Multi-Grained Vision Language Pre-Training: Aligning Texts with Visual Concepts

Yan Zeng, Xinsong Zhang, Hang Li
ByteDance AI Lab
{zengyan.yanne, zhangxinsong.0320, lihang.lh}@bytedance.com

Abstract

Most existing methods in vision language pre-training rely on object-centric features extracted through object detection, and make fine-grained alignments between the extracted features and texts. We argue that the use of object detection may not be suitable for vision language pre-training. Instead, we point out that the task should be performed so that the regions of ‘visual concepts’ mentioned in the texts are located in the images, and in the meantime alignments between texts and visual concepts are identified, where the alignments are in multi-granularity. This paper proposes a new method called X-VLM to perform ‘multi-grained vision language pre-training’. Experimental results show that X-VLM consistently outperforms state-of-the-art methods in many downstream vision language tasks.

1 Introduction

Vision language pre-training aims to learn vision language alignments from a large number of image-text pairs. A pre-trained Vision Language Model (VLM) fine-tuned with a small amount of labeled data has shown the state-of-the-art performances in many Vision Language (V+L) tasks such as visual question answering and image-text retrieval.

Existing methods in vision language pre-training fall into two approaches as shown in Figure 1 (a, b). Most of them detect objects in the image and align the text with fine-grained (object-centric) features. They either utilize pre-trained object detectors (Tan and Bansal, 2019; Lu et al., 2019; Li et al., 2019, 2020a; Chen et al., 2020; Li et al., 2020c; Gan et al., 2020) or conduct object detection on-the-fly in the pre-training process (Su et al., 2019; Xu et al., 2021). The other methods do not rely on object detection and only learn alignments between the text and coarse-grained (overall) features of the image (Huang et al., 2020, 2021; Kim et al., 2021; Li et al., 2021).

Both the fine-grained and coarse-grained approaches have drawbacks. Object detection identifies all possible objects in the image, and some of them might not be relevant to the text. Object-centric features cannot easily represent relations among multiple objects, e.g. “man crossing the street”. Moreover, it is challenging to pre-define the categories of objects suitable for the task. On the other hand, the coarse-grained approach cannot effectively learn fine-grained alignments between the text and image, and thus usually has lower performances in downstream tasks.

Ideally, we want the VLM to represent and learn multi-grained alignments between the images and texts. This is because the alignments between an image and a text are usually in multi-granularity, depending on their contents. Taking the image-text pair in Figure 1 as example, the text “a man wearing a backpack is crossing the street” is about the whole image, the text “man crossing the street” is only related to a region in the image, and the word “backpack” is only concerned with the referred object in the image. That is to say, the texts in fact describe different ‘visual concepts’ (Krishna et al., 2016;
Zhang et al., 2021; Changpinyo et al., 2021) in different granularities in the image. Unfortunately, existing methods either depend on object-centric features or overall features and cannot satisfactorily handle the alignments between texts and visual concepts.

In this paper, we propose performing multi-grained vision language pre-training. The key idea is to represent and learn multi-grained alignments between the texts and the images by aligning text descriptions of visual concepts with the corresponding regions in images. We re-formulate existing datasets\(^1\), so that an image may have multiple regions enclosed by bounding boxes, and a text is directly associated with the visual concept in each region. The visual concept may be an object, a region, or the image itself, as the example in Figure 1(c). The goal of the pre-training then becomes to learn a model that can represent visual concepts, texts, and the alignments between the visual concepts and texts.

Our multi-grained model, denoted as X-VLM, consists of an image encoder, a text encoder, and a cross-modal encoder which conducts cross-attention between the vision features and language features to learn vision-language alignments. X-VLM utilizes a simple mixed attention mechanism to implement the image encoder, so that the encoder produces an image representation with full attention and region/object representations with local attention. The learning of X-VLM is optimized by locating visual concepts in the image given the associated texts and in the meantime aligning the texts with visual concepts, e.g. by a contrastive loss, a matching loss, and a masked language modeling loss, where the alignments are in multi-granularity, as illustrated in Figure 1(c). In fine-tuning and inference, without bounding box annotations in the input images, X-VLM can still leverage the learned multi-grained alignments to perform the downstream V+L tasks.

We demonstrate the effectiveness of our approach on various downstream tasks. On image-text retrieval, our approach outperforms CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) which are pre-trained on datasets with larger orders of magnitudes, achieving absolute gains of 2.65% in terms of R@1 score on MSCOCO and 1.30% on Flickr30K. On visual reasoning tasks, it achieves absolute improvements of 0.79% on VQA and 1.06% on NLVR\(^2\) compared to VinVL (Zhang et al., 2021), with a much faster inference speed. On visual grounding (Refcoco+), it achieves absolute improvements of 7.24% compared to ALBEF (Li et al., 2021) on the weakly-supervised setting and outperforms UNITER (Chen et al., 2020) by 1.34% on the supervised setting. The code and pre-trained models will be available at \url{https://github.com/zengyan-97/X-VLM}.

The contributions of this work are as follows:

- We propose performing multi-grained vision language pre-training to handle the alignments between texts and visual concepts.
- We propose to optimize the model (X-VLM) by locating visual concepts in the image given the associated texts and in the meantime aligning texts with visual concepts, where the alignments are in multi-granularity.
- We empirically verify that our approach effectively leverages the learned multi-grained alignments in fine-tuning. X-VLM\(_{base}\) with 256 \(\times\) 256 image resolution achieves substantial improvements over existing state-of-the-art methods on many downstream V+L tasks.

2 Related Work

The existing work on vision language pre-training falls into two categories: fine-grained and coarse-grained.

Most existing methods belong to the fine-grained approach, which relies on object detection. An object detector first identifies all regions that probably contain an object, then conducts object classification on each region. An image is then represented by dozens of object-centric features of the identified regions. The challenge with the approach is that VLMs based on object-centric features cannot represent relations among multiple objects in multiple regions. Furthermore, it is not easy to define the categories of objects in advance, which are helpful for learning VLMs.

The VLMs in the fine-grained approach either utilize pre-trained object detectors (Tan and Bansal, 2019; Lu et al., 2019; Li et al., 2019, 2020a; Chen et al., 2020; Li et al., 2020c; Gan et al., 2020; Li et al., 2020b) or incorporate object detection into vision language pre-training (Su et al., 2019; Xu et al., 2021). Object detectors, such as Faster

---

\(^1\)Annotations of objects/regions are included.

\(^2\)Related Work
R-CNN (Ren et al., 2015), Bottom-Up and Top-Down Attention (BUTD) (Anderson et al., 2018), are trained on image annotations of common objects, e.g. COCO (Lin et al., 2014) (100K images) and Visual Genome (Krishna et al., 2016) (100K). VinVL (Zhang et al., 2021) has achieved SoTA performances on many V+L tasks by utilizing a powerful object detector pre-trained with a very large collection of image annotations (2.5M images). On the contrary, Xu et al. (2021) propose a unified Transformer encoder-decoder architecture for VLM which contains an end-to-end object detection module (Carion et al., 2020). However, end-to-end object detection suffers from slow convergence (Zhu et al., 2020). Therefore, the method has to address a challenging issue of simultaneously learning both tasks, and as a result, it does not yield performances comparable with the two-stage methods, e.g. VinVL.

The coarse-grained approach builds VLMs by extracting and encoding overall image features with convolutional network (Huang et al., 2020, 2021) or vision transformer (Kim et al., 2021; Li et al., 2021). The performances are usually not as good as fine-grained approach. Though object-centric features are only related to certain objects, the fine-grained features seem critical for learning VLMs. To cope with the problem, Huang et al. (2021) employ online clustering on overall image features to obtain more comprehensive representations, Kim et al. (2021) use a more advanced vision transformer, i.e. Swin-Transformer (Liu et al., 2021) for image encoding, and Li et al. (2021) incorporate contrastive learning and momentum distillation. However, the improvements still cannot close the gap with the fine-grained approach when the training corpus is of the same order of magnitude. Unlike previous work taking either the fine-grained or coarse-grained approach, we propose multi-grained vision language pre-training.

3 Method

3.1 Overview

X-VLM consists of an image encoder ($I$-trans), a text encoder ($T$-trans), and a cross-modal encoder ($X$-trans). All encoders are based on Transformer (Vaswani et al., 2017). The cross-modal encoder fuses the vision features with the language features through cross-attention at each layer.

We re-formulate widely used pre-training datasets (see Section 4.1) so that an image may have multiple regions enclosed by bounding boxes, and each of them is associated with a text that describes an object or a region, denoted as $(I, T, \{(V^j, T^j)\}_N)$. Note that a few images do not have associated texts, i.e., $T$ is NaN, and some images do not have bounding boxes, i.e., $N = 0$. Here, $V^j$ is an object or region in the image $I$ associated with a bounding box $b^j = (cx, cy, w, h)$ represented by the normalized center coordinates, width, and height of the box.\(^2\)

\(^2\)When the image itself represents a visual concept, $b = (0.5, 0.5, 1, 1)$. 
3.2 Image Encoder

We propose a simple mixed attention mechanism which is based on vision transformer (Dosovitskiy et al., 2020) as shown in Figure 2. With a low computational overhead, it produces multi-grained visual concept representations in an image.

The encoder first splits an image into non-overlapping 16x16 patches and linearly embeds all patches, yielding \{v^0, ..., v^N\}. For an image of 256x256 resolution, we have \(N^f = 256\). We prepend the embedding of [CLS] denoted as \(v_{cls}^0\) to represent the whole image. At the first L₁ layer, we apply bi-directional attention to all patches, i.e. full attention. If there are \(N\) regions in the image, we duplicate the contextualized patch representations \{\(v_{cls}^{l_1}, v_1^{l_1}, ..., v_N^{l_1}\)\} into \(N + 1\) copies, among which we apply full attention to one copy and we apply local attention to the others at the following \(L_2\) layers.

A region \(V^j\) corresponds to a set of image patches, and we denote the patches as \{\(p_1^j, ..., p_M^j\)\}. We incorporate self-attention mask in the \(L_1 + 1\) to \(L_1 + L_2\) Transformer layers to implement local attention:

\[
H = \text{softmax}(\frac{QK^T}{\sqrt{d}} + M)V, \tag{1}
\]

where the self-attention mask matrix \(M \in \mathbb{R}^{(N^f+1) \times (N^f+1)}\) (with \(M_{ij} \in \{0, -\infty\}\)) determines whether a position can attend to the other positions. Specifically, \(M_{ij} = 0\) allows the \(i\)-th position to attend to the \(j\)-th position and \(M_{ij} = -\infty\) prevents attention between the two positions. For \(V^j\), we only allow the positions in \{\(0, p_1^j, ..., p_M^j\)\} to attend to each other, and set \(M_{ij} = -\infty\) for the other positions. Note that we include [CLS] into local attention for all \(N\) copies.

The image encoder then creates \(N + 1\) concept representations in different granularities, represented as \(I - \text{trans}(V^j) = \{v_{cls}^j, v_1^j, ..., v_N^j\}, j \in [0, N]\), where we make \(I - \text{trans}(V^0)\) denote the image representation with full attention. Because of the mixed attention mechanism, the concept representations can model contextual information out of the boxes, unlike object-centric features. In the following section, we will describe how the representations are utilized in the learning of multi-grained alignments.

3.3 Cross-Modal Modeling

As shown in Figure 2, we locate visual concepts in the image given corresponding text descriptions and in the meantime align texts and visual concepts, where the alignments are in multi-granularity.

Bounding Box Prediction We let the model to predict the bounding box \(b^j\) of visual concept \(V^j\) given the image representation with full attention and the text representation, where \(b^j = (cx, cy, w, h)\). By locating different visual concepts in the same image, we expect that the model better learns fine-grained vision-language alignments. The bounding box is predicted by:

\[
\hat{b}^j(I, T^j) = \text{Sigmoid}(\text{MLP}(x_{cls}^j)), \tag{2}
\]

where Sigmoid is for normalization, MLP denotes multi-layer perceptron, and \(x_{cls}^j\) is the output [CLS] embedding of the cross-modal encoder.

For bounding box prediction, \(\ell_1\) is the most commonly-used loss. However, it has different scales for small and large boxes, even if their relative errors are similar. To mitigate this issue, we use a linear combination of the \(\ell_1\) loss and the generalized Intersection over Union (IoU) loss (Rezatofighi et al., 2019), which is scale-invariant. The overall loss is defined as:

\[
\mathcal{L}_{bbox} = \mathbb{E}_{(V^j, T^j) \sim I, I \sim D}[\mathcal{L}_{iou}(b^j, \hat{b}^j) + ||b^j - \hat{b}^j||_1] \tag{3}
\]

Meanwhile, we align texts and visual concepts by three objectives which are widely used in vision language pre-training (Chen et al., 2020; Radford et al., 2021; Li et al., 2021). We extend the objectives to incorporate multi-grained visual concepts in the images.

Contrastive Learning We predict (visual concept, text) pairs, denoted \((V, T)\), from in-batch negatives. Note that visual concepts include objects, regions, and images. We randomly sample a mini-batch of \(N\) pairs. Given a pair as positive example, similar to Radford et al. (2021), we treat the other \((N - 1)\) pairs within the mini-batch as negative examples. We define cosine similarity \(s(V, T) = g_v(v_{cls})^\top g_w(w_{cls}).\) \(v_{cls}\) is the output [CLS] embedding of the text encoder. \(g_v\) and \(g_w\) are transformations that map the [CLS] embeddings to normalized lower-dimensional (256-d) representations.

For a positive pair \((V, T)\), we calculate the in-batch vision-to-text similarity and text-to-vision
similarity as:
\[
p^{y_{V2T}}(V) = \frac{\exp(s(V, T)/\tau)}{\sum_{i=1}^{N} \exp(s(V, T_i)/\tau)},
\]
\[
p^{12V}(T) = \frac{\exp(s(V, T)/\tau)}{\sum_{i=1}^{N} \exp(s(V_i, T)/\tau)},
\]
where \(\tau\) is a learnable temperature parameter. Let \(y^{y_{V2T}}(V)\) and \(y^{12V}(T)\) denote the ground-truth one-hot similarity, where only the positive pair has a probability of 1. The contrastive loss is defined as the cross-entropy \(H\) between \(p\) and \(y\):
\[
L_{cl} = \frac{1}{2} \mathbb{E}_{(V,T)\sim D} [H(y^{y_{V2T}}(V), p^{y_{V2T}}(V)) + H(y^{12V}(T), p^{12V}(T))]
\]
Matching: We determine whether a pair of visual concept and text is matched. For each visual concept in a mini-batch, we sample an in-batch hard negative text by following \(p^{y_{V2T}}(V)\) in Equation 4. Texts that are more relevant to the concept have a larger probability to be sampled. We also sample one hard negative visual concept for each text. We use \(p_{cls}\), the output \([CLS]\) embedding of the cross-modal encoder, to predict the matching probability \(p^{match}\), and the loss is:
\[
L_{match} = \mathbb{E}_{(V,T)\sim D} H(y^{match}, p^{match}(V, T))
\]
where \(y^{match}\) is a 2-dimensional one-hot vector representing the ground-truth label.
Masked Language Modeling: We predict the masked words in the text based on the visual concept. We randomly mask out the input tokens with a probability of 25%, and the replacements are 10% random tokens, 10% unchanged, and 80% \([MASK]\). We use the cross-modal encoder’s outputs, and append a linear layer followed by softmax for prediction. Let \(T\) denote a masked text, and \(p^j(V, \hat{T})\) denote the predicted probability of the masked token \(t_j\). We minimize a cross-entropy loss:
\[
L_{mlm} = \mathbb{E}_{t_j\sim T; (V, \hat{T})\sim D} H(y^j, p^j(V, \hat{T}))
\]
where \(y^j\) is a one-hot distribution where the ground-truth token for \(t_j\) has a probability of one.
Finally, the pre-training objective of X-VLM is defined as:
\[
\mathcal{L} = L_{box} + L_{cl} + L_{match} + L_{mlm}
\]
\subsection{4.1 Pre-training Datasets}
Following UNITER (Chen et al., 2020) and other existing work, we construct our pre-training data using two in-domain datasets, COCO (Lin et al., 2014) and Visual Genome (VG) (Krishna et al., 2016), and two out-of-domain datasets, SBU Captions (Ordonez et al., 2011) and Conceptual Captions (CC) (Sharma et al., 2018). The total number of unique images is 4.0M, and the number of image-text pairs is 5.1M. Following ALBEF (Li et al., 2021), we also exploit a much noisier Conceptual 12M dataset (CC-12M) (Changpinyo et al., 2021), increasing the total number of images to 14M.

The image annotations are from COCO and VG, which contain 2.5M object annotations and 3.7M region annotations for 200K images. The existing approaches exploit the data by training object detectors or using region annotations under the assumption that they can describe the whole images. In contrast, we take the object labels as text descriptions of objects, and re-formulate the image annotations so that an image has multiple regions enclosed by bounding boxes and each region is associated with a text. The text describes the visual concept in the region, which could be an object, a region, or the image itself.

Since most downstream V+L tasks are built on top of COCO and VG, we exclude all images that also appear in the validation and test set of downstream tasks to avoid information leak. We also exclude all co-occurring Flickr30K (Plummer et al., 2015) images via URL matching, as COCO and VG are from Flickr, and there are some overlaps.

\subsection{4.2 Implementation Details}
The image encoder is a twelve-layer ViT\textsubscript{base} (Dosovitskiy et al., 2020) with 86M parameters, which is initialized with weights pre-trained on ImageNet-1k from Touvron et al. (2021). The text encoder and the cross-modal encoder are respectively a six-layer BERT\textsubscript{base} (Devlin et al., 2019) with a total of 124M parameters. The text encoder is initialized using the first six layers of BERT\textsubscript{base}, and the cross-modal encoder is initialized using the last six layers. For text input, we set the maximum number of tokens to 25.

X-VLM\textsubscript{base} takes images of 256 × 256 resolution as input. The image encoder splits an image into non-overlapping 16 × 16 patches. We apply full
attention to the last four layers and mixed attention to the first four layers. During fine-tuning, we increase the image resolution to $384 \times 384$ and interpolate the positional embeddings of image patches following Dosovitskiy et al. (2020).

We pre-train the model with mixed precision for 50 epochs on 32 NVIDIA V100 GPUs. The batch size is set to 3072, and we sample the data by making half of the images in a batch contain bounding box annotations. Training with 4M images takes three days. We use the AdamW (Loshchilov and Hutter, 2018) optimizer with a weight decay of $0.02$. The learning rate is warmed-up to $1e^{-4}$ from $1e^{-5}$ in the first 1/5 iterations and decayed to $1e^{-5}$ iterations following a cosine schedule.

4.3 Downstream Tasks

We adapt X-VLM to four downstream V+L tasks. We follow the settings in previous work on fine-tuning (see Appendix B).

**Image-Text Retrieval** There are two subtasks: text retrieval (TR) and image retrieval (IR). We evaluate X-VLM on the MSCOCO and Flickr30K (Plummer et al., 2015) benchmark datasets. We optimize $\mathcal{L}_{tr}$ and $\mathcal{L}_{match}$ for fine-tuning. In inference, we first compute $s(I, T)$ for all images and texts, and then take the top-5 candidates and calculate $p_{match}(I, T)$ for ranking. Following ALBEF, $k$ is set to 256 for MSCOCO and 128 for Flickr30K.

**Visual Question Answering** (Goyal et al., 2017) It requires the model to predict an answer given an image and a question. Following the previous work (Cho et al., 2021; Li et al., 2021), we use a six-layer Transformer decoder to generate answers based on the outputs of the cross-modal encoder. We also constrain the decoder to only generate from the 3,192 candidate answers during inference to make a fair comparison with existing methods.

**Natural Language for Visual Reasoning** (NLVR$^2$ (Suhr et al., 2018)) The task asks the model to determine whether a text describes the relations between two images. Following ALBEF, we extend the cross-modal encoder to enable reasoning over two images, and perform an additional pre-training step for one epoch using the 4M images: given two images and a text, the model assigns the text to either the first image, the second image, or none of them.

**Visual Grounding** We evaluate X-VLM on RefCOCO+ (Yu et al., 2016) in both supervised and weakly-supervised settings, which aims to locate the region in an image that corresponds to a specific text description. In the supervised setting with bounding box annotations, we fine-tune the model by optimizing $\mathcal{L}_{bbox}$. In the weakly-supervised setting where only image-text pairs are available, we fine-tune the model using $\mathcal{L}_{cl}$ and $\mathcal{L}_{match}$; in inference, following ALBEF, we apply GradCAM (Selvaraju et al., 2017) to acquire heatmaps and use them to rank the detected proposals provided by (Yu et al., 2018).

4.4 Evaluation on Image-Text Retrieval

Table 1 compares X-VLM with SoTA approaches on MSCOCO and Flickr30K. Under the 4M setting, X-VLM outperforms all the previous methods based on either object-centric features or overall image features by a large margin, showing the effectiveness of our multi-grained approach. Furthermore, when training with more instances (14M), X-VLM also achieves new SoTA results by a significant margin, outperforming all existing methods including ALIGN which is trained with an in-house 1.8B dataset. X-VLM (14M) also obtains considerable amount of improvement compared with X-VLM (4M), which shows that our approach is scalable with noisier web data.

4.5 Evaluation on Visual Reasoning

Table 2 shows experimental results on visual reasoning (VQA and NLVR$^2$). First, one can see that though ALBEF is comparable to VinVL on image-text retrieval, Pixel-BERT, SOHO, ViLT, and ALBEF, which do not use object detection, have much worse performance than VinVL in visual reasoning. Nevertheless, X-VLM (4M) without object detection is comparable to VinVL, achieving SoTA performance in visual reasoning. Meanwhile, as reported in Li et al. (2021), X-VLM, which utilizes vision transformer as image encoder, enjoys 10 times faster inference speed than VinVL. Furthermore, with extra training data, X-VLM (14M) achieves absolute improvements of 0.79% in VQA and 1.06% in NLVR$^2$ (average on metrics) over VinVL. The results indicate that our approach of X-VLM is both effective and efficient in visual reasoning.

4.6 Evaluation on Visual Grounding

Table 2 reports the performance of X-VLM in both supervised and weakly-supervised settings. In the weakly-supervised setting, following ALBEF, we only utilize image-text pairs in fine-tuning.
Table 1: Image-text retrieval results on MSCOCO and Flickr30K datasets. IR: Image Retrieval and TR: Text Retrieval.

| Method       | # Pre-train | MSCOCO (5K test set) | Flickr30K (1K test set) |
|--------------|-------------|-----------------------|-------------------------|
|              |             | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| UNITER 4M    |             | 64.4 | 87.4 | 93.1 | 50.3 | 78.5 | 87.2 | 85.9 | 97.1 | 98.8 | 72.5 | 92.4 | 96.1 |
| OSCAR 4M     |             | 70.0 | 91.1 | 95.5 | 54.0 | 80.8 | 88.5 | - | - | - | - | - | - |
| ALBEF 4M     |             | 73.1 | 91.4 | 96.0 | 56.8 | 81.5 | 89.2 | 94.3 | 99.4 | 99.8 | 82.8 | 96.7 | 98.4 |
| VinVL 5.6M   |             | 74.6 | 92.6 | 96.3 | 58.1 | 83.2 | 90.1 | - | - | - | - | - | - |
| X-VLM 4M     |             | 78.8 | 94.3 | 97.5 | 60.6 | 84.2 | 90.5 | 96.0 | 99.7 | 99.9 | 84.1 | 96.9 | 98.4 |
| ALIGN 1.8B   |             | 77.0 | 93.5 | 96.9 | 59.9 | 83.3 | 89.8 | 95.3 | 99.8 | 100.0 | 84.9 | 97.4 | 98.6 |
| ALBEF 14M    |             | 77.6 | 94.3 | 97.2 | 60.7 | 84.3 | 90.5 | 95.9 | 99.8 | 100.0 | 85.6 | 97.5 | 98.9 |
| X-VLM 14M    |             | 79.7 | 95.3 | 97.6 | 62.5 | 85.3 | 91.0 | 96.8 | 99.7 | 99.9 | 86.0 | 97.2 | 98.7 |

Table 2: Comparison on downstream V+L tasks. RefCOCO+ scores with * are in the weakly-supervised setting.

| Method               | VQA           | NLYR²          | RefCOCO+  |
|---------------------|---------------|---------------|------------|
|                     | test-dev      | test-std      | test-P     |
| VL-BERT (Lu et al., 2019) | 70.55 | 70.92 | - | 72.34 | 78.52 |
| ViLBERT (Su et al., 2019)       | 71.16 | - | - | 72.59 | 78.57 |
| LXMERT (Tan and Bansal, 2019)  | 72.42 | 72.54 | 74.90 | 74.50 | - | - |
| UNITER (Chen et al., 2020)     | 72.70 | 72.91 | 77.18 | 77.85 | 75.31 | 81.30 |
| 12-in-1 (Lu et al., 2020)      | 73.15 | - | - | 78.87 | - | - |
| OSCAR (Li et al., 2020c)       | 73.16 | 73.44 | 78.07 | 78.36 | - | - |
| Pixel-BERT (Huang et al., 2020) | 74.45 | 74.55 | 76.50 | 77.20 | - | - |
| VILLA (Gan et al., 2020)       | 73.59 | 73.67 | 78.39 | 79.30 | 76.05 | 81.65 |
| SOHO (Huang et al., 2021)      | 73.25 | 73.47 | 76.37 | 77.32 | - | - |
| ViLT (Kim et al., 2021)        | 70.94 | - | 75.24 | 76.21 | - | - |
| ALBEF (Li et al., 2021)        | 74.54 | 74.70 | 80.24 | 80.50 | 57.85* | 65.89* | 46.43* |
| ALBEF (14M)                   | 75.84 | 76.04 | 82.55 | 83.14 | 58.46* | 65.89* | 46.25* |
| VinVL (5.6M) (Zhang et al., 2021) | 75.95 | 76.12 | 82.05 | 83.08 | - | - | - |
| X-VLM (4M)                    | 76.20 | 76.23 | 82.40 | 82.42 | 75.81/64.95* | 82.13/72.07* | 68.26/54.84* |
| X-VLM (14M)                   | 76.77 | 76.89 | 83.40 | 83.84 | 76.31/64.79* | 82.78/72.46* | 67.40/55.08* |

The evaluation results are denoted with * in Table 2. Compared to ALBEF (14M), X-VLM (14M) achieves absolute improvements of 7.2% (average on metrics). In the supervised setting, X-VLM (4M) achieves absolute improvements of 0.9%, compared to VILLA. Note that X-VLM directly predicts the bounding boxes in the image, while VILLA relies on pre-trained object detectors to first localize all salient regions in the image and then formulate the task as a ranking problem, which appears to be trickier because the detectors have already seen the images during pre-training.

Figure 3 provides a few examples of images from the test set of RefCOCO+. For the supervised setting, we show the bounding boxes predicted by X-VLM given the text descriptions. For the weakly-supervised setting, following ALBEF, we provide the Grad-CAM visualization, which uses the cross-attention maps in the third layer of the cross-modal encoder. The visualization examples show that X-VLM has a strong ability of cross-modal understanding. It successfully predicts the correct regions in images, even though the text descriptions only differ in a single word. Furthermore, X-VLM can align each word in the text to the corresponding image region. We provide more examples in Appendix A, showing X-VLM’s superior performance in vision language alignment.

4.7 Ablation Study

Table 3 provides the results of ablation study on X-VLM (4M). We use R@1 as an evaluation measure in the retrieval tasks and Meta-Sum as a general measure. We report two Meta-Sum scores: one sums over all the scores, and the other marked by * excludes the results of supervised RefCOCO+.

First, we evaluate the effectiveness of visual concepts in different granularities, i.e. w/o object and w/o region. The results show that training without either of them hurts the performance, demonstrating the necessity of learning multi-grained alignments. Besides, we can observe that w/o region makes the performance drop more drastically than w/o object. Furthermore, the ablation study shows that bounding box prediction is a critical component of X-VLM, as w/o bbox loss leads to the lowest Meta-Sum. Especially for RefCOCO+, there is a decrease of 3.6% (average on metrics). The mixed attention mechanism also brings improvement on the performance as indicated in Table 3.
Figure 3: Grad-CAM visualization and bounding box prediction on unseen images.

Table 3: Ablation study on X-VLM: w/o object is training without concepts of object; w/o region is training without concepts of region; w/o bbox loss is without bounding box prediction; w/o mixed att is applying completely local attention to obtain fine-grained concept representations; w/o all represents removing all above components.

| Component       | Meta-Sum | MSCOCO | Flickr30K | VQA | NLVR \(^2\) | ReFCOCO+ |
|-----------------|----------|--------|-----------|-----|-------------|----------|
| X-VLM           | 755.4/605.0\(^*\) | 78.8 | 60.6 | 96.0 | 84.1 | 76.20 | 82.42 | 82.13/72.07\(^*\) | 68.26/54.84\(^*\) |
| w/o object      | 749.7/603.5\(^*\) | 77.4 | 60.4 | 95.0 | 83.7 | 75.87 | 82.10 | 81.19/73.37\(^*\) | 64.94/55.69\(^*\) |
| w/o region      | 742.0/596.0\(^*\) | 76.8 | 60.2 | 96.0 | 83.6 | 75.84 | 82.20 | 80.11/70.73\(^*\) | 64.12/50.60\(^*\) |
| w/o bbox loss   | -594.9\(^*\) | 77.5 | 60.2 | 95.7 | 83.5 | 76.77 | 81.49 | -69.32\(^*\) | -50.38\(^*\) |
| w/o mixed att   | 749.9/600.8\(^*\) | 77.0 | 60.1 | 95.4 | 83.2 | 75.87 | 82.55 | 82.08/72.16\(^*\) | 67.05/54.51\(^*\) |
| w/o all         | -580.6\(^*\) | 74.5 | 57.9 | 95.6 | 82.8 | 74.90 | 80.70 | -67.79\(^*\) | -46.43\(^*\) |

With mixed attention, object or region representations encoding the contextual information of images are more accurate. Mixed attention is also an efficient implementation, which saves 50% GPU memory. In general, though different components in X-VLM perform differently in various downstream V+L tasks, they all contribute to better overall performances and/or efficient implementations.

We also report the results of ‘w/o all’. Though in the 4M setting, only 5.0% of the images have dense annotations, X-VLM can leverage the data and substantially improve the performances in the downstream V+L tasks (Meta-Sum from 580.6\(^*\) to 605.2\(^*\)). As mentioned above, when training with more image-text pairs, X-VLM (14M) can still make a considerable amount of improvement over X-VLM (4M). It is worth noticing that in the 14M setting, only 1.4% of the images have annotations. The results show that our approach of X-VLM can scale up to large-scale data. One can also observe that fine-grained visual concepts, e.g. objects, are much less diverse than images. Still, they are components of images and have a significant impact on learning vision language alignment.

5 Conclusion

The existing vision language models utilize either fine-grained object-centric features of images or coarse-grained overall features of images. Although effective, both methods still have some drawbacks. In this paper, we have reconstructed existing datasets into image/region/object-text pairs, and proposed X-VLM, a novel approach to perform multi-grained vision language pre-training. Training of the model is driven by locating visual concepts in the image given the associated texts and aligning texts and visual concepts that are relevant, where the alignments are in multi-granularity. Experiments on many V+L tasks have shown that X-VLM achieves substantial improvements over existing SoTA methods.
Acknowledgements

We would like to thank Wenguan Huang and Xiu-jun Li at ByteDance for their generous assistance in data collection and insightful comments in technical discussions.

References

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, pages 6077–6086.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In European Conference on Computer Vision, pages 213–229. Springer.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu 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, pages 3558–3568.

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, pages 104–120. Springer.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. arXiv preprint arXiv:2102.02779.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186.

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.

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. arXiv preprint arXiv:2006.06195.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6904–6913.

Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. 2021. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12976–12985.

Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. arXiv preprint arXiv:2004.00849.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. arXiv preprint arXiv:2102.05918.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3128–3137.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. arXiv preprint arXiv:2102.03334.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. arXiv preprint arXiv:1602.07332.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020a. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 11336–11344.

Junnan Li, Ramprasaath R Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. arXiv preprint arXiv:2107.07651.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557.

Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2020b. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. arXiv preprint arXiv:2012.15409.
Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Xiujun Li, Xi Yin, Chunjuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020c. Oscar: Object-semantics aligned pre-training for vision-language tasks. In European Conference on Computer Vision, pages 121–137. Springer.

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, pages 740–755. Springer.

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.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In International Conference on Learning Representations.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in Neural Information Processing Systems, 32:13–23.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10437–10446.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. Advances in neural information processing systems, 24:1143–1151.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In Proceedings of the IEEE international conference on computer vision, pages 2641–2649.

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.

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, 28:91–99.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 658–666.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision, pages 618–626.

Piyou 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), pages 2556–2565.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vi-lbert: Pretraining of generic visual-linguistic representations. In International Conference on Learning Representations.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2018. A corpus for reasoning about natural language grounded in photographs. arXiv preprint arXiv:1811.00491.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5100–5111.

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, pages 10347–10357. PMLR.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008.

Haiyang Xu, Ming Yan, Chenliang Li, Bin Bi, Songfang Huang, Wenming Xiao, and Fei Huang. 2021. E2e-vlp: End-to-end vision-language pretraining enhanced by visual learning. arXiv preprint arXiv:2106.01804.

Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. 2018. Mattnet: Modular attention network for referring expression comprehension. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1307–1315.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In European Conference on Computer Vision, pages 69–85. Springer.
A Case Study

Figure 4 and 5 provide visualizations of some images from the test set of RefCOCO+. We show the bounding boxes predicted by X-VLM given the text descriptions. For the weakly-supervised setting, following ALBEF, we provide the Grad-CAM visualization which uses the cross-attention maps in the 3rd layer of the cross-modal encoder.

B Implementation Details of Downstream Tasks

We follow the settings in existing methods on fine-tuning. We describe how we implement fine-tuning on the downstream V+L tasks, and we also provide our fine-tuning scripts for more details. Note that we have cleaned our pre-training dataset to avoid data leak since downstream V+L tasks have overlaps in images with COCO and Visual Genome. However, the leak cannot be easily detected for the pre-trained object detectors.

Image-Text Retrieval We evaluate X-VLM on MSCOCO and Flickr30K (Plummer et al., 2015) benchmarks. We adopt the widely used Karpathy split (Karpathy and Fei-Fei, 2015) for both datasets. We optimize $L_{cl}$ and $L_{match}$ for fine-tuning. Since there are multiple ground-truth texts associated with each image in the datasets, we change the ground-truth similarity of contrastive learning, $y^{2t}(I)$ and $y^{12v}(T)$, to consider multiple positives, where each positive has a probability of $\frac{1}{n_{positives}}$. We fine-tune the model for 10 epochs. During inference, we first compute $s(I,T)$ for all images and texts. Then we take the top-$k$ candidates and calculate $p_{match}(I,T)$ for ranking. Following ALBEF, $k$ is set to 256 for MSCOCO and 128 for Flickr30K.

Visual Question Answering (VQA (Goyal et al., 2017)) Following existing methods (Tan and Bansal, 2019; Chen et al., 2020; Li et al., 2021), we use both train and validation sets for training, and include additional question-answer pairs from Visual Genome. The VQA model contains a 6-layer transformer-based decoder to generate answers based on the outputs of cross-modal encoder following previous work (Cho et al., 2021; Li et al., 2021). The decoder is initialized using the pre-trained weights from the cross-modal encoder. Then, it is fine-tuned by optimizing the auto-regressive loss for 8 epochs. During inference, we constrain the decoder to only generate from the 3,192 candidate answers to make a fair comparison with existing methods.

Natural Language for Visual Reasoning (NLVR$^2$ (Suhr et al., 2018)) Since the task asks the model to distinguish whether a text describes a pair of images, we follow ALBEF to extend the cross-modal encoder to enable reasoning over two images. We also perform an additional pre-training step for 1 epoch using the 4M images: given a pair of images and a text, the model needs to assign the text to either the first image, the second image, or none of them. Then, we fine-tune the model for 10 epochs.

Visual Grounding The task aims to localize the region in an image that corresponds to a specific text description (RefCOCO+ (Yu et al., 2016)). We evaluate our approach on both supervised and weakly-supervised settings. In the supervised setting with bounding box annotations, we perform an additional pre-training step for 1 epoch using $L_{box}$ only. Then, we fine-tune the model for 10 epochs. In the weakly-supervised setting where only image-text pairs are available, we fine-tune the model using $L_{cl}$ and $L_{match}$ for 5 epochs. During inference, following ALBEF, we apply Grad-CAM (Selvaraju et al., 2017) to acquire heatmaps and use them to rank the detected proposals provided by (Yu et al., 2018).
Figure 4: Locating visual concepts in unseen images given text descriptions. Since Grad-CAM visualization corresponds to individual words, we only show the figure of the subject word, e.g. "dog" for "brown dog".
Figure 5: Bounding box prediction and per-word visualization on unseen images. It shows that X-VLM can also align concept like “pulling” and “holding” to the corresponding regions in the images.