Achieving Human Parity on Visual Question Answering

MING YAN, HAIYANG XU, CHENLIANG LI, JUNFENG TIAN, BIN BI, WEI WANG, XIANZHE XU, JI ZHANG, SONGFANG HUANG, FEI HUANG, LUO SI, and RONG JIN,
Alibaba DAMO Academy, China

The Visual Question Answering (VQA) task utilizes both visual image and language analysis to answer a textual question with respect to an image. It has been a popular research topic with an increasing number of real-world applications in the last decade. This paper introduces a novel hierarchical integration of vision and language AliceMind-MMU (Alibaba’s Collection of Encoder-decoders from Machine Intelligence lab of Damo academy - MultiMedia Understanding), which leads to similar or even slightly better results than a human being does on VQA. A hierarchical framework is designed to tackle the practical problems of VQA in a cascade manner including: (1) diverse visual semantics learning for comprehensive image content understanding; (2) enhanced multi-modal pre-training with modality adaptive attention; and (3) a knowledge-guided model integration with three specialized expert modules for the complex VQA task. Treating different types of visual questions with corresponding expertise needed plays an important role in boosting the performance of our VQA architecture up to the human level. An extensive set of experiments and analysis are conducted to demonstrate the effectiveness of the new research work.

CCS Concepts: • Computing methodologies → Artificial intelligence; • Information systems → Multimedia information systems;

Additional Key Words and Phrases: Visual Question Answering, multi-modal pre-training, text and image content analysis, cross-modal interaction, visual reasoning

ACM Reference format:
Ming Yan, Haiyang Xu, Chenliang Li, Junfeng Tian, Bin Bi, Wei Wang, Xianzhe Xu, Ji Zhang, Songfang Huang, Fei Huang, Luo Si, and Rong Jin. 2023. Achieving Human Parity on Visual Question Answering. ACM Trans. Inf. Syst. 41, 3, Article 79 (April 2023), 40 pages. https://doi.org/10.1145/3572833

1 INTRODUCTION

Recent years have witnessed human-level performance reached or surpassed by well trained computer programs in tasks ranging from games such as Go [69] to classification of images in ImageNet [14] to natural language understanding on the GLUE benchmark [77]. Vision and language are two fundamental capabilities of human intelligence. We have seen dramatic progress in the area of representation learning across these two modalities. Inspired by the success of pre-training in
both computer vision (CV) [65] and natural language processing (NLP) [16], a number of vision-and-language (V&L) models [11, 35, 43, 74, 91, 93] have been proposed in the last couple of years to tackle challenges at the intersection of these two key areas of AI. Despite superhuman performance achieved respectively in some vision (e.g., ImageNet) and natural language tasks (e.g., GLUE), joint learning across these two modalities, which is essential to human cognition, has demonstrated limited human-level performance by prevalent V&L approaches.

A compelling reason to study vision and language jointly is the promise of language as a universal and natural interface for visual reasoning problems - useful both in specifying a wide range of problems and in communicating AI responses. Visual reasoning has long been recognized as most challenging for cross-modal learning owing to its requirement of higher-order cognition and commonsense reasoning intelligence. With the systematic design of our hierarchical integration framework, this research work unprecedentedly surpasses human performance in the popular Visual Question Answering (VQA) task [1]. This paper summarizes how we achieve the human parity in VQA. Most of the presented techniques are not specific to VQA and can be transferable to tackling other visual reasoning tasks.

A VQA system takes as input an image and a free-form, open-ended, natural-language question about the image, and produces a natural-language answer as the output. Example questions are shown in Figure 1. The open-ended questions require a potentially vast set of AI capabilities to answer, including question understanding, commonsense reasoning, visual recognition, relational reasoning, object counting, and visually-grounded language understanding, and so on. Therefore, achieving human performance in VQA would be an important milestone in artificial intelligence. Based on the previous studies [2, 11, 33, 35], we summarize that the main problems related to VQA lies in the following three aspects: (1) multi-granularity visual semantics: the real questions may focus on different semantic parts of the image, e.g., questions about object counting pay more attention on the local objects of the image, while global information and relative position information are needed for commonsense reasoning or relational reasoning questions, as shown in Figure 1. The current visual features cannot well capture the multi-granularity visual semantics.
within the image; (2) **cross-modal semantic gap**: the semantic gap between visual and textual modalities has always been treated as one of the most important challenges in cross-modality research. There is a lack of effective cross-modal fusion method, which considers the multi-granularity semantics between the visual and textual information; (3) **open-ended knowledge**: visual questions are open-ended and require various kinds of knowledge and capabilities to answer. There exist certain specialized questions that are difficult to address by the general-purpose VQA techniques, such as OCR-based text reading questions and clock-reading questions, as shown in Figure 1.

To address these problems, a new hierarchical framework for VQA is designed by improving and integrating the individual capabilities, including diverse visual semantics learning, enhanced **vision-language pre-training** (VLP) and knowledge-guided model integration. The overall framework is shown in Figure 2. Different from the previous methods that solely rely on region-based visual feature [2], grid-based visual feature [33] or patch-based visual feature [17], we systematically investigate the different characteristics of these visual features, and take all the three types of visual features into consideration. In this way, the extracted visual features can capture both local and global semantics of the image and complement with each other, so as to better support the subsequent cross-modal fusion. Then, to bridge the cross-modal semantic gap for effective cross-modal fusion, vision-language pre-training has been the most prevalent method in recent years. There mainly exist two kinds of architectures: **single-stream architecture** [11, 72] and **dual-stream architecture** [74, 91]. The single-stream architecture essentially treats the two input modalities equally, and thus does not make full use of the signal from each modality. On the other hand, the dual-stream architecture is insufficient to capture the fine-grained interaction between visual and textual hints. To address the limitations, based on the extracted multi-granularity features, we propose a new cross-modal fusion method with **modality adaptive attention** mechanism. The model is built upon the single-stream architecture to allow for fine-grained cross-modal interaction. The original self-attention in Transformer is replaced with a modality adaptive self-attention, where intra-modal attention and inter-modal attention are dynamically adjusted. This leads to effective alignment of the cross-modal semantics.

Finally, considering the open-ended knowledge of the questions, we first design a clustering-based knowledge mining method to automatically identify different types of the question, that are not well-addressed by the general-purposed vision-language models such as text-reading questions and clock-reading questions. Different from the existing methods for VQA that treat different types of visual questions in the same manner, we propose a two-step model integration method, which resolves each type of question with a specialized expert module and integrates model predictions for each question in a hierarchical structure. In the first step, we use a voting-based ensemble
method to obtain the prediction result for each expert module, by integrating the predictions of multiple homogeneous models. It is difficult to directly combine the heterogeneous results from different expert modules. Therefore, in the second step, we introduce a knowledge-guided Mixture-of-Experts (K-MoE) paradigm to learn to dispatch each question to the best-matched expert for final prediction. Beyond a patchwork of models like ensemble, the MoE paradigm learns which module to call upon by exploiting the expertise of each module, and thus intelligently delegates every question to a proper expert module. According to our quantitative analysis, the two-step model integration method plays an important role in boosting the performance of our VQA architecture up to the human level, significantly outperforming existing methods without explicit task decomposition by a large margin.

Our key contributions can be summarized as follows:

- We propose a hierarchical integration framework of vision and language, namely AliceMind-MMU, to systematically address the practical problems of VQA. It includes diverse visual semantics learning, enhanced vision-language pre-training and a knowledge-guided model integration in a hierarchical structure.
- For diverse visual semantics learning, we systematically investigate the different characteristics of three types of visual features, from which multi-granularity semantics of the image can be extracted. To bridge the cross-modal semantic gap, we introduce a novel modality adaptive attention mechanism, where intra-modal attention and inter-modal attention are dynamically adjusted. Finally, different from the naive model ensembling, we design a knowledge-guided model integration method to tackle the open-ended knowledge of different types of questions in VQA, with a two-step model integration at an expert level.
- AliceMind-MMU achieves new state-of-the-art performance on the popular VQA dataset. More surprisingly, to the best of our knowledge, this is the first time AI model has achieved human parity on the challenging VQA dataset. It sets a new milestone for the vision-and-language field, which involves both CV and NLP technologies.

The rest of this paper is organized as follows. Section 2 introduces the prior work related to VQA. Section 3 presents the design of our new hierarchical framework for VQA. The empirical evaluation and quantitative analysis are given in Section 4. Finally, the paper is concluded in Section 5 by discussing our findings and limitations.

2 RELATED WORK

2.1 Question Answering

Research on question answering (QA) dates back to 1960s, when the QA systems were basically based on the rule-based methods and a large number of hand-crafted rules are designed to conduct query understanding and database searching [70]. It was not until the release of QA track evaluations of the Text REtrieval Conference (TREC) [76] that text-based QA received gradually increasing attention. Traditional approaches typically use a pipeline method by considering question processing, document retrieval and answer reranking such as the famous IBM Watson system [19]. Such pipelined QA methods suffer from heavy engineering efforts and are not widely applicable. With the rise of deep learning and the release of the SQuAD dataset [62], researchers began to shift their attention from document-level understanding to a more fine-grained passage level, and studied the semantic-related machine reading comprehension (MRC) task. Different types of reading comprehension tasks have been proposed to make it practical, ranging from multi-choice reading comprehension [37, 81], span-based extraction [62, 83], multi-passage reading comprehension [56, 98], to multi-hop reasoning [90, 97].
Recently, open-domain QA \cite{53,58,78,96,97} has become one of the most popular topics in question answering. This task is originally proposed by Chen et al. \cite{9}, which introduces a two-step Retriever-Reader pipeline method. The Retriever utilizes a simple TF-IDF retriever module for selecting documents, while the Reader designs a span-based MRC module for extracting the final answer. Different from machine reading comprehension task that takes a single passage or document as input evidence, open-domain QA aims to address a more challenging and practical scenario by reading from a large collection of noisy documents. Current approaches to open-domain QA largely follow the two-stage Retriever-Reader framework, and focus on improving both the Retriever module \cite{34,80,97} and Reader module \cite{9,78}. Moreover, Guu et al. \cite{24} take one step further to jointly train the Retriever and Reader modules with language model pre-training, which provides strong end-to-end performance on open-domain QA tasks.

People perceive the world in a multi-modal way, and with the development of 5G technology, multi-modal content understanding such as visual question answering (VQA) has quickly evolved as a hot emerging topic. Given an image and a corresponding natural language question as input, the goal of VQA is to produce the answer by reasoning about the multi-modal content. Since first released in 2016, the research on VQA has attracted increasing interests in the past few years. A large number of new methods have been proposed to push the limit of this task. Existing approaches to the VQA task mostly put efforts on three key aspects: (1) better visual feature representation, (2) effective cross-modal fusion, and (3) task-specific expert optimization.

Region-based visual features have long been treated as the de-facto standard for vision and language tasks, where visual semantics and salient image objects are captured with bottom-up attention \cite{2}. The proposed bottom-up attention method with region features won the VQA Challenge of 2017, which has been largely adopted since then. On the other hand, Jiang et al. \cite{33} study the key factors that contributed to the effectiveness of existing bottom-up attention features, and found that the grid-based features from convolutional neural networks can yield comparable or even better results on the VQA task. Different from the previous work that purely relies on a single kind of feature, we design a feature fusion method and take the advantage of diverse visual features for improving the visual understanding ability.

Cross-modal fusion or interaction has always been treated as one of the most important challenges in cross-modality research. Traditional cross-modal fusion methods gradually improve from a simple linear model \cite{1} to bilinear pooling ones \cite{20,94}, and attention-based fusion ones \cite{52,89}, where questions and image features are fully interacted with each other. Recently, with the popularity of pre-train and fine-tune paradigm such as BERT \cite{16} in NLP, large-scale vision-and-language pre-training \cite{11,29,41,43,72,74,91} has been used to better align the vision-language representations. It has established the new state-of-the-art performance on the VQA task since 2020, by pre-training on a large amount of unlabeled image-text pairs. We also follow the pre-train and fine-tune paradigm, and propose a new modality adaptive attention-based fusion method which adaptively learns the intra-modal and inter-modal attention weights.

To further improve the VQA performance, some studies address the weakness of existing models on the VQA task with specific expert optimization. For example, MoVie \cite{55} revisit modulated convolutions for visual counting, which is more efficient and advances the state of the art on counting-specific VQA tasks. Recently, some researchers in the IR community have also begun to investigate specific problems in visual question answering. For example, \cite{23,44} address the Language-Prior problem in VQA and propose a reweighting-based method to reason based on both the image and text content. Qu et al. \cite{59} introduce external knowledge to conduct visual question answering, and design a dual-encoder passage retrieval method to get access to proper multi-modal content. In this work, we propose a novel knowledge mining framework to cluster the data points and identify two additional task-specific experts to further improve the VQA performance: i.e.,
OCR-based text reading and detection-based clock reading. Besides, a **Mixture-of-Experts (MoE)** model is designed to combine the mined knowledge experts.

### 2.2 Cross-modal Retrieval

Visual question answering involves the cross-modal fusion and alignment between image and text, which is related to the topic of cross-modal retrieval \([12, 26, 28, 32, 47, 60, 61, 84, 92]\). Cross-modal retrieval is used to implement a retrieval task across different modalities such as image-text and text-video retrieval. The main challenge is the semantic gap \([79]\) between different modalities of data, and the key lies in learning a common representation space, where content similarity between different modalities can be directly measured. Current approaches on cross-modal retrieval generally falls into two categories: (1) real-valued representation learning \([32, 61, 92]\), and (2) binary representation learning \([7, 12, 47]\), which is also called cross-modal hashing. The binary representation learning methods focus more on computational efficiency by projecting the input data into a common Hamming space encoded as binary codes.

On the other hand, real-valued representation learning aims to derive continuous vector representation for different types of data. Traditional approaches on this category usually follow the unsupervised subspace learning methods, by leveraging co-occurrence information, such as CCA \([63]\) and Deep CCA \([3]\). To learn more discriminative common representations, supervised methods exploit semantic category labels to encourage inner-class content similarity and decrease inter-class content similarity. For example, Sharma et al. \([66]\) extend CCA to a supervised version by leveraging the the semantic category labels to guide the representation learning. With the huge success of contrastive learning in image representation learning, the pairwise based methods have gained popularity by leveraging the aligned cross-modal data such as image-text pairs. For example, CLIP \([61]\) and ALIGN \([32]\) adopt a dual-encoder architecture and harness a large number of image-text pairs to learn common cross-modal representation with contrastive learning. Hong et al. \([26]\) propose a generative vision-language pre-training method to complete the missing modality for incomplete pairs, and then conduct cross-modal retrieval on the completed pairs of data.

The difference between visual question answering and cross-modal retrieval is that the VQA task focuses more on reasoning on both visual and textual information to answer a question, while cross-modal retrieval relies on cross-modal relevance matching. However, both tasks will benefit from vision-and-language pre-training models.

### 2.3 Vision-and-Language Pre-training

Inspired by the breakthrough of language pre-training in NLP, the research community begins to pay more attention to vision-language pre-training on large-scale image-text pairing data, which achieves state-of-the-art performance across a variety of **vision-language (VL)** tasks \([4, 30, 73]\).

Existing approaches to VLP \([11, 29, 41, 43, 72, 74, 91]\) mainly take a two-step training pipeline, which first extracts semantic visual features by specific object detector and then learns a cross-modal pre-training model to align text and visual features. Current research about this topic can be roughly divided into two lines. The first line adopts a single-stream Transformer architecture to model both image and text representations in a unified semantic space such as UNITER \([11]\) and OSCAR \([43]\). In contrast, the other line uses a two-stream Transformer architecture that first encodes the image and text modalities separately, and then fuses the cross-modal representations with another Transformer network, such as LXMERT \([74]\) and ERNIE-ViL \([91]\). In addition to image-text pairs, UNIMO \([42]\) also employed large scale of free text corpus and image collections for enhancing the cross-modal learning. These methods rely heavily on a task-specific bounding box (or region) based object detector, which impose unnecessary constraints on model designs and
limit potential applications of existing vision and language systems. Therefore, PixelBERT [29] and E2E-VLP [87] further proposed end-to-end VLP method, by jointly learning both the visual encoder and cross-modal Transformer simultaneously. To further improve the training and inference speed of VLP model, ViLT [35] removed both the region supervision and convolutional visual encoder, and conducted vision-language pre-training directly on image patch feature with linear projection. In this work, we take into consideration the region, grid, and patch visual features for VLP with a new cross-modal fusion method based on modality adaptive attention.

3 HIERARCHICAL VQA FRAMEWORK

3.1 Diverse Feature Representation Learning

Feature representation for vision and text is fundamental for cross-modal learning. Different types of features can help capture diverse data characteristics, which complements with each other.

3.1.1 Visual Features. To capture multi-granularity visual semantics within an image, three types of visual features are extracted: region feature, grid feature, and patch feature. These visual features are widely-used in previous vision-language models [2, 17, 33, 35, 87, 95], and they have different characteristics as shown in Figure 3. We can see that region feature focuses on the local semantics of the image and is good at locating the individual objects within the image. Grid feature is obtained by conducting deep convolution operation with CNN-based networks such as ResNet [25]. It can possess a global view of the image and capture more global semantics by deep layers of local convolution. Patch feature is a fixed-sized patch from a split of the image, so that it can reserve all the original information of the image and keep the relationship of relative position between patches. By conducting self-attention on the linear projected patches such as ViT [17], it can also capture global semantics of the image. Each type of visual feature has its own advantages for addressing different kinds of questions. For example, region feature can well identify the different objects within the image, which may be more suitable for object counting questions. Grid

| Feature Type         | Description                                                                 | Advantages                                           | Illustration                                                                 | Example Cases                |
|----------------------|-----------------------------------------------------------------------------|------------------------------------------------------|-----------------------------------------------------------------------------|-----------------------------|
| Region Feature       | obtained from an off-the-shelf object detector like Faster R-CNN             | Localization of individual objects with good explanations | ![Region Illustration](image) | How many oranges are in the photo? How many bears are dark brown? |
| Grid Feature         | grid convolutional features mapped from ConvNets                           | End2End and effective                               | ![Grid Illustration](image)   | What are the zebras moving? Is it raining? |
| Patch Feature        | operates on image patches using linear projection                           | End2End and computational efficiency                 | ![Patch Illustration](image)   | What is on the mountain tops? Where is the white powdered doughnut? |

Fig. 3. Three types of visual features to capture multi-granularity visual semantics within an image.
feature can better answer the reasoning-related questions such as yes/no type, since it can possess a global view of the whole image. Combining together all the visual features can lead to a better understanding of the multi-granularity visual semantics. For each type of visual feature, we adopt a certain pre-trained image encoder to generate the visual semantic feature, respectively. Example cases can be found in Figure 3, and we will describe more details of these visual features in the following.

**Region Feature.** With the discovery of ‘bottom-up’ attention [2], region-based visual features have been the de facto standard for vision and language tasks. Unlike normal ‘top-down’ attention that directly focuses on semantically irrelevant parts of visual input, bottom-up attention uses pre-trained object detectors [64] to identify salient regions based on the visual input. As a result, images are represented by a collection of region-based features, which provide better localization of individual objects and capture the detailed semantics within the image content. Generally, region-based visual encoders such as BUTD [2] are pre-trained with detection data like Visual Genome [36]. Recently, VinVL [95] has been built on a large-scale pre-trained object-attribute detection model with much larger amounts of data on four public object detection datasets, which helps better capture both coarse-level and fine-grained visual semantic information in images. The object detector from VinVL [95] is used in this work to extract a collection of object-level region features with more detailed visual semantics, where each object \( o_j \) is represented as a 2048-dimensional feature vector \( r_j \). To capture the spatial information of the object, box-level location features for each object are also encoded via a 4-dimensional vector \( l_j = (x_1, y_1, x_2, y_2) \) as in SemVLP [40] and LXMERT [74]. The \( r_j \) and \( l_j \) are concatenated to form a position-sensitive object feature vector, which is further transformed to a lower dimension of \( D \) using a linear projection to ensure that it has the same vector dimension as that of token embeddings.

Despite the superior performance obtained via region-based visual features, these kind of features suffer from several problems. Firstly, the region-based methods heavily rely on a pre-trained object detector, where the performance may be bounded by the capability of the object detector and its predefined visual vocabulary. Besides, only salient regions of the image are used in region-based methods, where the global or background information may be missing.

**Grid Feature.** To address the limitations of region-based features like locality, some works such as PixelBERT [29], E2E-VLP [87], and Grid-VLP [88] have been proposed to revisit grid-based convolutional features for multi-modal learning, skipping the expensive region-related steps. The advantage lies in that: (1) the grid-based feature allows to introduce more flexible architectural designs for vision and language tasks, which makes it possible to support end-to-end training and efficient online inference; (2) it operates on a full image instead of a collection of semantic regions, so it can better capture global information of an image such as the background information; (3) it does not rely on a pre-trained object detector with limited visual vocabulary. Specifically, starting from the raw image \( I_{img} \in R^{3 \times H_0 \times W_0} \) with three color channels, a fixed CNN-based image encoder such as ResNet [25] generates a lower-resolution activation map \( F_{img} \in R^{C \times H \times W} \), where \( C \) is the channel width and \( H = \frac{H_0}{32}, W = \frac{W_0}{32} \). As the cross-modal fusion network expects a sequence as input, the spatial dimensions of \( F_{img} \) are collapsed into one dimension, resulting in a \( HW \times C \) feature map. Finally, a linear projection layer is used to reduce the channel dimension of the high-level feature map from \( C \) to a smaller dimension \( D \) for matching the dimension of token embeddings. To distinguish between different modalities, the grid feature map is supplemented with a learnable modal type embedding which is added to the output of linear projection layer.

To generate good grid features, it is very important to pre-train a strong CNN-based image encoder, to which the visual semantic information is incorporated. In terms of the data used to
pre-train the image encoder, there are mainly two ways along this line: (1) **Supervised Pre-training**: the image encoder is pre-trained with image classification data such as ImageNet [14] or detection data such as Visual Genome [36]. As found in [33], the large-scale object and attribute annotations collected in the Visual Genome are very helpful to provide the grid feature with visual semantics incorporated; (2) **Unsupervised Pre-training**: the image encoder is pre-trained with a large amount of unlabeled image-text pairs without human supervision such as CLIP [61], where about 400M aligned image-text pairs are used. It belongs to the line of research that learns visual representations from natural language supervision [15, 32]. In this way, the image encoder is naturally aligned with the textual semantics to facilitate the cross-modal fusion. It is well recognized that fully supervised pre-trained CNN model shows promising performance on in-domain or near-domain datasets, while it may not yield best performance when coming to transfer learning on out-domain datasets. Features derived from different ways can well complement with each other, which adapts to different kinds of questions.

**Patch Feature.** **Vision Transformer (ViT)** [17] has achieved outstanding performance in various visual tasks [10, 49, 75]. It firstly splits an image into fixed-size patches, then uses a simple linear projection of a patch before feeding them into transformers. ViLT [35] is the first to explore patch-based features for multi-modal learning, and achieves up to dozens of times faster inference than previous region-based VLP methods. The advantage of patch-based features is that it is more effective in capturing the global structure and relationship of relative position of a full image with the self-attention mechanism, which can provide complementary visual features different from region-based and grid-based ones. Specifically, the 2D image \( I_{\text{img}} \in \mathbb{R}^{3 \times H_0 \times W_0} \) is reshaped into a sequence of flattened 2D patches \( x_p \in \mathbb{R}^{N \times (P^2 \cdot C)} \), where \((H_0, W_0)\) is the resolution of the original image, \(C\) is the number of channels, \((P, P)\) is the resolution of each image patch, and \(N = HW/P^2\) is the resulting number of patches and also serves as the input sequence length for the Transformer. Then, the patches are flattened and embedded to \(D\) dimensions with a trainable linear projection, and an appropriate position encoding is introduced to capture the geometric relationship among different patches. Finally, the sequence of patch embeddings serves as input of the visual transformer encoder to pretrain.

With the rapid development of various ViT variants, there are also different ways to generate diverse patch features as in grid feature extraction: (1) **Supervised Pre-training**: the image patch encoder is pre-trained with image classification data such as in ViT [17] or object detection data such as in Swin Transformer [49]; (2) **Unsupervised Pre-training**: the image patch encoder is pre-trained with a large number of unlabeled image-text pairs. For example, CLIP [61] pretrains the image patch encoder of ViT with 400M aligned image-text pairs and [8] provides a large dataset of 12M image-text pairs CC12M and conducts image-text pre-training to recognize long-tail visual concepts.

**Textual Features.** This research work utilizes the method in BERT [16] with the WordPiece tokenizer to tokenize the input text sentence into a sequence of sub-word tokens \(\{w_1, \ldots, w_m\}\). Then each token \(w_i\) is assigned three kinds of learnable embeddings: token, modal type, and position embeddings. The three embeddings are summed and layer-normalized to represent input sentence as a sequence of embedding vectors \(E_{\text{emb}} = \{e_{\text{CLS}}, e_1, \ldots, e_m, e_{\text{SEP}}\}\), where \([\text{CLS}]\) and \([\text{SEP}]\) are the special tokens in BERT.

To provide better textual features, text stream parameters were initialized with three different pre-trained language models: BERT [16], RoBERTa [48], and StructBERT [82]. RoBERTa trains on a larger corpus for more steps, and StructBERT incorporates more word ordering and sentence ordering information into pre-training language structures.
3.2 Vision-and-Language Pre-training with Modality Adaptive Attention

3.2.1 Vision-and-Language Pre-training (VLP). Vision-and-language pre-training is the most effective paradigm to bridge the cross-modal feature gap, by harnessing from abundant image-text aligned data. We base our VLP architecture on the single-stream Transformer, which is good at capturing more fine-grained interaction between cross-modal data. The input to Transformer is the image feature and its associated sentence (e.g., caption text). Each image is represented as a sequence of image features $\{o_1, \ldots, o_n\}$, and each sentence is represented as a sequence of word embeddings $\{e_1, \ldots, e_m\}$. The image and text embedding features are directly concatenated as input to the Transformer. We then pre-train the Transformer network for effective cross-modal fusion, by designing three types of pre-training tasks. A novel modality adaptive attention mechanism is proposed to further enhance the pre-training process by dynamically adjusting the intra-modal and inter-modal attention. The image representation for each kind of feature is described in Section 3.1.1, and we also concatenate all three types of image features together to construct a more comprehensive fusion feature. The overview of our enhanced VLP architecture is shown in Figure 4.

3.2.2 Modality Adaptive Attention. Traditional cross-modal attention either relies on the self-attention mechanism within a shared Transformer network by taking the concatenated image and text features as input, or leverages basic cross-attention from one modality to another for enhancing cross-modal understanding. However, they do not explicitly distinguish between different modalities when conducting self-attention or cross-attention. The proposed modality adaptive attention explicitly introduces modality prior with two learnable parameters in each layer of the self-attention block, so that the attention scores for inter-modal attention and intra-modal attention can be dynamically adjusted at a layer level. Reference [40] proposes SemVLP to learn the joint representation of vision and language, which aligns cross-modal semantics at multiple levels. It builds upon a shared Transformer encoder with specific self-attention masks for cross-modal interaction. However, the interaction between the two modalities are controlled by fixed self-attention mask with only two modes: interactive or non-interactive. Different from SemVLP [40], our VLP architecture uses two learnable self-attention weights for each layer to dynamically control the inter-modal and intra-modal interaction.

In single-stream models, the input to a Transformer layer is the concatenation of both modalities, $X = [X_L | X_V]$. As a result, in each single-stream attention head, the query matrix $Q$ and key matrix...
where \( (\cdot^L \cdot^V) \) are the language and visual sub-matrices of the input and the resulting output, \( W^Q \) and \( W^K \) are the projection matrices of the query and key in each Transformer attention head. As shown in Figure 4, the attention score matrix \( S \) can thus be divided in terms of four sub-matrices:

\[
S = QK^T = \begin{pmatrix} Q_L & S_{LL} & S_{LV} \\ V_L & S_{VL} & S_{VV} \end{pmatrix}
\]

(3)

Then, two learnable self-attention weights \( \varepsilon_1 \) and \( \varepsilon_2 \) are introduced for intra-modal attention score sub-matrices (diagonal of \( S \)) and inter-modal attention score sub-matrices (anti-diagonal of \( S \)), respectively. In each Transformer layer, the learnable weights are multiplied by the attention score matrix \( S \) to obtain the new attention score matrix \( S_\gamma \):

\[
S_\gamma = \begin{pmatrix} \varepsilon_1 & \varepsilon_2 \\ \varepsilon_1 & \varepsilon_2 \end{pmatrix} \odot \begin{pmatrix} Q_L & S_{LL} & S_{LV} \\ K_L & S_{VL} & S_{VV} \end{pmatrix}
\]

(4)

The following two methods are investigated to learn the self-attention weights \( \varepsilon_1 \) and \( \varepsilon_2 \):

- The weights are derived from a single-layer feed-forward network with the sigmoid activation function. \( V_{CLS} \) (the representation of [CLS]) is used as the input feature to reflect how well an image matches with text. It gives a useful signal to measure intra-modal and inter-modal interaction.

\[
(\varepsilon_1, \varepsilon_2) = FFN(V_{CLS})
\]

(5)

- The self-attention weights are learned directly as two parameters with specified initial values:

\[
(\varepsilon_1, \varepsilon_2) = \text{nn.Parameter}(\text{init\_value}_1, \text{init\_value}_2)
\]

(6)

For each layer of the Transformer block, we introduce a separate set of the two parameters \( \varepsilon_1 \) and \( \varepsilon_2 \) and learn the best parameters for different layers, so that the attention strength can be enhanced or weakened at a layer level. In this way, we can better adjust the attention weight of the intra-modal and inter-modal attention at each layer, since at different layers the focus on intra-modal or inter-modal attention may be different.

3.2.3 VLP with Modality Adaptive Attention. For each class of visual feature (i.e., region, grid, and patch) described in Section 3.1.1, we pre-train a VLP network and fuse it with the textual feature by adopting the novel modality adaptive attention mechanism, respectively. As a result, we obtain one pre-trained model for each class of feature, i.e., Region-VLP, Grid-VLP, and Patch-VLP. Besides, we also pre-train an additional enhanced variant with the fusion feature, namely Fusion-VLP.

**Fusion-VLP.** To capture the multi-granularity visual semantics of image and obtain diverse visual feature representation, we adopt a simple VLP variant, namely Fusion-VLP, which fuses different kinds of image features (region, grid, and patch) by concatenating them together. It is then combined with the text embedding features as input to the cross-modal Transformer with modality adaptive attention. Furthermore, to improve the efficiency of training, we employ a
two-stage strategy to pre-train Fusion-VLP, which first trains the region-grid fusion model initialized with Region-VLP, and then continues to train the Fusion-VLP model based on region-grid fusion model.

*Pre-training Tasks.* The pre-training tasks of the three types, i.e., language, vision, and cross-modality, are introduced in the pre-training stage, following LXMERT [74].

- **Masked LM Prediction.** The task setup is basically the same as in BERT [16]. The masked words are predicted by exploiting visual modality which helps to resolve ambiguity.
- **Masked Object Prediction.** Similarly, the vision side is pre-trained by randomly masking objects. In particular, 15% of image objects are randomly masked, and the model is then asked to predict properties of these masked objects with the output object representations $O^L$.
- **Image-Text Matching (ITM).** This task randomly samples 50% of mismatched image-text pairs and 50% matched pairs, and trains a classifier to predict whether an image and a sentence match with each other on the representation.
- **Image Question Answering (QA).** The image question answering task is cast as a classification problem where the model is pre-trained with image QA data as in LXMERT [74]. A classifier is then built on top of the representation $h^L_{CLS}$ in the model.

For region-based feature, the Region-VLP model is pre-trained with all the four pre-training tasks as in LXMERT [74], and the four objectives are jointly optimized with equal weights. For grid-based feature, the Grid-VLP model is pre-trained with the pre-training tasks except *masked object prediction*, since the *grid feature* does not capture explicit object-level semantics. Besides, to accelerate the pre-training process, a random sampling strategy is adopted to dynamically sample 100 image grids for each image following PixelBERT [29]. The *masked object prediction* task is also removed for Patch-VLP and Fusion-VLP.

During fine-tuning, the complete region/grid/patch features are used to retain all the extracted visual information. The hidden state $h^L_{CLS}$ of the last layer is used for cross-modality calculation.

### 3.3 Knowledge-guided Model Integration

Traditional VQA methods either design complicated cross-modal attention methods or improve vision-language pre-training from feature level, data level, and model size, but they seldom consider introducing extra expert knowledge for improving. Our method proposes to address the practical VQA questions at an expert level and design a knowledge-guided framework from expert mining to heterogeneous expert integration. Due to the complexity of the VQA task, there exist questions that are difficult to address by combining the diverse feature representation and V&L pre-training. To address these questions and enable the model to evolve constantly, we further propose a knowledge-guided integration framework, where a clustering-based knowledge mining method is first used to identify the key types of the questions, then a two-step model integration is conducted to obtain the final prediction result, as shown in Figure 5. Starting from a pre-trained V&L model (the base Vision Understanding Expert in our case), a knowledge mining module is introduced to automatically discover the types of the questions that are not well-addressed by the Vision Understanding Expert, such as text-reading questions and clock-reading questions. Then, we introduce two extra expert modules that are specially designed for these questions: Text Reading Expert and Clock Reading Expert, respectively. Different from simply ensembling all the results of these expert models, we first adopt a voting-based ensemble method to integrate the prediction results of homogeneous models from each expert module, and then a MoE-based routing method is used to combine the results of the three heterogeneous expert modules. The design of both the two steps of integration is guided by the expertise knowledge mined from the practical VQA data.
Fig. 5. Illustration of Knowledge-guided Model Integration. In Step 1, we ensemble different numbers of models for each expert so that the performance of each expert on the development set is almost saturated.

3.3.1 Knowledge Mining. On top of the comprehensive study of diverse feature representation and specific design of cross-modal interaction, we propose a continual learning framework to boost the power of the pre-trained V&L model. It includes two stages: (1) identify new sub-tasks which require extra knowledge to learn; and (2) learn expert models for these sub-tasks with the knowledge collected from domain experts or internet.

To identify new sub-tasks, we adopt a clustering-based method, which considers the low-confidence examples from a base model and discovers groups of these examples by their similarity to form new sub-tasks. Given a base model $M$ (i.e., Vision Understanding Expert in our case), we first collect examples which the model $M$ is difficult to give correct answers with high confidence. The model is unable to address these examples well with existing knowledge, which calls for specialized expert models with extra knowledge to handle them. Specifically, given an example $t$, the base model $M$ is designed to give a prediction with confidence score $s$. The output score on the predicted label of the Vision Understanding Expert is used to calculate this score $s$. The examples with low confidence scores ($s < 0.1$) thus indicate the cases that the base model finds difficult to handle. Then, the typical clustering algorithm K-Means [54] is used to partition the set of these low-confidence examples into sub-task clusters. Under our V&L circumstances, clustering is conducted on both the textual and visual content of examples. Therefore, the cross-modal representation of [CLS] in the last layer of the Transformer are used as input to the clustering algorithm.

Clustering the low-confidence examples allows us to identify new sub-tasks. In the VQA task, the clustering discovers the two types of visual questions (text-reading questions and clock-reading questions) that need OCR ability and clock-reading ability to deal with, respectively. Since the existing Vision Understanding Expert is incapable of resolving the two sub-tasks well, two extra expert modules: Text Reading Expert and Clock Reading Expert, are trained to deal with the low-confidence examples.

3.3.2 Text Reading Expert. Text-reading VQA is an important VQA sub-task, which focuses on the questions that require to read the textual content shown in an input image. The existing models
on the VQA dataset perform classification with frequent answers as labels. This classification modeling does not work well on text-reading samples where the answers are often not frequent enough to be included in the label set. Therefore, the StructuralLM model [39], a specially designed deep LM that aims to capture structure information from texts, is utilized on these text-reading samples to extract answers from the text recognized in images by OCR. StructuralLM introduces cell-level 2D-positional embeddings and a new pre-training objective that classifies cells’ positions. The pre-trained StructuralLM model is adopted to the text-reading VQA samples in the following way.

In order to adapt StructuralLM to our VQA scenario, we fine-tuned a pre-trained StructuralLM with text-reading samples. In particular, an OCR tool is first used to recognize text and serialize the cells (bounding boxes) from top-left to bottom-right in images. Each image is represented as a sequence of cells \( \{c_1, \ldots, c_n\} \), each of which contains a sequence of words \( c_i = \{w_{1i}, \ldots, w_{mi}\} \). A separator \([SEP]\) is added between every two bounding boxes to separate them, which gives an input sequence \( \{q_1, \ldots, q_e, [SEP], c_1, [SEP], c_2, \ldots, [SEP], c_n\} \). A token-level span prediction classifier is then built upon the token representation to perform an extractive QA task, as widely-used for machine reading comprehension [9, 62]. Finally, the added separator is removed from the predicted answer span.

3.3.3 Clock Reading Expert. With the powerful VQA features aforementioned, many questions can get satisfactory answers. However, it still suffers from reading precise time from clocks, as clock reading requires specific prior knowledge. Hence, a clock reading expert is introduced to address such kind of problems. The clock reading expert consists of a clock detector and a clock reader. The clock detector is used to detect clocks in images, which is essentially an object detection task. The Cascade-RCNN [6] is used as the backbone network for the clock detector. A binary classification loss and a bounding box regression loss are applied for training as the standard detection framework does [6]. The detected bounding boxes from the clock detector are fed into the clock reader, which reads the precise time in the clocks. Resnet50-IBN [57] is adopted as our clock reader backbone, and two specific branches are introduced for hour and minute prediction, respectively. The clock reading is modeled as both a classification task as well as a regression task. The training details of clock reading expert can be found in Appendix B.

3.3.4 Vision Understanding Expert. Vision-and-Language Pre-training (VLP) models with different visual features can help capture diverse visual semantics from images, which facilitates deep vision-and-language understanding. Therefore, we train multiple VLP models with different visual features, and all these homogeneous VLP models contribute to our Vision Understanding Expert.

3.3.5 Two-step Model Integration.

Step 1: Voting-based Ensemble with Homogeneous Models. For each expert module, we enhance the ability of the expert by ensembling multiple homogeneous models since the prediction scores from them can be directly compared. Since the abilities required to address the corresponding tasks for each expert are different, we ensemble more models for the expert that requires complicated vision and language understanding. Specifically, for the Vision Understanding Expert, to further improve the vision understanding ability, we use a diverse feature ensemble method, which ensembles multiple VLP models with different types of visual features: region feature, grid feature, and patch feature. For each kind of feature, the corresponding VLP models are trained separately. A simple maximum voting strategy is then utilized to ensemble the different VLP models based on

\footnote{We ensemble different numbers of models for each expert so that the performance of each expert on the development set is almost saturated.}
the prediction scores. Our experiments demonstrate the advantage of the diverse feature ensemble over a single class of visual features.

**Step 2: MoE-based Routing with Heterogeneous Experts.** To better integrate the prediction results from the heterogeneous experts, we further design a Knowledge-guided MoE (K-MoE) method for best expert selection. The methodology of Mixture of Experts (MoE) [31, 68] essentially decomposes a task into sub-tasks, and develops an expert on each sub-task. A gating network is then used to coordinate multiple experts for task completion. We follow the recent work of Switch Transformers [18], and adopt the simple routing strategy that the gating network routes to only a single expert. We use the simplest form to preserve model quality and reduce routing computation. In our framework, the VQA task can be decomposed into three sub-tasks according to the analysis of the visual questions via proposed knowledge mining. A multi-layer perception network is trained as the gating network, which performs three-class classification to determine which expert to choose for each instance.

Given an expert $M_t$ for sub-task $t$, the expert will give an answer and we compute a reward score $s_t$ between the predicted answer and the human annotated labels using Equation (9). The reward score $s_t$ is used as supervision for training, where the network is trained to route each instance to its best-match expert. At training time, we propose to maximize the Binary Cross Entropy (BCE) loss $L$ as follows:

$$L_{MoE} = \sum_t s_t \log \hat{s}_t + (1 - s_t) \log (1 - \hat{s}_t)$$  \hspace{1cm} (7)

where $s_t$ denotes the ground-truth reward score of sub-task $t$, $\hat{s}_t$ stands for the prediction score of the MoE network. At test time, we choose the single routed expert with the maximum prediction score $\hat{t} = \arg \max \hat{s}_t$. The prediction score $\hat{s}$ is calculated using a Multi-Layer Perception network (MLP) as follows:

$$\hat{s} = W_3 (\tanh(W_2 \tanh(W_1 x + b_1)) + b_2) + b_3$$  \hspace{1cm} (8)

where $x$ is the input feature and $W_i, b_i$ are the learnable parameters.

The following features are derived for training the gating network:

- **Each expert’s confidence**: for Vision Understanding Expert, the maximum prediction score is used for confidence score. For Text Reading Expert and Clock Reading Expert, the output score is used for confidence score, and if an image does not have text or any clock, the score is set to $-1$.
- **Question type**: A three-class classifier is trained to predict whether a question is asked about text reading, clock reading or vision understanding. To train the classifier, OCR-labeled data is collected from TextVQA [71] & STVQA [5], and clock-labeled data from the VQA dataset by retrieving the keywords *clock* and *what time*. Other cases from VQA data are sampled as negative samples by the ratio of 1 : 2. The prediction scores of these three classes are used as the input features.

Even though the current process is manually designed, in the future, we will be exploiting techniques that allow us to automatically discover subset of challenging cases, and incrementally add more experts to address the discovered cases.

4 EXPERIMENTS

4.1 Data

**Pre-training Data.** The same in-domain data is used as in LXMERT [74] for pre-training. It consists of the image caption data from MS COCO [46], Visual Genome [36], image question
Table 1. VQA Data Statistics

|        | Images | Questions | Yes/No | Number | Other | Answers |
|--------|--------|-----------|--------|--------|-------|---------|
| Training   | 80K    | 443K      | 169K   | 58K    | 219K  | 4.4M    |
| Validation | 40K    | 214K      | 81K    | 28K    | 106K  | 2.1M    |
| Test       | 80K    | 447K      | –      | –      | –     | –       |

Table 2. Text-reading VQA Data Statistics

| Dataset    | Images | Questions |
|------------|--------|-----------|
| VQA-Subset | 20K    | 21K       |
| TextVQA    | 25K    | 39K       |
| ST-VQA     | 19K    | 26K       |

answering data from VQA 2.0 [4], GQA balanced version [30], and VG-QA [99]. The total amount of the dataset is 9.18M image-and-sentence pairs on 180K distinct images. Also, additional out-of-domain data from Conceptual Captions [67] is used for model pre-training, which consists of about 3M image-text pairs on 3M images.

For Text Reading Expert, the StructuralLM [39] is used as the base model, which is pre-trained on the IIT-CDIP Test Collection 1.0 [38]. It is a large-scale scanned document image dataset containing more than 6 million documents, with more than 11 million scanned document images.

Fine-tuning Data. **Visual Question Answering (VQA)** is a dataset containing open-ended questions about images [4]. These questions require understanding of vision, language and commonsense knowledge to answer. It contains a large number of labeled question-image-answer triplets with 10 human annotators for each question. The detailed statistics for VQA training/validation/test data splits is shown in Table 1.

For training Text Reading Expert, three text-reading VQA datasets are used, including a subset of VQA data [4], TextVQA [71], and ST-VQA [5]. A classification model is trained to extract text-reading samples from VQA data. The questions of TextVQA and ST-VQA are treated as positive samples, and the questions on images without text in VQA are treated as negative samples. The detailed statistics for the three text-reading VQA datasets are shown in Table 2.

For training Clock Reading Expert, the images are collected from two sources. One is from open-access datasets. Specifically, a total of 4,863 images are collected from COCO2017\(^2\) and 2,691 images from ImageNet for clock labeling, both of which are widely used open-access datasets. Annotators are required to give the bounding boxes and the precision time of clocks in images. After labeling, 4,236 and 3,271 valid clock bounding boxes are obtained from COCO2017 and ImageNet, respectively. 785 clock bounding boxes are randomly sampled from COCO2017 images for validation. The other source is Internet images. To further increase the generalization and capacity of our clock reader, 2,878 images from internet with various clocks are collected. After careful annotation, 2,314 valid clocks are obtained. Note that this data is only used for the training of the clock reader.

Evaluation Metric. Following [4], an evaluation metric robust to inter-human variability is used in phrasing the answers:

\[
\text{Acc}(\text{ans}) = \min \left\{ \frac{\# \text{ human that said } \text{ans}}{3}, 1 \right\}
\]  \hspace{1cm} (9)

\(^2\)The COCO images used in the VQA test set are left unlabeled and excluded from our training data.
In order to be consistent with “human accuracies”, machine accuracies are averaged over all 10-choose-9 sets of human annotators.

### 4.2 Experimental Setup

**VLP.** The maximum sequence length for the sentence is set as 20. For the VLP models, the pre-trained Transformer encoder with 12 layers is used as our base architecture, and the one with 24 layers as the large architecture. The basic settings of the Transformer are the same as BERT [16], and the Transformer encoder is initialized with StructBERT [82] for its good performance. For the method of modality adaptive attention, the two learnable parameters are initialized with $\text{init}\_\text{value}_1 = 1.0$ and $\text{init}\_\text{value}_2 = L_s/L$, where $L$ is the number of Transformer layers and $L_s$ is the corresponding layer number. The base model is pre-trained with a total batch size of 512 for 30 epochs on 8 A100 GPUs and the AdamW optimizer with the initial learning rate of $1e^{-4}$. The 24-layer large architecture is pre-trained with the total batch size of 512 on 8 A100 GPUs. To deal with over-fitting, two-stage pre-training strategy is employed as in LXMERT [74]. Specifically, the model is first pre-trained without the question answering task with the initial learning rate of $5e^{-5}$ for 20 epochs, and then pre-trained with all the tasks together with the initial learning rate of $2e^{-5}$ for another 10 epochs. The detailed settings for the three VLP methods are listed as below:

- **Region-VLP:** The detection model is used in VinVL [95] to detect objects and extract region features. It is a large-scale object-attribute detection model based on the ResNeXt-152 C4 architecture. 100 objects are retained for each image to maximize the pre-training compute utilization by avoiding padding.

- **Grid-VLP:** It follows the basic settings in Grid-VLP [88]. ResNeXt is chosen to be the visual encoder with different sizes [86] as in [29, 33]. The shorter side of every input image is resized to 600, and the longer side is limited to at most 1,000. A fixed number of 100 grids are randomly selected each time during pre-training.\(^3\)

- **Patch-VLP:** It uses the visual Transformer encoder of the Swin detector [49] and CLIP [61]. The ViT-B/16 pre-trained model is chosen, which has 12 Transformers layers with input patches of size $16 \times 16$. Every input image is resized to $256 \times 256$ as CLIP does, resulting in $16 \times 16 = 256$ patches.

- **Fusion-VLP:** It takes the same settings as Region-VLP, Grid-VLP, Patch-VLP to pre-train with different kinds of image features.

**Fine-tuning on VQA.** Following [2], our architecture treats VQA as a multi-class classification task by picking an answer from a shared set of 3,129 answers. The hidden state of $h^{L}_{\text{CLS}}$ is used to map the representation into 3,129 possible answers with an additional MLP layer. The model is trained with a binary cross-entropy loss on the soft target scores. The pre-trained models are fine-tuned based on the three classes of features on the VQA training data for 3 epochs with the batch size of 32, and the BERT Adam optimizer is employed with the initial learning rate of $1e^{-4}$ for base models and $2e^{-5}$ for large models. At inference, a softmax function is used for prediction.

**Text Reading Expert.** The text reading expert follows the basic settings in StructuralLM [39] and uses the pre-trained StructuralLM-large as the backbone model. In particular, StructuralLM is pre-trained with a batch size of 16 for 50K steps. The question tokens and the OCR tokens of an image are concatenated as an input sequence, of which the maximum length is set as 128. For fine-tuning, the three kinds of text-reading VQA datasets are merged and split with 10-fold cross-validation. The StructuralLM is fine-tuned with the total batch size of 16 for 4 epochs, and the

\(^3\)We also tested with 64 and 128 selected grids. It did not lead to significantly different results.
Table 3. VQA Challenge Leaderboard

| Models                        | Overall | Yes/No | Number | Other |
|-------------------------------|---------|--------|--------|-------|
| Human                         | 80.83   | 95.48  | 81.29  | 67.97 |
| LXMERT ([74])                 | 74.34   | 89.45  | 56.69  | 65.22 |
| MCAN ([93])                   | 75.23   | 90.36  | 59.17  | 65.75 |
| VILLA ([21])                  | 75.85   | 91.30  | 59.23  | 66.20 |
| BGN ([22])                    | 75.92   | 90.89  | 61.13  | 66.28 |
| InterBERT ([45])              | 76.10   | 91.67  | 59.24  | 66.40 |
| GridFeat+MoVie ([33])         | 76.29   | 90.81  | 61.53  | 67.04 |
| VinVL ([95])                  | 77.45   | 92.38  | 62.55  | 67.87 |
| ROSITA ([13])                 | 78.34   | 92.66  | 63.24  | 69.33 |
| UNIMO ([42])                  | 78.40   | 93.10  | 63.06  | 69.12 |
| VQA Challenge 2021 winner     | 79.34   | 93.28  | 65.36  | 70.40 |
| SimVLM ([85])                 | 80.34   | 93.29  | 66.54  | 72.23 |
| **AliceMind-MMU**             | **81.26** | **93.55** | **72.01** | **72.67** |

AdamW optimizer is employed with the initial learning rate of 3e-5. Accuracy and ANLS (Average Normalized Levenshtein Similarity) are used as the metrics to evaluate the text reading expert.

Clock Reading Expert. The clock detector of the clock reading expert is trained following the basic settings of Cascade-RCNN [6]. The clock reader is trained with the batch size of 96 in 2 GPUs, and the initial learning rate is set as 0.02. It is trained for 150 epochs with the learning rate multiplied by 0.1 at 90-th and 120-th epochs. The data augmentation pipeline consists of 256×256 random resized cropping, random color jittering, random gray-scale conversion, Gaussian blurring and random rotation within ±45°.

Vision Understanding Expert. The vision understanding expert ensembles 46 models in total, including 14 Region-VLP models, 21 Grid-VLP models, 4 Patch-VLP models, and 7 Fusion-VLP models. Simple maximum voting is adopted to ensemble all the models based on their prediction scores.

Mixture of Experts. The MoE adopts Multi-layer Perceptron (MLP) as the gating network to determine experts for given questions. The MLP has two hidden layers of 100 neurons and 50 neurons, respectively. It uses tanh as the activation function, and the Adam optimizer with the initial learning rate of 1e-3. The network is trained for 5 epochs with the batch size of 256.

4.3 Comparison Methods

To demonstrate the effectiveness of AliceMind-MMU, we compare it with the top-ranked VQA submissions on VQA challenge leaderboard as shown in Table 3, which contains both single and ensemble models. Besides, to further examine the effectiveness of our method on Vision-Language Pre-training (VLP), we also compare with the state-of-the-art single VLP methods on VQA, which have been widely used in multi-modality field and dominated the image-text related tasks, as shown in Table 4. The details of the compared baseline models can be found in Appendix D.

4.4 Main Results

Table 3 presents our main results compared with all the previous public and unpublic best results on the VQA Challenge Leaderboard. From the results, it can be observed that: (1) Our VQA architecture AliceMind-MMU represents the first to achieve human parity on VQA Challenge.
Table 4. Performance Comparison with other Single Models

| Models          | Feature Type | BASE     | LARGE    |
|-----------------|--------------|----------|----------|
|                 |              | Params   | Test-dev | Test-std |
|                 |              |          |          |
| VLBERT ([72])   | Region       | 110M     | 71.16    | –        |
| GilBERT ([26])  | Region       | 110M     | –        | 71.45    |
| UNITER ([11])   | Region       | 110M     | 72.70    | 72.91    |
| OSCAR ([43])    | Region       | 110M     | 73.16    | 73.44    |
| UNIMO ([42])    | Region       | 110M     | 73.79    | 74.02    |
| VinVL ([95])    | Region       | 110M     | 75.95    | 76.12    |
| ViLBERT ([50])  | Region       | 221M     | 70.55    | 70.92    |
| 12-in-1 ([51])  | Region       | 221M     | 73.15    | –        |
| LXMERT ([74])   | Region       | 183M     | 72.42    | 72.54    |
| ERNIE-VIL ([91])| Region       | 250M     | 73.18    | 73.36    |
| PixelBERT ([29])| Grid         | 170M     | 74.45    | 74.55    |
| E2E-VLP ([87])  | Grid         | 94M      | 73.25    | 73.67    |
| ViLT ([35])     | Patch        | 110M     | 71.26    | –        |
| Region-VLP      | Region       | 110M     | 76.25    | –        |
| Grid-VLP        | Grid         | 110M     | 76.50    | –        |
| Patch-VLP       | Patch        | 110M     | 75.83    | –        |
| Fusion-VLP      | Region+Grid+Patch | 110M | 77.15 | 77.21 |

Leaderboard outperforming all the previous state-of-the-art methods, which demonstrates the effectiveness of our framework. (2) With regard to a breakdown of performance on different question types, AliceMind-MMU performs much better on the “Other” type than a human does, and gives comparable results on “Yes/No” questions. AliceMind-MMU performs worse than a human does on type “Number” for the two reasons: (a) in the “Number” type, there are many questions about reading OCR text, which are easier for a human to answer; and (b) there are many object counting questions that are more difficult for AliceMind-MMU to answer.

Table 4 presents the detailed results of our single VLP models compared with other state-of-the-art methods. From the results, it is observed that: (1) the proposed VLP model outperforms the others on each kind of visual feature (region/grid/patch), respectively. It demonstrates the effectiveness of the proposed cross-modal interaction with modality adaptive attention mechanism. (2) The methods with self-attention on patch feature perform worse than the region-based and grid-based methods do. There are two weaknesses of patch-based methods: (a) the visual semantic information is not well-captured in existing patch-based VLP methods. How to inject visual semantics into patch representation remains largely unexplored; (b) the image-text pre-training data is not enough for large-scale patch-based pre-training; (3) Fusion-VLP gives the best performance by fusing all the three classes of visual features as input, which validates the effectiveness of diverse feature representation.

4.5 Model Analysis by Modules

Hierarchical Framework. Here we present the ablation study to assess the importance of three key modules of the hierarchical framework on the VQA test-dev set. The results shown in Table 5 indicate that: (1) The integration framework based on diverse visual features achieves much better performance than the one based on only one kind of feature. We remove diverse visual features and keep only the region feature with Region-VLP models. The other settings are kept the same.
This ablation results in a significant drop from 81.26 to 79.95, demonstrating the importance of the diverse visual features in model integration; (2) To study the effect of the modality adaptive attention, we remove the modality adaptive attention mechanism in all the Region-VLP, Grid-VLP, Patch-VLP, and Fusion-VLP methods, which results in an obvious drop from 81.26 to 80.51; (3) Finally, we study the significance of knowledge-guided model integration. We remove both the Text Reading Expert and Clock Reading Expert, and just adopt a simple model ensemble on the VLP models based on diverse visual features. Over 1.5% of performance degradation resulted from ablating knowledge-guided model integration clearly demonstrates the power of the novel knowledge-guided integration method.

### Visual Feature Importance

Here we present the ablation study to assess the importance of different visual features for VLP on the VQA test-dev set. The results shown in Table 6 indicate that: (1) The VLP methods based on region and grid features achieve better performance than the ones based on patch feature, as stated in Section 4.4. When examining by individual question types, Region-VLP performs better on the “Number” type, while Grid-VLP does better on the “Yes/No” and “Other” types. The difference can be attributed to the fact that region feature captures more local information of an image at the object level, and thus is more effective in addressing the visual counting problem by identifying local objects in an image. On the other hand, grid feature captures globally visual context in an image, which helps to answer the “Yes/No” and “Other” questions; (2) by combining the three classes of features in the way of early fusion, Fusion-VLP performs the best among all the single models. It shows that the different kinds of features can complement well with each other.

### Modality Adaptive Attention

Here we present the ablation study to assess the importance of the modality adaptive attention mechanism on the VQA test-dev set. We conduct the ablation on each kind of VLP model, respectively. The 24-layer single VLP model is used as the baseline model, which is pre-trained and fine-tuned based on the original Transformer. The ablation applies the same pre-training and fine-tuning settings, and only modifies the self-attention block with the two ways of modality adaptive attention stated in Section 3.2.2. From Table 7, it can be seen that all the single VLP models of different visual features can obtain a consistent improvement of 0.4~0.6 points after adding the modality adaptive attention. In addition, the improvement is the most significant in Fusion-VLP, where the vision feature is more comprehensive and complicated, and the semantic gap between textual and visual features is largest. By adjusting the intra-modal

---

**Table 5. Ablation Study of Hierarchical Framework on the VQA Test-dev**

|                        | Overall | Yes/No | Number | Other |
|------------------------|---------|--------|--------|-------|
| AliceMind-MMU          | 81.27   | 93.67  | 72.55  | 72.60 |
| w/o diverse visual features | 79.95  | 92.91  | 71.59  | 70.69 |
| w/o modality adaptive attention | 80.51  | 93.3   | 72.22  | 71.39 |
| w/o knowledge-guided model integration | 80.05  | 93.67  | 66.78  | 71.40 |

**Table 6. Ablation Study of Visual Features on VQA Test-dev**

|         | Overall | Yes/No | Number | Other |
|---------|---------|--------|--------|-------|
| Fusion-VLP | 77.76  | 91.99  | 64.42  | 68.59 |
| Region-VLP | 77.17  | 91.62  | 63.69  | 67.84 |
| Grid-VLP  | 77.13   | 92.20  | 59.99  | 68.15 |
| Patch-VLP | 76.32  | 91.68  | 58.01  | 67.36 |

Visual Feature Importance. Here we present the ablation study to assess the importance of different visual features for VLP on the VQA test-dev set. The results shown in Table 6 indicate that: (1) The VLP methods based on region and grid features achieve better performance than the ones based on patch feature, as stated in Section 4.4. When examining by individual question types, Region-VLP performs better on the “Number” type, while Grid-VLP does better on the “Yes/No” and “Other” types. The difference can be attributed to the fact that region feature captures more local information of an image at the object level, and thus is more effective in addressing the visual counting problem by identifying local objects in an image. On the other hand, grid feature captures globally visual context in an image, which helps to answer the “Yes/No” and “Other” questions; (2) by combining the three classes of features in the way of early fusion, Fusion-VLP performs the best among all the single models. It shows that the different kinds of features can complement well with each other.

Modality Adaptive Attention. Here we present the ablation study to assess the importance of the modality adaptive attention mechanism on the VQA test-dev set. We conduct the ablation on each kind of VLP model, respectively. The 24-layer single VLP model is used as the baseline model, which is pre-trained and fine-tuned based on the original Transformer. The ablation applies the same pre-training and fine-tuning settings, and only modifies the self-attention block with the two ways of modality adaptive attention stated in Section 3.2.2. From Table 7, it can be seen that all the single VLP models of different visual features can obtain a consistent improvement of 0.4~0.6 points after adding the modality adaptive attention. In addition, the improvement is the most significant in Fusion-VLP, where the vision feature is more comprehensive and complicated, and the semantic gap between textual and visual features is largest. By adjusting the intra-modal
Table 7. Ablation Study of Modality Adaptive Attention on VQA Test-dev

|          | Overall | Yes/no | Number | Other |
|----------|---------|--------|--------|-------|
| Fusion-VLP (Baseline) | 77.13  | 91.64  | 63.73  | 67.74 |
| + Modality Adaptive Attention (FFN) | 77.62  | 91.75  | 64.02  | 68.23 |
| + Modality Adaptive Attention (Param) | **77.76**  | 91.99  | **64.42**  | **68.59**  |
| Region-VLP (Baseline) | 76.75  | 91.28  | 63.31  | 67.34 |
| + Modality Adaptive Attention (FFN) | 77.09  | 91.58  | 63.54  | 67.74 |
| + Modality Adaptive Attention (Param) | 77.17  | 91.62  | 63.69  | 67.84 |
| Grid-VLP (Baseline) | 76.65  | 92.15  | 59.62  | 67.94 |
| + Modality Adaptive Attention (Param) | **77.13**  | **92.20**  | **59.99**  | **68.15**  |
| Patch-VLP (Baseline) | 75.83  | 91.38  | 55.78  | 67.12 |
| + Modality Adaptive Attention (Param) | 76.32  | 91.68  | 58.01  | 67.36 |

and inter-modal attention weight dynamically, the proposed modality adaptive attention can better alleviate the cross-modal semantic gap problem. Among the two different ways, the one with two learnable parameters performs slightly better than the other one. The reason may lie in: (1) learning two unrestricted parameters for each layer allows for more parameter freedom, so as to better align cross-modal semantics, and (2) the discrepancy between pre-training and fine-tuning, where the representation of [CLS] is learnt to model the semantic relation between a caption and an image during pre-training, while it is repurposed for question answering in fine-tuning.

Knowledge-guided Model Integration. Figure 6 illustrates the clustering result. We choose the number of the clusters as 5, which gives the best performance on our quantitative test. For each cluster, we provide the probable identified topic, the most frequently asked question types, the proportion of the examples in this cluster, the key-word cloud, as well as two showcase examples. More examples can be found in Appendix C. From the results, we can see that the proposed knowledge mining can actually mine certain meaningful topic clusters, where similar examples are clustered together. For example, Cluster 1 is about asking questions about time and clocks. Cluster 2 is about counting problems. Cluster 3 is about reading texts from the wall or the clothes. Cluster 4 is about reading number texts on the vehicles. We found that there are two new tasks: clock-reading task (Cluster 1) and text-reading task (Cluster 3,4), both of which require specific prior knowledge. We also use a three-class classifier in Section 3.3.5 to classify the filtered candidate examples, so as to measure the consistency between the clustering result and classification result. Figure 7 gives t-SNE visualization of the clustering result and classification result. From the result, we can see high consistency between the clustering and classification result on the measured topics. The knowledge mining method can properly separate part of the clock-reading examples and OCR-reading examples from the other examples, although for the OCR-related examples there still exist limited examples mixed up in the common vision category.

To measure the consistency of the clustering result to the classification labels, we also provide detailed quantitative analysis on different clustering methods. We manually build a three-label classifier (OCR, clock, and vision) with 95% accuracy as in Section 3.3.5 and apply it to evaluate the consistence of each cluster. We project each cluster to the corresponding label heuristically. For example, in Figure 6, Cluster 1 is assigned to clock label, Cluster 3 and Cluster 4 are to OCR label, and Cluster 5 is to vision label. We then compare the assigned label of each cluster to that of the classification label (95% accuracy). We use accuracy, macro-precision, macro-recall and macro-f1 to measure how consistent the compared label in each cluster is. As listed in Table 8, K-Means ($K = 5$) achieves the best performance with 0.8448 accuracy and 0.8739 macro-F1, which shows that the clustering result is highly consistent with the assumed classification labels on OCR/clock/vision.
| Cluster | Topic (Proportion) | Question Types (First Three Words) | Word Cloud | Examples |
|---------|-------------------|-----------------------------------|------------|----------|
| 1       | Time & Clock (10.0%) | 1. What time is 59.59%  
2. What time does 23.14%  
3. What is the 11.56%  
4. What time was 0.88%  
5. What does the 0.62% | What time is?  
What time is it on the clock? |
| 2       | Counting (12.0%) | 1. How many people 12.81%  
2. How many windows 6.56%  
3. How many different 2.65%  
4. How many cars 2.41%  
5. How many animals 1.88% | How many people are in the picture?  
How many cows are there? |
| 3       | Reading text from the wall or the clothes (26.4%) | 1. What is the 25.81%  
2. What does the 20.65%  
3. What is written 5.23%  
4. What are the 3.37%  
5. Where is the 2.45% | What is the sign showing?  
What does the sign say? |
| 4       | Reading number texts on the vehicles (15.1%) | 1. What is the 28.44%  
2. What number is 27.88%  
3. How old is 3.48%  
4. What are the 3.45%  
5. What numbers are 2.34% | What is the number on the front of the train?  
What number is on the train? |
| 5       | Others (36.5%) | 1. What is the 13.08%  
2. Where is the 10.39%  
3. What kind of 7.61%  
4. What is on 4.19%  
5. What type of 4.01% | What is the cat sitting on?  
Where is the cat’s head? |

Fig. 6. Illustration examples of clustering results.

Fig. 7. The t-SNE visualization of clustering results. Figure (a) shows the clustering results and the label 1/2/3/4/5 is cluster id. Figure (b) shows the classification results and the label ocr/clock/vision is classification label. The classifier is manually built and the accuracy of it is 95.0%.
Achieving Human Parity on Visual Question Answering

Table 8. Quantitative Analysis on the Clustering Results of Different Clustering Methods

| Method                      | Acc    | P      | R      | Macro-F1 |
|-----------------------------|--------|--------|--------|----------|
| DBSCAN (eps = 0.5)          | 0.1544 | 0.4605 | 0.3861 | 0.1466   |
| K-Means (K = 3)             | 0.4969 | 0.6487 | 0.662  | 0.6163   |
| K-Means (K = 4)             | 0.7894 | 0.8239 | 0.8219 | 0.8195   |
| K-Means (K = 5)             | 0.8448 | 0.8659 | 0.8898 | 0.8739   |
| K-Means (K = 6)             | 0.8443 | 0.8668 | 0.8918 | 0.8740   |

Table 9. Ablation Study of Knowledge-guided Model Integration on the VQA Test-dev

| Expert Importance                          | Overall | Yes/No | Number | Other |
|--------------------------------------------|---------|--------|--------|-------|
| AliceMind-MMU                              | 81.27   | 93.67  | 72.55  | 72.60 |
| w/o Text Reading Expert                    | 80.26   | 93.67  | 68.91  | 71.31 |
| w/o Clock Reading Expert                   | 81.00   | 93.67  | 69.75  | 72.69 |
| Integration Method                         |         |        |        |       |
| Vision Understanding Expert (Baseline)     | 80.05   | 93.67  | 66.78  | 71.40 |
| w/ Knowledge-guided MoE                    | 81.27   | 93.67  | 72.55  | 72.60 |
| w/ One-step Voting-based Ensemble          | 80.37   | 93.67  | 69.05  | 71.52 |

The ablation study of knowledge-guided model integration is shown in Table 9. We examine two key points in knowledge-guided model integration: (1) the importance of each expert module, and (2) the importance of knowledge-guided MoE for two-step model integration. With only vision understanding expert, the model gives a strong performance of accuracy 80.05 on the VQA Test-dev set. To study the importance of each expert, we ablate text reading expert and clock reading expert, respectively. The text reading expert proves to be critical with significant drops on the overall accuracy for more than 1% after the ablation. The performance on the "Number" type drops by more than 3.5% after the clock reading expert is removed. The gating network of MoE mimics human who is able to identify domain experts based on the nature of tasks. As shown in Table 9, the traditional One-step Voting-based Ensemble method has a small improvement, because the number of models for different experts are quite different, and the confidence score output by the different experts also cannot be compared directly. A naive ensemble of different models may not integrate the prediction results properly. Our proposed Knowledge-guided MoE can achieve greater improvements, which demonstrate the effectiveness of the proposed method for integrating the prediction results of heterogeneous experts. This knowledge-guided model integration can also be easily extended to incorporate more specialized experts for continual self-evolution.

4.6 Model Analysis by Effectiveness and Efficiency

Effectiveness and Efficiency of VLP models. For the real-world applications, except for the model performance, the training cost and inference speed are both important. As the data volume and model size continually increase, it will consume huge computation resources to train an effective AI model and it is also a trend for the current AI development, especially for the big pre-trained foundation model. It is very important to decrease the training cost of an AI model. Besides, to better serve the online request for real-world deployment, it is also necessary for the AI model to increase the inference speed. Here we first test the training time, inference time and performances of different VLP models as well as the simplest version of our AliceMind-MMU framework. Table 10
Table 10. Effectiveness and Efficiency of Different Single VLP Models and the Simplest Version of AliceMind-MMU Framework on VQA Test-dev

| Model          | Overall Accuracy | # Pre-train Data | Training Time (Stage 1) | Inference Time (Stage 1) | Inference Time (Stage 2) | Inference Time (Total) |
|----------------|-----------------|------------------|-------------------------|--------------------------|--------------------------|------------------------|
| VinVL          | 76.52           | 8.8M             | –                       | 746ms                    | 28ms                     | 774ms                  |
| SimVLM†        | 80.03           | 1.8B             | 49152h                  | –                        | 1540ms                   | 1540ms                 |
| Fusion-VLP     | 77.76           | 12M              | 1440h                   | 940ms                    | 208ms                    | 1148ms                 |
| Region-VLP     | 77.17           | 12M              | 864h                    | 746ms                    | 32ms                     | 778ms                  |
| Grid-VLP       | 77.13           | 12M              | 1008h                   | 148ms                    | 158ms                    | 306ms                  |
| Patch-VLP      | 76.32           | 12M              | 1080h                   | 46ms                     | 150ms                    | 196ms                  |
| Vision Expert* | 79.31           | 12M              | 4392h                   | 940ms                    | 548ms                    | 1488ms                 |
| AliceMind-MMU* | 80.35           | 12M              | 4422h                   | 1512ms                   | 618ms                    | 2130ms                 |

Stage 1: feature extraction, Stage 2: cross-modal fusion. For Vision Expert*, we integrate the results of four diverse models with one for each type of the visual features. For AliceMind-MMU*, we integrate four diverse VLP models as a vision understanding expert, a text reading expert and a clock reading expert with knowledge-guided model integration on the three expert modules. For each example, we only need to extract the three types of visual features for once. † denotes that the inference time of SimVLM is not reported in the paper, and we build the model ourselves to infer the computation cost.

shows that: (1) most of the proposed single VLP models can outperform a state-of-the-art region-based VLP model VinVL with comparable computation cost, which even uses far more training steps and larger object detection data; and (2) The simplest version of AliceMind-MMU with six diverse models can achieve better performance than a recent state-of-the-art model SimVLM (++) with acceptable inference speed, which requires 11 × more training time pre-trained on 100 × more aligned image-text pairs than ours. Furthermore, although the total inference time of AliceMind-MMU is a bit longer than SimVLM, since different models in our framework are separated trained and inferred, AliceMind-MMU can further speed up the training and inference through parallel computing; (3) The diverse Vision Expert with different visual features (Region, Grid, Patch, Fusion) can outperform the single Fusion-VLP remarkably by 1.5%, which proves the importance to capture multi-granularity semantics from the image.

The trend chart w.r.t. number of models. To better understand the effectiveness and efficiency trade-off of the proposed framework, we also conduct a series of experiments to test the training time, inference time, and performances with respect to the number of models used in our framework. The results shown in Figure 8 indicate that: (1) as the number of homogeneous models increases in our framework, the performance also gradually improves and it can improve rapidly at the early stage of adding models especially before six diverse models, and the performance improvement slows down gradually. (2) The training and inference time may also increase linearly as the number of models increases. However, the training cost is still much less than the SimVLM model, which trains a larger model on 100 × more pre-training data than ours. (3) The inference time is longer than SimVLM, especially when the number of used models is large. This is also a limitation of the proposed framework, but we could also improve the online inference time of AliceMind-MMU by removing the heavy region-based models and conducting parallel inference, by concurrent requests to different VLP experts. This can be a trade-off between effectiveness and efficiency for real-world deployment and we will further explore it in the future work.

4We only list the number of pre-training image-text pairs here, and in addition the total pre-training data also contains another 5.4M labeled object detection data.
5For SimVLM, the whole model is trained end-to-end without a separate feature extraction process.
Fig. 8. Analysis of the impact on training time, inference time and performances w.r.t. the number of models on the VQA test-dev set. (The unit of training time is GPU hours on the A100-40G card, and the unit of inference time is ms).

Fig. 9. The distribution of abilities in each answer type.

4.7 VQA Dataset Analysis

This subsection provides a more detailed analysis of our VQA results. To gain an understanding of the types of questions and answers, 1,000 examples are randomly sampled from the validation set for analysis. The 1,000 examples are classified into the categories listed below by manual examination based on the abilities required. The categorization is multi-label in the sense that every example is classified into all applicable categories. Figure 9 provides an estimate of the proportion for each category. Commonsense Knowledge, Relational Reasoning and Object Counting are the top three categories in the overall distribution. Commonsense Knowledge accounts for over 80% of the Yes/No type. In the Number type, Object Counting and Textual Recognition are the two most popular categories compared with the other two types. The type Other has a similar distribution as that of the overall distribution. Figure 10 presents representative examples from each category.

- **Commonsense Knowledge.** This category contains questions inquiring commonsense knowledge from our daily life, such as colors, weather, food, and furniture.
### Visual Recognition
This category requires the ability to acquire specialized knowledge with visual recognition to answer questions in this category.

### Relational Reasoning
This category requires understanding and reasoning over certain relationships of objects in an image, such as positional relationship, comparison relationship, and so on.

### Textual Recognition (OCR)
This category requires the ability to recognize and utilize text together with the positions or visual information in an image (e.g., road signs, ads on a bus).

### Object Counting
This category contains the examples that test the ability of counting objects in an image.

### Clock Reading
This category contains the examples that test the ability of reading a clock.

### Other
This category contains the questions that are ambiguous or cannot be answered based on given images.

#### 4.8 AliceMind-MMU vs. Human

A comparative study of AliceMind-MMU and human on visual question answering has been conducted. Tables 11 and 12 show the overall and per-category performance of AliceMind-MMU and human on the val split, respectively, from which there are the following observations: (i) AliceMind-MMU outperforms human annotators on the two largest categories, Commonsense Knowledge and Relational Reasoning. It shows AliceMind-MMU’s superiority of identifying common scene objects in daily life and leveraging commonsense knowledge such as colors and weather. This result also demonstrates the power of AliceMind-MMU in reasoning over relative positions, such as the left sign on a wall, to answer a spatial reasoning question. Besides, it is surprising that AliceMind-MMU can reason over simple comparison, such as which object is the tallest. (ii) The

| Category                      | Examples                          | Category                      | Examples                          |
|-------------------------------|-----------------------------------|-------------------------------|-----------------------------------|
| Common Sense Knowledge        | ![Common Sense Knowledge Example](image1) | Visual Recognition             | ![Visual Recognition Example](image2) |
| Q: Is there snow on the ground? A: yes | Q: What kind of bear is this? A: grizzly | Q: What type of flowers are those? A: daffodils |
| Object Counting               | ![Object Counting Example](image3)   | Relational Reasoning           | ![Relational Reasoning Example](image4) |
| Q: How many couches? A: 2    | Q: How many trees are in the picture? A: 3? | Q: Which elephant is tallest? A: left |
| Textual Recognition (OCR)     | ![Textual Recognition Example](image5) | Clock                          | ![Clock Example](image6) |
| Q: What number is the player? A: 46 | Q: What does the sign say? A: one way | Q: What time is it? A: 10:20 |

Fig. 10. Representative examples from each category.
Table 11. The Overall Performance of AliceMind-MMU and Human on Val Split

|       | Test-std | Val        | Overall | Yes/no | Number | Other |
|-------|----------|------------|---------|--------|--------|-------|
| VLP   | 81.26    | 79.54      | 92.47   | 70.63  | 72.00  |
| Human | 80.83    | 78.69      | 94.87   | 78.79  | 66.34  |

Table 12. The Performance of AliceMind-MMU and Human by Category

| Category                  | Commonsense Knowledge | Relational Reasoning | Object Counting | Visual Recognition | Textual Recognition (OCR) | Clock Reading | Other |
|---------------------------|-----------------------|----------------------|-----------------|--------------------|---------------------------|--------------|-------|
| Num                       | 767                   | 159                  | 103             | 70                 | 74                        | 7            | 5     |
| VLP                       | 83.60                 | 71.19                | 77.76           | 68.14              | 52.03                     | 86.00        | 70.00 |
| Human                     | 80.04                 | 70.20                | 81.29           | 59.76              | 76.62                     | 60.66        | 49.52 |

questions in the Object Counting category seem rather difficult for AliceMind-MMU to answer. AliceMind-MMU is found to be good at counting a small number (<10) of objects. It would give an incorrect count when encountering a large number of small objects and/or requiring reasoning over them. (iii) AliceMind-MMU significantly surpasses human performance on Visual Recognition which requires specialized knowledge. It is expected that AliceMind-MMU, as a machine learner trained with large data, is skilled in memorizing specialized/professional knowledge with visual recognition, compared with non-professional human annotators. (iv) AliceMind-MMU is more capable of reading time shown in a clock than human, as demonstrated by the result of Clock Reading. On text reading, however, there is still a big gap between AliceMind-MMU and human in recognizing and understanding text in an image, as shown by the result of Textual Recognition. Some research progress has been made on text-reading VQA tasks, such as TextVQA [71].

Figures 11, 12, and 13 present AliceMind-MMU’s predictions together with ground truth on each category for case study. In particular, a couple of representative examples are listed for each category, each containing a question, an image and the answer predicted by AliceMind-MMU. The scores of human annotators and AliceMind-MMU, as well as the top three ground-truth annotations are also given for comparison. The scores of Human and AliceMind-MMU are calculated based on Equation (9). These examples are studied by category as follows:

Commonsense Knowledge. AliceMind-MMU is knowledgeable in many aspects of daily life. As shown in Figure 11, AliceMind-MMU is able to tell not only weather and the sentiments of people, but also classic sports and electronic products as an ordinary person does. Also, it is skilled in geography and understands the food around the world. For example, AliceMind-MMU recognizes the small English flag and Big Ben in the image, by which the country is identified as England. As another example, based on the rice and dishes from the image, Chinese food can be identified by AliceMind-MMU and people familiar with Chinese cuisine. The other people may not tell cuisine of the food. Our experiments show that AliceMind-MMU trained on adequate data can capture the commonsense knowledge (the largest category in the VQA dataset) in our daily life.

Visual Recognition. Except for Clock Reading, AliceMind-MMU shows much better performance than an ordinary person does in this category. As shown in Figure 11, it is relatively easy for an AI model trained adequately to memorize specialized knowledge, while rather difficult for people unfamiliar with the specific domain. For example, AliceMind-MMU can better identify the specific categories of the animals, such as dog and bird, and the historical style of the furniture, which requires specialized knowledge. By locating and recognizing the barely visible logo from the
Fig. 11. Case Study for Commonsense Knowledge and Visual Recognition. The scores of Human and AliceMind-MMU are calculated with Equation (9). Ground Truth gives the top three annotations.

motor bike, AliceMind-MMU correctly recognizes its brand, while human may miss the details in the image and give an answer based on their best guesses. On the other hand, AliceMind-MMU has a slim chance of being fooled by the activity present in the image. It incorrectly identifies that the boy is playing baseball based on his motion, while it is actually a Wii game. As a result, by incorporating relevant specialized knowledge and visual recognition capability, AliceMind-MMU can outperform human by a large margin in this category.

Object Counting. As shown in Figure 12, AliceMind-MMU achieves human parity when counting a small number of objects, but fails in more complicated cases where there are occlusions or a great quantity of objects. It surprises us that AliceMind-MMU can give a count very close to the correct answer in the example of counting trees. However, the object counting ability is still quite limited compared with ordinary people. One reason may lie in that the visual detection is too weak to detect all the objects in an image when the number of objects is large. There are few cases with more than 10 objects in the training set, and thus the model is not fully trained with sufficient data. This is shown in the third example where AliceMind-MMU fails to identify all people from the image. Another possible reason is that the object detector is difficult to count in the presence of
Fig. 12. Case Study for Object Counting and Relational Reasoning.

occlusion. The second example shows that ALICEMIND-MMU counts the racing people incorrectly due to the occluded person.

Relational Reasoning. Figure 12 shows that ALICEMIND-MMU has the abilities to reason over the relationship of positions, comparison, and exclusion. It is observed that ALICEMIND-MMU may be more capable than human in precise position identification and knowledge reasoning for relational reasoning questions. Specifically, (1) position: the first two examples show the power of ALICEMIND-MMU in distinguishing the positions of the left-right and front-back, and conducting one-step reasoning over the positional relationship; (2) comparison: the third example demonstrates that ALICEMIND-MMU can even compare the colors of two objects, which is a simple one-step reasoning over the comparison between attributes of two objects; (3) exclusion: the last example shows that ALICEMIND-MMU is able to identify the exclusion relationship, and reason over it with commonsense knowledge.

Textual Recognition (OCR). As shown in Figure 13, the text reading expert (StructuralLM) is able to identify text and layout in simple cases. In the first example, the model correctly answers with the words displayed on the man’s shirt. In addition, StructuralLM is capable of learning the interactions between text and layout, which makes StructuralLM aware of location of text present in an image. This is shown in the second example, where the model predicts the answer correctly when asked about the sign on the left. However, the model fails in the two cases: (1) OCR errors; (2) answering complex questions which requires visual feature and reasoning abilities. As shown
in the third example, when asked the words on the man’s shirt, the model can predict only “3” because the OCR tool cannot recognize the word “cardinals”. In the fourth example, given the question about the number present on the white shirt, the model answers incorrectly due to the lack of visual feature of colors and the reasoning ability. Currently, the text reading expert utilizes only the layout and textual information to answer a text-reading question, without leveraging visual signals. There is an urgent need for deep interaction between visual information and OCR textual information in images, which is left for future work.

Clock Reading. As shown in Figure 13, the clock reading expert is able to read the clock time accurately at a five-minute level. One important problem to be addressed is distinguishing the hour hand (generally shorter) from the minute hand (generally longer). The clock reading expert is trained well on this objective. Therefore, in the first example, the model predicts the correct time “8:05”, while some human annotators misread the hour hand and minute hand, and thus give the wrong time reading “1:40”. In the second example, the clock reading expert can tell the time accurately even when the hour hand and minute hand overlap. There also exist limitations for the current clock reading expert, that it can’t tell more accurate time at a minute level. In the fourth example, the model only recognizes the time is about “12:10”, but cannot tell the exact time of “12:12”. The reason comes from casting the problem as detection and then classification, of which the clock training data is not adequate to support training of 1-minute level clock reading.

5 CONCLUSION

This paper describes our new research work on improving the full pipeline of the VQA task, which has achieved human parity on this challenging task for the first time. The key to the breakthrough...
lies in three aspects: (1) diverse visual semantics learning for comprehensive image content understanding, (2) more effective cross-modal interaction with modality adaptive attention, and (3) a knowledge-guided model integration with three specialized expert modules for the complex VQA task. It demonstrates the power of an AI model to achieve human parity on the challenging cross-modal understanding task under a closed-set scenario, which requires AI to be able to understand both visual and textual information. This makes it possible to further conduct higher-order cognition and commonsense reasoning intelligence.

Despite the success, the current AI technology on V&L understanding still has notable limitations, and bridging the gap between machine intelligence and human intelligence still has a long way to go. For the VQA task, there still exist certain weaknesses for our model. In terms of specific task capabilities, (1) object counting is still a very difficult problem for the current VLP model, especially in the cases where a large number of small objects exist. Besides, some objects can even overlap with each other, which introduces more complexity; (2) composite reading of both OCR text and visual content is still challenging for the current VLP model. In terms of model design and training, (3) the size of data used for pre-training ALICE MIND-MMU is rather limited, only the public 12M image-text pairs are used for multi-modal pre-training, which is relatively small compared with that used in NLP and one of the recent VLP model SimVLM. Therefore, the current model can only work well in a closed-set scenario, inapplicable to the open-domain or unseen set. Large-scale V&L pre-training on billion-scale image-text data provides great potentials to further push the limit of VQA task; (4) While the current MoE framework learns to mix experts in response to different types of questions, it has not applied to ensemble of models in the vision understanding expert.

In the future, we will work on improving ALICE MIND-MMU from the following perspectives:

- We will scale up the model size and image-text data used for ALICE MIND-MMU pre-training, e.g., working towards a larger ALICE MIND-MMU model pre-trained with billion-scale image-text data.
- We are working on designing a unified model architecture in a more elaborate early-fusion way, which can well address the different subtasks in VQA such as OCR-based text reading, object counting, and visual understanding.
- We are also interested in applying our ALICE MIND-MMU pre-training method to other V&L understanding and generation tasks, such as image-text retrieval and image captioning.

We hope to witness more breakthrough on open-set learning and more intelligent AI models, which can evolve itself by acquiring and reasoning about new knowledge.

APPENDICES

A  TEXT READING EXPERT

Ablation Study. As stated in Section 3.3.2, the pre-trained StructuralLM model is adapted on text-reading VQA samples in different ways. Table 13 shows the ablation results on the VQA test-dev set. The visual understanding expert is used as the baseline method, upon which all the ablation experiments for text reading expert is conducted. First, it is observed that text reading expert greatly improves the performance on the “Number” type by over 6%, where many questions are asked about reading numbers from OCR text, such as a bus number and a football player number. On the “Other” type, the performance can be improved by over 1%. Answering many questions of this type requires the ability to reason with both visual and textual OCR information. Adding the

---

6 The size of pre-training data we use is more than 100 times smaller than SimVLM.
7 An early version of visual understanding expert for VQA Challenge 2021 is used as the baseline.
### Table 13. Ablation Study of Text Reading Expert on the VQA Test-dev

|                         | Overall | Yes/No | Number | Other | ANLS |
|-------------------------|---------|--------|--------|-------|------|
| Visual Understanding Expert | 79.44  | 93.31  | 65.70  | 71.16 | –    |
| + Text-reading VQA data  | 80.35  | 93.31  | 69.81  | 71.49 | 79.85|
| + add separator         | 80.41  | 93.31  | 69.82  | 71.64 | 79.96|
| + continue pre-training | 80.63  | 93.31  | 69.97  | 72.01 | 80.33|

A separator between textual bounding boxes and continual pre-training on domain data can lead to further improvement, demonstrating the effectiveness of adapting the pre-trained StructuralLM for text-reading VQA.

### B CLOCK READING EXPERT

**Training Network.** In the clock reader, Resnet50-IBN [57] is adopted as our backbone, and two specific branches are introduced for hour and minute prediction, respectively. Furthermore, as the hour and minute hands in the clock are the keys to predict the time, attention modules were introduced to force the focus of the model on the hands. A SE-layer [27] is employed after the backbone for channel-wise attention, and a spatial attention module which consists of convolution layers and ReLU activation is employed in the beginning of hour and minute branch, respectively, for spatial-wise attention. Such a corporation of channel and spatial wise attention is able to adapt to the individual bias of hour and minute prediction. The feature outputs from two branches are listed as follows:

\[
\begin{align*}
    f_m &= E_m(\text{Attn}_{sp}(F) \ast F), \\
    f_h &= E_h(\text{Attn}_{sp}(F) \ast F)
\end{align*}
\]

where \( I \) is the image, and \( E, E_h, E_m \) are the backbone, hour branch, and minute branch, respectively. \( F \) is the feature map from the backbone after SE-layer, \( F = \text{Attn}_{se}(E(I)) \).

As the clock reader is formulated as both a classification task and a regression task, it introduces loss from two perspectives. From the classification aspect, a 12-category classification loss is used for both hour and minute\(^8\) prediction. The cross-entropy loss is adopted as follows:

\[
L_{cls} = - \sum_i^N g_i \times \log p_i
\]

where \( N \) is the number of categories and set to 12, \( g_i \) and \( p_i \) is the ground-truth probability and prediction probability of the corresponding category, respectively.

As 2:00 is closer to 3:00 than to 9:00, it is also important to solve the problem from a regression perspective. With the predicted classification probability \( p_i \) above, we are able to obtain a continuous prediction \( p_{reg} \) by soft-argmax, which is formulated as Equation (12). \( d \) means discrete numbers of hour or minute. Then the continuous prediction is used to form a regression loss as Equation (13):

\[
p_{reg} = \sum_i d_i \times p_i
\]

\[
L_{reg} = \cos \left( \frac{2\pi}{C} \times (p_{reg} - g_{reg}) - \pi \right) + 1
\]

The cosine formulation is used for the periodicity constraint of the clock prediction. \( C \) is the periodicity of hour or minute, \( g_{reg} \) is the regression ground truth, which is exactly the ground truth category.

\(^8\)the minute task is divided into 12 bins by 5 moves per bin.
Table 14. Ablation Study of Clock Reading Expert

| Clock Detector | Clock Reader | VQA Test-dev |
|----------------|-------------|-------------|
| Detection(mAP) | Clock Accuracy | Number | Overall |
| Baseline | - | - | - | 59.93 | 76.51 |
| 79.30 | ✓ | 72.5 | 62.52 | 76.79 |
| | ✓ ✓ | 73.0 | 62.59 | 76.80 |
| | ✓ ✓ ✓ | 74.7 | 62.65 | 76.81 |

We are able to give a continuous prediction of hour, due to the soft-argmax. As the decimal part of hour is strongly related to the minute in clock, e.g., when the prediction of hour is 2.5, the prediction of minute should around 30. A self-supervised loss is introduced according to this prior to improve the generalization of clock reader. We note the prediction of minute and hour as $p_h$ and $p_m$, respectively.

$$L_{self} = \text{Smooth}_{L1} (C \times (p_h - \lfloor p_h \rfloor) - p_m)$$  \hspace{1cm} (14)

Finally, the total loss is:

$$L = L_{cls} + L_{reg} + \lambda L_{self}$$  \hspace{1cm} (15)

where $\lambda$ is used to weight the self-supervised loss, and set to 0.01.

**Ablation Study.** The ablation study of clock reading expert is shown in Table 14, where only the results on the “Number” and “Overall” types are given, because questions on reading clocks are only present in the “Number” type. Adding clock reading expert results in more than 4.5% performance improvement on the “Number” type, which demonstrates the effectiveness of proposed ideas in the clock reading expert. Specifically, the proposed regression loss is prone to provide a larger gradient when there is a bigger difference between the predicted time and the ground truth, which benefits prediction of the clock reader. Moreover, it can be observed that the self-supervised loss boosts the performance significantly, as the relationship prior constrains hour and minute branches both, which eliminates the confusion of hour and minute hands.
C  MORE CLUSTER EXAMPLES

To identify the specific semantic information of each cluster, we present more examples in Figure 14.

| Cluster 1         | Cluster 2                  | Cluster 3                | Cluster 4                | Cluster 5                |
|-------------------|---------------------------|--------------------------|--------------------------|--------------------------|
| What time is it?  | How many people are in the picture? | What is the sign showing? | What is the number on the front of the train? | What is the cat sitting on? |
| What time is it on the clock? | How many cows are there? | What does the sign say? | What number is on the train? | Where is the cat's head? |
| What time does the clock say? | How many red objects on the shelves? | What word is on the boy's shirt? | What number is on the license plate? | What kind of instrument is the man playing? |
| What time is it on the clock? | How many people are holding flags? | What is written on the red suitcase? | What is the number of the train? | What kind of food is shown? |
| What time does the clock show? | How many birds are shown? | What is the cross street? | What is the number in front of the train? | Where is the storm drain? |
| What time is it? | How many cars are pictured? | What does the sign under the clock say? | What is the number on this bus? | What type of sign is shown? |

Fig. 14. More examples of each cluster are illustrated to identify its semantic information.

D  COMPARED BASELINES

- **MCAN** [93]: proposes a deep modular co-attention network, which models the self-attention of questions and images, as well as the question-guided-attention of images jointly in a deep cascaded manner. It is a popular non-pretraining method, which won the champion in VQA Challenge 2019.
- **BGN** [22]: develops bilinear graph networks to model the context of the joint embeddings of words and objects, where an image-graph and a question-graph are used to capture more important relationships between words and objects, respectively. It won the VQA Challenge Runner up in 2020.
- **GridFeat+MoVie** [33]: proposes to utilize the grid-based convolutional features for VQA instead of the prevalent region-based object features [2]. This makes it possible to train the VQA models end-to-end and enables more flexible network design. It is the state-of-the-art non-pretraining method, which won the champion in VQA Challenge 2020.
• **LXMERT** [74]: is the pioneering work to pre-train a two-stream multi-modal Transformer, which consists of an object relationship encoder, a language encoder and a cross-modality encoder. Five diverse pre-training tasks are used to pre-train the LXMERT model. It is widely used as a baseline method for VLP models.

• **VILLA** [21]: proposes the idea of large-scale adversarial training for vision-language representation learning, which adds embedding-based adversarial training on both stages of task-agnostic pre-training and task-specific fine-tuning. It can further improve the performance in a wide range of VL tasks over the pre-training baselines.

• **InterBERT** [45]: designs specific model architectures to enhance interactions between the information flows of different modalities, where single-stream interaction is first applied on each modality respectively, and then the two-stream module is added on top to model cross-modal interactions.

• **VinVL** [95]: pre-trains a large-scale object-attribute detection model with much larger amounts of supervised data on four public object detection datasets for extracting better region-based visual feature, and creating new state-of-the-art results on seven public benchmarks.

• **ROSITA** [13]: different from other VLP methods that only exploit the intra-modal knowledge from each modality, it encodes the object-level cross-modal knowledge and intra-modal knowledge in a unified scene graph, so as to better learn the fine-grained semantic alignments across modalities.

• **UNIMO** [42]: proposes a unified-modal pre-training architecture with cross-modal contrastive learning, which can effectively adapt to both single-modal and multi-modal understanding and generation tasks. Except for limited image-text pairs, it utilizes large amounts of single-modal data such as text or image for pre-training.

• **SimVLM** [85]: different from previous VLP methods that only use limited (4M-10M) image-text pairs for pre-training, it proposes a simple VLP model with a single prefix language modeling objective, which pre-trains on an extremely large aligned cross-modal data of about 1.8B noisy image-text pairs. This is also a latest state-of-the-art method on image captioning.

• **VLBERT** [72]: is a pioneering work to pre-train a single-stream multi-modal Transformer, which jointly trains both the Transformer-based cross-modal fusion and Fast R-CNN image feature extractor in both pre-training and fine-tuning phases. It is widely used as a baseline method for VLP models.

• **GilBERT** [26]: proposes a generative visual-linguistic pre-training method, which simultaneously learns generic representations of image-text data and complete the missing modality for incomplete pairs. It is mainly designed to tackle the modality incompleteness problem in image-text retrieval, and can effectively adapt to both VQA and image captioning tasks.

• **UNITER** [11]: proposes an improved single-stream VLP method, by designing two new pre-training strategies: (1) it uses conditional masking on pre-training tasks instead of random masking strategy, (2) it designs a new word-region alignment pre-training task via the use of optimal transport to explicitly encourage fine-grained alignment between words and image regions.

• **OSCAR** [43]: proposes to use object tags detected in images as anchor points to ease the learning of cross-modal alignments, where the input to the Transformer is a combination of image, text, and object tags.
- ViLBERT [50]: proposes one of the first work that extend the BERT architecture to a multi-modal two-stream VLP model, which processes both visual and textual inputs in separate streams that interact through co-attentional transformer layers.
- 12-in-1 [51]: leverages the task relationships among different VL tasks, and develops a large-scale, multi-task training regime, which jointly trains 12 related VL tasks in one unified model.
- ERNIE-ViL [91]: proposes a knowledge-enhanced VLP method, by pre-training with three additional scene graph prediction tasks on textual side: i.e., object prediction, attribute prediction, and relationship prediction. It obtains the state-of-the-art performance of two-stream VLP method on five cross-modal downstream tasks.
- PixelBERT [29]: provides the first end-to-end VLP method with grid-based convolutional features, and adopts a simple random pixel sampling mechanism to enhance the robustness of visual representation during pre-training.
- E2E-VLP [87]: proposes the first end-to-end VLP method for both V+L understanding and generation, with a unified Transformer encoder-decoder architecture. Except for the popular cross-modal pre-training tasks, it incorporate the tasks of object detection and image captioning into pre-training for enhancing visual learning.
- ViLT [35]: proposes the first end-to-end VLP method without convolution and region supervision. It follows the architecture of ViT [17], and pre-trains directly with patch-based features. It can be much more efficient than the previous VLP methods, with comparable or better downstream task performance.

REFERENCES

[1] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. 2017. VQA: Visual question answering. *Int. J. Comput. Vision* 123, 1 (May 2017), 4–31.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6077–6086.

[3] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *International Conference on Machine Learning*. PMLR, 1247–1255.

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*. 2425–2433.

[5] Ali Furkan Biten, Rubén Tito, Andrés Mafla, Lluís Gómez, Marçal Rusiñol, Ernest Valveny, C. V. Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering. *CoRR* abs/1905.13648 (2019). arXiv:1905.13648 http://arxiv.org/abs/1905.13648.

[6] Zhaowei Cai and Nuno Vasconcelos. 2018. Cascade R-CNN: Delving into high quality object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6154–6162.

[7] Yue Cao, Mingsheng Long, Jianmin Wang, Qiang Yang, and Philip S. Yu. 2016. Deep visual-semantic hashing for cross-modal retrieval. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1445–1454.

[8] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radiu Soricut. 2021. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3558–3568.

[9] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. arXiv preprint arXiv:1704.00051 (2017).

[10] Haiting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. 2021. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12299–12310.

[11] 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. In *European Conference on Computer Vision*. Springer, 104–120.
Achieving Human Parity on Visual Question Answering

79:37

[12] Miaomiao Cheng, Liping Jing, and Michael K. Ng. 2020. Robust unsupervised cross-modal hashing for multimedia retrieval. ACM Transactions on Information Systems (TOIS) 38, 3 (2020), 1–25.

[13] Yuhao Cui, Zhou Yu, Chunqi Wang, Zhongzhou Zhao, Ji Zhang, Meng Wang, and Jun Yu. 2021. ROSITA: Enhancing Vision-and-Language Semantic Alignments via Cross- and Intra-modal Knowledge Integration. arXiv:2108.07073 [cs.CV].

[14] 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. IEEE, 248–255.

[15] Karan Desai and Justin Johnson. 2021. VirTex: Learning visual representations from textual annotations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 11162–11173.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations.

[18] William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. arXiv preprint arXiv:2101.03961 (2021).

[19] David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John Prager, et al. 2010. Building Watson: An overview of the DeepQA project. AI Magazine 31, 3 (2010), 59–79.

[20] Dalu Guo, Chang Xu, and Dacheng Tao. 2019. Bilinear graph networks for visual question answering. arXiv preprint arXiv:1907.09815 (2019).

[21] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. In NeurIPS.

[22] Peng Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 7132–7141.

[23] Zicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. Pixel-BERT: Aligning image pixels and alleviating the language prior problem in visual question answering. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 75–84.

[24] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-augmented language model pre-training. arXiv preprint arXiv:1907.09815 (2019).

[25] Kaiqing He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 770–778.

[26] Weixiang Hong, Kaixiang Ji, Jiajia Liu, Jian Wang, Jingdong Chen, and Wei Chu. 2021. GilBERT: Generative vision-language pre-training for image-text retrieval. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 1379–1388.

[27] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 7132–7141.

[28] Peng Hu, Liang Li, and Jifeng Dai. 2019. Scalable deep multimodal learning for cross-modal retrieval. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 635–644.

[29] Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. 1991. Adaptive mixtures of local experts. Neural Computation 3, 1 (1991), 79–87.

[30] Drew A. Hudson and Christopher D. Manning. 2019. QQA: A new dataset for real-world visual reasoning and compositional question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 6700–6709.

[31] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multi-modal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847 (2016).

[32] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. Pixel-BERT: Aligning image pixels and alleviating the language prior problem in visual question answering. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 75–84.

[33] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Jiannong Cao. 2020. Pixel-BERT: Aligning image pixels and alleviating the language prior problem in visual question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 11162–11173.

[34] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In Empirical Methods in Natural Language Processing (EMNLP).
[35] Wonjae Kim, Bokyung Son, and Ildeok Kim. 2021. ViLT: Vision-and-language transformer without convolution or region supervision. arXiv preprint arXiv:2102.03334 (2021).

[36] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision 123, 1 (2017), 32–73.

[37] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale reading comprehension dataset from examinations. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 785–794.

[38] D. Lewis, G. Agam, S. Argamon, O. Friered, D. Grossman, and J. Heard. 2006. Building a test collection for complex document information processing (SIGIR’06). Association for Computing Machinery, New York, NY, USA, 665–666. https://doi.org/10.1145/1148170.1148307

[39] Chenliang Li, Bin Bi, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. 2021. StructuralLM: Structural pre-training for form understanding. CoRR abs/2105.11210 (2021). arXiv:2105.11210 https://arxiv.org/abs/2105.11210.

[40] Junyang Lin, An Yang, Yichang Zhang, Jie Liu, Jingren Zhou, and Hongxia Yang. 2020. InterBERT: Vision-and-language interaction for multi-modal pretraining. arXiv preprint arXiv:2012.15409 (2020).

[41] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In European Conference on Computer Vision. Springer, 121–137.

[42] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language by cross-modal pre-training. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34. 11336–11344.

[43] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visual language representation learning. In The IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[44] Zujie Liang, Haifeng Hu, and Jiaying Zhu. 2021. LPF: A language-prior feedback objective function for de-biased visual question answering. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval.

[45] Junyang Lin, An Yang, Yichang Zhang, Jie Liu, Jingren Zhou, and Hongxia Yang. 2020. InterBERT: Vision-and-language interaction for multi-modal pretraining. arXiv preprint arXiv:2003.13198 (2020).

[46] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In European Conference on Computer Vision. Springer, 740–755.

[47] Song Liu, Shengsheng Qian, Yang Guan, Jiawei Zhan, and Long Ying. 2020. Joint-modal distribution-based similarity hashing for large-scale unsupervised deep cross-modal retrieval. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 1379–1388.

[48] Yinhua Li, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTA: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019).

[49] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030 (2021).

[50] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Advances in Neural Information Processing Systems. 13–23.

[51] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[52] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. Advances in Neural Information Processing Systems 29 (2016), 289–297.

[53] Sean MacAvaney, Franco Maria Nardini, Raffaele Perego, Nicola Tonellotto, Nazi Goharian, and Ophir Frieder. 2020. Training curricula for open domain answer re-ranking. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 529–538.

[54] James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1. Oakland, CA, USA, 281–297.

[55] Duy-Kien Nguyen, Vedanuj Goswami, and Xinlei Chen. 2020. MoVie: Revisiting modulated convolutions for visual and language processing. arXiv preprint arXiv:2004.11883 (2020).

[56] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In CoCo@ NIPS.

[57] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. 2018. Two at once: Enhancing learning and generalization capacities via IBN-Net. In Proceedings of the European Conference on Computer Vision (ECCV). 464–479.
Achieving Human Parity on Visual Question Answering

[58] Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W. Bruce Croft, and Mohit Iyyer. 2020. Open-retrieval conversational question answering. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 539–548.

[59] Chen Qu, Hamed Zamani, Liu Yang, W. Bruce Croft, and Erik Learned-Miller. 2021. Passage retrieval for outside-knowledge visual question answering. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval.

[60] Leigang Qu, Meng Liu, Jianlong Wu, Zan Gao, and Liqiang Nie. 2021. Dynamic modality interaction modeling for image-text retrieval. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 1104–1113.

[61] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021).

[62] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250 (2016).

[63] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert R. G. Lanckriet, Roger Levy, and Nuno Vasconcelos. 2010. A new approach to cross-modal multimedia retrieval. In Proceedings of the 18th ACM International Conference on Multimedia. 251–260.

[64] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems 28 (2015), 91–99.

[65] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN features off-the-shelf: An astounding baseline for recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 806–813.

[66] Abhishek Sharma, Abhishek Kumar, Hal Daume, and David W. Jacobs. 2012. Generalized multiview analysis: A discriminative latent space. In 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2160–2167.

[67] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image-alt-text dataset for automatic image captioning. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2556–2565.

[68] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538 (2017).

[69] David Silver, Aja Huang, Christopher J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. 2016. Mastering the game of Go with deep neural networks and tree search. Nature 529 (2016), 484–503.

[70] Robert F. Simmons. 1965. Answering English questions by computer: A survey. Commun. ACM 8, 1 (1965), 53–70.

[71] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA models that can read. CoRR abs/1904.08920 (2019). arXiv:1904.08920 http://arxiv.org/abs/1904.08920.

[72] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Luewi Lu, Furu Wei, and Jifeng Dai. 2019. VL-BERT: Pre-training of generic visual-linguistic representations. arXiv preprint arXiv:1908.08530 (2019).

[73] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Hua Jain, and Yoav Artzi. 2018. A corpus for reasoning about natural language grounded in photographs. arXiv preprint arXiv:1811.00491 (2018).

[74] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490 (2019).

[75] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In International Conference on Machine Learning. PMLR, 10347–10357.

[76] Ellen M. Voorhees and Dawn M. Tice. 2000. Building a question answering test collection. In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 200–207.

[77] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Association for Computational Linguistics, Brussels, Belgium, 353–355.

[78] Bingning Wang, Ting Yao, Qi Zhang, Jingfang Xu, Zhixing Tian, Kang Liu, and Jun Zhao. 2019. Document gated reader for open-domain question answering. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 85–94.
[79] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. 2016. A comprehensive survey on cross-modal retrieval. arXiv preprint arXiv:1607.06215 (2016).
[80] Shuohang Wang, Mo Yu, Xiaoxiao Guo, Zhiguo Wang, Tim Klinger, Wei Zhang, Shiyu Chang, Gerry Tesauro, Bowen Zhou, and Jing Jiang. 2018. R 3: Reinforced ranker-reader for open-domain question answering. In Thirty-Second AAAI Conference on Artificial Intelligence.
[81] Shuohang Wang, Mo Yu, Xiaoxiao Guo, Zhiguo Wang, Tim Klinger, Wei Zhang, Shiyu Chang, Gerry Tesauro, Bowen Zhou, and Jing Jiang. 2018. R 3: Reinforced ranker-reader for open-domain question answering. In Thirty-Second AAAI Conference on Artificial Intelligence.
[82] Wei Wang, Bin Bi, Ming Yan, Chen Wu, Zuyi Bao, Jiangnan Xia, Liwei Peng, and Luo Si. 2019. StructBERT: Incorporating language structures into pre-training for deep language understanding. arXiv preprint arXiv:1908.04577 (2019).
[83] Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated self-matching networks for reading comprehension and question answering. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 189–198.
[84] Yongxin Wang, Zhen-Duo Chen, Xin Luo, and Xin-Shun Xu. 2021. High-dimensional sparse cross-modal hashing with fine-grained similarity embedding. In Proceedings of the Web Conference 2021. 2900–2909.
[85] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 21–29.
[86] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2369–2380.
[87] Fei Yu, Jiji Tang, Weiichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE-ViL: Knowledge enhanced vision-language representations through scene graphs. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35. 3208–3216.
[88] Tan Yu, Yi Yang, Yi Li, Lin Liu, Hongliang Fei, and Ping Li. 2021. Heterogeneous attention network for effective and efficient cross-modal retrieval. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 1146–1156.
[89] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 6281–6290.
[90] Xinyu Zhang, Ke Zhan, Enrui Hu, Chengzhen Fu, Lan Luo, Hao Jiang, Yantao Jia, Fan Yu, Zhicheng Dou, et al. 2021. Answer complex questions: Path ranker is all you need. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 481–490.
[91] Yukun Zheng, Ping Nie, Arun Ramamurthy, and Le Song. 2021. Answering any-hop open-domain questions with iterative document reranking. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 481–490.
[92] Yukun Zheng, Ping Nie, Arun Ramamurthy, and Le Song. 2021. Answering any-hop open-domain questions with iterative document reranking. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 481–490.
[93] Yuekun Zheng, Jiaxin Mao, Yiqun Liu, Zixin Ye, Min Zhang, and Shaoping Ma. 2019. Human behavior inspired machine reading comprehension. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 425–434.
[94] Yueke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 4995–5004.

Received 23 January 2022; revised 28 September 2022; accepted 9 November 2022