Graph Attention Networks (GATs)** are based on attention mechanisms, which have the ability of handling variable sized inputs. Therefore attention mechanisms could be applied successfully for the representations of graphs containing nodes with different edge degrees. Indeed, self-attention allows each node to attend to its neighbors, thus computing the hidden encoding. This process can be parallelized for neighboring pairs, independently of graph structure (Veličković et al., 2018).

Despite GNNs being able to learn representation on fixed and homogeneous graphs, they are not so powerful on graphs containing heterogeneous information on nodes and edges. In such cases, embedding extraction mechanisms should be adjusted appropriately to encode heterogeneous information. Various types of relationships create meta-paths between nodes, containing diverse and rich semantic information. GATs are extended by applying a hierarchical attention mechanism with node-level and semantic-level attention, assigning different importances to nodes and meta-paths. Neighboring features are aggregated in a hierarchical fashion forming the final node embeddings (Wang et al., 2021b).

**Graph Transformer Networks (GTNs)** can also handle heterogeneity and generate reasonable meta-paths connecting nodes, as well as effective node representation. Meta-paths may be of any length up to the number of Graph Transformer layers and contain arbitrary edge types. Multiple generated meta-paths can simultaneously be considered, defining multiple learned graphs. Due to continuously generating new graph structures from adjacency matrices contained in data, more powerful node representations are learned through convolutions, as there are higher chances of finding more useful meta-paths. A GCN is applied on each meta-path, yielding an ensemble of GCNs when considering all the generated meta-paths (Yun et al., 2020).

The Heterogeneous Graph Transformer (HGT) is designed to model web-scale heterogeneous graphs using heterogeneous attention. Node and edge-level dependent parameters are exploited in order to obtain dedicated node and edge representations. HGT extracts all related pairs from a sampled heterogeneous subgraph, so that each pair contains a source node and a target node. The information from source nodes is aggregated to provide a contextualized representation for a target node. This process can be decomposed in three parts, namely Heterogeneous Mutual Attention, Heterogeneous Message Passing and Target-Specific Aggregation. The weight matrices for the heterogeneous mutual attention, message passing, and propagation steps are parameterized using the meta-relations of the graph. Dynamic representation graphs are handled using the relative temporal encoding technique, which captures arbitrary long dynamic structural dependencies (Hu et al., 2020).

### 4.2.1 Graph representations in KVL models

Apparently, one important aspect towards KVL models is how knowledge is incorporated with vision and language modalities. First attempts to incorporate knowledge stored in KGs were based on exact matching between extracted visual and textual concepts from the given input and existing KG nodes. However, exact matching will result in errors even when there is a small discrepancy between KG entities and extracted concepts, limiting the contributions of additional knowledge. In order to better exploit semantically rich structured knowledge sources, more refined strategies, such as the usage of GNN representations are explored in recent literature. The variability of GNN implementations as described previously allows dedicated representations for different graphs, with incorporation of node, edge and feature information. Therefore retrieved knowledge is more accurate and informative, ultimately capable of boosting VL models.

### 5 Knowledge

In recent years, there has been an ever increasing interest about incorporating external knowledge $K$ in VL models. We can identify the merits of such an approach: additional knowledge can offer performance boosting, extendability and potentially explainability of existing VL tasks. Most common senses of additional knowledge offering those benefits are analyzed below.

**Hierarchical knowledge** refers to is-a relationships forming a tree structure, with the root serving as the most generic concept and parent node of all the rest, while leaves constitute the most specific concepts. For example, cat is-a mammal is an instance that represents such hierarchical relationships.

**Lexical knowledge** serves as a structured dictionary, offering linguistic rules, while being able to resolve issues such as word sense disambiguation. Lexical knowledge
can be combined with hierarchical knowledge, providing hypernym/hyponym relationships.

**Named entities** cover a variety of proper names as instances of entities, and include names of people, locations, companies, organizations etc (Grishman and Sundheim, 1996). For example, the sentence *Joe Biden is the president of the United States* contains the named entities *Joe Biden* and *United States*.

**Factual knowledge** includes encyclopedic information of the world, such as the historical fact *WW2 lasted from 1939 to 1945*. Such knowledge can also refer to more specific scientific facts, like knowledge in medical, biological, chemistry domains and many more. Facts can also be combined with named entities, forming statements like *Zebras live in Africa (Africa is a named entity)*.

**Commonsense knowledge** is the self-evident perception of the world according to humans; *sugar is sweet* and *if I go out in the rain I’ll get wet* are obvious commonsense statements. We can identify several discrete senses of commonsense knowledge, affecting aspects of the world a human experiences. Such subcategories refer to *similarity/dissimilarity* concepts, like synonyms and antonyms of words. Another commonsense variant includes *part-whole* (part-of) relationships representing concepts belonging to more generic ones or consisting members of a group, for example *the bark is a part of a tree, the tree is part of the forest*. Part-whole in terms of lexical knowledge is expressed via *meronyms* (part) and *holonyms* (whole). *Utility* relationships describe usage scenarios, such as *a fork is used for eating* or *capability (wheels can rotate)*. *Spatial* information offers knowledge about usual locations of objects in the physical world, for example *boats are situated near water*, or even geographic information, such as *Italy is located at Europe*, which sits on the intersection with *factual knowledge* and *named entities*. *Comparative knowledge* provides rules of comparison between objects, for example *leopards are larger than cats*. Such statements are crucial towards learning logical reasoning scenarios. *Numerical knowledge* addresses common enumerations in real life, providing facts such as *humans have two eyes*. *Intents, desires and plans* constitute another sense of commonsense knowledge, including facts like *hungry people want to eat* and *a hungry person cooks to eat*. *Behavioral knowledge* results from logical reasoning over commonsense facts, forming rules like *a child cannot drink 10 liters of water in one day*. Statements like *a song is created by a musician or bread is made from wheat* belong to *creator knowledge*.

*Event/temporal knowledge* contains chronological information and order of events, blending factual and commonsense knowledge. Events can refer to a large number of chronologically distinct time periods from widely known events such as *world wars, significant political events, sports, social/scientific movements* and a lot more, to more specialized events known to smaller audiences. Temporal sequences can contain chronologically ordered events. For example, *COVID-19 started in 2019. Vaccines for COVID-19 were developed during 2020* is a factual sequence of events. *Commonsense sequence* of events could contain information such as *Spring comes after winter*. Sequences can also refer to *causal* relationships with the cause preceding the event, such as *the boy dropped a glass of water and then the glass broke*, which can also be transformed to *hypothetical if/then* statements, for example *if a boy drops the glass of water, the glass will break*, or even *counterfactual* statements expressing what would have happened if an alternative scenario occurred, like *if the boy had not dropped the glass of water, the glass would not have been broken*.

**Visual knowledge** contains images and possibly additional annotations to connect *visual perception* with *commonsense*. Attributes of objects, such as shape, color, texture and others can be connected with their visual counterpart, visualizing commonsense situations such as *tomatoes are red and round*. Visual knowledge is ideal for learning instances of the world about object relationships and attributes, paving the way for more complex reasoning required in several multimodal tasks. *Spatial* relationships are naturally combined with images; for example *apples placed inside a bowl, bowl placed on a table*. More types of relationships can be further visualized, including *actions* between visual entities such as *a girl is holding a tennis racket, object details like black and white striped hat, part-whole has-a relationships such as woman has long blonde hair or scene text like a truck with Coca-Cola logo*. Those rather obvious statements can be extended to commonsense *assumptions*, such as *the temperature is low, when an image of an icy landscape is provided*. More complex visual instances can provide information about *intents* (a customer enters a restaurant to eat, a person holding a suitcase and a passport plans to travel), *causes/effects* (a biker cycling out in the rain will get wet), *factual* instantiations (an ancient Greek temple of the 5th century BC, girls with Japanese kimono dresses), *similarity reasoning* (the dog’s toy looks similar to a plate), similarity including *named entities* (A man looking similar to Brad Pitt), *creator knowledge* (the painting was created by a person holding a paintbrush), *capability* (a cat can jump on the tree branch).
5.1 Types of knowledge sources

We divide external knowledge in three categories: implicit knowledge, present in non-symbolic form, explicit knowledge, typically stored in structured knowledge bases, and web-crawled knowledge, acquired from various online sources, usually in unstructured form.

Moreover, we can recognize the category of internal knowledge or self-knowledge, which does not rely on external sources, but rather obtains extra knowledge from the existing data.

Implicit knowledge refers to information stored in a non-symbolic form, such as neural network weights. The indisputable popularity of neural architectures in recent deep learning literature has led to numerous relevant contributions, even if their primary goal deviates from knowledge representation. Unsupervised or self-supervised pre-training of transformer models can provide implicit knowledge in several linguistic (Safavi and Koutra, 2021) or multimodal tasks. Therefore, incorporating large-scale linguistic and visual data in the pre-training stage can seemingly form unstructured knowledge bases.

Insightful studies have attempted to discover the optimal pre-training regime and whether it serves as a necessary prior for performance. The right dataset and design choices are crucial for achieving successful representations, which can become a prerequisite for the overall success of the final task, after proper fine-tuning. First, the amount and type of pre-training data need to be examined. Specifically, the relevance of the selected pre-training dataset with respect to the downstream task has been proven more influential rather than dataset size, an observation that remains valid even when generated datasets are used over out-of-domain natural ones. Only scaling up in-domain data seems to positively impact the model, encouraging the scalability of implicit knowledge bases. In some cases, especially when selected pre-training datasets are not diverse enough, pre-training transferability towards downstream tasks is low, so that in fact pre-training knowledge is deemed insufficient. In the same fashion, even though the pre-training scenario which demonstrates the lowest losses can be regarded as the best prior, it can be proven suboptimal without proper fine-tuning. Data relevance seems to be more significant than model size, even though deeper single stream models mitigate the semantic gap between images and text in later layers. However, for double-stream models earlier layers present more narrow semantic gap, disagreeing with the single-stream observation (Cao et al., 2020; Singh et al., 2020).

Other investigations question if all participating modalities contribute equally to the learned representations, proving that during inference, the contribution of language is more prominent than vision in both single-stream and double-stream encoder architectures. Nevertheless, rich visual knowledge is encoded in pre-training, effectively capturing visual relationships (Cao et al., 2020). Such observations can offer valuable insights in the complex pre-training procedure and provide enhancements on the knowledge acquired, towards more high-quality, flexible and robust implicit knowledge bases.

Pre-training holds the advantage that it can exploit unlabelled data for achieving a generic understanding of language (Devlin et al., 2019; Brown et al., 2020), which can serve as an initial point for linguistic or VL tasks. In most multimodal cases however, a level of supervision is required, such as the need for paired images and captions. The current abundance of paired image-text samples resolves the labelling issue up to some point, at least for general-purpose domains. The raw nature of linguistic, visual and paired image-text data that are typically used for pre-training alleviates the need for a strict representation, which would limit the flexibility of incorporated knowledge in many aspects, including storage, expressivity and accessibility (Safavi and Koutra, 2021). Additionally, the automatic incorporation of extra information by repeating pre-training or performing fine-tuning can help extending and refining the required implicit knowledge, contrary to handcrafted knowledge bases which are hard to be extended at scale. Even though such pre-training procedures are very expensive computationally, luckily pre-trained transformer models are offered to the research community in ready-to-use models. Transfer learning can then effectively leverage existing implicit knowledge sources achieving impressing results in various tasks, which accounts for many advancements in VL tasks and transformer-based implementations in general.

Nevertheless, implicit knowledge is not always sufficient to answer questions requiring general, factual and commonsense knowledge, especially when rare information is requested. Additionally, its black box nature raises concerns about how and what a pre-trained model has learned; for example, biased or erroneous data received during pre-training will be reflected in all later stages, resulting in decreased performance of the downstream VL model. Tracing back the source of such a problem is not possible due to the lack of interpretability tied to implicit knowledge bases.

Explicit knowledge is based on clear, structured facts in the form of a knowledge graph and it is able to explicitly fill the gaps that cannot be covered via transfer learning. Even though most contemporary multimodal approaches, including transformer-based ones, have acquired a cer-
tain understanding of language, visual concepts and their in between relationships, they cannot effectively handle concepts and relationships they have never seen during training (Ilievski et al., 2021) (excluding implementations that present zero-shot capabilities (Radford et al., 2021; Wang et al., 2021c)). Consequently, even the best VL models will fail in cases the data distribution is significantly different from the one they have been trained on. The same discrepancy may apply even when an implicit knowledge source is used, if the implicit distribution remains rather distant. For example, a model trained on pairs of generic images and corresponding captions will inevitably present much lower metrics when asked to infer from medical images accompanied by relevant captions with scientific vocabulary. The same limitation is prevalent when there is a shortage of training data (Ilievski et al., 2021). Although an intuitive scenario would suggest to repeat the pre-training procedure, so that this extra information will be reflected via updated neural weights, the pre-training cost is in reality computationally unaffordable (Sharir et al., 2020) for the majority of research institutions. Even in that case, repeated occurrences of out-of-distribution data would demand from scratch pre-training or at least fine tuning each time, preventing the scalability of related tasks.

Another issue strongly interconnected with massive pre-training is the lack of explainability (Kafle et al., 2019), as it is difficult to track what and how a pre-trained model has learned from data. This black box nature poses questions regarding how rare concepts or rare combinations present in data are handled and whether they can be represented with equal success as the more common ones. At the same time, possible data biases, errors and inconsistencies will be reflected in learned representations, with those issues being hard to be captured and resolved beforehand.

On the contrary, in the case of explicit knowledge bases, the degree of contribution of the knowledge base can be measured and evaluated, offering valuable transparent insights regarding the role of knowledge. Such out-of-distribution information is well represented in structured knowledge graphs. Large scale knowledge can complement pre-trained models by extending their understanding to previously unseen concepts, either by substituting the need for extra training, or by enriching existing datasets to achieve more informative, fair and high quality representations, if (re-)training is necessary. Even in that case, pre-training demands can be reduced, achieving similar representation capabilities to larger models pre-trained without additional explicit knowledge. The quality of such representations is somehow controllable, a benefit which can be attributed to the explicit and transparent nature of KGs: issues regarding biases, errors, concept drifts and inconsistencies can be captured and resolved easily, exploiting automatic techniques or manual interventions. In any case, KGs should contain relevant information to the downstream task in order to be beneficial (Ilievski et al., 2021).

There are some downsides regarding the usage of explicit knowledge in the form of KGs. First, many KGs may require manual labor for data collection and curation. The same disadvantage also applies on the construction and maintenance of the graph itself. In certain cases, such as in the medical domain, experts are necessary in order to design and construct dedicated KGs. Moreover, there are difficulties regarding alignment and co-operation between different KGs (Ilievski et al., 2021), thus sometimes limiting in practice the improvements they offer.

Combining implicit and explicit information can offer advanced capabilities to downstream tasks, as implicit sources can fuse large-scale general knowledge to a model, while explicit sources can fix errors, enrich existing knowledge and increase a model’s transparency.

**Web-crawled knowledge** refers to unstructured knowledge obtained from the web, which is able to combine benefits present in implicit and explicit knowledge bases. There is no need for labelled data, but also no need for expensive pre-training, which is one major limitation of implicit knowledge. Online sources can be accessed easily, while the amount and the content of retrieved knowledge is its quality, as it is hard to validate any available web source. Low-quality data may deteriorate the final performance of the model, therefore time and effort has to be invested in techniques that ensure high-quality data automatically. Web knowledge can offer some amount of transparency, as a sentence leading to the final prediction can be tracked, even if reasoning is not as fully explicit as in cases of structured graphs.

**Internal or self-knowledge** is a knowledge type that does not rely on any external source, as it can be obtained from existing textual and visual data themselves. For example, producing a scene graph can enable learning more fine-grained representations compared to just utilizing VL data in their original format (Yu et al., 2021a). Self-knowledge has demonstrated improvements in downstream model performance, especially when detailed disambiguation is necessary, as it enables better associations between existing data. However, self-knowledge does not extend the knowledge a model has already acquired from the data it has been trained on. Furthermore, it is prone to errors associated with the knowledge acquisition process,
such as scene graph generation errors.

An overview of the available types of knowledge sources is provided in Figure 4.

5.2 Widely used Knowledge graphs

Widely used knowledge graphs in literature are analyzed below. WordNet (Fellbaum, 1998) is a large-scale lexical database which provides cognitive synonyms, called synsets, for words (nouns, verbs, adjectives and adverbs) of the English language, representing their conceptual and lexical relationships. In total there are 117K WordNet synsets that form a tree-structure through their relationships. Verb and noun synsets are interlinked with transitive hierarchical (is-a) relationships, bringing semantically related synsets close together. Therefore, distance on the WordNet graph provides a measure of concept similarity. There is also a distinction between types and instances, with types (common nouns) expressing more specific meanings of a concept (cat is a type of animal), while instances are specific persons, countries and geographic entities (Rome is an instance of a city). The root node of the WordNet tree belongs to the {entity} synset, the most generic concept and parent of all other synsets. Traversing from root to leaves leads to more and more specific concepts, with instances always being located at the leaf level. Synsets easily offer word sense disambiguation, as certain words that have many distinct meanings are mapped to a different synset, so that finally all synsets represent different meanings. Part-whole relationships are also present and parts of a current node can be inherited from parent nodes, but not vice versa. Adjectives are linked via semantic similarity and dissimilarity links, with antonyms and synonyms directly related.

ConceptNet (Speer et al., 2017) is a multilingual commonsense knowledge graph, containing a variety of relationships that express many dimensions of commonsense knowledge. In its current version of ConceptNet 5, it is comprised of ~8 million nodes interconnected by 21 million edges. Specifically, it contains hierarchical relationships (IsA), lexical (DerivedFrom, FormOf), path-whole (PartOf, HasA, MadeOf), similarity (Synonym, SimilarTo), dissimilarity/distinctness (Antonym, DistinctFrom), spatial (LocatedNear), utility (CapableOf, UsedFor), creation (CreatedBy), quality (HasProperty, SymbolOf), desire/goal (CausesDesire, ObstructedBy), temporal (Causes, Entails, HasSubevent) and relational-other (RelatedTo, HasContext) (Ilievski et al., 2021). ConceptNet data are aggregated from different sources, with initial versions based on crowdsourcing via the Open Mind Common Sense website. Supplemental sources include other knowledge bases, such as WordNet for lexical information, DBPedia for Wikipedia infoboxes and OpenCyc ontology for high-level commonsense facts, as well as multilingual sources serving similar knowledge content. Additionally, knowledge about intuitive linguistic associations is gathered via so-called ‘games with a purpose’.

DBPedia (Auer et al., 2007) is a multi-lingual knowledge base that extracts factual structured information from Wikipedia. It covers a variety of domains and evolves together with Wikipedia, thus handling concept drifts. DBPedia also provides SPARQL endpoints to enable queries.

Wikidata (Vrandečić and Krötzsch, 2014) is a free collaborative multilingual knowledge base of more than 6.7k relationships and more than 97 million data items, available to everyone for editing. It provides links of entities to their sources and other databases, endorsing the verifiability of contents.

WebChild (Tandon et al., 2014) is a large commonsense knowledge base automatically collected from the web. It presents high accuracy of statements, with fine-grained entries involving part-whole, comparative, property, activity and spatial relationships. All entries are disambiguated by mapping on WordNet synsets. In total, WebChild 2.0 (Tandon et al., 2017) contains more than 2 million concepts, interconnected by 18 million relationships.

HasPartKB (Bhakthavatsalam et al., 2020) contains part-whole statements extracted from a large generic corpus, covering numerous common terms. Due to the huge possible instances of part-whole relationships in real world, salient parts are preferred, referring to instances that are most probably useful to be stored. to all Entities are linked with WordNet and Wikipedia.

YAGO4 (Tanon et al., 2020) is a general purpose knowledge base storing knowledge about people, places, movies, organizations and others. It consists of 2 billion triples and 64 million entities automatically extracted from Wikipedia. WordNet is leveraged for entity and relationship disambiguation. YAGO4 actually forms a cleaner and more human-readable version of Wikidata, satisfying logical constrains, therefore allowing automated reasoning on the data.

ATOMIC (Hwang et al., 2021) is a commonsense knowledge graph with 1.33M tuples regarding events and entities, incorporating both social and physical senses of everyday experiences. In total, it contains 23 relationship types, with 9 types referring to social interactions, 7 types concerning physical entities and 7 types representing event relationships.

Visual Genome (Krishna et al., 2016) can be also viewed as a knowledge base, thanks to the scene graph annotations that explicitly showcase entities, relationships
and attributes present in a scene, accompanied by the actual visual information of each image. Mappings to WordNet synsets offer disambiguation as well as hierarchical relationships between visual instances. In general, scene graphs can act as spatial and visual knowledge bases.

6 Multimodal Tasks with knowledge

Recent efforts of integrating internal and external knowledge sources in VL tasks are analyzed in this section, together with datasets and evaluation methods they follow. As mentioned in Section 2, we can divide existing KVL approaches in single-task and multi-task models. Starting from single-task models, we first present works in the most developed field of discriminative KVL tasks, and more specifically understanding tasks such as visual question answering (VQA), visual reasoning (VR), visual commonsense reasoning (VCR), followed by generative tasks, such as image captioning (IC), visual storytelling (VIST), story visualization (SV), conditional (text-to-image) generation (cIG). Finally, multi-task models, targeting more than one downstream tasks at once are analyzed.

In Figure 5 an overview and taxonomy of KVL tasks is presented. Some tasks such as visual referring expressions (VRE), visual question answering (VQA), visual commonsense reasoning (VCR) and visual dialog (VD) can be either discriminative or generative, even though their discriminative variants are more widespread, being comparatively easier. Single-task models are focusing either on generative or understanding tasks, while there is no single-task model focusing exclusively on a retrieval task such as cross-modal retrieval (TIR/ITR) and visual referring expressions (VRE). Those two tasks are exclusively tackled -among with others- by multi-task models. In the same fashion, visual entailment (VE) is only addressed by multi-task models. Finally, a couple of tasks such as visual-language navigation (VLN) and multimodal machine translation (MMT) still lack a knowledge-enhanced counterpart.

The upcoming sections are primarily organized starting from single-task approaches as per task, followed by multi-task approaches in the end. Datasets, methods and evaluation metrics are provided for each independent task. A more detailed division of methods is achieved based on the knowledge type (external/internal) and the language representation scheme (sequential models/transformers) per approach.

6.1 Knowledge in Visual Question Answering (K-VQA)

6.1.1 Datasets

The following datasets have been used in implementations that have integrated external knowledge sources in order to provide an answer. Many of those, including VQA (Agrawal et al., 2016; Goyal et al., 2016), VQA-E (Li et al., 2018), COCO-QA (Ren et al., 2015) are also used in knowledge-free versions of VQA, and external knowledge is not required in order for their questions to be answered.

VQA (Agrawal et al., 2016) was the first dataset introducing the task of Visual Question Answering. It contains approximately 204K images with diverse and complex scenes, with at least 3 open-ended free-form questions per image and a total of 760K questions in the whole dataset. Many of those questions are commonsense related, such as *Is this a vegetarian pizza?*. For each question 10 ground truth answers are suggested from different annotators, with a total of almost 10 million answers in the dataset. Most answers are short, with the majority of them consisting of a single word. The answers can
Figure 5: A taxonomy of VL tasks with knowledge

be evaluated either in open-ended or in multiple-choice settings. Open-ended answers should be validated by 3 annotators agreeing on exactly the same answer for a given question. The multiple-choice scenario regards 18 unique candidate answers per question. Such candidates can be correct answers, obtained from the 10 matched answers per question, plausible answers, i.e. possibly incorrect answers provided by annotators without viewing the image, popular answers, i.e. the most frequently appearing answers in the dataset, and random answers sampled from other random questions within the dataset.

**VQA with Explanations (VQA-E)** (Li et al., 2018) pursues the tractability of the reasoning process leading to an answer. In total, it contains around 108k images, and more than 269k explanations assigned to an equal number of QA pairs. Based on VQAv2 (Goyal et al., 2016), it automatically constructs explanations with the help of COCO captions (Lin et al., 2014), as they are connected with VQA v2 images. Caption embeddings and question/answer embeddings are coupled, forming pairs of highest cosine similarities, thus assigning captions to images. Resulting explanations are highly diverse, with more than 171k unique instances, although they cannot cover images with subjective questions, such as emotional (Do you think this pony is cute?), commonsense (Can you cross the street?) or behavioral (Could you eat all these bananas by yourself?) ones. Human evaluators assess the quality of explanations, measuring is they are fluent, correct, relevant and complementary to the answer.

**DAQUAR** (Malinowski and Fritz, 2014) is a dataset of real world indoor scenes containing fine-grained object categories. Questions and answers related to the images are very rich regarding the objects they express: 573 unique nouns are mentioned within the whole corpus of questions and answers. Questions requiring commonsense knowledge such as Which object on the table is used for cutting? are included in DAQUAR, while even spatial questions such as What is above the desk in front of scissors? can be benefited from additional knowledge.

**COCO-QA** (Ren et al., 2015) addresses shortcomings of the DAQUAR (Malinowski and Fritz, 2014) datasets, such as its small size in terms of train/test samples and the limited number of object classes. COCO-QA contains 123,287 images, together with 78,736 train and 38,948 test questions obtained from COCO image descriptions (Lin et al., 2014). Questions are divided in 4 types with
various numbers of questions in each of them: Object, Number, Color and Location questions.

**KB-VQA** (Wu et al., 2016b) has been constructed in order to evaluate VQA models on questions that need visual information as well as external knowledge to explicitly infer the right answer. It includes images from COCO (Lin et al., 2014) containing approximately 150 object classes and 100 scene classes, question-answer pairs following pre-defined templates and question labels. The questions involved in KB-VQA are divided in three categories: visual questions can be answered by extracting information from the image (such as *Is there a dog in this image?*); common-sense questions rely on external knowledge contained in commonsense knowledge bases (*How many road vehicles are in this image?*); finally, KB-knowledge questions require information from Wikipedia or similar sources (*When was the home appliance in this image invented?*).

**Factual VQA (FVQA)** (Wang et al., 2018a) is a dataset addressing factual VQA, based on images sampled from COCO (Lin et al., 2014) and ImageNet (Deng et al., 2009) which form three types of visual content (object, scene and action classes), together with structured visual-related knowledge extracted from DBpedia (Auer et al., 2007), ConceptNet (Speer et al., 2017) and WebChild (Tandon et al., 2014). All this information is stored in a graph of RDF triplets. Annotators construct questions and answers which require both selected visual content and associated facts. In total, FVQA contains 2,190 images of 326 object classes and 221 scene classes, 5,826 questions of 32 categories, which correspond to 4,216 unique facts.

**Knowledge-aware VQA (KVQA)** (Shah et al., 2019) targets world knowledge-aware VQA by filling the gap of named entities knowledge. It relies on knowledge present in Wikidata (Vrandečić and Krötzsch, 2014) knowledge graph, resulting in 183K question-answer pairs which involve more than 18K named entities and 24K images.

**Outside-knowledge VQA (OK-VQA)** (Marino et al., 2019) contains more than 14K diverse and difficult questions of 10 mutually exclusive categories which cannot be answered without external knowledge. More than 14K images were sampled and filtered from COCO (Lin et al., 2014). In contrast with previous related works, OK-VQA does not consult a fixed knowledge graph to guide answer prediction, but dynamically recognizes what knowledge is needed, either structured or unstructured.

**Text-KVQA** (Singh et al., 2019) is a very large dataset addressing scene-text recognition for knowledge enabled VQA. It contains images from book covers (Iwana et al., 2017) and movie posters (mov), as well as Google scraped images of 1000 business brands. All images are evaluated so that they contain for sure scene text relevant to the content. Knowledge bases corresponding to each of those 3 scene types were constructed based on Wikidata (Vrandečić and Krötzsch, 2014) for business scenes, IMDb (imd) for movie posters and Iwana et al. (2017) for book covers. The train/validation/test splits enable zero-shot capabilities, as there is no entity overlap between them. The supporting facts are not tied with their corresponding entities, but instead are dynamically mined from the knowledge bases.

**Visual7W+KB** (Yu et al., 2020) is an extension of the Visual7W (Zhu et al., 2016) test split which further contains knowledge-based visual questions guided from ConceptNet (Speer et al., 2017). However, the dataset is not tied with a specific knowledge graph, even though ConceptNet is indeed preferred in practice. In total it consists of 16,850 open-domain question-answer pairs and 8,425 images from Visual Genome (Krishna et al., 2016). The questions belong to 7 categories (*what, where, when, who, why, which and how*), while the answers are provided in multiple-choice format.

One of the major challenges in knowledge-enhanced VQA is that questions should encourage exploitation of all participating modalities, therefore data-related weaknesses arise in existing benchmarks. In the meanwhile, information leakage between train and test set answers often promotes guessing rather than reasoning. **S3VQA** (Jain et al., 2021) is a dataset aiming to tackle those issues, by containing questions that can be answered only with the help of a knowledge graph, as well as visual and textual information from the image.

**Zero-shot Fact VQA (ZS-F-VQA)** (Chen et al., 2021c) extends F-VQA (Wang et al., 2018a) for zero-shot learning settings. It considers the image-question-answer triples whose answers belong among the 500 most frequent ones. The filtered dataset is split in train (seen) and test (unseen) triples which contain non-overlapping answers. In total, 5 splits of the original F-VQA dataset are performed, yielding on average 2,732 train and 2,760 test triples.

**Art QUestion Answering (AQUA)** (Garcia et al., 2020) is a visual reasoning dataset for the art domain. There are many challenges tied with analyzing and reasoning over artworks. First, there are different levels of abstraction regarding common objects and entities, as many paintings deviate from realism. Therefore, recognizing objects and reasoning about them is much harder compared to scenes existing in most datasets. Moreover, domain knowledge regarding artists, art movements, historical periods and other cultural influences can only be recognized with the help of a knowledge source. This
information also affects the interpretation of a painting. QA pairs are generated automatically based on paintings and descriptions of the SemArt (Garcia and Vogiatzis, 2018) dataset, which form the knowledge source. In total, AQUA contains more than 69K QA training pairs after cleansing, from which around 29K pairs are visual and 40K pairs are knowledge oriented.

6.1.2 Methods

6.1.2.1 Keyword-based explicit KG querying

First attempts target the construction of a scalable multimodal knowledge base which aims to answer visual queries that require real-world knowledge. Image classes, attributes and actions are extracted from the images, forming logical rules. The knowledge base built upon those rules contains nodes of visual and textual entities, as well as edges of diverse types between the entities. (Zhu et al., 2015) However, this constructed knowledge base remains limited to the visual information present specifically in the SUN (Xiao et al., 2010) dataset. Most subsequent methods utilize already constructed large knowledge bases, targeting a broader range of concepts, commonsense knowledge and more complex questions to be answered.

Towards this direction, early approaches focus on handling open ended questions regarding contents of a scene with the assistance of provided external knowledge. Attributes extracted from images using a fine-tuned VGG-16 (Simonyan and Zisserman, 2015) model act as SPARQL queries to knowledge bases such as DBPedia (Auer et al., 2007), and contribute to caption generation. Retrieved knowledge embedded via Doc2Vec (Le and Mikolov, 2014), together with attributes and LSTM-based caption representations are fed in another LSTM model which generates the final answer. (Wu et al., 2016b) An improvement of this version followed in (Wu et al., 2016a), extending the framework to two more datasets, namely DAQURA-ALL (Malinowski and Fritz, 2014) and its reduced version DAQUAR-REDUCED.

Even from early works in knowledge-enhanced VQA, explainability is addressed as an important topic to define how a model actually learns from visual content and external knowledge towards concluding to an answer. Therefore, Wang et al developed a knowledge-enhanced VQA framework that provides the reasoning path from which the answer is inferred. Objects are detected using Fast-RCNN (Girshick, 2015) object detectors trained on ImageNet (Deng et al., 2009) and MS-COCO (Lin et al., 2014), scene classes are extracted from a VGG-16 (Simonyan and Zisserman, 2015) pre-trained on MIT-Places (Zhou et al., 2014), and scene attributes are captured via a VGG-16 pre-trained on ImageNet and fine-tuned on MS-COCO. All those visual concepts form RDF triples and are linked with corresponding DBPedia (Auer et al., 2007) entities. Questions are parsed so that key-phrases are extracted and mapped to the knowledge base entities. The same work introduced the KB-VQA dataset. (Wang et al., 2017)

Consequent works further proceed towards avoiding SPARQL queries on the knowledge graph, but rather fully utilize embedding representations for fact selection and reasoning to provide an answer.

6.1.2.2 Sequential language models for question encoding

Instead of SPARQL querying from plain keyword extraction, vector representation of involved modalities set the basis for improved performance and state of the art results in knowledge-enhanced VQA. Initially, fact ranking based on embedding similarity metrics paved the path of successful approaches, upon which graph neural network reasoning further advanced the contribution of external knowledge and overall performance.

Embedding-based fact retrieval from KG

Traditional VQA models utilizing RNNs for language encoding, focus on learning a question-answer mapping. Due to the limited and opaque reasoning capabilities of this approach over diverse answers, a more scalable solution that involves learning the mapping between questions and KB-queries using LSTMs was proposed. This new approach is explainable, as the fact connecting a question and an answer reveals the reasoning procedure. (Wang et al., 2018a)

Projecting question-image pairs and facts on a common embedding space poses advantages over previous approaches, such as extendability to different knowledge bases and error elimination by avoiding explicit query generation. Images and questions are embedded using CNNs (for objects, scenes and actions) and an LSTM respectively, and they are projected in a common space using a Multi Layer Perceptron. Another LSTM is used to retrieve facts from the knowledge base, which are then encoded in GloVe (Pennington et al., 2014) vectors. The dot similarity between the question-image representation and fact embeddings provides a fact ranking, from which the final answer is inferred. (Narasimhan and Schwing, 2018)

Narasimhan et al, building upon (Narasimhan and Schwing, 2018), argue that considering multiple relevant facts instead of a single top-ranked fact at a time leads to better generalization. During the fact retrieval stage, a
subset of highly-relevant facts is obtained with the help of LSTM-based question embeddings. In the answer prediction stage, each node is represented by concatenating the selected entity representation from the previous stage, visual features from the image and the question embedding. The subgraph formed from all the relevant facts is jointly assessed by a graph convolutional network (GCN) (Kipf and Welling, 2016), followed by a multi-layer perceptron that decides if each entity constitutes the final answer or not. (Narasimhan et al., 2018)

Based on the KVQA dataset, a memory network (memNet) framework sets the baseline for VQA with knowledge of named entities. Specifically, entities extracted from the question and the image are used to obtain facts from Wikidata (Vrandečić and Krötzsch, 2014) knowledge graph. Retrieved facts together with corresponding entity coordinates from the image are used to produce memory embeddings via a BiLSTM network, and a similar procedure is followed for the question embeddings. Both representations contribute to the final answer, which is defined by a multi-layer perceptron. (Shah et al., 2019)

Text present on an image can provide further information towards inferring the correct answer. Extracted text and image areas are fused together with the given question to retrieve relevant facts during the fusion stage, and a multi-relational graph is constructed based on all those components. The text recognition part relies on word proposals assisted by the knowledge graphs accompanying the text-KVQA dataset. Scene proposals were created with the help of Places dataset for scene recognition (Zhou et al., 2018) and a fine-tuned VGG-16 (Simonyan and Zisserman, 2015). A gated graph neural network (GGNN) performs one-hop reasoning on this graph to derive the final answer. (Singh et al., 2019)

Multiple feature spaces Addressing the zero-shot setting of knowledge-enhanced VQA, a knowledge graph can help capturing semantics outside training data. Multiple feature spaces are used for independent alignment between image/question semantics and KG entities. The semantic space focuses on the linguistic information of the input (image, question) pair, representing a feature space of relationships; the object space acts as a support entity feature space, capturing visual and textual salient features; finally, knowledge space is dedicated to answer representation. (Chen et al., 2021c)

6.1.2.3 Transformer-based models
Transformer based approaches form end-to-end architectures that utilize single-modality or joint representations rather than creating queries to knowledge bases and then
injecting the retrieved entities. Thus, we can classify the transformer-based approach in two categories: the first includes architectures which use transformer architectures for text encoding, while the second utilizes multimodal transformers to jointly encode vision and language.

**Transformer architectures for language encoding**

Similarly to (Yu et al., 2020; Zhu et al., 2020), the usage of dedicated graphs for different modalities is also followed in (ZiaeeFard and Lécué, 2020), attempting to represent relationships between visual objects and semantic entities present in a scene graph and a knowledge graph respectively. The scene graph is constructed from visual and question embeddings which form the graph nodes and relationships. In the meanwhile, joint image and question embeddings select the most relevant knowledge graph node embeddings to construct the concept graph. Both image-question and knowledge representations are obtained via pre-trained language models for sentence similarity, such as sentence-BERT (SBERT) (Reimers and Gurevych, 2019) and Universal Sentence Encoder (USE) (Cer et al., 2018). The most relevant nodes of both scene and concept graphs are selected via a Graph Attention Network (GAT) (Veličković et al., 2018), which decides the edge weights with respect to the question. A joint embedding incorporates the question embeddings together with the scene graph and knowledge graph outputs.

Even though explicit and structured knowledge bases constitute the majority of the approaches analyzed so far, implicit and unstructured knowledge can also boost the VQA task. GPT-3 (Brown et al., 2020) can retrieve knowledge based on text prompts and effectively reason over it in a few-shot manner: no fine-tuning is required and instead only a few examples during inference are provided. Captions are first extracted from images using VinVL (Zhang et al., 2021a) to form GPT-3 inputs. Regarding sample selection for the few-shot inference stage, both improving the quality and increasing the number of samples have been explored. The top-n most similar prompt examples comparing to the inference-time question to be answered are defined by CLIP (Radford et al., 2021), thus maximizing sample relevance. On the other hand, multiple queries corresponding to one inference-time example can be used to retrieve answers from GPT-3 using n example prompts each time, and their ensembling results in the final answer. Contrary to most works on knowledge-enhanced VQA, inferring the answer is a generative task and not a discriminative one among pre-selected candidate answers or graph nodes. (Yang et al., 2021)

Unimodal pre-trained transformers yield better generalization capabilities over multimodal approaches of comparable size when external knowledge is necessary. Language models employed for the knowledge-enhanced VQA task are sufficient even to compensate for the limitations of image captioning models, which often fail to fully capture visual semantics. To this end, a pre-trained image captioning system, in this case the multi-task OSCAR (Li et al., 2020c) transformer, is used to extract linguistic information from an image, while a language model such as BERT, acting as an implicit knowledge source, receives the caption and the question to infer an answer. Moreover, text-only and multimodal approaches have complementary capabilities, therefore their combination can yield even more powerful models. (Salaberria et al., 2021)

VIKING is a framework accompanying AQUA dataset (Garcia et al., 2020) for visual QA on the artistic domain. As questions in AQUA may either be visual or knowledge oriented, a modality selector first decides the right category by receiving the encoded image and question. Visual oriented questions do not require external knowledge in order to be answered. For knowledge-oriented questions, a two stage fact retrieval strategy is followed, pairing the given question with the most relevant painting description, which corresponds to the external knowledge fact needed. The first stage utilizes TF-IDF to rank descriptions according to the question, and in the second stage re-ranking is performed using BERT. Finally, a fine-tuned XLNet (Yang et al., 2019b) model provides the final answer.

**Joint multimodal encoding with attention-based fusion**

A slightly different technique is followed in (Zheng et al., 2021), where the visual modality is not captioned, but fused with the BERT-embedded question. More specifically, a knowledge graph for artistic VQA is constructed based on YAGO (Tanon et al., 2020) knowledge graph and AQUA (Garcia et al., 2020) dataset. A Hierarchical-Knowledge Embedding module is responsible of retrieving relevant relationships $r$ from the knowledge graph that can form a $(h, r, t)$ triple with question related entities serving as the head $h$ of the triple, and answer related entities as the tail $t$. A Network-Based Representation Learning module extracts visual and textual features and fuses them together in order to obtain a VL representation. The fusion part first applies local attention per modality, and then global attention on both text and image, where locally ‘attended’ text features form the query, and visual features the key and value. Query, key and value are inserted in a multi-head attention unit which further promotes the joint representation to consequent layers, until a global representation is obtained. Then, a Knowledge-Based Representation Learning module injects hierarchical-knowledge embeddings to the network-
based representation. This representation is inserted into a relational module, which performs meta-training: a representation learned on the training data is transferred to a support set of disjoint labels. Finally, the relational module derives the answer.

**Joint multimodal encoding with VL transformers**

ConceptBERT is one of the first attempts towards the end-to-end transformer-based direction, where all modalities are jointly exploited for learning. The first step includes obtaining representations for each individual modality. Visual features are extracted using a pre-trained Faster R-CNN network (Ren et al., 2017) and BERT (Devlin et al., 2019) provides the question representation. ConceptNet (Speer et al., 2017) acts as the commonsense knowledge source, and is encoded using the ConceptNet embedding (Malaviya et al., 2019) method, a Graph Convolutional Network (Kipf and Welling, 2016) variant that relies on message passing from node to node in order to obtain the ConceptNet graph representation. Two modules receive the embedded inputs: A vision-language module consisting of two streams in a ViLBERT (Lu et al., 2019) fashion, and a concept-language module based on the bidirectional Transformer architecture are proposed to model the interactions between the relevant modalities. Both outputs of these modules are joined to form a concept-vision-language representation, which finally concludes to the answer via a classifier. (Gardères et al., 2020)

Knowledge obtained from the web according to the given question and respective answer can act as a large external implicit knowledge source for the OK-VQA dataset, covering knowledge ‘gaps’ in several domains without manual human effort. The proposed weakly-supervised framework consists of two phases: the first one (Retriever) retrieves relevant knowledge, which guides answer prediction in the second stage (Reader). Two different approaches are followed for representing the question-image pair inputs: either the question and image are encoded using an LXMERT (Tan and Bansal, 2019) transformer, resulting in a multimodal representation, or the image caption and the question are encoded via BERT (Devlin et al., 2019), leading to an exclusively linguistic representation. The BERT-based linguistic encoding can contribute to both a neural based retriever and a term based (Robertson and Zaragoza, 2009) retriever. A similarity score defines the relevant knowledge for the neural based retriever, which is further concatenated with the question representation, and consequently with the image using again an LXMERT model. (Luo et al., 2021)

KRISP framework addresses the scenario where essential external knowledge is absent during training, as well as test time. Both implicit and explicit knowledge sources are utilized: Explicit knowledge combines DBpedia (Auer et al., 2007), ConceptNet (Speer et al., 2017), Visual Genome (Krishna et al., 2016) and hasPart KB (Bhakthavatsalam et al., 2020) in a knowledge graph of 36,000 edges and 8,000 nodes after filtering out irrelevant concepts, while pre-training using BERT can offer implicit knowledge. Explicit visual symbols from images are extracted to constrain the knowledge graph entities corresponding to image-related concepts, including objects, parts of objects, attributes and places. Likewise, symbols are extracted from question to contribute to the formation of a graph for all explicit symbols. A Relational Graph Convolutional Network (RGCN) is used for graph representation, allowing dedicated processing for different edge types and directions. After reasoning, a symbolic prediction vector is returned. Regarding the implicit information stream, a multimodal BERT (MMBERT) model incorporates visual and textual embeddings to produce an implicit prediction vector. Finally, the top-ranked prediction from both vectors defines the answer. (Marino et al., 2021)

The presence of scene text can offer valuable information for properly predicting the correct answer. The detected text, among with the image, the relevant knowledge from Google Knowledge Base (GKB) and the question representation are fed into a multimodal transformer, enabling interaction through attention mechanisms between the different modalities. The OCR extracted text act as a query to GKB to retrieve candidate entities, which are then disambiguated based on the visual context. External knowledge not only boosts the understanding of scene-text even in unseen instances, but also tackle biases present in training data. (Dey et al., 2021)

While employing an abundance of knowledge sources can cover more visual topics, a lot of noise may be introduced, as more irrelevant information is retrieved. MAVEx utilizes multi-granular queries to retrieve external knowledge with the purpose of validating and correcting predicted answers among suitable candidates with the help of various knowledge sources. Specifically, a fine-tuned ViLBERT (Lu et al., 2019) model creates a pool of candidate answers, and together with the corresponding question, extracted keywords and phrases from the question and the possible answers are utilized to query external knowledge. Wikipedia, ConceptNet and Google images act as knowledge sources regarding different views of knowledge. Finally, retrieved knowledge instances are matched with the queries to acquire the highest ranked supporting fact, which returns the degree of agreement with respect to candidate answers, guiding decision to-
Wikipedia by aligning Wikipedia2Vec embeddings with
within the language encoder’s architecture, while the vi-
while BERT encodes the passage, and a dot similarity
wards the most trustworthy knowledge source. (Wu et al.,
2021a)

Passage retrieval can serve as an answer selection tech-
tique to VQA instead of choosing among pre-defined
candidate answers. Both sparse and dense retrieval are
investigated. For sparse retrieval, given a question and
an image, visual clues such as object names and captions
are extracted from the image and BM25 is used to re-
turn the k most relevant passages. For dense retrieval,
questions and images are jointly encoded in dense vectors
using LXMERT. In any case, retrieved external knowledge
can be integrated dynamically from diverse and generic
sources, without using a fixed knowledge base. As posi-
tive passages are considered the ones containing exactly
the ground truth answer. LXMERT (Tan and Bansal,
2019) is used to encode the question and the image jointly,
while BERT encodes the passage, and a dot similarity
between them defines the k most relevant passages. (Qu
et al., 2021)

Based on the E-BERT (Poerner et al., 2020) strategy
of knowledge injection without expensive re-training,
LXMERT (Tan and Bansal, 2019) language encoder in-
put is modified to incorporate factual knowledge from
Wikipedia by aligning Wikipedia2Vec embeddings with
BERT wordpiece vectors. No other change is required
within the language encoder’s architecture, while the vi-
ual encoder remains entirely intact. Only fine-tuning is
required for achieving advanced accuracy due to knowl-
edge injection. In the meanwhile explainability regarding
visual and textual modalities is also enhanced. For this
purpose, BM-GAE (Chefer et al., 2021) is employed to
extract visual and token explanations that help identify
in which parts knowledge injection was helpful. (Garcia-
Olano et al., 2021)

S3, presented with the S3VQA dataset, targets to an-
swer visual question based on all the participating modal-
ities simultaneously. Entity spans from the question are
selected to be matched with objects of scene graphs cor-
responding to images. This match can be often guided by
external knowledge sources, which enables answering
more complex questions that require multi-hop reasoning.
BERT identifies those appropriate question spans, while
object detectors propose the objects that most likely fill
the spans. Wordnet (Fellbaum, 1998) synsets are mapped
to the objects, and their hierarchical positions are repre-
sented via structural embedding methods. Finally, Google
search is used to retrieved the top results of the enriched
question representation. Alternatively, the answer can be
provided via classification of possible candidate answers.
(Jain et al., 2021)

6.1.3 Evaluation

Classification/Ranking metrics are widely used in K-
VQA, following the paradigm of knowledge-free VQA,
with most works rely on the top-1 accuracy metric for
comparison. Accuracy metric is further decomposed to
explain the contribution of subcomponents in many works:
object, counting, color, and location accuracies, as well as
accuracy per question type are reported (Wu et al., 2016b;
Wang et al., 2017). Other accuracy reportings include
performance as per selected knowledge source, per vi-
sual concept and per answer source (image or knowledge)
(Wang et al., 2018a); the individual accuracies of involved
stages of the reasoning process, which together contribute
to answer prediction (Narasimhan and Schwing, 2018;
Narasimhan et al., 2018); accuracies as per question cat-
egory (Shah et al., 2019; Yu et al., 2020; Ziaeefard and
Lécué, 2020; Gardères et al., 2020). Moreover, precision@k and recall@k have appeared in fewer works, as
well as ranking metrics such as MRR. Some early works
(Wu et al., 2016b; Wang et al., 2018a) utilize WUPS with
typically used thresholds of 0.0 and 0.9.

Human evaluation provides further insights compar-
ing to solid accuracy-based scores that fail to fully de-
scribe the success or the shortcomings of a metric. There-
fore, many authors propose human evaluation experiments
to grade the model’s response comparing to human per-
ception, and count the number of agreement instances
over all results (Wang et al., 2017, 2018a; Narasimhan
et al., 2018).

Evaluation of reasoning paths is followed by employ-
ing human judgement, as being one of the most trustwor-
thy indicators (Wang et al., 2017). Due to the transpar-
ent reasoning process, failure cases can be traced down,
revealing the exact stage where the prediction deviated
from the intended one. Thus, shortcomings can be at-
tributed to architectural stage, encoding techniques, or
even incorrect data annotations. Other metrics regarding
explainable reasoning include top-k fact retrieval accu-
rbacy for different k values, a crucial step for returning
the correct answer in several approaches (Wang et al., 2018a;
Narasimhan and Schwing, 2018; Narasimhan et al., 2018;
Zhu et al., 2020). Fact recall can also assess the frac-
tion of relevant facts retrieved for a given question (Singh
et al., 2019).

Generally, benchmarking knowledge-enhanced VQA
approaches is not trivial. The plethora of combinations
between available knowledge based datasets and external
knowledge sources is rather large comparing of the num-
ber of available implementations in the field. Additionally,
various choice of metrics in literature makes comparisons
of model performance even harder. Only recently imple-
ments started becoming more consistent, focusing on evaluating results with plain accuracy and leveraging OK-VQA as a widely used dataset.

6.2 Knowledge in Visual Reasoning (K-VR)

6.2.1 Datasets

High-order Visual Question Reasoning (HVQR) (Cao et al., 2019) is a knowledge-based dataset endorsing interpretable visual reasoning using commonsense knowledge. Given an image and a question, an answer is inferred, as well as a reasoning path as explanation. Even thought this is similar to rationales used in knowledge-free datasets for VCR (Zellers et al., 2019) in fact the format of HVQR explanations differ. Instead of textual rationales, rules for the whole reasoning path are returned, combining visual and knowledge oriented triples, derived from the scene graph and the commonsense knowledge graph respectively. HVQR contains questions that require multi-step reasoning to infer an answer. Moreover, each knowledge triplet appears only once per question, in order to avoid frequency-based biases. An evaluation scheme validates each step of the reasoning process based on the commonsense and scene graphs provided. More than 157K QA pairs comprise the dataset, from which 289,720 pairs are unique, together with approximately 32K images and corresponding scene graphs from Visual Genome (Krishna et al., 2016). Based on the reasoning steps required for the answer, first-order and second-order questions can be recognized, corresponding to 68,448 and 88,753 questions respectively. Another split defines 87K KB-related questions and 70K KB-not-related questions. Additionally, 193,449 facts from WebChild (Tandon et al., 2014), ConceptNet (Speer et al., 2017), and DBpedia (Auer et al., 2007) formulate the knowledge base. Scene graphs per image are combined with related entities from the knowledge base, constituting image-specific knowledge graphs.

Compositional Language and Elementary Visual Reasoning - CLEVR (Johnson et al., 2017) is a synthetic dataset of 3D objects which contain annotations regarding their position and attributes. Those attributes describe the size (small, large), color (red, brown, yellow, green, blue, cyan, purple, gray), shape (cube, cylinder, sphere) and material (rubber, metallic) of each object. Positions can belong in 4 types, namely left, right, behind, in front. Highly compositional questions form 5 question categories: Exist, Count, Compare Integer (equal, less, greater), Query Attribute (size, color, material, shape) and Compare Attribute (size, color, material, shape). Moreover, CLEVR contains 90 question families following different program templates, as well as text templates, so that natural language questions can be derived. The questions are translated in natural language by filling the template with template parameters. CLEVR is not connected to some external knowledge, although due to the limited semantics and the nature of the task they target.

CLEVR CoGenT is a benchmark derived from CLEVR (Johnson et al., 2017) that assesses the ability to capture unseen combinations of attributes during testing, thus showcasing a model’s generalization capabilities.

6.2.2 Methods

6.2.2.1 Sequential language models

External knowledge KM-net (knowledge routed modular network) is introduced in the same work with the HVQR dataset (Cao et al., 2019), addressing multi-step (compositional) reasoning using visual and commonsense knowledge. Each question is decomposed into consecutive subqueries via LSTM encoder-decoder schemes, passed to a visual reasoning module and a commonsense reasoning module to extract different types of knowledge accordingly. The subqueries form a query layout, i.e. a tree structure revealing the relationships of subqueries, with leaf nodes belonging to distinct words of the queries. A bottom-up attention R-CNN provides visual features for the image. The subqueries are processed sequentially starting from the most specific ones, driven by the KM-net reasoning module. First, the knowledge reasoning module receives question entities from the query layout and returns the most probable candidate entities from the knowledge base. Then, the visual reasoning module receives entities from the scene graph, together with the candidate entities of the knowledge module and fuses the candidate entities, image features and query embedding to derive the answer.

Internal knowledge Self-knowledge includes the usage or construction of a scene graph based on the detected objects, relationships and attributes. In the same time, the question can be parsed in a structured program via an LSTM, producing subqueries that form a tree structure. Given those two graph representations, an encoding is derived for the query. Node attention and edge attention are calculated based on the query embedding. Combining a node attention vector and an edge attention matrix, new objects can inferred due to the graph structure; basically starting from the attended node vector and traversing over an attended edge, a new node vector will be provided. The same procedure can be followed for all subqueries, respecting the structure of the query tree. Logically relating subqueries, results in logical operations (such as and, or, not) over attended scene graphs. Finally, based on the
question type and the final scene graph, the answer can be provided. (Shi et al., 2018)

6.2.3 Evaluation
Classification metrics are commonly used for benchmarking, with answer accuracy providing a general measure of performance (Cao et al., 2019; Shi et al., 2018). Compositional commonsense reasoning heavily relies on the evaluation of reasoning paths that provide the final answer (Cao et al., 2019). The accuracy score is further decomposed to present KB-related and KB-not-related accuracies depending on the need for external knowledge; those can be decomposed into first-order and second-order accuracies, regarding the number of reasoning steps required; finally, a more fine grained categorization provides question-type accuracy, based on the template the query components follow.

Ranking metrics such as average recall are used to evaluate the retrieval success of supporting facts for explanations. Average recall is further decomposed to KB-related and KB-not-related fact recall.

6.3 Knowledge in Visual Commonsense Reasoning (K-VCR)
Various rich in information external knowledge sources can provide insights of unseen concepts that humans would effortlessly infer from the information provided in a scene. This missing commonsense knowledge is able to guide answer explanation towards the right rationale, revealing if a more accurate reasoning process is followed by VCR models when knowledge is added.

6.3.1 Datasets
Visual Commonsense Reasoning (VCR) (Zellers et al., 2019) is the dataset which introduced the task and serves both knowledge-free and knowledge-enhanced versions of VCR. It contains 110k unique images from movie scenes, 290k multiple choice challenging questions, with 290k correct answers and rationales. Images contain annotations which are anchored over questions, answers and rationales. The technique of adversarial matching is chosen for the answers in order to minimize biases; each correct answer appears four times in the whole dataset, once as a positive answer and three times as negative answer. Therefore, a VCR model will not favor more frequently appearing answers which would endorse guessing rather than reasoning. The questions are classified in non-mutually exclusive categories according to their purpose, with categories being Explanations (Why is [person1] wearing sunglasses inside?), Activity (What are [person1] and [person2] doing?), Temporal (What will [person6] do after unpacking the groceries?), Mental (What is [person3] thinking while [person5] shakes his hand?), Role (What is [person1]’s relation to [person4]?), Scene (Where is [person1] now?), Hypothetical (What would happen if [person3] fell asleep?). This dataset originally is knowledge-free, therefore not necessarily requiring external knowledge, nor is it associated with any knowledge base. Nevertheless, the questions existing in the various VCR (Zellers et al., 2019) question categories can be greatly benefited by the introduction of external knowledge sources which can explicitly incorporate senses like the ones described in section 5.

Visual Commonsense Graphs (VCG) (Park et al., 2020) is a large scale dataset that provides information regarding temporal commonsense relationships, such as what may have happened before, what may happen in the near future and what are the intents of the people present based on static images. In total it contains more than 59K images and more than 139K textual descriptions of event at present. Additionally, around 295K intents at present, as well as more than 584k events before and 586k events after complete the dataset, resulting in more than 1,4 million commonsense inferences. People and locations appearing in the images are grounded with their mentions in the textual descriptions.

6.3.2 Methods
6.3.2.1 Transformer-based models
External knowledge Some of the first knowledge-enhanced transformer-based attempts stepped upon BERT (Devlin et al., 2019) framework to introduce knowledge-vision-language (KVL) learning as an instance of multimodal learning. In the KVL-BERT architecture (Song et al., 2021), ConceptNet (Speer et al., 2017) is leveraged to enrich sentences with relevant commonsense information. The knowledge-enriched linguistic input will be inserted in a BERT-like multimodal transformer. The preservation of semantic structure is achieved by using relative position embeddings. However, injected information should be only visible to their corresponding textual entities of the sentence and not other tokens or visual features, a need that is satisfied via a ‘weakening’ visible matrix. Moreover, it is possible that different enriched textual tokens in the sentence share the same relative position embeddings, which would make unrelated tokens obtain high self-attention scores, implying that they are related. This contradiction is resolved by imposing a mask-self-attention mechanism via the visible matrix, with the purpose to restrict the area a token can attend. After those treatments, the input is in a form suitable to be fed in a VL transformer, in this case VL-BERT (Su et al., 2020). It was observed that KVL-BERT outperforms its multi-
task baselines, as well as models dedicated to the VCR task, even though it cannot trespass the performance of knowledge-free VL transformers that invest on additional pre-training.

A somehow different strategy is employed in the case of Vision–Language–Knowledge Co-Embedding (ViLaKC) (Lee and Kim, 2020): the three modalities are first embedded independently and afterwards are fused together. Initially, a knowledge extraction module (KEM) retrieves relevant knowledge from ConceptNet based on concepts appearing on the image, question and candidate answers. The encoding of modalities is performed in the two-stage knowledge vector representation. The knowledge based reasoning module consists of the co-embedding submodule which aligns and integrates the three vectors via a multi-head self-attention mechanism. The co-embedder is pre-trained in two phases, the first being task-agnostic, such as in several VL transformer models, and the second task-specific, utilizing significantly less data (∼200K samples) coming from all three modalities. The task-specific pre-training stage introduces novel pre-training tasks, such as masked language modeling with image and knowledge (MLMIK), masked object classification with text and knowledge (MOCTK), and vision-language-knowledge matching (VLKM), in order to enforce co-learning. This joint embedding is then inserted in an answer determination module (ADM) consisting of a fully connected layer followed by a softmax.

The CKRM framework (Wen and Peng, 2021) consists of two stages, the first used for knowledge retrieval and the second one for reasoning. SWAG (Zellers et al., 2019), is a dataset containing pairs of situations which describes a situation (context) and possible endings, serves as the commonsense knowledge source, aiming to transfer knowledge regarding everyday situations to the target task of VCR. A source and a task encoder are responsible of receiving (context, ending) pairs and (question, answer) pairs respectively to perform knowledge transfer in different granularity layers. The encoders first use BERT (Devlin et al., 2019) followed by a BiLSTM structure to model temporal interactions of words. Cell-level knowledge transfer refers to the most fine-grained information fusion from source to target task, with layer-level and attention-level knowledge corresponding to coarser aspects of information. This strategy offers acquiring knowledge from various perspectives for a more enriched representation. The knowledge based reasoning module incorporates the multi-level knowledge from the previous stage together with visual features in the Knowledge-enriched visual attention module. Finally, a reasoning composition module combines all aspects of knowledge derived from the multi-level transfer procedure and the enriched visual representations to derive the answer.

### 6.3.3 Evaluation

**Classification metrics**, especially **classification accuracy** is employed for evaluating K-VCR results that follow the multiple-choice format for answers (A) and rationales (R). Accuracy is decomposed by evaluating independently each of the following aspects:

1. $Q \rightarrow A$: given a question $Q$, choose as A one of the 4 available answers and compare if it matches the real answer or not.
2. $QA \rightarrow R$: given a question $Q$ and the correct answer A, select as R one out of the 4 rationales and compare if it matches the real rationale or not.
3. $Q \rightarrow AR$: given a question $Q$ select as A one of the 4 answers, and depending on selected A choose one of the 4 rationales. The result is regarded to be correct if and only if both right A and R are chosen.

### 6.4 Knowledge in Image Captioning (K-IC)

#### 6.4.1 Datasets

There are no dedicated knowledge-enhanced or knowledge demanding datasets for K-IC. Knowledge-enhanced models are using **COCO captions** (Lin et al., 2014) as described in section 3.4.3. Moreover, **Flickr30k** (Young et al., 2014), a dataset containing 31,783 scene images accompanied by 5 human annotated sentences each is widely employed for K-IC.

#### 6.4.2 Methods

**External knowledge** First attempts for knowledge-enhanced image captioning propose the extension of existing implementations by injecting commonsense knowledge from external sources. Specifically, the backbone image captioning architecture extracts visual features from images via a CNN, which are then inserted in an LSTM to images a knowledge-free answer. To enhance this baseline with knowledge, objects extracted from the image are used as queries to ConceptNet (Speer et al., 2017). Related ConceptNet entities, either regarding individual objects (direct terms) or the remaining image areas (indirect terms), are fed to a pre-trained LSTM which provides semantic representations for each of those two points of view. Then, visual features, direct terms representations
and indirect terms representations are concatenated to form the initial state of another LSTM model, which finally generates the knowledge-enhanced caption (Zhou et al., 2019c).

Both visual and commonsense knowledge for image captioning are used in (Hou et al., 2019). The first step includes dense region sampling from images in order to acquire visual and knowledge mappings. Dense visual feature extraction includes the definition of candidate regions, which are clustered together to provide a more concrete representation: the cluster center points for each dense region cluster serve as the corresponding visual feature. Consequently, the knowledge mapping receives visual features and knowledge embedding vectors from Visual Genome (Krishna et al., 2016) and returns a knowledge-related representation per region cluster. Both visual and knowledge embeddings resulting from the two mapping procedures are concatenated and then inserted in a commonsense reasoning module. This module projects the two inputs in the same semantic space, from which a semantic graph is constructed under the guidance of commonsense knowledge. In the relational reasoning module a GCN (Kipf and Welling, 2016) operates on the semantic graph to obtain relation-aware node features. Finally, a LSTM receiving as inputs the knowledge-aware node embeddings generates the caption.

Inferring words not appearing in the image remains a challenge in image captioning, as there is no guidance regarding how those unseen words should be inferred to be used in captions. Such unmatched elements can be solved with internal self-knowledge based on more fine-grained alignments between individual words and image regions, which is achieved by attention mechanisms, and external commonsense knowledge to capture implicit information that cannot be derived from the existing data. Objects detected on the image are used to retrieve knowledge from ConceptNet (Speer et al., 2017). Region features extracted from the image via a region proposal network and word-level attention on the sentence part co-operate towards attending to the most salient features of the image. This visual attention guided by language attention, together with the corresponding word embedding are inserted in a LSTM, which feedbacks each previous hidden state to update the word-level attention signal that contributes to the visual attention in every round. The external knowledge is incorporated in a later stage, when the answer is generated; therefore, it can tune the probabilities of LSTM-generated words to be added in the sentence towards more meaningful results. A reinforcement learning training strategy is followed by setting the LSTM as an agent, the words and visual features as the environment, and the generation of the best next word from the captioning model as the policy. (Huang et al., 2020)

Even though local information is well-represented based on detected objects, image captioning is generally not interpretable and therefore not explicitly controllable. An external knowledge source can help in grounding detected objects with semantic entities from the graph, which in turn provides enriched semantic labels for the objects present in the image. In order to control objects appearing in the caption, an attention-based human-interpretable mask is introduced, which assists in diverse caption generation. This masked can be dynamically tuned by a human to influence the resulting caption. (Aditya Mogadala, 2020)

Off-the-self object detectors have served several image captioning architectures. However, some tough situations such as very small objects, occlusion or rare object classes can result in error propagation and negatively impact all consequent components until the final caption generation. Commonsense constrains and semantic correlations extracted from Visual Genome (Krishna et al., 2016) can act as priors to guide a more accurate representation. A semantic graph is constructed upon extracted image regions, allowing GCN-based (Kipf and Welling, 2016) reasoning which. Specifically, visual semantics such as objects, attributes, relationships are captured by extracting candidate region proposals. CNN-based region features can satisfy object and attribute representations, while features from regions union areas provide relationship representation. Visual features are projected in the same high-level semantic space as Knowledge embedding derived from Visual Genome. Therefore, knowledge-enhanced visual triplets are formed, respecting rules imposed by knowledge. The semantic graph is built upon those triples. Then, relational reasoning is performed on the semantic graph using a GCN, the output of which is inserted in the LSTM module that generates the answer. (Hou et al., 2020)

**Internal knowledge** Visiolinguistic priors are naturally connected with describing images, in the sense that humans logically infer unseen entities given a partial description of a visual situation. Obtaining such priors from existing images and captions is a way of ‘creating’ knowledge and facilitate reasoning of image caption models without adding external sources.

Scene graph generation is a widely used technique for self-augmentation of information present in the dataset. Both images and text need to be represented in graph structures to bridge the two modalities. The Scene Graph Auto-Encoder (SGAE) framework utilizes this graph conversion to instill language priors into the encoder-decoder
image captioning structure. More specifically, a learnable
dictionary maps the relationships between a sentence and
its corresponding scene graph iteratively, reconstructing
the initial text from the generated graph in each round.
For scene graph generation from text, a pre-trained scene
graph parser is utilized, while for the reverse procedure,
a trainable RNN decoder converts the dictionary back
to text. During this procedure, the dictionary achieves to
capture the necessary language prior to be transferred for
captioning. The learned dictionary can then be inserted in
the image-involving pipeline: a scene graph parser con-
verts the image to a scene graph, which is then passed
to the dictionary encoded by a GCN (Kipf and Welling,
2016). Finally, the decoding of the dictionary provides
the final caption. (Yang et al., 2018)

Attention mechanisms are able to identify such struc-
tured visiolinguistic priors and highlight connections be-
tween text and images, therefore augmenting image cap-
tioning implementations. Conditional Latent Topic Atten-
tion (CLTA) in combination with sentence prior are able to
fuse the model with prior knowledge without the need of
constructing scene graphs. Latent topic models are able to
recognize semantically significant topics which are
 driving attention mechanisms to capture local and global
dependencies in images. Thus, salient visual features
emerge through words, and also more candidate salient
regions are discovered and re-weighted accordingly, if
they are associated with a topic contributing to an existing
salient region. CLTA implements this re-weighting proce-
dure to construct a context vector. Moreover, a sentence
autoencoder acting as the sentence prior encourages the
extraction of more context information and enhances gen-
eralization. Both the context vector and the sentence prior
are inserted in a LSTM that generates the answer. (Goel
et al., 2020)

6.4.2.2 Transformer-based models

Recent knowledge-enhanced image captioning models
are implemented based on Transformers as an expected
substitution of sequential models.

External knowledge Named entities and event knowl-
edge has not been studied in previous image captioning
works. This type of information is widely available in
news articles, with raw sources being too complicated
for language models to infer the right semantics. Special
datasets are crafted for this purpose, providing an appro-
riate form of information for named entity/event-aware
image captioning. The heart of the proposed method is
the cross-modal entity matching, which incorporates in-
formation from various sources. Sub-graphs are extracted
from the image and the article text descriptions forming
structure representations for the input. The nodes of the
text sub-graph belong to named entities, and the edges to
their in-between relationships, while the image sub-graph
is more generic, by representing objects present on the
image. The two sub-graphs are linked via similarity be-
tween image sub-graph objects and text sub-graph named
entities in the cross-modal entity matching module. This
module is trained with the help of multimodal external
knowledge from Wikipedia. As a result, a multimodal
knowledge graph is produced containing visual, textual
and knowledge information. Embedding representations
are obtained for each modality: a GAT (Veličković et al.,
2018) produces a multimodal knowledge graph embed-
ding, RoBERTa (Liu et al., 2019b) encodes news captions
and image features are derived from a pre-trained ResNet-
152 (He et al., 2016). An entity-aware captioning model
receives the visual, textual and multimodal knowledge
graph representations, feeding them to a Transformer
(Vaswani et al., 2017) decoder to produce the caption.
(Zhao et al., 2021)

BART transformer (Lewis et al., 2019) can provide
further advancements towards the refined task of Visual
Commonsense Generation (VCG) (Park et al., 2020) ly-
ing on the intersection of the generative image caption-
ting task and the non-generative visual commonsense rea-
soning task. To this end, knowledge-enhanced Multi-
omodal BART (KM-BART) was developed, able to incor-
porate both visual and linguistic information with the
help of modality and task-relevant tokens in the trans-
former input. More specifically, task-relevant tokens are
added in the beginning of the input sequence denoting
the task type. For example, for VCG <before>, <after>,
or <intent> tokens, representing temporal sequence of
events (what happened before, what may happen next)
and intents of people present in the image. Furthermore,
the pre-training task of Knowledge-based Commonsense
Generation (KCG) fuses commonsense knowledge from
structured sources early in the pipeline, actually achiev-
ing in implicitly integrating explicit knowledge. Also
Attribution Prediction (AP) and Relation Prediction (RP)
pre-training tasks are used for the first time in knowledge-
enhanced VL learning. COMET (Bosselut et al., 2019)
is a transformer model trained on knowledge bases such as
ATOMIC (Hwang et al., 2021) and ConceptNet (Speer
et al., 2017) that generates commonsense descriptions,
and acts as a knowledge source for KM-BART. Two possi-
bile settings are examined for KM-BART, one containing
the image and the event description (i.e. some textual
information about the image that provides context of the
Internal knowledge Transformer based captioning poses some challenges, one of those attributed to the AR training procedure which is based on the maximum likelihood estimation (MLE). The main issue stemming from MLE is that when the generated sequence does not match the ground truth one, there is no discrimination between different ‘failed’ predictions. Therefore, words that are totally unrelated to the ground truth match are treated the same as semantically similar words. For this reason, a KL divergence term is added to weight semantic relationships between generated words, with respect to their ground truth match. Moreover, a knowledge graph is used to enrich the transformer input embeddings infusing contextual information from neighboring entities in the graph. This knowledge graph is constructed from the linguistic information itself, by leveraging cosine similarity between embedded words to position them within a vector space. The original Transformer (Vaswani et al., 2017) architecture is leveraged for the task, with image features word embeddings representing the visual modality. (Zhang et al., 2021b)

6.4.3 Evaluation
Language metrics such as BLEU, ROUGE, METEOR, CIDEr are used for evaluation as in most language generation tasks. (Zhou et al., 2019c; Hou et al., 2019; Goel et al., 2020; Huang et al., 2020a) SPICE is also used in (Aditya Mogadala, 2020; Zhang et al., 2021b).

Human Evaluation can qualitatively evaluate generated sentences. The human evaluation experiment in (Yang et al., 2018) compares the quality of generated captions from different models according to the perception of 30 evaluators. Even though such an experiment is rather subjective, it indicates the importance of language priors. In (Zhao et al., 2021) human preference is measured comparing with the previous best-performer in the before, after, intent generated sentences.

6.5 Knowledge in Visual Dialog (K-VD)
6.5.1 Datasets
VisDial (Das et al., 2016) is a dataset used in both knowledge-free and knowledge-enhanced versions of VD. It consists of 133k dialogs and an equal number of images from COCO, with train and validation splits (125k dialogs) assigning 10-round dialogs -QA pairs- per image. In the test split (8k dialogs) random rounds are paired with each image. Some important aspects of this dataset is the presence of coreferences, endorsing the coherence of the conversation in linguistic level, and temporal continuity in topics, which supports the preservation and consistency of semantic meaning across the dialogs. The questions mostly follow a concrete and rather exploratory pattern: starting from asking about entities involved in COCO captions, then diving into details, trying to define a categorization of the whole scene or the most appropriate setting description, questioning about the weather of the scene, exploring key semantics not mentioned previously and finally validating and expanding the understanding of elements provided in the answers.

VisDialCK (Zhang et al., 2022) is an extension of VisDial containing 940 history-required and commonsense-required dialogs.

6.5.2 Methods
6.5.2.1 Transformer-based models
Internal knowledge Very recently a knowledge-enhanced implementation for visual dialog was introduced, inspired from the fact that commonsense related questions are ignored. A visual dialog model requires two necessary inputs: an image and dialog history. Visual graphs have assisted the task by providing object relationships explicitly, even though this knowledge is not adequate for commonsense inferences. The integration of commonsense knowledge can be well-represented with graph-level facts and sentence-level facts. Then, facts from a commonsense knowledge graph such as ConceptNet (Speer et al., 2017) can be extracted based on the calculation of cosine similarity between their word embedding representation comparing to embedding representations of the words in the sentences and the detected objects. Those graph level facts can complement entities from the visual graph. Therefore, an enriched vision-fact graph can be produced after individual graphs are purified by removing redundant information. The sentence-level facts are derived from the dialog sentences in the form of (subject, relation, object) triples, forming a graph structure. Similarly to the visual stream, the sentence graph is cleaned and enriched with commonsense knowledge. Finally, a transformer-based fusion module receives the enriched graphs, as well as the question embedding to provide the answer, exploiting a generative and a discriminative decoder. (Zhang et al., 2022)

6.5.3 Evaluation
Ranking metrics such as NDCG, MRR, R@1, R@5, R@10, Mean position provide the quality of answer retrieval for visual dialog for both generative and discriminative answer prediction. (Zhang et al., 2022)

Human evaluation is used in the generative setting of
(Zhang et al., 2022). Specifically, two metrics are provided: the first one indicates the percentage of responses passing the Turing test, thus providing the amount of generated sentences that could be perceived as human-written; the second metric measures the number of generated responses that were perceived as of equal or better quality compared to specific human responses as baselines.

6.6 Knowledge in Visual Storytelling (K-VIST)

Visual Storytelling presents many situations where hypothetical concepts can be driven from commonsense and temporal reasoning. Unseen events can enrich or even be necessary for appropriate and coherent textual stories. For example, some sequential inferences were presented in the event/temporal knowledge analysis 5, providing knowledge such as the boy dropped a glass of water and then the glass broke. Not all concepts mentioned in this sentence may be explicitly apparent on a frame of the visual sequence. However, a knowledge graph can guide inference by searching for possible connections between concepts appearing on images, and thus acquire imaginary concepts.

6.6.1 Datasets

There are no dedicated datasets for K-VIST. On the contrary, relevant literature relies on datasets used for the knowledge-free version of the task, such as VIST (Huang et al., 2016a). This dataset contains more than 81K unique photos in around 20K sequences with corresponding textual stories. Textual stories are following a narrative style imposing more high level inference capabilities comparing to literal visual descriptions. This requirement is an extension against the majority of visual description tasks, which do not directly focus on sequential coherence and even abstract meanings. Two extra descriptions are provided per frame in order to bridge literal description with narratives: descriptions of images-in-isolation (DII) and images-in-sequence (DIS).

6.6.2 Methods

6.6.2.1 Sequential language models

External knowledge A two-stage structure was proposed in (Yang et al., 2019a), consisting of a reasoning and a generation module. The vision aware commonsense reasoning module is responsible of extracting the most relevant knowledge from an external knowledge base. Objects detected on all images of a sequence are fed in a GRU which provides a semantic and temporal representation. In the same time, candidate ConceptNet entities are fetched based on the detected objects. Attention modules finally select the most relevant ConceptNet candidates, which after passing through a GRU provides the final commonsense representation. The knowledge augmented generation module receives the extracted commonsense knowledge together with the visual information, as well as the previously generated sentences.

A prevalent issue in VIST is the monotonous and repetitive generated stories. This can be attributed to the limited vocabulary of the VIST dataset. In KG-Story (Hsu et al., 2019) the first stage (distill) gathers words from images using object detection and GRUs for word prediction. Potential relationships between pairs of concepts throughout images are searched on external knowledge graphs, and if multiple candidates occur, a scoring function is used to rank their relevancy. This is the enrich stage. Finally, the generate stage utilizes a Transformer which imposes a repetition penalty to mitigate redundant narration. Further modifications in the default Transformer structure is the introduction of an anaphoric expressions generator to enhance coreferences and usage of pronouns, as well as positional encodings of variable length to enable representing stories of different lengths.

Addressing again the coherence and novelty of generated stories, authors of (Xu et al., 2021) propose a three-stage structure corresponding to imagination, reasoning and writing capabilities of human. The first stage (imagine) focuses on the sequential consistence by extracting the visual topic of a frame through the combination of the current visual features and the sentence generated in the previous step. The knowledge part targets the content of narratives and consists of three graph types: a general commonsense knowledge graph, a scene graph and an event graph. A GCN applied on each graph selects the most suitable knowledge parts, which are combine to form the second stage (reason). Both imagine and reason outputs are fed to the third stage (write), which is responsible for generating the story.

6.6.2.2 Transformer-based models

External knowledge Towards informative and more diverse stories, (Chen et al., 2021a) is the first knowledge-enhanced approach that utilizes a generative transformer to produce the story output. The concept enrichment stage connects concepts present in images with ConceptNet. Then, a graph attention network (GAT) operates on the graph and image features in order to integrate information of the most appropriate candidate concept nodes, which will be passed in the next selection module. The concept selection module utilizes two different selection methods: a Sequential Selection Module (SSM) that operates in an encoder-decoder fashion, outputting selected concepts
after encoding the embedded candidate concepts; a Maximal Clique Selection Module (MCSM) outputs a maximal clique containing all appropriate concepts for story generation given the concept graph. Finally, the concept to story module uses either an RNN structure or a BART language model, with BART demonstrating more diverse stories while preserving quality.

6.6.3 Evaluation

Language generation metrics BLUE, ROUGE, METEOR and CIDEr are widely used automatic metrics that evaluate the linguistic quality of generated stories.

Distinct-n (Dist-n) is measured via the Diversity of generated stories is measured via the Dist-n (Dist-n) score (Li et al., 2015). This metric indicates the originality of generated text by calculating the frequency of n-grams throughout the whole corpus of generated stories. Higher Dist-n scores represent more diverse stories. (Yang et al., 2019a; Chen et al., 2021a)

Human Evaluation is very important for generative tasks, as automatic evaluation metrics cannot assess the full range of linguistic capabilities, especially when it comes to evaluating sequential quality. However, different implementations perform varying human evaluation experiments, which somehow impedes the direct comparison of models.

In (Yang et al., 2019a) four aspects were examined: Fluency checks the linguistic quality, relevance measures the success of textual description in describing visual concepts, informativeness measures the diversity of produced stories and coherence evaluates the semantic continuity of stories in a sequence. Each aspect receives a score from 1 (worse) to 5 (best) from three evaluators, and their average values serve as the final results. Similarly, in (Xu et al., 2021) relevance, coherence and informativeness are regarded, receiving scores from 0 (worse) to 2 (best) from five evaluators. A different human evaluation strategy is followed in (Hsu et al., 2019): comparative experiments between VIST models are performed, asking users to rank generated stories from different models either with or without revealing the corresponding images. This is an indirect evaluation of linguistic quality and coherence, when only text is regarded, and also semantic relevance, when corresponding images are provided. The comparative approach is also used in (Chen et al., 2021a), with two evaluators declaring their preference (or tie) between two models regarding three aspects: relevance and informativeness similar to (Yang et al., 2019a; Xu et al., 2021), together with logicality which measures the logical coherence over story sequences. Additionally, overall indicates the evaluator’s preference in general between the two models.

6.7 Knowledge in Image Generation

6.7.1 Datasets

A variety of datasets have been used in visual generation, which however do not contain some certain sense of knowledge, and are widely used in knowledge-free settings. Datasets used in conditional image synthesis are ImageNet (Deng et al., 2009), CIFAR (Krizhevsky, 2009), FFHQ (Karras et al., 2018), Oxford Flowers (Nilsback and Zisserman, 2008), CUB (He and Peng, 2020) and many others.

Sequential synthesis (story visualization) greatly utilizes Pororo-SV cartoon dataset (Kim et al., 2017). It contains more than 16k pairs of scenes and dialogs extracted from 20 hours of video, 27,328 fine-grained scene descriptions in natural language provided by human annotators, and 8,913 QA multiple-choice pairs related to the story. In total, 10 main characters appear in the frames. Questions are divided in 11 types: Action, Person, Abstract, Detail, Method, Reason, Location, Statement, Causality, Yes/No, Time. FlinstonesSV (fli) is also based on cartoon frames. It is composed of 25184 densely annotated videos, each of which containing 75 frames. The annotations include bounding boxes with labels for characters and items of the frames, as well as segmentation masks. Another emerging dataset for Story Visualization is DiDeMo-SV (Maharana et al., 2022; Hendricks et al., 2017), a dataset based on video captions that contains 10,000 with more than 40,000 temporally localized textual descriptions.

6.7.2 Methods

6.7.2.1 Knowledge in Conditional Image Generation (K-cIG)

Internal knowledge Even though GANs have been powerful in synthesizing novel images, they cannot handle combinations of attributes they have not encountered in the training data. Therefore, if the textual condition refers to such unseen combinations, the synthesized image has sacrificed some of the semantics, in order to produce a result that remains within the learned distribution. The insertion of additional knowledge can expand the generated distribution to enhance the consistency on the condition without sacrificing resulting fidelity. This can be translated in two needs regarding a GAN model: the generator should become more flexible, and the generator more tolerant. KG-GAN meets those requirements by introducing a second generator, trained on domain knowledge by utilizing a novel knowledge loss. This second generator shares parameters with the original one, which is responsible of synthesizing images conditioned on text. A regression network receives the synthesized
images from the seen-image generator and the ones from the knowledge generator, imposing constrains regarding the plausibility of unseen combinations. The semantic vector produced by the knowledge generator is redirected to the seen-image generator to guide generation outside the predefined classes. KG-GAN does not exploit external knowledge sources, but with this simple distribution enhancement it achieves some preliminary zero-shot capabilities. (Chang et al., 2019)

6.7.2.2 Knowledge in Story Visualization (K-SV)

External knowledge Story Visualization is another task with limited contributions in knowledge-enhanced settings. Structured information from text can be obtained via parse trees which can permit hierarchical encoding of longer phrases. Missing information regarding visual details in text can be filled out with external knowledge from ConceptNet (Speer et al., 2017). Moreover, conceptually similar sentences that are phrased in a different way need to be placed closed in an embedding space, an issue that external knowledge can again effectively resolve. Spatial knowledge is also underrepresented in most sentences, even though scene synthesis needs detailed information of object positions. Dense captioning as a form of self-augmenting knowledge provided detailed positioning information due to the usage of region bounding boxes. The combination of internal spatial and external semantic knowledge is able to better guide sequential synthesis, resolving all involved aspects such as text-image consistency, visual quality and sequential continuity. A Memory-Augmented Recurrent Tree-Transformer (MARTT) encodes the parse trees for the text, while a Graph Transformer (Yun et al., 2020) embeds the commonsense knowledge. Both embeddings are inserted in the story encoder, which outputs contextualized embeddings for the image generator. The generated images are passed to image and story discriminators, which redirect the model to continue the visual story in a consistent way. Story Visualization has been a task lacking sufficient datasets, due to the increased effort needed to construct appropriate ones, either manually or automatically. To this end, external unstructured knowledge obtained from pre-trained DALL-E (Ramesh et al., 2021) enables even zero-shot sequential synthesis based on input 'story' text.

6.7.3 Metrics

Image generation metrics (section 3.6.5) such as FID for seen and unseen classes were used in KG-GAN (Chang et al., 2019). FID is also used to evaluate quality of generated frames independently in (Maharana and Bansal, 2021; Maharana et al., 2022). R-precision indicates quality by measuring the retrieval capabilities of generated frames over ground truth captions comparing to retrieval using the real frames. (Maharana and Bansal, 2021)

Classification metrics, such as Character F1 score measure the quality of generated characters in predicted images. Also, frame accuracy checks the exact match between semantics of the ground truth and generated frames. (Maharana and Bansal, 2021; Maharana et al., 2022)

Language metrics are also relevant: viewing SV frames as a video, captions for generated frames can be produced using video captioning techniques. BLEU scores evaluate the quality of captions as an indirect measure of visual quality, based on the idea that well designed semantics will be captured in captions better than low quality concepts. (Maharana and Bansal, 2021)

Human Evaluation can reveal the human perception over quality, as in most generative tasks. Specifically for SV, evaluators need to assess results over visual quality, consistence and relevance comparing to the previous state-of-the-art model on the same task. (Maharana and Bansal, 2021; Maharana et al., 2022)

6.8 Multi-task transformers with knowledge

6.8.1 Methods

Multi-task models can easily be built using multimodal transformer backbones. Instead of utilizing external knowledge graphs as in previous methods, many implementations employ self-knowledge exclusively, by obtaining more structured representations from the existing visual and textual data.

External knowledge A natural unification of multiple tasks under the same model would incorporate tasks moving in the same direction such as cross-modal reasoning tasks or cross-modal-retrieval tasks. Indeed, VQA, VCR and VE were unified in Rationale VT transformer
(Marasović et al., 2020), a framework that utilizes visual and linguistic clues to generate free-text rationales. Two knowledge sources attempt to provide reasoning information regarding scenes: a grounded situation recognizer (Pratt et al., 2020) describes activities on scenes, entities involved and draws bounding boxes for entities to visually ground them; and Visual Commonsense Graphs (Park et al., 2020) to fuse commonsense inferences about events and intents so that a temporal perspective of a scene is also considered. Rationales are generated for VQA-E (visual question answering) (Li et al., 2018), E-SNLI-VE (visual entailment) (Do et al., 2020) and VCR (visual commonsense reasoning) (Zellers et al., 2019) datasets. Visual recognition of objects is the first step for visual understanding, followed by capturing their in-between relationships utilizing the knowledge provided by the grounded situation recognizer (Pratt et al., 2020). Higher-level cognition is achieved using knowledge from VisualCOMET (Park et al., 2020), which receives the knowledge stored in Visual Commonsense Graphs to generate commonsense inferences. VisualCOMET is built upon GPT-2 (Radford et al., 2019), therefore a unimodal purely linguistic input can be provided, utilizing object labels, textual question/answers and inferences. Alternatively, GPT-2 can be adapted, resulting in a hybrid implementation: visual features and bounding box coordinates act as visual embeddings, combined with VisualCOMET token embeddings indicating the beginning of before, after, intent inferences.

Targeting again reasoning tasks, (Shevchenko et al., 2021) builds on top of LXMERT (Cho et al., 2020) to address the knowledge-enhanced versions of the VQA, VCR and VE tasks on the OK-VQA (Marino et al., 2019), FVQA (Wang et al., 2018a), NLVR2 (Suhr et al., 2019), SNLI-VE (Xie et al., 2019) datasets. External knowledge is provided from ConceptNet (Speer et al., 2017) and Wikidata (Vrandečić and Krötzsch, 2014). Knowledge-rich expressions are created by matching embedded knowledge with training sentences from the datasets. Moreover, a training objective targeting the alignment of knowledge embeddings and knowledge-rich expressions encourages learning a global representation structure. Utilizing this objective is proven beneficial during both pre-training and fine-tuning. It is also observed that the introduction of this knowledge-oriented objective smooths the embedding space, which facilitates similarity matching between words.

KB-VLP (Chen et al., 2021b) utilizes knowledge embeddings based on Wikidata (Vrandečić and Krötzsch, 2014) entities, which are concatenated with the visiolinguistic instances as inputs of a VL transformer. Specifically, entity recognition on text is performed to extract relevant Wikidata entries, which are embedded via Wikipedia2vec to form text-related knowledge embeddings. Object tags obtained from the image are used to obtain image-related knowledge embeddings from relevant Wikidata entities. The input vector consists of 5 components: word embeddings for text, text-related knowledge embeddings, word embeddings sequences for object tags per image, visual features and image-related knowledge embeddings. Two specialized pre-training objectives are used: sentence-level objective substitutes elements from the input vector with other random elements, while token-level objective extends text-image masking to text-related knowledge embedding - image-related knowledge embedding masking. Task specific datasets for KB-VLP are VQA (Agrawal et al., 2016), GQA (Hudson and Manning, 2019) and OK-VQA (Marino et al., 2019) for visual question answering, and NLVR2 (Suhr et al., 2019) for visual reasoning.

Internal knowledge OSCAR (Li et al., 2020c) is one of the models that effortlessly transit from knowledge-free to knowledge-enhanced learning utilizing self-acquired knowledge in its simplest form. Instead of -rather naively-letting the model infer the correct image-text alignments in an exhaustive way, OSCAR facilitates the procedure with the usage of object tags, as intermediaries between text and image instances. This procedure is endorsed from the observation that salient objects in the image will most probably also appear in text. The input to the VL transformer module consists of word tokens, object tag embeddings and visual features. The intermediary object tags form separate semantic spaces, depending on whether they are paired with text or image, yielding two dedicated pre-training objectives. The masked token loss objective views text and tag word representations in the same space, randomly masking each of them and letting reconstruct the missing parts through the visual modality. Conversely, contrastive loss views tags paired with visual features, and randomly replaces the real tag sequence with another one sampled from the dataset, learning to pull apart mismatched tag sequences and bring close together the matching ones. OSCAR succeeds in both understanding tasks, such as cross-modal retrieval (ITR/TIR), visual question answering (on VQA (Agrawal et al., 2016) and GQA (Hudson and Manning, 2019)) and visual reasoning (on NLVR2 (Suhr et al., 2019)), as well as in generation tasks, such as image captioning and novel object captioning.

ERNIE-ViL (Yu et al., 2021a) leverages structured visual knowledge from scene graphs to bridge detailed semantics across vision and language. Such fine-grained representations are important to differentiate between con-
ceptually similar scenes. Scene graph prediction tasks (object, attribute and relationships prediction) encourage learning those fine-grained differences. Even though not using external knowledge, ERNIE-ViL internally constructs structured knowledge during the cross-modal pre-training. Nevertheless, this self-knowledge is sufficient to boost performance in 5 VL tasks, especially in those that fine-grained associations are required, such as visual referring expressions (VRE). Other tasks benefited from this approach are VCR, VQA and cross-modal retrieval (ITR/ITR).

ROSITA (Cui et al., 2021b) extends the self-knowledge idea by employing both cross-modal and intra-modal knowledge in the same time. Given an image-text pair, the first step is to construct intra-modal graphs, i.e an image graph and a text graph. The image graph consists of regions (defined by a pre-trained object detector) as nodes, with IoU scores of paired regions acting as edge weights between those regions. Similarly for the text graph, objects, attributes and relationships are extracted from text to fill the nodes of the text graph, while edge weights are defined by the co-occurrence frequency between pairs of nodes. In both graphs, zero similarity scores between nodes indicate absence of edge. A cross-modal scene graph is derived from the image and text graph by aligning predicted region tags from the image side and words from the text side by comparing their textual semantic similarity. By calculating this similarity score for all possible tag-word pair, edge weights between cross-modal nodes are defined. Nodes connected via cross-modal edges, named anchor nodes, form subgraphs which maintain intra-modal and cross-modal edges, as well as two-hop connections that contain paths of cross-modal followed by intra-modal edges. ROSITA leverages this enhanced representation to boost three downstream tasks: VQA, VRE and ITR.

6.8.2 Evaluation

Human Evaluation is useful in cases of language generation tasks, such as the rationales generation of (Marasović et al., 2020). In this case, the need for human evaluation arises from the observation that certain discrete rationales, even though not being paraphrases of each other, can be suitable. The following aspects were evaluated: visual plausibility referring to how well the generated rationales support the answer (in VQA and VCR) or the entailment (VE) given the image, and visual fidelity measuring the appearance of irrelevant information within more plausible generated rationales. By excluding images, textual Plausibility evaluates generated rationales based on their support on the answer (in VQA and VCR) or the entailment (VE) exclusively.

Classification metrics such as accuracy serve as the golden standard for non-generative models on cross-modal reasoning tasks. In (Shevchenko et al., 2021), OKVQA accuracies per question types are also reported, in order to validate improvements in commonsense-oriented categories attributed to the injection of commonsense knowledge.

Ranking metrics provide valuable insights in cases when retrieval tasks are performed (Yu et al., 2021a; Cui et al., 2021b), Recall@k for k=1, 5, 10 is reported.

Language metrics such as BLEU (Papineni et al., 2002), CIDER (Vedantam et al., 2014), SPICE are used for language generation tasks, such as image captioning and novel object captioning (Li et al., 2020c).

7 The future of knowledge in VL

7.1 Explainability and biases

Some early works in KVL tasks widely addressed the need for boosting explainability via knowledge graphs. The complex opaque reasoning accompanying many state-of-the-art VL models indeed renders explainability a significant aspect. However, as development of more advanced models emerged, interpretability and fairness became incidental to performance. The pursuit of impressive results often promotes models with more vulnerabilities, which have been significantly underexplored. For example, the leading field of NLP has experienced some non-negligible failures, such as producing completely wrong statements based on a false conditioning\(^3\). It is unknown how many such vulnerabilities exist in state-of-the-art models, as there is no systematic way to capture and recognize them, nor is there any guarantee they will not occur. Those issues question the trust of humans over black-box models and even make such models susceptible to misuse. Explainability and robustness of VL models are even less explored comparing to NLP, leaving lots of space for the development of transparent models and post-hoc explainability methods. Generally, given the interwoven nature of knowledge graphs and explainability, it is expected that sooner or later research interests will resume towards this direction.

7.2 Zero-shot learning

Previous works regarding zero-shot classification tasks (Wang et al., 2018b; Nayak and Bach, 2020; Geng et al., 2020, 2021) leverage knowledge graphs to fuse feature information from seen to unseen classes. Little work has
been done so far towards the more complex task of multi-modal zero-shot learning with external knowledge (Chen et al., 2021c), leaving numerous unexplored directions open for future research.

7.3 Exploitation and integration of more knowledge senses

Despite the increasing interest towards knowledge-enhanced multimodal learning, there are some significantly underexplored knowledge aspects that could be leveraged in various tasks. For example, factual knowledge does not have a noticeable presence outside of VQA applications. Named entities and events are only addressed in a couple of applications. Temporal knowledge or even hypotheses and counterfactual thinking could reveal new aspects of existing tasks with the potential of interesting implementations.

7.4 Datasets

Even though dedicated knowledge-based datasets have been developed for VQA (Shah et al., 2019; Wang et al., 2018a; Marino et al., 2019; Singh et al., 2019; Yu et al., 2020; Jain et al., 2021), lack of corresponding ventures have been observed in other VL tasks. Such datasets could either incorporate external knowledge from certain knowledge bases in the first place, or be more flexible and dynamically retrieve knowledge to satisfy more challenging inputs. Suitable and high-quality datasets are the first step towards the evolution of the knowledge-enhanced multimodal learning field. Furthermore, knowledge-based datasets could be combined with explanation-oriented datasets, like VQA-E (Li et al., 2018) and e-SNLI-VE (Do et al., 2020) to address the issues mentioned in section 7.1.

7.5 Knowledge-enhanced generative tasks

The field sitting on the intersection of knowledge graphs and generative models has been significantly underexplored, despite the major success both fields have experienced in recent years. Until now, visual knowledge has been attempted in image synthesis, forming the task of image generation from scene graphs and layouts with interesting results and improvements on complex scene synthesis (Johnson et al., 2018; Li et al., 2019b; He et al., 2021b). Also domain knowledge (Chang et al., 2019) in GANs has demonstrated some insightful preliminary observations regarding knowledge-guided synthesis of unseen attributes, without the need of massive pre-training. A handful of aforementioned generative multimodal approaches (Yang et al., 2019a) incorporate knowledge graphs for the tasks of visual storytelling and story visualization. More compelling results could unfold from the combination of various knowledge graphs with multimodal generative models, enabling conditioning on commonsense, hierarchical, factual knowledge, and also enforcing interpretable insights into the generation process.

7.6 The need for multi-task learners

An abundance of multi-task knowledge-free transformer-based VL models have emerged in recent literature, presenting impressive results on a variety of downstream tasks by utilizing the same pre-trained body each time. On the other hand, knowledge-enhanced VL models usually target a single task, and only a few knowledge-enhanced multi-task models (Marasović et al., 2020; Shevchenko et al., 2021; Yu et al., 2021a; Cui et al., 2021b; Chen et al., 2021b) have been developed. Even in the case of multi-task transformers, almost half of them utilize only self-knowledge without exploiting the benefits of additional external sources. In the same time, the harder venture of integrating external knowledge limits the range of tasks that multi-task models target. Specifically, current implementations have covered only reasoning tasks. Therefore, as a first future direction, retrieval tasks can also be attempted. Going one step further, some VL tasks addressed in multi-task knowledge-free VL transformers have never been explored in literature; hence, unified multi-task architectures could possibly explore their knowledge-enhanced capabilities, without the need of developing individual non-reusable approaches. In any case, multi-task knowledge-enhanced models would unlock the full potential of the contributions of knowledge, with competitive architectures pushing the state-of-the-art even further.

8 Conclusion

Introducing external knowledge in multimodal learning has demonstrated promising research directions, targeting performance, explainability and extendability of existing tasks. In this survey paper, we analyzed the meeting point of visiolinguistic representation learning and knowledge assisted learning, focusing on the contribution of existing knowledge graphs and unstructured knowledge sources. The presented taxonomy of knowledge-enhanced datasets, tasks and models provides one of the first attempts towards structuring the field of knowledge-enhanced VL learning, with the aim to guide future research and address prospects and challenges of this upcoming field.

References

Imdb. https://www.imdb.com/.
Meet the flintstones dataset.  http://flintstones-dataset-dev-rev.
s3-website-us-west-2.amazonaws.com/.

Movie genre from its poster.  https://www.kaggle.com/neha1703/
movie-genre-from-its-poster.

Dietrich Klakow, Aditya Mogadala, Xiaoyu Shen. 2020. Integrating rule-based entity masking into image captioning.

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. 2016. Vqa: Visual question answering.

Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. 2019. Fusion of detected objects in text for visual question evaluation.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sunderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. 2007. Dbpedia: A nucleus for a web of open data. In ISWC/ASWC.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2017. Multimodal machine learning: A survey and taxonomy.

Santanjee Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Lisa Bauer, Lingjia Deng, and Mohit Bansal. 2021. Ernie-nli: Analyzing the impact of domain-specific external knowledge on enhanced representations for nli. In DEELIO.

Sumithra Bhakhthavatsalam, Kyle Richardson, Niket Tandon, and Peter Clark. 2020. Do dogs have whiskers? a new knowledge base of haspart relations.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5:135–146.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In Advances in Neural Information Processing Systems, volume 33, pages 1877–1901. Curran Associates, Inc.

Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. 2020. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In ECCV.

Qingxing Cao, Bailin Li, Xiaodan Liang, and Liang Lin. 2019. Explainable high-order visual question reasoning: A new benchmark and knowledge-routed network.

Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder.

Che-Han Chang, Chun-Hsien Yu, Szu-Ying Chen, and Edward Y. Chang. 2019. Kg-gan: Knowledge-guided generative adversarial networks.

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 397–406.

Hong Chen, Yifei Huang, Hiroyo Takamura, and Hideki Nakayama. 2021a. Commonsense knowledge aware concept selection for diverse and informative visual storytelling. In AAAI.

Kezhen Chen, Qiujuan Huang, Yonatan Bisk, Daniel J. McDuff, and Jianfeng Gao. 2021b. Kb-vlp: Knowledge based vision and language pretraining.

Qian Chen, Xiaodan Zhu, Zhenhua Ling, Diana Inkpen, and Si Wei. 2018a. Neural natural language inference models enhanced with external knowledge. In ACL.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Diana Inkpen, and Si Wei. 2018b. Neural natural language inference models enhanced with external knowledge. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2406–2417, Melbourne, Australia. Association for Computational Linguistics.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning.

Zhuo Chen, Jiaoyan Chen, Yuxia Geng, Jeff Z. Pan, Zonggang Yuan, and Huajun Chen. 2021c. Zero-shot visual question answering using knowledge graph. In SEMWEB.
Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation.

Jaemin Cho, Jiasen Lu, Dustin Schwenk, Hannaneh Hajishirzi, and Aniruddha Kembhavi. 2020. X-1xmrct: Paint, caption and answer questions with multi-modal transformers. In EMNLP.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches.

Leyang Cui, Yu Wu, Shujie Liu, and Yue Zhang. 2021a. Knowledge enhanced fine-tuning for better handling unseen entities in dialogue generation. ArXiv, abs/2109.05487.

Yuhao Cui, Zhou Yu, Chunqi Wang, Zhongzhou Zhao, Ji Zhang, Meng Wang, and Jun Yu. 2021b. Rosita: Enhancing vision-and-language semantic alignments via cross- and intra-modal knowledge integration. Proceedings of the 29th ACM International Conference on Multimedia.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2016. Visual dialog.

Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2016. Guesswhat?! visual object discovery through multi-modal dialogue.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. ArXiv, abs/1810.04805.

Arka Ujjal Dey, Ernest Valveny, and Gaurav Harit. 2021. External knowledge enabled text visual question answering.

Prafulla Dhariwal and Alex Nichol. 2021. Diffusion models beat gans on image synthesis.

Virginie Do, Oana-Maria Camburu, Zeynep Akata, and Thomas Lukasiewicz. 2020. e-snli-ve: Corrected visual-textual entailment with natural language explanations.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations.

Yifan Du, Zikang Liu, Junyi Li, and Wayne Zhao. 2022. A survey of vision-language pre-trained models.

Shiv Ram Dubey. 2021. A decade survey of content based image retrieval using deep learning. IEEE Transactions on Circuits and Systems for Video Technology, page 1–1.

Alaaeldin El-Nouby, Shikhar Sharma, Hannes Schulz, Devon Hjelm, Layla El Asri, Samira Ebrahimi Kahou, Yoshua Bengio, and Graham W. Taylor. 2019. Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction.

D. Elliott, S. Frank, K. Sima’an, and L. Specia. 2016. Multi30k: Multilingual english-german image descriptions. In Proceedings of the 5th Workshop on Vision and Language, pages 70–74.

Ko Endo, Masaki Aono, Eric Nichols, and Kotaro Funakoshi. 2017. An attention-based regression model for grounding textual phrases in images. In IJCAI.

Patrick Esser, Robin Rombach, and Björn Ommer. 2020. Taming transformers for high-resolution image synthesis.

Christiane Fellbaum. 1998. Wordnet: An electronic lexical database.

Difei Gao, Ruiping Wang, S. Shan, and Xilin Chen. 2019. Cric: A vqa dataset for compositional reasoning on vision and commonsense.

Noa Garcia and George Vogiatzis. 2018. How to read paintings: Semantic art understanding with multi-modal retrieval.

Noa Garcia, Chentao Ye, Zihua Liu, Qingtao Hu, Mayu Otani, Chenhui Chu, Yuta Nakashima, and Teruko Mitamura. 2020. A dataset and baselines for visual question answering on art.

Diego Garcia-Olano, Yasumasa Onoe, and Joydeep Ghosh. 2021. Improving and diagnosing knowledge-based visual question answering via entity enhanced knowledge injection.

Francois Gardères, Maryam Ziaeefard, Baptiste Abelos, and Freddy Lécué. 2020. Conceptbert: Concept-aware representation for visual question answering. In FINDINGS.

Yuxia Geng, Jiaoyan Chen, Zhuo Chen, Jeff Z. Pan, Zhiquan Ye, Zonggang Yuan, Yantao Jia, and Huajun Chen. 2021. Ontosil: Ontology-enhanced zero-shot learning. Proceedings of the Web Conference 2021.

Yuxia Geng, Jiaoyan Chen, Zhuo Chen, Zhiquan Ye, Zonggang Yuan, Yantao Jia, and Huajun Chen. 2020. Generative adversarial zero-shot learning via knowledge graphs.

Ross Girshick. 2015. Fast r-cnn.

Arushi Goel, Basura Fernando, Thanh-Son Nguyen, and Hakan Bilen. 2020. Injecting prior knowledge into image captioning. In ECCV Workshops.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In Advances in Neural Information Processing Systems, volume 27.
Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Making the v in vqa matter: Elevating the role of image understanding in visual question answering.

Ralph Grishman and Beth Sundheim. 1996. Design of the muc-6 evaluation. In Proceedings of a Workshop on Held at Vienna, Virginia: May 6–8, 1996, TIPSTER ’96, page 413–422, USA. Association for Computational Linguistics.

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Wenzhong Guo, Jianwen Wang, and Shiping Wang. 2019. Deep multimodal representation learning: A survey. IEEE Access, 7:63373–63394.

William L. Hamilton. Graph representation learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, 14(3):1–159.

William L. Hamilton, Rex Ying, and Jure Leskovec. 2018. Representation learning on graphs: Methods and applications.

Feijuan He, Yaxian Wang, Xianglin Miao, and Xia Sun. 2021a. Interpretable visual reasoning: A survey. Image and Vision Computing, 112:104194.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778.

Sen He, Wentong Liao, Michael Yang, Yongxin Yang, Yi-Zhe Song, Bodo Rosenhahn, and Tao Xiang. 2021b. Context-aware layout to image generation with enhanced object appearance. In CVPR.

Xiangteng He and Yuxin Peng. 2020. Fine-grained visual-textual representation learning. IEEE Transactions on Circuits and Systems for Video Technology, 30(2):520–531.

Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. 2016. Generating visual explanations.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. Neural computation, 9:1735–80.

Jingyi Hou, Xinxiang Wu, Yuyun Qi, Wentian Zhao, Jiebo Luo, and Yunde Jia. 2019. Relational reasoning using prior knowledge for visual captioning. ArXiv, abs/1906.01290.

Jingyi Hou, Xinxiang Wu, Xiaoxun Zhang, Yuyun Qi, Yunde Jia, and Jiebo Luo. 2020. Joint commonsense and relation reasoning for image and video captioning. In AAAI.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification.

Chao-Chun Hsu, Zi-Yuan Chen, Chi-Yang Hsu, Chih-Chia Li, Tzu-Yuan Lin, Ting-Hao ’Kenneth’ Huang, and Lun-Wei Ku. 2019. Knowledge-enriched visual storytelling.

Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. 2020. Heterogeneous graph transformer. In Proceedings of The Web Conference 2020, WWW ’20, page 2704–2710, New York, NY, USA. Association for Computing Machinery.

Feicheng Huang, Zhixin Li, Shengjia Chen, Canlong Zhang, and Huifang Ma. 2020a. Image captioning with internal and external knowledge. Proceedings of the 29th ACM International Conference on Information & Knowledge Management.

Ting-Hao K. Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Aishwarya Agrawal, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016a. Visual storytelling. In 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2016).

Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016b. Visual storytelling. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1233–1239, San Diego, California. Association for Computational Linguistics.

Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. 2021. Seeing out of the box: End-to-end pre-training for vision-language representation learning.

Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020b. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering.

Yuqi Huo, Manli Zhang, Guangzhen Liu, Haoyu Lu, Yizhao Gao, Guoxing Yang, Jingyuan Wen, Heng Zhang, Baogui Xu, Weihao Zheng, Zongzheng Xi, Yueqian Yang, Anwen Hu, Jinming Zhao, Ruichen Li, Yida Zhao, Liang Zhang, Yuqing Song, Xin Hong, Wanqing Cui, Danyang Hou, Yingyan Li, Junyi Li, Peiyu Liu, Zheng Gong, Chuhao Jin, Yuchong Sun, Shizhe Chen, Zhiwu Lu, Zhicheng Dou, Qin Jin, Yanan Lan, Wayne Xin Zhao, Ruihu Song, and Ji-Rong Wen. 2021. Wenlan: Bridging vision and language by large-scale multi-modal pre-training.
Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. In AAAI.

Filip Ilievski, Alessandro Oltramari, Kaixin Ma, Bin Zhang, Deborah L. McGuinness, and Pedro Szekely. 2021. Dimensions of commonsense knowledge.

Brian Kenji Iwana, Syed Tahseen Raza Rizvi, Sheraz Ahmed, Andreas Dengel, and Seiichi Uchida. 2017. Judging a book by its cover.

Aman Jain, Mayank Kothiyari, Vishwajeet Kumar, Preethi Jyothi, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2021. Select, substitute, search: A new benchmark for knowledge-augmented visual question answering. Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.

Shaoxiong Ji, Shirui Pan, E. Cambria, P. Marttinen, and Philip S. Yu. 2021. A survey on knowledge graphs: Representation, acquisition and applications. IEEE transactions on neural networks and learning systems, PP.

Zizheng Ji, Lin Dai, Jin Pang, and Tingting Shen. 2020. Leveraging concept-enhanced pre-training model and masked-entity language model for named entity disambiguation. IEEE Access, 8:100469–100484.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision.

Justin Johnson, Agrim Gupta, and Li Fei-Fei. 2018. Image generation from scene graphs. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1219–1228.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In CVPR.

Kushal Kafle, Robik Shrestha, and Christopher Kanan. 2019. Challenges and prospects in vision and language research.

Tero Karras, Samuli Laine, and Timo Aila. 2018. A style-based generator architecture for generative adversarial networks.

Kyung-Min Kim, Min-Oh Heo, Seong-Ho Choi, and Byoung-Tak Zhang. 2017. Deeppstory: Video story qa by deep embedded memory networks.

Wonjae Kim, Bok jung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.

Satwik Kottur, José M. F. Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2019. Clevr-dialog: A diagnostic dataset for multi-round reasoning in visual dialog.

Ranjay Krishna, Ines Chami, Michael Bernstein, and Li Fei-Fei. 2018. Referring relationships.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations.

Alex Krizhevsky. 2009. Learning multiple layers of features from tiny images.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents.

JaeYun Lee and Incheol Kim. 2020. Vision–language–knowledge co-embedding for visual commonsense reasoning.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Chunye Li, Liya Kong, and Zhiping Zhou. 2020a. Improved-storygan for sequential images visualization. Journal of Visual Communication and Image Representation, 73:102956.

Guohao Li, Xin Wang, and Wenwu Zhu. 2020b. Boosting visual question answering with context-aware knowledge aggregation. Proceedings of the 28th ACM International Conference on Multimedia.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Goyal, Shafiq Joty, Calmimg Xiong, and Steven Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation.

Liumian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019a. Visualbert: A simple and performant baseline for vision and language.

Qing Li, Qingyi Tao, Shafiq Joty, Jianfei Cai, and Jiebo Luo, 2018. Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions.

Xiujuan Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yeqin Choi, and Jianfeng Gao. 2020c. Oscar: Object-semantics aligned pre-training for vision-language tasks.

Yikang Li, Tao Ma, Yeqi Bai, Nan Duan, Sining Wei, and Xiaogang Wang. 2019b. Pastegan: A semi-parametric method to generate image from scene graph. NeurIPS.
Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. 2019c. Storygan: A sequential conditional gan for story visualization. pages 6322–6331.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In Text Summarization Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Junyang Lin, An Yang, Yichang Zhang, Jie Liu, Jingren Zhou, and Hongxia Yang. 2021. Interbert: Vision-and-language interaction for multi-modal pretraining.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microcosm coco: Common objects in context. In Computer Vision – ECCV 2014, pages 740–755, Cham. Springer International Publishing.

Zachary C. Lipton, John Berkowitz, and Charles Elkan. 2015. A critical review of recurrent neural networks for sequence learning.

Linlin Liu, Xin Li, Ruidan He, Lidong Bing, Shafiq R. Joty, and Luo Si. 2021a. Knowledge based multilingual language model. ArXiv, abs/2111.10962.

Runtao Liu, Chenxi Liu, Yutong Bai, and Alan Yuille. 2019a. Clevr-ref4+: Diagnosing visual reasoning with referring expressions.

Xingchao Liu, Chengyue Gong, Lemeng Wu, Shujian Zhang, Hao Su, and Qiang Liu. 2021b. Fusedream: Training-free text-to-image generation with improved clip+gan space optimization.

Yinhua Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenha Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furui Wei, and Baining Guo. 2022. Swin transformer v2: Scaling up capacity and resolution. In International Conference on Computer Vision and Pattern Recognition (CVPR).

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021c. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

Robert Logan, Nelson F. Liu, Matthew E. Peters, Matt Gardner, and Sameer Singh. 2019. Barack’s wife hillary: Using knowledge graphs for fact-aware language modeling. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5962–5971, Florence, Italy. Association for Computational Linguistics.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning.

Yinquan Lu, H. Lu, Guirong Fu, and Qun Liu. 2021. Kelm: Knowledge enhanced pre-trained language representations with message passing on hierarchical relational graphs. ArXiv, abs/2109.04223.

Man Luo, Yankai Zeng, Pratyay Banerjee, and Chitta Baral. 2021. Weakly-supervised visual-retriever-reader for knowledge-based question answering. In EMNLP.

Adyasha Maharana and Mohit Bansal. 2021. Integrating visuospatial, linguistic, and commonsense structure into story visualization. ArXiv, abs/2110.10834.

Adyasha Maharana, Darryl Hannan, and Mohit Bansal. 2021. Improving generation and evaluation of visual stories via semantic consistency. ArXiv, abs/2105.10026.

Adyasha Maharana, Darryl Hannan, and Mohit Bansal. 2022. Storydall-e: Adapting pretrained text-to-image transformers for story continuation.

Chaitanya Malaviya, Chandra Bhagavatula, Antoine Bosselut, and Yejin Choi. 2019. Commonsense knowledge base completion with structural and semantic context.

Mateusz Malinowski and Mario Fritz. 2014. Towards a visual turing challenge.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. Introduction to Information Retrieval. Cambridge University Press, USA.

Ana Marasović, Chandra Bhagavatula, Jae Sung Park, Ronan Le Bras, Noah A. Smith, and Yejin Choi. 2020. Natural language rationales with full-stack visual reasoning: From pixels to semantic frames to commonsense graphs. In FINDINGS.

Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Kumar Gupta, and Marcus Rohrbach. 2021. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 14106–14116.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 3190–3199.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2018. Learned in translation: Contextualized word vectors.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality.

Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets.
Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016d. Generative adversarial text to image synthesis.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Exploring models and data for image question answering.

S. Ren, K. He, R. Girshick, and J. Sun. 2017. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 39(6):1137–1149.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3:333–389.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-resolution image synthesis with latent diffusion models.

Nataniel Ruiz, Yuazhuan Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2022. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation.

Tara Safavi and Danai Koutra. 2021. Relational world knowledge representation in contextual language models: A review. *ArXiv*, abs/2104.05837.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic text-to-image diffusion models with deep language understanding.

Ander Salaberria, Gorka Azkune, Oier Lopez de Lacalle, Aitor Soroa Etxabe, and Eneko Agirre. 2021. Image captioning for effective use of language models in knowledge-based visual question answering. *ArXiv*, abs/2109.08029.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans.

Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2017. Modeling relational data with graph convolutional networks.

Mike Schuster and Kuldip K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.*, 45:2673–2681.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. Kvqa: Knowledge-aware visual question answering. In *AAAI*.

Or Sharir, Barak Peleg, and Yoav Shoham. 2020. The cost of training nlp models: A concise overview.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*.

Violetta Shevchenko, Damien Teney, Anthony R. Dick, and Anton van den Hengel. 2021. Reasoning over vision and language: Exploring the benefits of supplemental knowledge. *ArXiv*, abs/2101.06013.

Jiaxin Shi, Hanwang Zhang, and Juanzhi Li. 2018. Explainable and explicit visual reasoning over scene graphs.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition.

Ajeet Kumar Singh, Anand Mishra, Shashank Shekhar, and Anirban Chakraborty. 2019. From strings to things: Knowledge-enabled vqa model that can read and reason. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4601–4611.

Amanpreet Singh, Vedanuj Goswami, and Devi Parikh. 2020. Are we pretraining it right? digging deeper into visio-linguistic pretraining.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2021. Flava: A foundational language and vision alignment model.

Dandan Song, Siyi Ma, Zhanchen Sun, Sicheng Yang, and Lejian Liao. 2021. Kvl-bert: Knowledge enhanced visual-and-linguistic bert for visual commonsense reasoning. *Knowl. Based Syst.*, 230:107408.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Ciscanelli, Giuseppe Fiameni, and Rita Cucchiara. 2021. From show to tell: A survey on deep learning-based image captioning.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vi-bert: Pre-training of generic visual-linguistic representations.
