

## Seeing Out of tHe bOx: End-to-End Pre-training for Vision-Language Representation Learning

Zhicheng Huang<sup>1,2\*</sup>, Zhaoyang Zeng<sup>3\*</sup>, Yupan Huang<sup>3\*</sup>, Bei Liu<sup>4</sup>, Dongmei Fu<sup>1,2</sup>, Jianlong Fu<sup>4</sup>

<sup>1</sup>School of Automation and Electrical Engineering, University of Science and Technology Beijing

<sup>2</sup>Beijing Engineering Research Center of Industrial Spectrum Imaging

<sup>3</sup>Sun Yat-sen University, <sup>4</sup>Microsoft Research Asia

### Abstract

We study joint learning of Convolutional Neural Network (CNN) and Transformer for vision-language pre-training (VLPT) which aims to learn cross-modal alignments from millions of image-text pairs. State-of-the-art approaches extract salient image regions and align regions with words step-by-step. As region-based visual features usually represent parts of an image, it is challenging for existing vision-language models to fully understand the semantics from paired natural languages. In this paper, we propose **SOHO** to “Seeing Out of tHe bOx” that takes a whole image as input, and learns vision-language representation in an end-to-end manner. SOHO does not require bounding box annotations which enables inference 10 times faster than region-based approaches. In particular, SOHO learns to extract comprehensive yet compact image features through a visual dictionary (VD) that facilitates cross-modal understanding. VD is designed to represent consistent visual abstractions of similar semantics. It is updated on-the-fly and utilized in our proposed pre-training task Masked Visual Modeling (MVM). We conduct experiments on four well-established vision-language tasks by following standard VLPT settings. In particular, SOHO achieves absolute gains of 2.0% R@1 score on MSCOCO text retrieval 5k test split, 1.5% accuracy on NLVR<sup>2</sup> test-P split, 6.7% accuracy on SNLI-VE test split, respectively.

### 1. Introduction

With the success of Transformer and self-supervised learning, we have recently witnessed a boosting number of research works on cross-modal learning, especially on vision-language pre-training (VLPT) [7, 22, 23, 27, 34, 37, 48]. VLPT models learn better cross-modal representation with large-scale easy-accessible image-text pairs. They have established state-of-the-art results in many vision-

\*Equal Contribution. This work was performed when Zhicheng Huang, Zhaoyang Zeng and Yupan Huang were visiting Microsoft Research Asia as research interns.

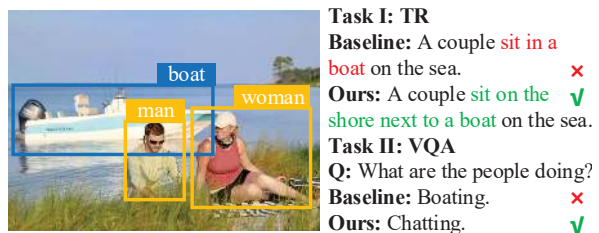


Figure 1: Comparisons of SOHO and region-based methods by top-1 image-to-text retrieval (TR) and visual question answering (VQA) results. Baselines lack global context and fail to understand the image. SOHO discovers visual clues out of region boxes and infers correct human activities. [Best viewed in color.]

language tasks, such as visual question answering (VQA) [3], image-text retrieval [25], natural language for visual reasoning (NLVR) [35], etc.

Visual representation plays an important role in VLPT models. The recent success of VLPT models has been accompanied by the usage of region-based image features, which are extracted by object detectors pre-trained on the Visual Genome dataset [2]. However, there are three challenges to directly utilize region-based image features for vision-language understanding. Firstly, regions focus on objects inside bounding boxes while neglecting the contextual information out of the boxes, which is important for relation understanding and reasoning. For example in Figure 1, we can easily detect “man”, “woman” and “boat” in the image. However, without the contextual information out of these boxes, a model will misunderstand the relation as “people boating” and result in an incorrect answer for either text retrieval or VQA. Secondly, visual understanding of images will be limited to the pre-defined categories for regions. Thirdly, most region-based image features are extracted by a detection model, which will suffer from low quality, noise, and over-sampling [2] and rely on large-scale boxes annotation data. Although some works try to train detection model[38, 46] with weakly-supervised, the performance is far below the requirements. Recently, some

works challenge that grid-based convolutional features are also effective to learn visual representations [9, 16, 17, 32]. Among them, Jiang *et al.* show that grid features can be equally effective as region features for VQA [17]. Sariyildiz *et al.* and Desai *et al.* use image-text data to train visual backbone for recognition tasks (e.g., image classification) [9, 32]. These models are designed either for specific vision-language task [17] or vision task [9, 32]. In this paper, we focus on VLPT and propose one of the first end-to-end VLPT model without relying on region features.

To overcome the limitation of region-based image features and better utilize image-text pairs for cross-modal understanding, we propose **SOHO**, an end-to-end vision-language pre-training framework to directly learn image embedding, language embedding, and their semantic alignment from image-text pairs. Compared with existing VLPT works, SOHO adopts a simple pipeline that does not require a complicated visual backbone for pre-training and releases the design effort for VLPT tasks. Without the requirement of laborious annotated categories or boxes, SOHO can enrich visual semantics by directly optimizing visual representations by a wider range of image-text data.

End-to-end learning for vision and language raises challenges by different representations of the two modalities. Visual representation at pixel-level is much more diverse and dense than language embedding. And the lack of explicit supervision for pixel-level language adds the difficulty to alignment learning. To tackle the above problems, we introduce a visual dictionary (VD) which represents more comprehensive and compact semantics in visual domain. To learn the visual dictionary, we design a moving-averaged encoder to group visual pixels with similar visual semantics. VD can be dynamically updated through our trainable CNN backbone directly from visual-language data during pre-training. For pre-training tasks, we propose a novel Masked Vision Modeling (MVM) based on the learned visual dictionary besides two commonly used tasks, Masked Language Modeling (MLM) and Image-Text Matching (ITM).

Our contributions are summarized as follows: (i) We propose SOHO, one of the first end-to-end VLPT models to learn cross-modal representation directly with image-text pairs. Without the need of extracting bounding boxes, our model can achieve at least 10 times speedup for inference. (ii) To better align visual features and language tokens, we propose a novel dynamic-updated visual dictionary that represents a visual abstraction of similar semantics in images. (iii) We conduct extensive experiments on four well-established downstream tasks. Experimental results show that SOHO can improve the SOTA performance with absolute gains of 2.0% R@1 score on MSCOCO text retrieval 5k test split, 1.5% accuracy on NLVR<sup>2</sup> test-P split, 6.7% accuracy on SNLI-VE test split, and 0.56% VQA score on VQA2.0 test-std split. We will release both model and code

to facilitate the research community<sup>1</sup>.

## 2. Related Work

### 2.1. Visual Representation for Vision-Language

Visual representation learning for vision-language understanding is a long-standing research topic. Early works use CNN classification models pre-trained on ImageNet to extract visual features [8, 24, 26, 28, 44, 45]. Later on, Anderson *et al.* propose a Bottom-Up and Top-Down Attention (BUTD) detection model [2] pre-trained on Visual Genome dataset to extract salient region features as visual inputs for VQA and image captioning tasks. The BUTD features are adopted by many vision-language works [2, 20, 33, 35] and pre-training works [7, 19, 37]. Recently, some works propose to directly learn visual representations in the form of grid features with convolutional networks in specific vision-language tasks [17] or vision recognition tasks [9, 32]. Our work shares a similar format of visual representation with [17] while we focus on the area of vision-language pre-training and propose the first end-to-end VLPT model without relying on the box annotations.

VideoBERT [36] and the bag of words [13] literature also use vector quantization to represent visual information. The key difference between VD and related works is that we dynamically update the VD-based embedding with the output of a trainable visual encoder, instead of pre-computed input features. The dynamic updating mechanism for VD can capture text-guided semantics from the vision-language dataset. Thus the model can be directly optimized with high-level semantics for VL understanding and alignment.

### 2.2. Pre-training for Vision-Language

Many vision-language pre-training (VLPT) works have been proposed to learn cross-modal representations [7, 22, 23, 27, 34, 36, 37, 48]. They can be categorized as two-stream or single-stream models. The two-stream models process visual and language information respectively and fuse them afterward by another Transformer layer [27, 37]. Contrarily, the single-stream models use BERT [10] to learn a bi-directional joint distribution over the detection bounding box feature and text embedding feature [1, 7, 22, 23, 34, 48]. Both types use the Transformer-based model to learn vision-language joint embedding features. While they neglect that visual representation learning is also important to vision-language tasks.

The key differences between our SOHO and existing VLPT works are 1) SOHO adopts a simple VLPT pipeline. Our vision backbone only uses ImageNet pre-trained parameters, and achieves even higher performance than existing VLPT works using VG features on five downstream tasks. 2) SOHO uses the least annotations to achieve SOTA

<sup>1</sup><https://github.com/researchm/soho>

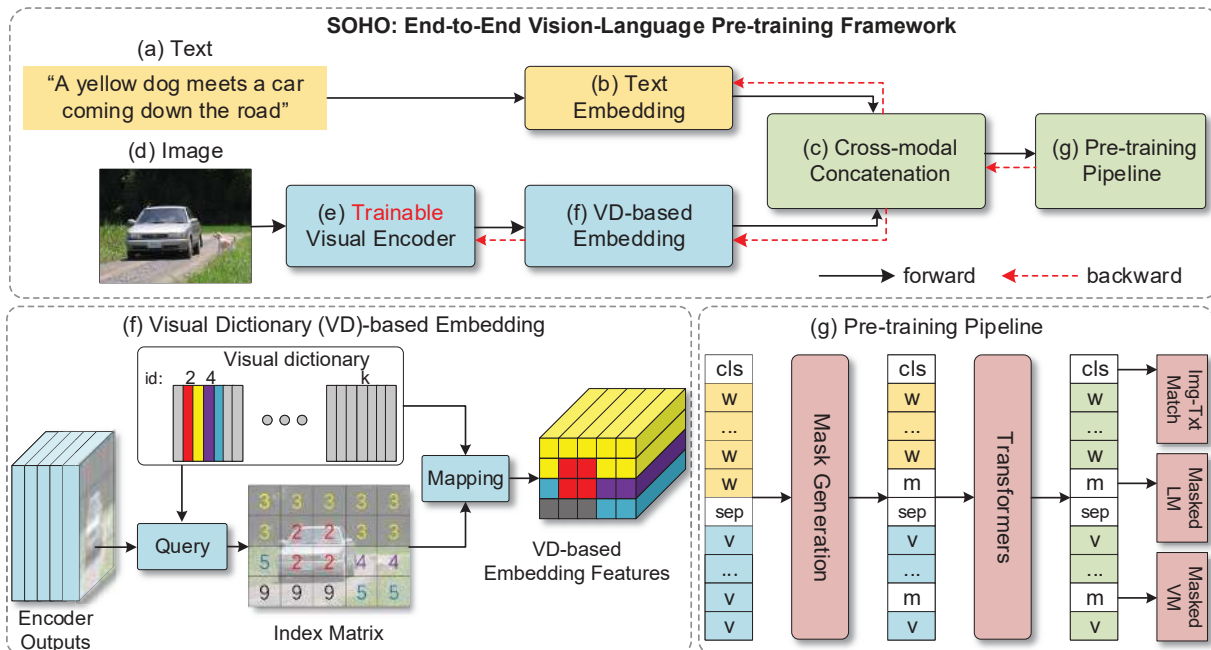


Figure 2: The framework of the proposed end-to-end pre-training model SOHO. For an input text (a), we use the text embedding operation (b) to extract the textual embedding features. For an input image (d), we propose to use a trainable CNN-based encoder (e) to extract visual representations. To further transform image features to consistent semantics, we apply a visual dictionary-based image embedding (f) to the image encoder outputs. Finally, we apply multi-layer Transformers to the output of multi-modal concatenation (c) with three pre-training tasks. Note that the index matrix in (f) will be used as labels in the masked VM task in (g). [Best viewed in color.]

performances. 3) SOHO enriches visual semantics by directly optimizing visual inputs for target language tasks.

### 3. Approach

The overall architecture of our proposed vision-language pre-training framework SOHO is shown in Figure 2. SOHO is an end-to-end framework, which consists of a trainable CNN-based visual encoder, a visual dictionary (VD) embedding module, and a multi-layer Transformer. The visual encoder takes an image as input and produces the visual features. VD embedding module is designed to aggregate diverse visual semantic information into visual tokens with a proposed visual dictionary. The Transformer is adopted to fuse features from visual and language modalities, and produce task-specific output. SOHO can be end-to-end pre-trained by Masked Vision Modeling (MVM), Masked Language Modeling (MLM), and Image-Text Matching (ITM) tasks. SOHO can also be easily adapted to several downstream tasks including Image-Text Retrieval, VQA, NLVR, and Visual Entailment.

#### 3.1. Trainable Visual Encoder

Most recent vision-language researches follow Bottom-up and Top-Down attention [2] to extract region-level visual features by a Faster R-CNN [31] detector which is pre-trained on the Visual Genome [20] dataset. The represen-

tation ability of such extracted region-based features will be limited by the pre-defined object and attribute categories (i.e. 1,600 objects and 400 attributes). Besides, some contextual information out of the regions is important for VL understanding, while being neglected because they are out of the pre-defined categories. Even though considering the whole image as a region and extract its feature as global representation is an improved solution, the detector can not guarantee the feature quality because such global regions are unseen in the training stage. Despite that, most recent VLPT models adopt pre-extracted region-level visual features because it is complicated to end-to-end fine-tune an object detector in VL tasks. Besides, the extracted region-level visual features have a semantic gap with the language domain, while existing works try to bridge such domain gap by only one or several fully-connected layers.

To keep all visual information, we propose to use a trainable CNN visual encoder, which takes the whole image as input and produces image-level visual features instead of region-level features. Without the limitation of bounding boxes, the visual encoder can be end-to-end learned and updated from pre-training losses or downstream task-specific losses, and further optimize the cross-modal learning in turn. Given an input image  $\mathcal{I}$ , we get its feature  $\mathcal{V}$  by:

$$\mathcal{V} = E(\mathcal{I}, \theta) \in \mathbf{R}^{l \times c}, \quad (1)$$

where  $E(\cdot, \theta)$  is the visual feature encoder with parameter

$\theta$ .  $l$  denotes the number of embedded feature vectors, and  $c$  is the embedded dimension. We adopt ResNet [15] pre-trained on ImageNet [8] followed by a  $1 \times 1$  convolutional layer and a  $2 \times 2$  max pooling layer as the architecture of the encoder  $E$ . For simplicity, we use  $v_i$  to denote the  $i^{th}$  feature vector of  $\mathcal{V}$  for the rest of this paper.

### 3.2. Visual Dictionary

The visual feature  $\mathcal{V}$  extracted by visual feature encoder is more diverse and dense than language word tokens, which will bring difficulty to the learning of cross-modal understanding. To bridge its representation gap from language tokens, we propose a visual dictionary (VD) to tokenize the visual features by aggregating similar visual semantic into the same image feature.

**Visual Dictionary Embedding.** We define a visual dictionary (VD) as a matrix  $\mathcal{D} \in \mathbf{R}^{k \times c}$  which contains  $k$  embedding vectors with  $c$ -dim. The  $j^{th}$  embedding vector is denoted as  $d_j$ . For visual feature  $v_i$ , we compute its mapping index by searching nearest neighbor in  $\mathcal{D}$ , denoted as:

$$h_i = \operatorname{argmin}_j \|v_i - d_j\|_2. \quad (2)$$

We define visual dictionary embedding as a mapping function  $f$ , which maps  $v_i$  to  $\mathcal{D}$  by:

$$f(v_i) = d_{h_i}, \quad (3)$$

which uses the nearest embedding vector to represent the visual feature. We denote  $f^{-1}(j)$  as an inverse mapping function, which maps the index  $j$  back to a group of visual features. We use  $|f^{-1}(j)|$  to represent the inverse mapping group size, and use  $f(\mathcal{V})$  to represent the encoding features.

**Momentum Learning for Visual Dictionary Update.** The visual dictionary  $\mathcal{D}$  is randomly initialized, and further updated by a moving average operation in one mini-batch, which is denoted as:

$$\hat{d}_j = \gamma * d_j + (1 - \gamma) * \frac{\sum_{h_i=j} v_i}{|f^{-1}(j)|}, \quad (4)$$

where  $\hat{d}_j$  indicates the updated embedding vector of  $d_j$ , and  $\gamma$  is a momentum coefficient whose value range is  $[0, 1]$ . Note that Eqn. 4 will only be applied when  $|f^{-1}(j)| \neq 0$ .

**Gradient Back Propagation.** Since the argmin operation is not differentiable, the gradient back propagation will be stopped by the visual dictionary. To make the visual feature encoder trainable, we follow [39] to update  $f(v_i)$  by:

$$f(v_i) = \operatorname{sg}[d_{h_i} - v_i] + v_i, \quad (5)$$

where  $\operatorname{sg}[\cdot]$  is the stop gradient operator.

The visual dictionary performs an online clustering on visual feature maps based on feature similarity, and represents each feature vector by its cluster center. Feature vectors sharing similar semantics will be aggregated into the same cluster, and the clustered index can be considered as a virtual visual semantic label. Since the clustering can be affected by the vision-language learning tasks, the learned

semantics of each embedding vector is more suitable for cross-modal understanding and alignment.

The visual dictionary faces a cold-start problem, where directly copying the gradient from randomly initialized embedding vectors to visual feature maps will lead to incorrect model optimization direction (i.e., mode collapse). Therefore, we freeze the parameters of ResNet in the visual feature encoder in the first 10 training epochs.

### 3.3. Pre-training Pipeline

We apply a multi-layer Transformer to learn cross-modal representations with the fusion of visual and language features. In order to learn a universal representation for vision and language-related tasks, we apply the self-supervised method to pre-train the model on a large aggregated dataset. We follow the existing works [7, 22, 27, 34, 37, 48] to adopt Masked Language Modeling (MLM) and Image-Text Matching (ITM) pre-training tasks. Besides, we propose a novel Masked Visual Modeling (MVM) pre-training task based on the virtual visual semantic labels produced by the visual dictionary.

**Cross-Modal Transformer.** For visual representation, we utilize 2-D position embedding computed by sine function to encode spatial information of visual tokens following other works [6, 11, 29]. For the input sentence, we follow BERT [10] to tokenize it, and then represent the tokens by embedding vectors  $\mathcal{W}$ . We use  $w_i$  to denote the  $i^{th}$  embedding vector in  $\mathcal{W}$ . The word embedding and the VD embeddings share the dimension  $c$  on their outputs. We concatenate the VD embeddings and word embedding vectors together to form an input sequence for cross-modal learning. Similar to other VLPT models, we add two special tokens  $[CLS]$  and  $[SEP]$  into the input sequence to indicate classification position and the end of a text, respectively. A multi-layer Transformer is adopted to take the joint vision-language input, and outputs the attended features.

**Masked Language Modeling.** We follow [7] and adopt Masked Language Modeling (MLM) to encourage the model to build the mapping between language tokens and visual contents. The goal of MLM is to predict the masked word tokens based on other word tokens  $\mathcal{W}_{\setminus i}$  and all image features  $f(\mathcal{V})$  by minimizing the negative log-likelihood. The learning target can be formulated as:

$$\mathcal{L}_{\text{MLM}} = -\mathbb{E}_{(\mathcal{W}, f(\mathcal{V})) \sim D} \log p(w_i | \mathcal{W}_{\setminus i}, f(\mathcal{V})), \quad (6)$$

where  $D$  indicate hereinafter the whole training dataset. We adopt the same masking strategy used in BERT [10].

**Masked Visual Modeling.** We propose Masked Visual Modeling (MVM) by visual dictionary, which is a symmetry to the MLM. We randomly mask the image features before feeding them into the Transformer. The learning target of MVM is denoted as:

$$\mathcal{L}_{\text{MVM}} = -\mathbb{E}_{(\mathcal{W}, f(\mathcal{V})) \sim D} \log p(f(v_j) | \mathcal{W}, f(\mathcal{V})_{\setminus j}). \quad (7)$$

The goal of MVM is to predict the masked image features based on their surrounding image features  $f(\mathcal{V})_{\setminus j}$  and

Table 1: Statistics of different tasks. Notation “\*” denotes Karpathy split [18]. Notation “-” denotes not applicable. Detailed train/test image and text numbers can be found in the supplementary material.

Task	Dataset	Train Split	Test Split	Metric
Pre-training	VG [20]	train	-	-
	MSCOCO [25]	train+restval*	-	-
Image-Text Retrieval	MSCOCO [25]	train+restval*	test*	Recall@1,5,10
	Flickr30K [30]	train	test*	
Visual Question Answering	VQA2.0 [14]	train+val	test-dev/test-std	VQA-score [14]
Visual Reasoning	NLVR <sup>2</sup> [35]	train	dev/test-P	Top-1 Accuracy
Visual Entailment	SNLI-VE [42]	train	val/test	Top-1 Accuracy

all language tokens  $\mathcal{W}$  by minimizing the negative log-likelihood. MVM can encourage the model to infer visual knowledge from the contextual visual information as well as language. When an image feature  $v_i$  is masked, its mapping index  $h_i$  in VD is considered as its label. In visual feature maps, neighbor features may have similar values, and thus share the same mapping index. This will cause the model to directly copy the label from surrounding features as predictions in a lazy way. To prevent this, in the masking stage, we first randomly select an existing label index  $j$ , then replace all visual embedding vectors in  $f^{-1}(j)$  with the special [MASK] token embedding vector.

**Image-Text Matching.** To enhance the cross-modal matching, we adopt Image-Text Matching (ITM) task for pre-training as in previous works [7]. We apply a binary classifier  $\phi(\cdot)$  on the joint embedding feature of [CLS] token to predict whether the input image and text are matched or not. ITM task is driven by the following loss function:

$$\mathcal{L}_{\text{ITM}} = -\mathbb{E}_{(\mathcal{W}, f(\mathcal{V})) \sim D} \log p(y | \phi(\mathcal{W}, f(\mathcal{V}))), \quad (8)$$

where  $y \in \{0, 1\}$  indicates whether the image and text is matched ( $y = 1$ ) or not ( $y = 0$ ).

The visual feature encoder, VD-based image embedding module and the cross-modal Transformer is end-to-end jointly trainable. We assign equal loss weight to the three pre-training objectives, and thus the full pre-training objective of SOHO is:

$$\mathcal{L}_{\text{Pre-training}} = \mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{MVM}} + \mathcal{L}_{\text{ITM}}. \quad (9)$$

### 3.4. Pre-training Datasets

Several large-scale datasets have been proposed to facilitate VL pre-training. According to typical settings in UNITER [7], these datasets can be categorized into two classes: “in-domain” and “out-domain”. In our work, we use “in-domain” as a pre-training dataset as most VL pre-training tasks are built on them [7, 23, 37]. We construct our pre-training datasets with MSCOCO [25] and VG [20].

To avoid data leak, we only use the *train* and *restval* splits of MSCOCO dataset, and the *train* and *val* splits of VG dataset in the training stage. The detailed statistic of our pre-training datasets can be found in Table 1. Detailed comparisons of pre-training dataset usage of most VLPT works, including our train/test image and text numbers, are included in our supplementary material.

## 4. Experiment

### 4.1. Implementation Details

For the **language processing**, we follow BERT to use the WordPiece tokenizer [41] to split each text into language tokens. For the **visual processing**, as most previous works adopt feature extractor which uses  $600 \times 1000$  as input resolution, we also adopt setting to resize the shorter edge of input images to 600, and limit the longer edge to be lower than 1000 for a fair comparison. We use pre-trained models based on public accessible ImageNet [8] and BERT [10] to initialize the parameters of our **visual backbone** and **Transformer architecture**, respectively. Specifically, we adopt the widely-used ResNet-101 backbone and 12-layer Transformer to fairly compare with other works, while we adopt a lightweight ResNet-18 backbone and 3-layer Transformer in our ablation studies to reduce experiment cost. We will use  $R_X$  to denote X-layer ResNet architecture in the rest of this paper for simplicity (e.g. R101 denotes ResNet-101). Since the visual backbone and Transformer favor different kinds of optimizers [47], we follow the suggestion of Zhang *et al.* [47] to use SGD and AdamW optimizers for visual backbone and Transformer respectively. We use SGD with learning rate  $1e-2$  and weight decay  $5e-4$  for the visual backbone, and apply AdamW with learning rate  $1e-4$  and weight decay  $1e-2$  for Transformer. We pre-train our model with 32 NVIDIA Tesla V100 GPUs with a batch size of 4,096 image-text pairs. The training process takes 40 epochs until convergence and we empirically decay the learning rate by 10 times at  $25^{th}$  and  $35^{th}$  epoch.

We adopt mixed-precision training to reduce memory cost and speed up training procedure. An image will be paired with four texts in each batch during pre-training, including two positive pairs and two negative pairs. We only apply MLM and MVM on the positive image-text pairs.

### 4.2. Downstream Tasks and Results

We test the performance of SOHO on four well-established downstream tasks, include image-text retrieval, visual question answering (VQA), natural language for visual reasoning (NLVR), and fine-grained visual reasoning (Visual Entailment, or VE). Image-text retrieval task includes two sub-tasks, i.e., image-to-text retrieval (TR)

Table 2: Evaluation of image-to-text retrieval (TR) and text-to-image retrieval (IR) task on MSCOCO Dataset. ”-” indicates the detail is not reported.

Model	Backbone	TR			IR			TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
1K Test set													
VSE++[12]	R152	64.6	90.0	95.7	52.0	84.3	92.0	41.3	71.1	81.2	30.3	59.4	72.4
SCAN[21]	R101	72.7	94.8	98.4	58.8	88.4	94.8	50.4	82.2	90.0	38.6	69.3	80.4
Unicoder-VL[22]	-	84.3	97.3	99.3	69.7	93.5	97.2	62.3	87.1	92.8	46.7	76.0	85.3
UNITER[7]	R101	-	-	-	-	-	-	64.4	87.4	93.1	50.3	<b>78.5</b>	<b>87.2</b>
SOHO (ours)	R101	<b>85.1</b>	<b>97.4</b>	<b>99.4</b>	<b>73.5</b>	<b>94.5</b>	<b>97.5</b>	<b>66.4</b>	<b>88.2</b>	<b>93.8</b>	<b>50.6</b>	78.0	86.7

Table 3: Evaluation of image-to-text retrieval (TR) and text-to-image retrieval (IR) on Flickr30K dataset. ”-” indicates the detail is not reported.

Model	Backbone	TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10
VSE++[12]	R152	52.9	80.5	87.2	39.6	70.1	79.5
SCAN[21]	R101	67.4	90.3	95.8	48.6	77.7	85.2
ViLBERT[27]	R101	-	-	-	58.2	84.9	91.5
Unicoder-VL[22]	-	86.2	96.3	99.0	71.5	90.9	94.9
UNITER[7]	R101	85.9	97.1	98.8	<b>72.5</b>	92.4	<b>96.1</b>
SOHO (ours)	R101	<b>86.5</b>	<b>98.1</b>	<b>99.3</b>	<b>72.5</b>	<b>92.7</b>	<b>96.1</b>

and text-to-image retrieval (**IR**), and are conducted on Flickr30K [43] and MSCOCO [25] datasets. The tasks of VQA, NLVR, and VE are conducted on datasets of VQA 2.0 [14], NLVR<sup>2</sup> [35] and SNLI-VE [42] respectively. Table 1 summarizes the statistics of all our downstream tasks.

We compare our approach with several task-specific methods and pre-training models. Most pre-training models adopt Transformer-like architectures [40] with BERT-like objectives [10] to learn cross-modal representations [7, 22, 23, 27, 34, 37, 48]. For downstream tasks, we find that using input features of the VD module for visual representation is better than directly applying VD embedding. We adopt the former setting in our experiment. This shows that VD suits visual representation learned with a very large scale of semantics while dense features provide more details in a relatively small dataset.

#### 4.2.1 Task I: Image-Text Retrieval

Image-text retrieval requires a model to retrieve the most relevant caption from candidate images, or vice versa. It is one of the most typical tasks in the field of vision-language learning which enables a broad range of applications (e.g., image searching). Image-text retrieval includes two sub-tasks of image-to-text retrieval (**TR**) and text-to-image retrieval (**IR**). During training, we construct *aligned* and *unaligned* pairs inside of a mini-batch like most image-text retrieval models. We randomly sample  $t$  aligned image-caption pairs from ground truth annotations to form a mini-batch. All the other  $t - 1$  captions are used as the unaligned captions for each image. To encourage the model to predict the right labels for both the aligned and unaligned pairs, we

Table 4: Evaluation of VQA on VQA 2.0 dataset. ”-” indicates the detail is not reported. X101 denotes ResNeXt-101 architecture [43].

Model	Backbone	test-dev	test-std
MUTAN[4]	R152	60.17	-
BUTD[2]	R101	65.32	65.67
Unified VLP [48]	X101	70.50	70.70
ViLBERT[27]	R101	70.55	70.92
VisualBERT[23]	R152	70.80	71.00
VLBERT[34]	R101	71.79	72.22
LXMERT[37]	R101	72.42	72.54
UNITER[7]	R101	72.70	72.91
SOHO (Ours)	R101	<b>73.25</b>	<b>73.47</b>

consider the retrieval task as a binary classification problem.

In our implementation, we use the joint embedding representation of the  $[CLS]$  token from Transformers to predict whether an image-caption pair is aligned or not. Since the objective of image-text retrieval task is consistent with the image-text matching (ITM) task in pre-training stage, the pre-trained parameters can well be inherited for fine-tuning. We adopt AdamW optimizer with  $1e-4$  learning rate and  $1e-2$  weight decay. The mini-batch size  $t$  is set to 24. We train 20 epochs until convergence and decay the learning rate by half at  $3^{rd}$ ,  $5^{th}$ ,  $9^{th}$  and  $13^{th}$  epoch empirically.

We conduct experiments on MSCOCO [25] and Flickr30k [30], and the results are shown in Table 2 and Table 3 respectively. It worth noting that UNITER additionally uses out-of-domain datasets and the results are expected to be better than merely use in-domain datasets as they reported [7]. Unicoder-VL [22] adopts merely out-of-domain datasets, which is also not directly comparable to our SOHO. Nevertheless, SOHO outperforms the most recent VLPT works under most metrics on both MSCOCO and Flickr30k. The performance improvements indicate that SOHO learns better image-text embeddings by our end-to-end pre-training architecture, and exploits comprehensive yet compact visual semantic abstraction by the proposed visual dictionary.

#### 4.2.2 Task II: Visual Question Answering

Visual Question Answering (VQA) requires a model to take an image and a question as input and output an answer.

Table 5: Evaluation of Visual Reasoning on NLVR<sup>2</sup> dataset.

Model	Backbone	dev	test-P
Image Only[35]	R152	51.60	51.90
CNN+RNN[35]	R152	53.50	52.40
MaxEnt[35]	R152	54.10	54.80
VisualBERT[23]	R152	67.40	67.00
LXMERT[37]	R101	74.90	74.50
UNITER[7]	R101	75.85	75.80
SOHO (Ours)	R101	<b>76.37</b>	<b>77.32</b>

This task requires machines to act like humans and reason across vision and language, which is approaching intelligent AI. We model VQA as a classification problem by learning multi-layer perception from the [CLS] token. We follow [19] to treat it as a 3,192-way classification problem, and optimize the model via binary cross-entropy loss. We fine-tune for 18 epochs with a batch size of 256 until convergence. We set the optimizer the same as in the pre-training stage. The initial learning rates are also set the same as pre-training, and we decay the learning rate by 10 at the 12<sup>th</sup> and 16<sup>th</sup> epochs empirically.

Results are presented in Table 4. The most direct comparable baseline to our SOHO is LXMERT [37], which adopts the same backbone and pre-training dataset as our SOHO. SOHO obtains 0.83% and 0.93% absolute improvements on test-dev and test-std split over LXMERT respectively. It is worth noting that SOHO outperforms UNITER [7] even under an inferior experimental setting, where UNITER additionally uses out-domain datasets in the pre-training stage. The promising results of SOHO on VQA demonstrate that our end-to-end pre-training approach enables intelligent question answering on visual contents.

#### 4.2.3 Task III: Visual Reasoning

Visual Reasoning with Natural Language (NLVR) requires a model to predict whether a text is related to a given pair of images. Compared with VQA, NLVR addresses the challenge of compositional visual reasoning on relations, comparisons, and quantities. We conduct this task on NLVR<sup>2</sup> dataset [35]. In our implementation, we follow LXMERT [37] and UNITER [7] to input two pairs of image-text to Transformer and get two embedding vectors from [CLS] tokens. Then we learn a classifier on the concatenation of the embedding vectors over “true” or “false” by a cross-entropy loss. The settings of the optimizer, epoch number, and learning rate are the same as VQA settings. Since the number of input images for NLVR<sup>2</sup> is twice as VQA, the batch size of NLVR<sup>2</sup> is half of VQA.

We mainly compare with the SOTA results provided by LXMERT [37] and UNITER [7] under the same settings for fair comparisons. From the results shown in Table 5, we observe 0.52% and 1.52% absolute gains of SOHO against UNITER on dev and test-P split respectively. This result

Table 6: Evaluation of Visual Entailment on SNLI-VE.

Model	Backbone	val	test
EVE-Image[42]	R101	71.56	71.16
UNITER[37]	R101	78.59	78.28
SOHO (Ours)	R101	<b>85.00</b>	<b>84.95</b>

validates that SOHO also has advantages when applying to compositional visual reasoning tasks.

#### 4.2.4 Task IV: Visual Entailment

Visual Entailment (VE) is a fine-grained visual reasoning task to predict whether an image semantically entails a text. In pursuit of visual intelligence, the relationship between an image and a text pair in the VE task is more fine-grained than VQA and NLVR, which can be true (entailment), false (contradiction) or neutral. SNLI-VE dataset [42] is proposed for the VE task and is constructed based on Stanford Natural Language Inference (SNLI) [5] and Flickr30K [30] datasets. We follow UNITER [7] to treat the VE task as a three-way classification problem and predict the scores of each class by a fully-connected layer on the representation of the [CLS] token from the output of the Transformer. We fine-tune the model for 6 epochs with batch size 128 until convergence. The learning rate is initialized as 1e-4, and decay to 1e-5 after four epoch empirically.

We compare SOHO with a VLPT work UNITER [7] and a task-specific method EVE-Image [42]. As reported in Table 6, SOHO achieves 85.00% and 84.95% accuracy on val and test split respectively. The results significantly outperform the SOTA results provided by UNITER [7], where 6.41% and 6.67% absolute accuracy improvements are obtained on the val and test split respectively. The results indicate the advantage of our end-to-end framework for refining the CNN backbone together with the cross-modal Transformer to facilitate thorough vision-language alignment.

#### 4.3. Ablation Study

We perform ablation studies to validate the effectiveness of the visual dictionary (VD) on all downstream tasks. We first establish a baseline model without VD, then incorporate VD with the baseline and further evaluate the influence of the embedding vector size (VD size)  $k$ .

Results are presented in Table 7. Generally, we observe that for most tasks, a VD size of 2048 or 4096 achieves the best results among four sizes ranging from 1024 to 8192. This is reasonable as VD is designed to aggregate similar visual semantics into the same image feature. With such design, the bigger VD could learn to group more fine-grained and complete visual semantics, which benefits the VL alignment as expected. However, too fine-grained visual semantics being grouped into different image features may deteriorate the abstraction of visual semantics, which consequently is harmful to VL alignment. We empirically find

Table 7: Ablation study on the effectiveness of Visual Dictionary (VD) and the embedding vector size of VD. Results are obtained under the settings of a ResNet-18 backbone and a 3-layer Transformer architecture. Image-text Retrieval is conducted on the MSCOCO 1k test set. The top-1 and top-2 results of each metric are highlighted in bold and underlined respectively. Notation  $\Delta$  indicates the performance gains of 2048 VD size results over baseline results without VD.

VD size	Text Retrieval			Image Retrieval			VQA		NLVR <sup>2</sup>		SNLI-VE		
	R@1	R@5	R@10	R@1	R@5	R@10	test-dev	test-std	dev	test-P	val	test	
w/o VD	-	72.80	<u>93.20</u>	96.90	58.22	88.32	94.40	66.08	66.33	62.62	62.61	82.28	82.16
w/ VD	1024	<u>73.40</u>	92.10	97.00	<u>58.55</u>	88.84	94.70	<u>66.75</u>	66.95	63.32	64.60	<u>82.47</u>	<u>82.55</u>
	2048	<b>75.50</b>	<b>93.50</b>	<b>97.30</b>	<b>59.03</b>	<u>88.88</u>	<u>94.84</u>	66.69	<u>67.09</u>	<b>64.62</b>	<b>65.32</b>	<b>82.56</b>	82.54
	4096	71.20	<u>93.20</u>	<b>97.30</b>	58.50	<b>88.92</b>	<b>94.96</b>	<b>66.76</b>	66.91	<u>63.60</u>	<u>64.80</u>	82.53	<u>82.55</u>
	8192	72.10	92.30	96.50	58.01	88.08	94.70	66.65	<b>67.10</b>	63.15	64.49	82.29	<b>82.69</b>
$\Delta$	2048	2.70 $\uparrow$	0.30 $\uparrow$	0.40 $\uparrow$	0.81 $\uparrow$	0.56 $\uparrow$	0.44 $\uparrow$	0.61 $\uparrow$	0.76 $\uparrow$	2.0 $\uparrow$	2.71 $\uparrow$	0.28 $\uparrow$	0.38 $\uparrow$

that  $k = 2048$  works the best in most cases, thus we adopt  $k = 2048$  as our default setting.

When compared with the baseline without VD, our proposed method with VD enjoys better performances under almost all metrics with a wide range of  $k$  (i.e., 1024, 2048, and 4096). It validates the effectiveness of VD and shows that VD is generally applicable to a broad range of tasks.

#### 4.4. Visualization of Visual Dictionary

To share insights on what the proposed Visual Dictionary (VD) learned, we visualize some representative VD indices in Figure 3. As introduced in Sec 3.2, a VD index is correlated with many visual features, where each visual feature corresponds to an image patch. We randomly sample some indices from VD and visualize their corresponding image patches. As shown in Figure 3, the VD groups meaningful and consistent image patches into different indices, which reflects an abstraction of visual semantics. The visualization shows the strong capability of the learned VD. More cases can be found in supplementary materials.

#### 4.5. Inference Time

BUTD-based methods mainly include three inference stages: CNN forwarding, region feature generation, and Transformer forwarding [2]. In contrast, SOHO only includes two inference stages of CNN and Transformer forwarding. To compare the inference efficiency of SOHO and BUTD-based methods, we set up an experiment on a V100 GPU with  $600 \times 1000$  input resolution, a ResNet-101 backbone, a 12-layer Transformer, 100 boxes, 16 sentence padding length. The average inference time for extracting BUTD features on ResNet-101 is 21ms. The input sequence length of the Transformer for BUTD-based methods and SOHO are  $100 + 16 = 116$  and  $\lceil 600/64 \rceil * \lceil 1000/64 \rceil + 16 = 176$ , respectively. Thus the inference time of Transformer is 17ms and 23ms for BUTD-based methods and SOHO, respectively. For BUTD-based methods, in addition to a 420ms time cost of region feature generation, the main time cost, however, comes from the non-maximum suppression which is required to be applied to all 1,600 categories. Consequently, the 44ms time cost of SOHO for an inference step is about 10 times faster than the 464ms time cost



Figure 3: Visualization of VD. The left and right indices reflect the semantic of “head” and “building” with consistent visual patterns, respectively.

of BUTD-based methods. Therefore, our highly-efficient SOHO could be better applied to real applications.

### 5. Conclusion

In this paper, we show a new perspective for vision-language model design. Particularly, we propose SOHO, one of the first end-to-end vision-language pre-training models that learns comprehensive yet compact visual representation for cross-modal understanding. To generate visual features that can be fused with language tokens, we propose a novel visual dictionary to transform an image to concrete semantics. Three pre-training tasks are conducted to build connections between images and languages. Performances on four downstream tasks show the superiority of SOHO over pre-training models with region-based image features. Moreover, we relieve the requirement for bounding box annotations, and reduce heavy human labeling costs. This end-to-end framework also shows the merit of accelerating the inference time in vision-language tasks about 10 times, which enables more online vision-language applications. In the future, we will further explore vision-language generation tasks, and study the utilization of large-scale unpaired multi-modal data for cognition-level visual understanding.

### 6. Acknowledgments

This work is supported by Ministry of Science and Technology International Exchange Project (No. 43-17).



## References

- [1] Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. Fusion of detected objects in text for visual question answering. In *EMNLP*, pages 2131–2140, 2019. [2](#)
- [2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086, 2018. [1](#), [2](#), [3](#), [6](#), [8](#)
- [3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *ICCV*, pages 2425–2433, 2015. [1](#)
- [4] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *ICCV*, pages 2612–2620, 2017. [6](#)
- [5] Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. In *EMNLP*, pages 632–642, 2015. [7](#)
- [6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *ECCV*, 2020. [4](#)
- [7] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: Universal image-text representation learning. In *ECCV*, pages 104–120, 2020. [1](#), [2](#), [4](#), [5](#), [6](#), [7](#)
- [8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. [2](#), [4](#), [5](#)
- [9] Karan Desai and Justin Johnson. VirTex: Learning Visual Representations from Textual Annotations. In *CVPR*, 2021. [2](#)
- [10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *ACL*, pages 4171–4186, 2019. [2](#), [4](#), [5](#), [6](#)
- [11] Alexey Dosovitskiy, Lucas Beyer, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929*, 2020. [4](#)
- [12] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: Improving visual-semantic embeddings with hard negatives. In *BMVC*, 2017. [6](#)
- [13] Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, volume 2, pages 524–531, 2005. [2](#)
- [14] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *CVPR*, pages 6904–6913, 2017. [5](#), [6](#)
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. [4](#)
- [16] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv:2004.00849*, 2020. [2](#)
- [17] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *CVPR*, pages 10267–10276, 2020. [2](#)
- [18] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015. [5](#)
- [19] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *NeurIPS*, pages 1564–1574, 2018. [2](#), [7](#)
- [20] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017. [2](#), [3](#), [5](#)
- [21] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, pages 201–216, 2018. [6](#)
- [22] Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, pages 11336–11344, 2019. [1](#), [2](#), [4](#), [6](#)
- [23] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. VisualBERT: A simple and performant baseline for vision and language. *arXiv:1908.03557*, 2019. [1](#), [2](#), [5](#), [6](#), [7](#)
- [24] Qing Li, Jianlong Fu, Dongfei Yu, Tao Mei, and Jiebo Luo. Tell-and-answer: Towards explainable visual question answering using attributes and captions. In *EMNLP*, 2018. [2](#)
- [25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pages 740–755, 2014. [1](#), [5](#), [6](#)
- [26] Bei Liu, Jianlong Fu, Makoto P Kato, and Masatoshi Yoshikawa. Beyond narrative description: Generating poetry from images by multi-adversarial training. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 783–791, 2018. [2](#)
- [27] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, pages 13–23, 2019. [1](#), [2](#), [4](#), [6](#)
- [28] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *NeurIPS*, volume 29, 2016. [2](#)
- [29] Niki Parmar, Ashish Vaswani, et al. Image transformer. In *ICML*, pages 4055–4064. PMLR, 2018. [4](#)
- [30] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, pages 2641–2649, 2015. [5](#), [6](#), [7](#)
- [31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, pages 91–99, 2015. [3](#)
- [32] Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. Learning visual representations with caption annotations. In *ECCV*, 2020. [2](#)

- [33] Amanpreet Singh, Vedanuj Goswami, Vivek Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia-a platform for vision & language research. In *NeurIPS, SysML Workshop*, volume 2018, 2018. [2](#)
- [34] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: Pre-training of generic visual-linguistic representations. In *ICLR*, 2019. [1](#), [2](#), [4](#), [6](#)
- [35] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *ACL*, pages 6418–6428, 2019. [1](#), [2](#), [5](#), [6](#), [7](#)
- [36] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. VideoBERT: A joint model for video and language representation learning. In *ICCV*, pages 7464–7473, 2019. [2](#)
- [37] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP*, pages 5103–5114, 2019. [1](#), [2](#), [4](#), [5](#), [6](#), [7](#)
- [38] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *CVPR*, pages 2843–2851, 2017. [1](#)
- [39] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NeurIPS*, pages 6306–6315, 2017. [4](#)
- [40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. [6](#)
- [41] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv*, 2016. [5](#)
- [42] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment task for visually-grounded language learning. In *NeurIPS ViGIL workshop*, 2018. [5](#), [6](#), [7](#)
- [43] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, pages 1492–1500, 2017. [6](#)
- [44] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *CVPR*, pages 21–29, 2016. [2](#)
- [45] Dongfei Yu, Jianlong Fu, Tao Mei, and Yong Rui. Multi-level attention networks for visual question answering. In *CVPR*, pages 4709–4717, 2017. [2](#)
- [46] Zhaoyang Zeng, Bei Liu, Jianlong Fu, Hongyang Chao, and Lei Zhang. Wsod2: Learning bottom-up and top-down objectness distillation for weakly-supervised object detection. In *ICCV*, pages 8292–8300, 2019. [1](#)
- [47] Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank J Reddi, Sanjiv Kumar, and Suvrit Sra. Why ADAM beats SGD for attention models. *arXiv*, 2019. [5](#)
- [48] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and VQA. In *AAAI*, pages 13041–13049, 2020. [1](#), [2](#), [4](#), [6](#)