DiMBERT: Learning Vision-Language Grounded Representations with Disentangled Multimodal-Attention

FENGLIN LIU, ADSPLAB, School of ECE, Peking University
XIAN WU and SHEN GE, Tencent
XUANCHENG REN, MOE Key Laboratory of Computational Linguistics, School of EECS, Peking University
WEI FAN, Tencent
XU SUN, School of EECS, Peking University and Center for Data Science, Peking University
YUEXIAN ZOU, ADSPLAB, School of ECE, Peking University and Peng Cheng Laboratory

Vision-and-language (V-L) tasks require the system to understand both vision content and natural language, thus learning fine-grained joint representations of vision and language (a.k.a. V-L representations) is of paramount importance. Recently, various pre-trained V-L models are proposed to learn V-L representations and achieve improved results in many tasks. However, the mainstream models process both vision and language inputs with the same set of attention matrices. As a result, the generated V-L representations are *entangled* in *one common latent space*. To tackle this problem, we propose DiMBERT (short for **Di**sentangled **M**ultimodal-Attention **BERT**), which is a novel framework that applies separated attention spaces for vision and language, and the representations of multi-modalities can thus be disentangled explicitly. To enhance the correlation between vision and language in disentangled spaces, we introduce the visual concepts to DiMBERT which represent visual information in textual format. In this manner, visual concepts help to bridge the gap between the two modalities. We pre-train DiMBERT on a large amount of image–sentence pairs on two tasks: bidirectional language modeling and sequence-to-sequence language modeling. After pre-train, DiMBERT is further fine-tuned for the downstream tasks. Experiments show that DiMBERT sets new state-of-the-art performance on three tasks (over four datasets), including both generation tasks (image captioning and visual storytelling) and classification tasks (referring expressions). The proposed DiM (short for **Di**sentangled **M**ultimodal-Attention) module can be easily incorporated into existing pre-trained V-L models to boost their performance, up to a 5% increase on the representative task. Finally, we conduct a systematic analysis and demonstrate the effectiveness of our DiM and the introduced visual concepts.

CCS Concepts: • **Computing methodologies** → **Scene understanding**; **Image representations**; *Natural language generation*;

Additional Key Words and Phrases: Vision-and-language tasks, pre-training, vision-language representations, disentangled attention, visual concepts

Authors' addresses: F. Liu, ADSPLAB, School of ECE, Peking University, 2199 Lishui Road, Nanshan District, Shenzhen, Guangdong, 100871, China; email: fenglinliu98@pku.edu.cn; X. Wu, S. Ge, and W. Fan, Tencent, Yinke Building, 38 Haidian St, Haidian District, Beijing, 100080, China; emails: {kevinxwu, shenge, Davidwfan}@tencent.com; X. Ren, MOE Key Laboratory of Computational Linguistics, School of EECS, Peking University, No.5 YiHeYuan Road, Haidian District, Beijing, 100871, China; email: renxc@pku.edu.cn; X. Sun, School of EECS, Peking University, Center for Data Science, Peking University, No.5 YiHeYuan Road, Haidian District, Beijing, 100871, China; email: xusun@pku.edu.cn; Y. Zou, ADSPLAB, School of ECE, Peking University, Peng Cheng Laboratory, 2199 Lishui Road, Nanshan District, Shenzhen, Guangdong, 100871, China; email: zouyx@pku.edu.cn.

1556-4681/2021/06-ART1 $15.00
https://doi.org/10.1145/3447685

ACM Reference format:
Fenglin Liu, Xian Wu, Shen Ge, Xuancheng Ren, Wei Fan, Xu Sun, and Yuexian Zou. 2021. DiMBERT: Learning Vision-Language Grounded Representations with Disentangled Multimodal-Attention. *ACM Trans. Knowl. Discov. Data* 16, 1, Article 1 (June 2021), 19 pages.
https://doi.org/10.1145/3447685

1 INTRODUCTION

Recently, there is a surge of research interests in **vision-and-language (V-L)** tasks, such as image captioning [11] and visual storytelling [28]. In V-L tasks, it is vital to learn the alignments and relationships between V-L modalities and generate fine-grained V-L representations [12, 53, 58, 72]. However, many existing systems are task-specific models, focusing on individual tasks only. As a result, learning universal V-L representations and empowering models with the ability to adapt to a wide range of downstream V-L tasks are now becoming the critical topics in current V-L research.

Inspired by the successful pre-training models like ResNet [24] in **computer vision (CV)** and BERT [15] in natural language processing, several attempts [1, 58, 72, 76, 99] have been conducted to learn such universal V-L representations. Table 1 summarizes some representative works in image domain. As we can see, most systems adopt the Transformer framework [15, 77] as the backbone, feeding both visual and textual features into the same stacked transformers. In the pre-train stage, these systems use large-scale image-sentence pairs to adapt the transformers to the V-L scenarios; In the fine-tune stage, the transformers are further optimized for downstream tasks. Although these pre-training V-L systems receive performance gains across multiple downstream tasks, there are still some potential directions for further improvement:

- **Entangled Attention:** Existing works feed both the visual and textual features into the same set of stacked transformers. In this manner, the same set of attention matrices are used to transform both visual and textual embeddings. We denote such type of attention mechanism as *Entangled Attention*, because the visual and textual embeddings are projected to *one common latent space*. Despite the pre-training on image-sentence pairs, the initial parameters of current systems [12, 39, 44, 72, 99] are usually directly inherited from BERT [15] or BERT-based models, e.g., UniLM [16], which is optimized for language modeling only,[1] to boost the performance. On one hand, it could be in-appropriate to apply the parameters that are trained in language modality to the features in visual modality; on the other hand, the ability of language modeling brought by the original BERT could be somewhat affected by the introduced visual modality. Besides, during parameter optimization, the model needs to consider the intra-relationships of both visual and textual embeddings, as well as the inter-relationships across V-L embeddings. Modeling these three kinds of relations with only one shared set of attention matrices could be insufficient.
- **Modality Gap:** Most V-L systems only use the **region-of-interests (RoIs)**/video frames as the visual features. Although the transformers are proved to be effective in mining correlations, there are still huge gaps between the visual and language modalities.
- **Generation Task:** Most systems can only be fine-tuned directly on classification tasks, lacking the capability to handle generation tasks. Although VideoBERT [74] and CBT [73] have been proposed to support the generation tasks in video domain, they have to train a

[1]Some works, e.g., VL-BERT [72], attempt to pre-train the V-L systems on a large amount of text-only datasets in the initial pre-training stage.

Table 1. Comparison Between Our DiMBERT and Other Works on Learning V-L Representations

Method	Basic Module	Visual Features	Textual Features	Pre-train Captioning Datasets	Pre-training Tasks	Downstream Tasks
B2T2 [1]	ESA-based Transformer	Image RoIs	Sentence Words	Conceptual Captions [70]	ISRP + BLM	Classification Task
VisualBERT [44]				MSCOCO caption [11]	ISRP + BLM	
Unicoder-VL [39]				Conceptual Captions [70]	ISRP + BLM + MOP	
VL-BERT [72]				Conceptual Captions [70]	BLM + MOP	
UNITER [12]				Conceptual Captions [70] + VG Captions [37] + MSCOCO caption [11] + SBU Captions [65]	ISRP + BLM + MOP	
VLP [99]				Conceptual Captions [70]	BLM + S2SLM	Classification Task + **Generation Task**
DiMBER [**Ours**]	**DiM**-based Transformer	Image RoIs + **Visual Concepts**	Sentence Words	Conceptual Captions [70]	BLM + S2SLM	

The **red** colored texts indicate differences from most existing works. ESA and DiM stands for the Entangled Self-Attention and Disentangled Multimodal-Attention. ISRP, BLM, MOP, and S2SLM are short for Image-Sentence Relationship Prediction, Bidirectional Language Modeling, Masked Object Prediction, and Seq-to-Seq Language Modeling, respectively.

separate video-to-text decoder to perform the generation tasks, because they pre-train the V-L systems only as encoders.

To tackle the above three concerns, in this article, we present the **Disentangled Multimodal-Attention BERT (DiMBERT)**, which takes both visual features (i.e., RoIs and visual concepts) from images and textual features (i.e., sentence words) from sentences as input, and then applies a single cross-modal Transformer to learn vision-language grounded representations. In particular, we propose **Disentangled Multimodal-Attention (DiM)** module to explicitly disentangle visual and textual modalities. In implementations, DiM module introduces separate projection matrices to project visual and textual modalities into their corresponding visual and textual latent spaces. Following common practice [1, 12, 39, 44, 72, 99], the weights of textual projection matrices are initialized with the pre-trained parameters from BERT [15], while the weights of visual projection matrices are trained from scratch.

To enhance the correlation between visual and textual modalities, we introduce the visual concepts [18], which capture a wide range of high-level visual semantic information from images [41, 66, 91]. The visual concepts transform visual features to a set of words describing *object* (e.g., *cat*), *attribute* (e.g., *small*), and *relationship* (e.g., *standing*) of images, which (1) provide a more semantic representation of visual information and thus help shorten the gap between vision and language modalities; and (2) contain rich visual semantics and thus help understand vision and language effectively.

Inspired by the work of [99], we pre-train the proposed DiMBERT on the Conceptual Captions [70] with two unsupervised language modeling objectives: **bidirectional language modeling (BLM)** [15] and **sequence-to-sequence language modeling (S2SLM)** [16], where the latter enables the direct fine-tuning of DiMBERT on generation tasks. We conduct comprehensive

experiments and systematic analysis on three tasks: image captioning [11], visual storytelling [28], and referring expressions [33]. The proposed DiMBERT sets new state-of-the-arts on three tasks (over four benchmark datasets) and the DiM module boosts the performance of various pre-trained V-L models on referring expressions task, which validate our motivation and corroborate the effectiveness and universality of our approach.

2 RELATED WORK

Our work relates to the V-L problems, the joint representations of vision and language (a.k.a. V-L representations), and the efforts in developing pre-trained models.

2.1 V-L Problems

V-L problems, which including image captioning [11], visual storytelling [28], referring expression [33], and image caption retrieval [64], and others, have drawn remarkable attention in both natural language processing and CV. These tasks combine image and language understanding together at the same time, are tough yet practical. However, current works usually design a task-specific model to deal with one single task at one time. In this article, we propose a generic framework to conduct various V-L tasks and further improve the performance of each task.

2.2 V-L Representations

For a variety of V-L problems, an important goal is to understand the image and language despite their different application scenarios, which justifies the acquisition of fine-grained V-L representations. In the literature, to represent the images, visual features extracted by **convolutional neural networks (CNNs)** or Region-CNNs are most-widely used [3, 86], while visual concepts consisting of a set of textual words are also proposed [18]. To represent the language, textual features extracted by recurrent neural networks or off-the-shelf **natural language processing (NLP)** models are most-widely used [15, 17, 25]. Therefore, in previous task-specific V-L models, the features derived from off-the-shelf CV and NLP models are combined in an ad-hoc way to acquire the V-L representations for specific tasks. Model training is performed on the dataset for the specific task only, without any generic V-L pre-training. As a result, these task-specific models, which are directly trained for the specific target task, may well suffer from overfitting when the data for the target task is scarce.

2.3 Pre-trained Models

In CV, pre-trained models, such as ResNet [24], VGG [71] and GoogLeNet [75], pre-trained on ImageNet [14], have achieved early successes in promoting various downstream CV tasks. Transformer-based pre-trained NLP models, such as BERT [15], XLNet [89], and RoBERTa [16], have also achieved great success in advancing the state-of-the-arts for a wide range of NLP tasks. Recently, UniLM [16], which is adapted from the BERT architecture, has been proposed to enable BERT to work for both natural language understanding and generation tasks.

Most recently, several pre-trained V-L models [1, 6, 7–9, 10, 12, 13, 19–23, 26, 29, 30, 34, 35, 38, 39, 42–47, 58, 59, 61–63, 72–74, 76, 79, 83, 85, 93, 96, 99, 100] have been proposed to learn V-L representations for various V-L tasks (Table 1 summarizes some representative works). However, most existing works do not consider to learn such representations by explicitly disentangling multi-modalities and incorporating visual concepts, and are unable to perform downstream generation tasks directly. It is worth noticing that, VideoBERT [74], CBT [73], and VLP [99] proposed recently, are capable of performing generation tasks, and are thus the most relevant works to our approach. However, they were still entangling visual and textual modalities, and did not attempt to take the visual concepts into consideration. Besides, for VideoBERT [74] and CBT [73], they pre-train the

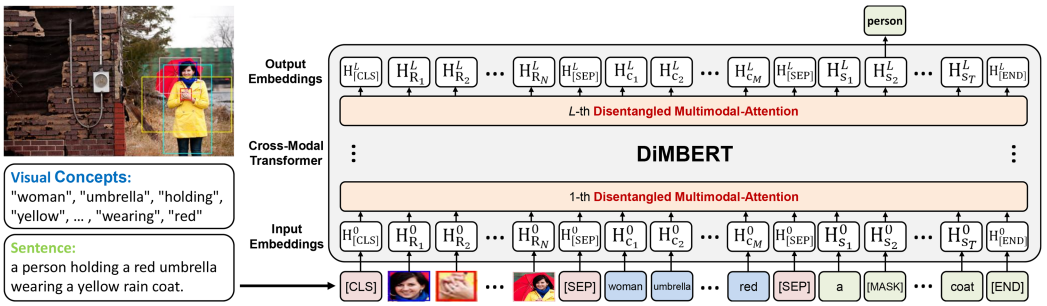

Fig. 1. Illustration of the proposed DiMBERT, which consists of a single cross-modal Transformer [72, 77] and takes visual features (i.e., RoIs and visual concepts), textual features (i.e., sentence words), and four special tokens (i.e., [CLS], [SEP], [MASK], and [END]) as input. In particular, we introduce the Disentangled Multimodal-Attention to implement the Transformer.

systems only as encoders to learn V-L representations, so they have to train a separate decoder for the generation tasks. For VLP [99], our DiMBERT outperforms it on image captioning tasks and our superiority is further validated on a long text generation task, i.e., visual storytelling [28].

3 APPROACH

In this section, we first introduce our model in detail. Next, we describe the pre-training tasks to learn V-L grounded representations. Figure 1 gives an overview of our DiMBERT.

3.1 Model Overview

As shown in Figure 1, our DiMBERT consists of three parts: (1) the embeddings of input visual features from images and textual features from sentences; (2) a single cross-modal transformer to learn the alignments and relationships between visual and textual modalities; and (3) the embeddings of learned V-L representations.

3.1.1 Input Embeddings. There are four types of input embeddings: RoIs, visual concepts, sentence words, and four special tokens.

RoI Embedding. In our approach, we use $N = 36$ RoIs for each input image. RoIs are extracted by a variant of Faster R-CNN [68] with ResNeXt-101 FPN backbone [84], which is pre-trained on Visual Genome [37], following [3].

Specifically, the appearance feature $\boldsymbol{r}_i \in \mathbb{R}^{d_r}$ is the extracted region feature, which is the output of *fc6* layer. The visual geometry feature $\boldsymbol{g}_i \in \mathbb{R}^{d_g}$ is used to encode the geometry location of the RoI in the input image, where $\boldsymbol{g}_i = \left(\frac{x_{\text{TL}}}{W}, \frac{y_{\text{TL}}}{H}, \frac{x_{\text{BR}}}{W}, \frac{y_{\text{BR}}}{H}, A_r\right)$, in which $(x_{\text{TL}}, y_{\text{TL}})$ and $(x_{\text{BR}}, y_{\text{BR}})$ denote the coordinates of the top-left and bottom-right corner, respectively, of the region bounding box; W and H are the width and height of the input image; and A_r represents the relative area, i.e., the area ratio of RoI bounding box to the entire image. Besides, following [60, 98], to enrich region features, we inject the region class information $\boldsymbol{c}_i$ into $\boldsymbol{g}_i$, defined as: $g_i^* = \left[\text{LN}\left(W_g g_i\right); \text{LN}\left(W_c c_i\right)\right]$, where [;] and LN stand for concatenation and layer normalization [4], respectively; W_g and W_c are learnable parameters; $\boldsymbol{c}_i \in \mathbb{R}^{d_c}$ is the prediction scores (probabilities) of region object label, where $d_c = 1,600$ is the number of object categories. The segment embedding is used to indicate which input segment it belongs to. For RoIs, we adopt the embedding of [RoI] token $\mathbb{E}_{[\text{RoI}]}$ as segment embedding. Finally, the RoI embedding $H_{R_i} \in \mathbb{R}^{d_{\text{model}}}$ is calculated by a weighted sum of these aforementioned embeddings. **Visual Concept Embedding.** Visual concepts contain rich visual

semantics, and have been used to provide explicit high-level semantic information of an image [81]. Following [18], we adopt a weakly-supervised approach of Multiple Instance Learning [95] to build the visual concepts extractor, which is trained on the MSCOCO caption dataset for 1,000 visual concepts. For each image, only the top $M = 20$ visual concepts are selected. We sort these extracted visual concepts by prediction scores, which means that the position embedding of each visual concept indicates its relevance to the image. Following BERT, the visual concepts use the Word-Piece embeddings [82]. They will then be further combined with position embeddings and a segment token [CEP]. The result is denoted as the visual concept embedding $H_{c_i} \in \mathbb{R}^{d_{\text{model}}}$.

Sentence Word Embedding. Similar as in the visual concept embeddings, we tokenize the sentence into WordPieces [82]. After that, a position embedding and a segment token [SEN] embedding are assigned to each sentence word, where the position embedding indicates its order in the input sentence. We denote each sentence word embedding as $H_{s_i} \in \mathbb{R}^{d_{\text{model}}}$.

At last, following BERT [15], we define a small set of special tokens: [CLS], [END], [SEP], and [MASK]. [CLS] and [END] are inserted at the first and last position, representing the start and the end of the sentence, respectively. [SEP] token is added as the boundary between two different input segments, and [MASK] token indicates the random masked-out word, which need to be predicted by DiMBERT based on all the other available elements, including RoIs, visual concepts and available sentence words. Thus, the input embeddings can be written as:

$$H^0 = \left\{H_{[\text{CLS}]}, H_R, H_{[\text{SEP}]}, H_c, H_{[\text{SEP}]}, H_s, H_{[\text{END}]}\right\}, \tag{1}$$

where the H_R, H_c and H_s are the sets of related vectors.

3.1.2 Single Cross-Modal Transformer. Our DiMBERT is adapted from $\text{BERT}_{\text{BASE}}$, thus the backbone of our approach is the 12-layer (L) Transformer [77], with 768 hidden units (d_{model}) in each layer. In implementation, we propose the DiM to replace the **Self-Attention (SA)** in Transformer. So we hereby describe the difference between DiM and SA. First, we denote the intermediate representations of l-th layer as $H^l = \{H^l_R, H^l_c, H^l_s\}$ (the representations of four special tokens are omitted for conciseness). And we find that H^l_R extracted from the image belongs to the visual modality H^l_{V}, while H^l_c extracted from the image and H^l_s extracted from the sentence belong to the textual modality H^l_{T}, which we write as $H^l = \{H^l_{\text{V}}, H^l_{\text{T}}\}$.

SA. In current pre-trained V-L systems [12, 39, 44, 72, 99], to learn the alignments and relationships between the visual modality H^l_{V} and textual modality H^l_{T}, they adopt SA, which consists of $n = 12$ parallel heads with each head SA_i defined as:

$$\text{SA}_i(H_{\text{V}}, H_{\text{T}}) = \text{softmax}\left(\begin{bmatrix} H_{\text{V}}W^q_{\text{T}} \\ H_{\text{T}}W^q_{\text{T}} \end{bmatrix} \begin{bmatrix} H_{\text{V}}W^k_{\text{T}} \\ H_{\text{T}}W^k_{\text{T}} \end{bmatrix}^T\right) \begin{bmatrix} H_{\text{V}}W^v_{\text{T}} \\ H_{\text{T}}W^v_{\text{T}} \end{bmatrix}, \tag{2}$$

where $W^q_{\text{T}}, W^k_{\text{T}}, W^v_{\text{T}}$ are parameters initialized with pre-trained parameters from BERT, which are pre-trained on text data only. Besides, the divisor $\sqrt{d_k}$ ($d_k = d_{\text{model}}/n = 64$) is omitted in equations for conciseness, please see [15, 77] for details. As we can see, SA first projects visual and textual modalities into *one common latent space*, resulting in the entanglements between the two modalities. After SA, they use a position-wise **feed-forward network (FFN)** [15, 77], which keeps on processing the features with mixed modalities and thus loses the capability to learn the relationships between the two modalities [97]. After that, the intermediate representations of (l+1)-th layer H^{l+1} are obtained by: $H^{l+1}_{\text{V}}, H^{l+1}_{\text{T}} = \text{FFN}(\text{SA}(H^l_{\text{V}}, H^l_{\text{T}}))$.

DiM. As we can see, due to the entanglements between the visual and textual modalities in SA, the models have to devote most of its capability on disentangling them, which makes it hard for systems to learn the relationships between visual and textual modalities efficiently. To this end, we

propose DiM to explicitly disentangle visual and textual modalities. In implementation, the DiM also consists of n parallel heads but with each head DiM_i defined as:

$$\text{DiM}_i(H_\text{V}, H_\text{T}) = \text{softmax}\left(\begin{bmatrix} H_\text{V}W_\text{V}^q \\ H_\text{T}W_\text{T}^q \end{bmatrix}\begin{bmatrix} H_\text{V}W_\text{V}^k \\ H_\text{T}W_\text{T}^k \end{bmatrix}^T\right)\begin{bmatrix} H_\text{V}W_\text{V}^v \\ H_\text{T}W_\text{T}^v \end{bmatrix}, \tag{3}$$

where $W_\text{T}^q, W_\text{T}^k, W_\text{T}^v$ are initialized with UniLM [16] parameters pre-trained on text data only; and the $W_\text{V}^q, W_\text{V}^k, W_\text{V}^v$ are new learnable parameters (randomly initialized). After that, following SA, we obtain the intermediate representations of (l+1)-th layer H^{l+1} by: $H_\text{V}^{l+1}, H_\text{T}^{l+1} = \text{FFN}(\text{DiM}(H_\text{V}^l, H_\text{T}^l))$.

The reason that we adopt the proposed DiM is to learn the alignments and relationships between visual and textual modalities in a more efficient way.

3.1.3 Output Embeddings. After multiple DiM layers, we use the output of last layer as the output embeddings. (H_R^L, H_c^L, and H_s^L denote RoIs, concepts, and sentence representations, respectively)

3.2 Pre-Training Tasks

In our work, we pre-train DiMBERT on the training split [99] of a large scale image–sentence dataset: i.e., Conceptual Captions dataset [70], which contains around 3.3M image–sentence pairs. To pre-train DiMBERT, we introduce two unsupervised language modeling tasks, which are adapted from the **masked language modeling (MLM)** task: (1) BLM [15], which learns to predict the randomly masked sentence words based on all available input information, i.e., the visual and textual features; and (2) S2SLM [16], which learns to predict the randomly masked sentence words based on partial input information, i.e., all the visual features and the sentence words on the left side of the word to be predicted in the sentence, which satisfies the auto-regressive property and enables our DiMBERT to perform downstream generation tasks. In this section, we will describe these pre-training task in detail.

MLM. Following BERT, during the pre-training stage, we randomly mask out the input sentence words with 15% probability, replacing the word with 80%, 10%, and 10% probabilities of [MASK] token, random word, and original word, respectively. Thus, the objective of MLM is to predict the randomly masked sentence word based on the available information. We denote trainable parameters as θ, and a pair (w_m, e_a) of input with the masked word as w_m, and all available elements as e_a, which are sampled from the training set D. The MLM is trained via minimizing negative log likelihood, defined as:

$$\mathcal{L}_\text{MLM}(\theta) = -E_{(w_\text{m}, e_\text{a})\sim D} \log\left(p_\theta\left(w_\text{m}|e_\text{a}\right)\right), \tag{4}$$

where the masked tokens are predicted as a classification problem.

BLM and S2SLM. The main difference between two language modeling tasks is the different portion of available information that can be used to predict the masked word w_m. For BLM, as in BERT [15], the model is allowed to use all the input embeddings on both left and right side of the [MASK] token. Thus e_a consists of all RoIs $\boldsymbol{R}$, all visual concepts $\boldsymbol{c}$ and all other sentence words $\boldsymbol{s}_{\setminus w_\text{m}}$. The loss function is defined as:

$$\mathcal{L}_\text{BLM}(\theta) = -E_{(w_\text{m}, e_\text{a})\sim D} \log\left(p_\theta\left(w_\text{m}|\boldsymbol{R}, \boldsymbol{c}, \boldsymbol{s}_{\setminus w_\text{m}}\right)\right). \tag{5}$$

For S2SLM, in order to enable the encoder-decoder and auto-regressive properties in SA layer, each visual elements (i.e., RoIs and visual concepts) in the first two segments is only allowed to attend to other visual elements within the first two segments, which constitutes a visual encoder; for predicting the masked sentence word, only the left side elements of the [MASK] token can be used, which constitutes a sentence decoder. If we denote the position of [MASK] token in the input

sentence as t, then the usable elements are all the RoIs $\boldsymbol{R}$, all the visual concepts $\boldsymbol{c}$ and all the left side of the [MASK] token in the sentence $\boldsymbol{s}_{1:t}$, and the loss function is:

$$\mathcal{L}_{\text{S2SLM}}(\theta) = -E_{(w_{\text{m}}, e_{\text{a}})\sim D} \log \left(p_\theta \left(w_{\text{m}} | \boldsymbol{R}, \boldsymbol{c}, \boldsymbol{s}_{1:t}\right)\right). \tag{6}$$

In implementation, the two language modeling tasks, which are alternated with random sampling, participate in the pre-training stage at a ratio of 25% and 75%, respectively. We pre-train DiMBERT on 8 GPUs (Tesla V100) with a batch size of 512 for 30 epochs. We use the Adam optimizer [36] with initial learning rates of 3e-4. Through pre-training on Conceptual Captions dataset, our model is capable to learn V-L grounded representations. Following VisualBERT [44], to let our DiMBERT better adapt to the downstream target domains, we further pre-train DiMBERT using the data from downstream tasks. Eventually, we get the final pre-trained model by averaging the last 20 checkpoints.

4 EXPERIMENTS

We evaluate DiMBERT on three representative V-L tasks, i.e., two generation tasks (image captioning [11] and visual storytelling [28]) and a classification task (referring expressions [33]).

4.1 Image Captioning

The task of image captioning aims to generate a descriptive sentence for an input image and has received extensive research interests.

Datasets and Metrics. We use the popular Flickr30k [92] and MSCOCO [11] datasets to evaluate our reported results. The datasets contain 31,783 images and 123,287 images, respectively, with 5 sentences paired to each image. To make fair comparisons [27, 52, 55, 56, 57, 88, 99], we use the widely-used splits in the work of Karpathy and Li [32] to report our results. As a result, there are 5,000 images each in the validation set and the test set for MSCOCO, and 1,000 images as for Flickr30k.

We test the model performance with MSCOCO captioning evaluation toolkit [11], which reports the widely-used automatic evaluation metrics SPICE [2], CIDEr [78], ROUGE [49], METEOR [5, 48], and BLEU [67]. SPICE is based on scene graph matching and CIDEr is based on n-gram matching. They both incorporate the consensus of a reference set for an example. These two metrics are specifically designed for the evaluation of image captioning systems. ROUGE is proposed for automatic evaluation of the extracted text summarization. METEOR and BLEU are originally designed for machine translation evaluation.

Fine-Tuning and Inference. Figure 2 illustrates the details of fine-tuning. As we can see, we apply the S2SLM task to fine-tune (cross-entropy optimization) the pre-trained DiMBERT on the image captioning task. The pre-trained DiMBERT is fine-tuned on 8 GPUs with a batch size of 512 for 30 epochs. We use the learning rate of 3e-5 and 1e-4 for parameter optimization on Flickr30k and MSCOCO datasets, respectively. Furthermore, for fair comparisons with state-of-art works [27, 88] on the MSCOCO dataset, we further perform CIDEr-based training objective using reinforcement training [69] with a learning rate of 1e-6.

In the inference stage, we initialize the input of model with {[CLS], RoIs, [SEP], Concepts, [SEP], [MASK]}, then the model will generate a $word_1$ from the position of [MASK] token. Next, DiMBERT takes the {[CLS], RoIs, [SEP], Concepts, [SEP], $word_1$, [MASK]} as input. The entire inference process repeats such generation until DiMBERT outputs an [END] token. Following common practice [27, 54, 55, 88, 90, 99], we apply beam search with beam size = 3 during inference.

Results. We compare our DiMBERT with two types of existing works: (1) the state-of-the-art task-specific models like GVD [98] on Flickr30k and AoANet [27] on MSCOCO image captioning datasets. (2) The pre-training based models, like VLP [99]. The results are shown in Table 2. As we

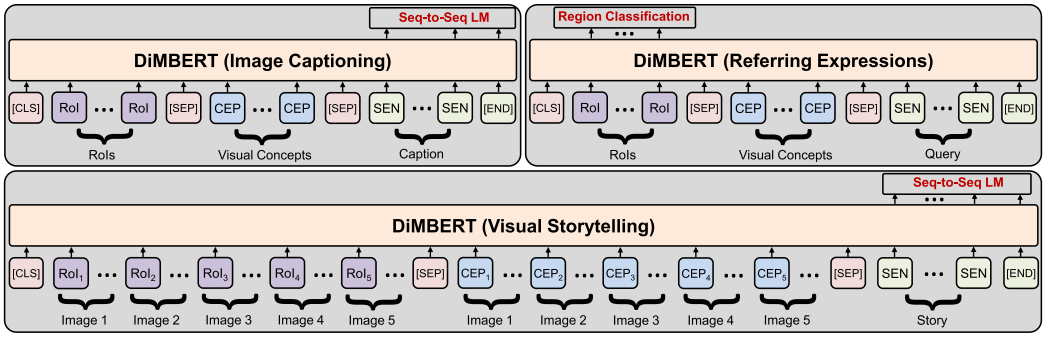

Fig. 2. Illustration of fine-tuning the pre-trained DiMBERT on various V-L downstream tasks, including two generation tasks, i.e., image captioning and visual storytelling, and a classification task, i.e., referring expressions.

Table 2. Comparisons with the State-of-the-art Task-specific Models and Pre-training Models on various Downstream Tasks, i.e., Image Captioning (Flickr30k and MSCOCO Datasets (with CIDEr Optimization)), Visual Storytelling (VIST Dataset), and Referring Expression (RefCOCO+ Dataset)

	Methods	RefCOCO+						VIST				Flickr30k Image Captioning					MSCOCO Image Captioning				
		val	testA	testB	val^d	testA^d	testB^d	B-4	M	R	C	B-4	M	R	C	S	B-4	M	R	C	S
Task-Specific Models	MAttNet [94]	71.0	75.1	66.1	65.3	71.6	56.0	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	AREL [80]	-	-	-	-	-	-	14.1	35.0	29.5	9.4	-	-	-	-	-	-	-	-	-	-
	VSCMR [41]	-	-	-	-	-	-	14.3	35.5	30.2	9.0	-	-	-	-	-	-	-	-	-	-
	GVD [98]	-	-	-	-	-	-	-	-	-	-	27.3	22.5	-	62.3	16.5	-	-	-	-	-
	SGAE [88]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	39.0	28.4	58.9	129.1	22.2
	AoANet [27]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	38.9	29.2	58.8	129.8	22.4
Pre-Training Models	ViLBERT [58]	-	-	-	72.3	78.5	62.6	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	VL-BERT$_{\text{LARGE}}$ [1]	80.3	83.6	75.5	72.6	78.6	62.3	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	UNITER$_{\text{LARGE}}$ [12]	84.0	85.9	78.9	74.9	81.4	65.4	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	XGPT [83]	-	-	-	-	-	-	-	-	-	-	31.8	23.6	-	70.9	17.6	37.2	28.6	-	120.1	21.8
	VLP [99]	-	-	-	-	-	-	-	-	-	-	31.1	23.0	-	68.5	17.2	39.5	29.3	-	129.3	23.2
	DiMBERT [Ours]	**84.6**	**86.0**	**79.7**	**75.6**	**81.6**	**66.9**	**15.3**	**36.0**	**31.3**	**13.8**	**32.4**	**24.1**	**50.7**	**72.3**	**17.9**	**40.7**	**29.7**	**59.6**	**135.3**	**23.7**

B-4, M, R, C, and S are short for BLEU-4, METEOR, ROUGE-L, CIDEr, and SPICE, respectively. For referring expression task, following common practice [12, 72], we evaluate on both ground-truth RoIs (val, testA, testB) and detected boxes (val^d, testA^d, testB^d) provided by [94]. The models of referring expression are evaluated in terms of accuracy (%).

can see, the proposed DiMBERT outperforms all baselines across all metrics over the board, which demonstrates the capability of DiMBERT to achieve consistent performance gains over different datasets.

We also evaluate our DiMBERT on the online MSCOCO evaluation server,[2] where the ground truth captions are not available. We compare with the top-performing entries on the leaderboard whose methods are published, which including AoANet [27], SGAE [88], and ETA [40]. For online evaluation, nearly all of the recent submitted systems use model ensemble [3, 27, 31, 88, 90]. From Table 3, we can find that our DiMBERT is able to outperform these state-of-the-art models across

[2] https://competitions.codalab.org/competitions/3221#results.

Table 3. Leaderboard Performance on the Online MSCOCO Image Captioning Evaluation Server

Methods	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE-L		CIDEr	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
SCST [69]	78.1	93.7	61.9	86.0	47.0	75.9	35.2	64.5	27.0	35.5	56.3	70.7	114.7	116.7
Up-Down [3]	80.2	95.2	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
CAVP [51]	80.1	94.9	64.7	88.8	50.0	79.7	37.9	69.0	28.1	37.0	58.2	73.1	121.6	123.8
ETA [40]	81.2	95.0	65.5	89.0	50.9	80.4	38.9	70.2	28.6	38.0	58.6	73.9	122.1	124.4
RFNet [31]	80.4	95.0	64.9	89.3	50.1	80.1	38.0	69.2	28.2	37.2	58.2	73.1	122.9	125.1
GLIED [54]	80.1	94.6	64.7	88.9	50.2	80.4	38.5	70.3	28.6	37.9	58.3	73.8	123.3	125.6
GCN-LSTM [90]	-	-	65.5	89.3	50.8	80.3	38.7	69.7	28.5	37.6	58.5	73.4	125.3	126.5
SGAE [88]	81.0	95.3	65.6	89.5	50.7	80.4	38.5	69.7	28.2	37.2	58.6	73.6	123.8	126.5
AoANet [27]	81.0	95.0	65.8	89.6	51.4	81.3	39.4	71.2	29.1	38.5	58.9	74.5	126.9	129.6
DiMBERT [Ours]	**81.7**	**96.1**	**66.4**	**91.0**	**51.8**	**83.0**	**39.7**	**73.1**	**29.5**	**39.2**	**59.3**	**75.0**	**129.9**	**133.1**

c5 means comparing to 5 references and c40 means comparing to 40 references. As we can see, our DiMBERT outperforms all state-of-the-art models across all metrics over the board, including AoANet [27] that uses 4-model ensemble, in a single model submission.

all metrics over the board, in a single model submission. It further demonstrates the effectiveness of the proposed DiMBERT.

4.2 Visual Storytelling

Visual storytelling task belongs to long text generation tasks. The goal is to generate a reasonable and coherent paragraph-level story based on the image stream.

Datasets and Metrics. In visual storytelling, our reported results are evaluated on the VIST dataset [28], which contains 210,819 unique images in 10,117 Flickr albums. Each sample contains one story, describing five selected images. We follow the standard split [80] for fair comparisons. There are 40,098/4,988/5,050 images for training/validation/testing, respectively. Following common practice [41, 80, 87], we adopt four evaluation metrics, including BLEU, ROUGE, METEOR, and CIDEr, for the evaluation of our approach.

Fine-Tuning and Inference. For the visual storytelling task, we adopt the same fine-tuning and inference strategy as the image captioning task. These two tasks differ solely in the input format. As illustrated in Figure 2, the visual storytelling task takes five images as input. For each image, we first extract corresponding RoIs and concepts. Then we feed the embeddings of these RoIs and concepts into DiMBERT from Image 1 to Image 5.

Results. We choose two recently proposed state-of-the-art models AREL [80] and VSCMR [41] for comparison. As shown in Table 2, the DiMBERT outperforms the AREL and VSCMR by a large margin of 46.8% and 53.3% in terms of CIDEr scores, respectively, and sets a new state-of-the-art, which shows that the proposed DiMBERT also works well on generating long texts.

4.3 Referring Expressions

Referring Expressions belongs to classification tasks which aims to locate a target image region given a textual query.

Datasets and Metrics. We evaluate our model on RefCOCO+ dataset [33], which consists of 141k expressions for 50k referred objects in 20k images in the MSCOCO dataset [50]. The referring expressions in RefCOCO+ are forbidden from using absolute location words, e.g., right car. Therefore the referring expressions focus on purely appearance-based descriptions. RefCOCO+ is split into train, validation and two test sets (testA and testB). Specifically, images containing multiple

people are in testA set, while images containing multiple objects of other categories are in testB set. Following common practice [12, 72], we evaluate on both ground-truth RoIs (val, testA, testB) and detected boxes (val^d, testA^d, testB^d) provided by [94]. We choose MAttNet [94], ViLBERT [58], VL-BERT [72] and UNITER [12] for comparison, where the first one MAttNet is the state-of-the-art task-specific model, while the rest are pre-training models. The models are evaluated in terms of accuracy (%).

Fine-Tuning and Inference. As illustrated in Figure 2, we add a linear layer on the H_V^L to output the classification scores for all the input RoIs, and take the RoI with the highest score as the final prediction. DiMBERT is fine-tuned under a binary cross-entropy loss on 8 GPUs with a batch size of 256 for 20 epochs.

Results. Table 2 shows that our DiMBERT outperforms all models, which validates the effectiveness of DiMBERT in referring expressions task. In particular, DiMBERT outperforms VL-BERT$_{\text{LARGE}}$ and UNITER$_{\text{LARGE}}$ which are developed from a larger model BERT$_{\text{LARGE}}$ [15], while DiMBERT is adapted from BERT$_{\text{BASE}}$ [15]. In other word, DiMBERT can achieve higher accuracy with much less parameters. Furthermore, in addition to Conceptual Captions dataset [70], UNITER$_{\text{LARGE}}$ uses extra a large amount of image-sentence pairs (see Table 1), e.g., VG Captions [37] and SBU Captions [65], to pre-train the model. This proves the efficiency of DiMBERT in pre-training datasets and parameters.

In all, these results suggest that the DiMBERT can be applied to a wide range of downstream tasks, no matter what the type of task is (e.g., classification task and generation task). More encouragingly, our approach outperforms existing published state-of-the-art models, including the task-specific models and pre-training models, across all aforementioned tasks, which further confirms the effectiveness and universality of the proposed DiMBERT.

5 ANALYSIS

In this section, we conduct a series of analysis on a generation task, i.e., image captioning (MSCOCO), and a classification task, i.e., referring expressions (RefCOCO+), to provide some insights and answer the following questions: (1) What is the contribution of each component in DiMBERT? (2) What is the effect of DiM module? (3) What is the contribution of visual concepts? (4) Why DiMBERT can adapt effectively to a wide range of downstream tasks? (5) Where does the actual improvement in the evaluation scores comes from? Please note that the performance on MSCOCO image captioning dataset reported in this section is different from the ones reported in Section 4: as for fair comparisons with state-of-art works [27, 88], we further perform CIDEr-based training objective using reinforcement training [69]. While in this section, for simplicity, we directly do fine-tuning with cross entropy loss.

5.1 Ablation Study

In this section, we conduct the ablation analysis on RefCOCO+ and MSCOCO image captioning datasets. To analyze the effect of each component, we evaluate from the following three perspectives:

- **Parameter Initialization:** the initial parameters inherited from BERT[15] or UniLM [16].
- **Language Model Pre-train Strategies:** the pre-train on the textual part of DiMBERT: BLM and S2SLM.
- **Language-Vision Pre-train Strategies:** (1) Image-Sentence Relationship Prediction: predict whether an image and a sentence match each other, this task is introduced by LXMERT [76]; (2) Masked Object Prediction: predict the label of masked region; and (3) Masked Visual Concept (ViCo) Prediction task, predict the masked visual concept.

Table 4. Ablation Study of Our Proposed DiMBERT which is Performed on MSCOCO (with Cross-entropy Optimization Only) and RefCOCO+

Components	Init. from BERT [15]	Init. from UniLM [16]	BLM	S2SLM	Image-Sentence Relationship Prediction	Masked Obj. Prediction	Masked ViCo Prediction	DiM	RefCOCO+	MSCOCO			
									val^d	B-4	M	C	S
Base									68.5	35.5	28.0	113.7	20.8
(1)	√								68.7	35.2	28.2	113.9	20.9
(1)		√							68.8	36.0	28.4	115.7	21.4
(2)		√	√						72.3	37.5	28.4	118.2	21.5
(2)		√		√					71.5	37.8	28.5	118.8	21.7
(2)		√	√	√					73.4	38.0	28.5	119.3	21.7
(3)		√	√	√	√				72.6	36.3	28.3	116.4	21.6
(3)		√	√	√		√			73.3	35.8	28.2	114.9	21.0
(3)		√	√	√			√		71.5	36.4	28.4	115.6	21.3
DiMBERT		√	√	√				√	**75.6**	**38.1**	**28.7**	**123.4**	**22.0**

Table 4 shows the results in which the base model denote the one trained from scratch on downstream tasks without parameter initialization and pre-train, DiMBERT denotes the full model proposed in this article. We can find that:

—Initializing the parameters from existing models like BERT is beneficial, as the pre-trained language models could better capture the contextual representations and the structure of sentences.
—Comparing the results of (1) and (2) in Table 4, both pre-training tasks BLM and S2SLM in DiMBERT can facilitate referring expressions and image captioning. Pre-training only on the textual part of DiMBERT is helpful for downstream tasks.
—The introduction of Image-Sentence Relationship Prediction, Masked Object Prediction, and Masked ViCo Prediction pre-training tasks actually hurts the performance. For Masked Object Prediction, it may due to the introduction of noise when there exists overlapped regions or wrongly detected labels [99]. For Masked ViCo Prediction, it may due to masking and predicting wrongly extracted visual concepts, which mislead the model to learn relationships between visual and textual features. For Image-Sentence Relationship Prediction, the unmatched image-sentence training pairs could hinder the training of other pre-training tasks [72, 99].

5.2 Effect of DiM Module

In this subsection, we evaluate the effectiveness of the proposed DiM mechanism. We compare DiM with **Entangled Self-Attention (ESA)** which applies the same set of attention matrices to sentences, RoIs, and concepts. As shown in Table 5, the DiMBERT equipped DiM outperforms ESABERT equipped with ESA on both image caption and referring expressions tasks. Specifically, the DiM promotes the performance of DiMBERT from 73.4% to 75.6% and from 119.3 to 123.4 in terms of accuracies for RefCOCO+ and CIDEr for MSCOCO, respectively.

This performance increase may attribute to disentangled attention. DiM explicitly use different attention matrices to model the visual and textual modalities, enabling better usage of the pre-trained BERT. As to the ESA, the same set of attention matrices are used to model inner-vision, inner-language and mutual vision-language relations. Such multiple target optimization could affect language model's capability of the per-trained BERT.

Table 5. Impact of DiM

Methods	RefCOCO+	MSCOCO			
	val^d	B-4	M	C	S
ESABERT	73.4	38.0	28.5	119.3	21.7
w/ DiM (DiMBERT)	**75.6**	**38.1**	**28.7**	**123.4**	**22.0**

ESABERT and DiMBERT represent the Vanilla BERT Model and "BERT w/ DiM," respectively.

Table 6. Analysis about the Effectiveness and Universality of DiM

Methods	RefCOCO+			Methods	RefCOCO+		
	val^d	testA^d	testB^d		val^d	testA^d	testB^d
UNITER$_{\text{BASE}}$ [12]	71.5	77.0	60.1	VL-BERT$_{\text{BASE}}$ [72]	70.7	76.8	60.3
w/ DiM	**72.7**	**78.6**	**62.1**	w/ DiM	**73.2**	**78.9**	**63.2**
ViLBERT [58]	72.4	78.3	62.5	ESABERT	73.4	79.3	63.7
w/ DiM	**74.3**	**80.1**	**64.7**	w/ DiM (DiMBERT)	**75.6**	**83.2**	**66.9**

We perform the analysis on the RefCOCO+ dataset.

Table 7. Impact of Visual Concepts (ViCo)

Methods	RefCOCO+	B-4	M	C	S				
	val^d				All	Objects	Attributes	Relations	Color
Base	**68.5**	**35.5**	**28.0**	**113.7**	**20.8**	**37.9**	**10.2**	**6.2**	**10.2**
w/o ViCo	66.6	35.4	27.9	112.9	20.4	37.3	9.3	5.8	9.5
DiMBERT	**75.6**	**38.1**	**28.7**	**123.4**	**22.0**	**40.0**	**11.1**	**7.3**	**11.5**
w/o ViCo	74.3	37.9	28.5	120.2	21.5	39.4	9.9	5.7	10.4

We further list the breakdown of SPICE F-scores [2], for a better understanding of the contribution of visual concepts.

DiM can also be easily integrated into existing pre-trained models. In this section, we further equip UNITER$_{\text{BASE}}$ [12], VL-BERT$_{\text{BASE}}$ [72], and ViLBERT [58] with DiM. Specifically, we pre-train these models on the Conceptual Captions dataset and evaluate the accuracy on the referring expression task. As shown in Table 6, DiM can successfully boost all baselines, with the most significant improvement up to relatively 4%, 5%, and 5% for val^d, testA^d, and testB^d, respectively. The significant improvements demonstrate the effectiveness and universality of the proposed DiM module.

5.3 Effect of Visual Concepts

We conduct some experiments to investigate the contribution of visual concepts (ViCo) in our model. Table 7 shows that the visual concepts promote the baselines over all metrics, especially in *Attributes* and *Color*. The reason is that the visual concepts contain more *attribute* words and *color* words than the sentence. The abundant visual semantics in visual concepts can greatly enrich semantic representations of image regions. It can be confirmed in Figure 3(c), which illustrates the image representations refined by DiMBERT w/o ViCo. Compared with the image representations refined by DiMBERT (Figure 3(b)), it is obvious that the DiMBERT w/o ViCo is insufficient in providing suitable semantic information for image regions. It is worth noticing that given the absence of input sentence, most existing models will not work well, but our model can still refine

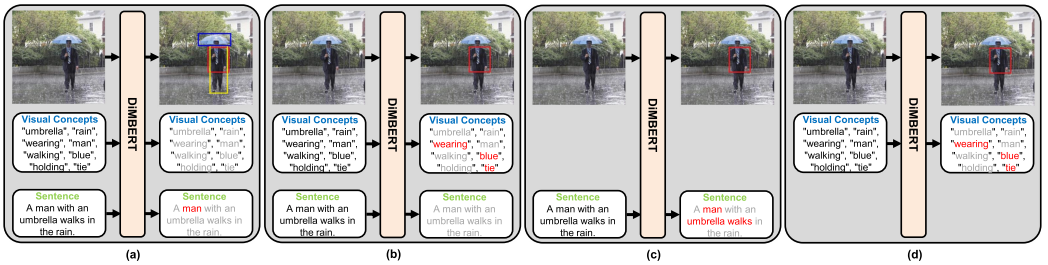

Fig. 3. Average attention weights of all heads in the last Disentangled Multimodal-Attention layer of DiMBERT. Please view in color. We show the alignments of a typical sentence word, i.e., *man*, with top-3 image regions, i.e., text-to-RoI attention, in (a), and the the alignments of a typical image region, i.e., the **Red** region, with top-3 semantic words, i.e., RoI-to-text attention, in (b), (c), and (d). Specifically, (a) and (b) show that DiMBERT can provide visual-referred sentence representations and semantic-grounded image representations, respectively; (c) shows that the model is insufficient in providing suitable semantic information for regions without visual concepts; and (d) shows that DiMBERT still works well even without input sentence.

Table 8. Analysis about the Effect of the Number of Visual Concepts

Number	Precision (%)	Recall (%)	F1 (%)	RefCOCO+	MSCOCO			
				val^d	B-4	M	C	S
M=0	-	-	-	67.4	35.0	27.7	112.7	20.4
M=10	**72.6**	29.8	44.3	68.1	35.3	27.9	113.4	20.6
M=20	52.9	45.1	**49.5**	**68.3**	**35.5**	**28.0**	**113.7**	**20.8**
M=30	42.2	54.2	47.4	67.9	35.1	27.8	113.5	20.6
M=40	35.1	**59.9**	43.8	67.7	35.2	27.8	113.0	20.5

All variants are conducted on the base model. We also report the performance of visual concept extractor in terms of the Precision, Recall, and F1 scores. As we can see, when the number of visual concepts M is 20, the visual concept extractor and the base model get the highest F1 score and highest performance.

the image representations. Figure 3(d) shows the semantic-grounded image representations [53] refined by our DiMBERT without input sentence.

Table 8 shows that when the number of visual concepts M is 20, the visual concept extractor and the model get the highest F1 score and highest performance, respectively, which is the reason why the value of M is set to 20 in our DiMBERT. For other variants, we speculate that when M is set to small values (lower recall score), the model will suffer from the inadequacy of information. When M is set to large values (lower precision score), the module will introduce more noisy information.

5.4 Visualization of DiMBERT

To evaluate the effectiveness of DiMBERT, we visualize the alignments between visual regions and semantic words according to the attention weights in the last DiM layer. Figure 3(a) and(b) shows the examples of V-L representations learned by our DiMBERT. As we can see, the model provides visual references for the input sentence, e.g., the sentence word *man* is aligned to the **blue**, **yellow**, and **red** regions. The visual-referred sentence representations play an important role in understanding the sentence correctly, because the visual references help alleviate semantic ambiguity, e.g., the word *bank* can either refer to a financial organization or the side of a river, and the word *mouse* can either refer to a mammal or an electronic device. Similarly, the model can provide a clearer semantic information for image regions, e.g., the **red** region is aligned to the words *wearing*, *blue*, and *tie*. Thus, the original image representations are refined to semantic-grounded image

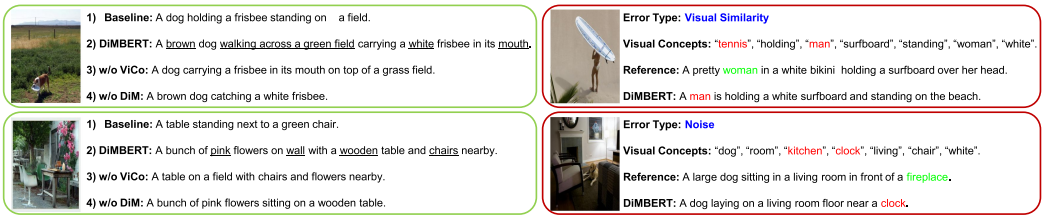

Fig. 4. Examples of the generated captions. The left plot and right plot show the correct examples and the error analysis of our approach, respectively. The ViCo and DiM represent the Visual Concepts and Disentangled Multimodal-Attention, respectively. The color **Green** denotes desirable results, while **Red** denotes unfavorable results.

representations [53]. Note that there is no suitable semantic word in the sentence to align with the **red** region, thus DiMBERT does not assign too much attention weights for any sentence word, which also indicates the effectiveness of our model when it comes to insufficient sentence words.

The refined semantic-grounded image representations and visual-referred sentence representations are beneficial for understanding both image and sentence, providing a solid bias for V-L tasks.

5.5 Examples and Bad Case Analysis

In the left plot of Figure 4, we list some intuitive examples on MSCOCO image captioning task to find the the actual improvement. While comparing the 1^{st} and 2^{nd} lines, we find that the pre-training procedure helps the base model to generate more complete and accurate captions. While comparing the 2^{nd}, 3^{rd}, and 4^{th} lines, we find that the visual concepts are bringing more details in colors and attributes, such as *brown*, *white*, *pink*, and *wooden* than the "w/o ViCo" model, and the DiM is bringing more comprehensive in objects, such as *field*, *mouth*, *wall*, and *chairs* than the "w/o DiM" model, which corroborate the effectiveness of our approach.

We also analyze some bad cases to provide insights on how the DiMBERT may be improved. We find that there are mainly two types of errors, i.e., visual similarity and noise. In the right plot of Figure 4, we give some examples. In the first example, our model dis-identify the *woman* as a *man* due to their visual similarity. However, humans can find that the person in the image wearing a bra. In the second example, the DiMBERT mistakes the incorrect visual concept, i.e., *clock*, for an appropriate one when it generates caption. A more powerful ViCo extractor may be helpful in solving the problem, but it is unlikely to be completely avoided.

6 CONCLUSIONS

We present a visual concept and proposed DiM based pre-training model, i.e., DiMBERT, which is pre-trained on a large amount of image–sentence pairs to learn V-L grounded representations. The pre-trained DiMBERT can be fine-tuned on various V-L tasks, including generation tasks and classification tasks. Extensive experiments and systematic analysis validate our motivations and corroborate the effectiveness and the universality of our DiM and DiMBERT, where the latter sets new state-of-the-arts on three downstream tasks (over four datasets).

ACKNOWLEDGMENTS

Special acknowledgements are given to Aoto-PKUSZ Joint Lab for its support. We thank all the anonymous reviewers for their constructive comments and suggestions. Xu Sun and Yuexian Zou are the corresponding authors of this paper.

REFERENCES

[1] Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. 2019. Fusion of detected objects in text for visual question answering. In *EMNLP*.

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic propositional image caption evaluation. In *ECCV*.

[3] 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 VQA. In *CVPR*.

[4] Lei Jimmy Ba, Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).

[5] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL*.

[6] Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. 2020. Multimodal pretraining unmasked: Unifying the vision and language BERTs. *arXiv preprint arXiv:2011.15124* (2020).

[7] Andrea Burns, Donghyun Kim, Derry Wijaya, Kate Saenko, and Bryan A. Plummer. 2020. Learning to scale multilingual representations for vision-language tasks. In *ECCV*.

[8] Ozan Caglayan, Menekse Kuyu, Mustafa Sercan Amac, Pranava Madhyastha, Erkut Erdem, Aykut Erdem, and Lucia Specia. 2021. Cross-lingual visual pre-training for multimodal machine translation. In *EACL*.

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

[10] Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. 2021. VisualGPT: Data-efficient image captioning by balancing visual input and linguistic knowledge from pretraining. *arXiv preprint arXiv:2102.10407* (2021).

[11] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325* (2015).

[12] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: Learning universal image-text representations. In *ECCV*.

[13] Jaemin Cho, Jiasen Lu, Dustin Schwenk, Hannaneh Hajishirzi, and Aniruddha Kembhavi. 2020. X-LXMERT: Paint, caption and answer questions with multi-modal transformers. In *EMNLP*.

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

[16] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *NeurIPS*.

[17] Jeffrey L. Elman. 1990. Finding structure in time. *Cognitive Science* 14, 2 (1990), 179–211.

[18] Hao Fang, Saurabh Gupta, Forrest N. Iandola, Rupesh Kumar Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. 2015. From captions to visual concepts and back. In *CVPR*.

[19] Zhangyin Feng, Daya Guo, Duyu Tang, Nan Duan, Xiaocheng Feng, Ming Gong, Linjun Shou, Bing Qin, Ting Liu, Daxin Jiang, and Ming Zhou. 2020. CodeBERT: A pre-trained model for programming and natural languages. In *EMNLP (Findings)*.

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

[21] Dehong Gao, Linbo Jin, Ben Chen, Minghui Qiu, Peng Li, Yi Wei, Yi Hu, and Hao Wang. 2020. FashionBERT: Text and image matching with adaptive loss for cross-modal retrieval. In *SIGIR*.

[22] François Gardères, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lécué. 2020. ConceptBert: Concept-aware representation for visual question answering. In *EMNLP (Findings)*.

[23] Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. 2020. COOT: Cooperative hierarchical transformer for video-text representation learning. In *NeurIPS*.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.

[25] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[26] Haoyang Huang, Lin Su, Di Qi, Nan Duan, Edward Cui, Taroon Bharti, Lei Zhang, Lijuan Wang, Jianfeng Gao, Bei Liu, Jianlong Fu, Dongdong Zhang, Xin Liu, and Ming Zhou. 2020. M3P: Learning universal representations via multitask multilingual multimodal pre-training. *arXiv preprint arXiv:2006.02635* (2020).

[27] Lun Huang, Wenmin Wang, Jie Chen, and Xiaoyong Wei. 2019. Attention on attention for image captioning. In *ICCV*.

[28] Ting-Hao (Kenneth) Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross B. Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual storytelling. In *HLT-NAACL*.

[29] 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* (2020).

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

[31] Wenhao Jiang, Lin Ma, Yu-Gang Jiang, Wei Liu, and Tong Zhang. 2018. Recurrent fusion network for image captioning. In *ECCV*.

[32] Andrej Karpathy and Fei-Fei Li. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.

[33] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. 2014. ReferItGame: Referring to objects in photographs of natural scenes. In *EMNLP*.

[34] Aisha Urooj Khan, Amir Mazaheri, Niels da Vitoria Lobo, and Mubarak Shah. 2020. MMFT-BERT: Multimodal fusion transformer with BERT encodings for visual question answering. In *EMNLP (Findings)*.

[35] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. ViLT: Vision-and-language transformer without convolution or region supervision. *arXiv preprint arXiv:2102.03334* (2021).

[36] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *ICLR*.

[37] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV* 123, 1 (2017), 32–73.

[38] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. 2021. Less is more: ClipBERT for video-and-language learning via sparse sampling. In *CVPR*.

[39] Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. 2020. Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*.

[40] Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. 2019. Entangled transformer for image captioning. In *ICCV*. IEEE, 8927–8936.

[41] Jiacheng Li, Haizhou Shi, Siliang Tang, Fei Wu, and Yueting Zhuang. 2019. Informative visual storytelling with cross-modal rules. In *ACM MM*.

[42] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. HERO: Hierarchical encoder for video+language omni-representation pre-training. In *EMNLP*.

[43] Linjie Li, Zhe Gan, and Jingjing Liu. 2020. A closer look at the robustness of vision-and-language pre-trained models. *arXiv preprint arXiv:2012.08673* (2020).

[44] 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* (2019).

[45] Liunian Harold Li, Haoxuan You, Zhecan Wang, Alireza Zareian, Shih-Fu Chang, and Kai-Wei Chang. 2020. Weakly-supervised VisualBERT: Pre-training without parallel images and captions. *arXiv preprint arXiv:2010.12831* (2020).

[46] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2020. UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409* (2020).

[47] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*.

[48] Chin-Yew Lin and Eduard H. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *HLT-NAACL*.

[49] Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *ACL*.

[50] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *ECCV*.

[51] Daqing Liu, Zheng-Jun Zha, Hanwang Zhang, Yongdong Zhang, and Feng Wu. 2018. Context-aware visual policy network for sequence-level image captioning. In *ACM-MM*.

[52] Fenglin Liu, Meng Gao, Tianhao Zhang, and Yuexian Zou. 2019. Exploring semantic relationships for image captioning without parallel data. In *ICDM*.

[53] Fenglin Liu, Yuanxin Liu, Xuancheng Ren, Xiaodong He, and Xu Sun. 2019. Aligning visual regions and textual concepts for semantic-grounded image representations. In *NeurIPS*.

[54] Fenglin Liu, Xuancheng Ren, Yuanxin Liu, Kai Lei, and Xu Sun. 2019. Exploring and distilling cross-modal information for image captioning. In *IJCAI*.

[55] Fenglin Liu, Xuancheng Ren, Yuanxin Liu, Houfeng Wang, and Xu Sun. 2018. simNet: Stepwise image-topic merging network for generating detailed and comprehensive image captions. In *EMNLP*.
[56] Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 2020. Federated learning for vision-and-language grounding problems. In *AAAI*.
[57] Fenglin Liu, Xian Wu, Shen Ge, Xiaoyu Zhang, Wei Fan, and Yuexian Zou. 2020. Bridging the gap between vision and language domains for improved image captioning. In *ACM-MM*.
[58] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*.
[59] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In *CVPR*.
[60] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2018. Neural baby talk. In *CVPR*.
[61] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Xilin Chen, and Ming Zhou. 2020. UniViLM: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353* (2020).
[62] Kazuki Miyazawa, Tatsuya Aoki, Takato Horii, and Takayuki Nagai. 2020. lamBERT: Language and action learning using multimodal BERT. *arXiv preprint arXiv:2004.07093* (2020).
[63] Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. 2020. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. In *ECCV*.
[64] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual attention networks for multimodal reasoning and matching. In *CVPR*.
[65] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2Text: Describing images using 1 million captioned photographs. In *NIPS*.
[66] Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei. 2017. Video captioning with transferred semantic attributes. In *CVPR*.
[67] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *ACL*.
[68] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*.
[69] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *CVPR*.
[70] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*.
[71] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
[72] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: Pre-training of generic visual-linguistic representations. In *ICLR*.
[73] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. 2019. Contrastive bidirectional transformer for temporal representation learning. *arXiv preprint arxiv:1906.05743* (2019).
[74] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. VideoBERT: A joint model for video and language representation learning. In *ICCV*.
[75] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *CVPR*.
[76] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *EMNLP*.
[77] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.
[78] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *CVPR*.
[79] Jianfeng Wang, Xiaowei Hu, Pengchuan Zhang, Xiujun Li, Lijuan Wang, Lei Zhang, Jianfeng Gao, and Zicheng Liu. 2020. MiniVLM: A smaller and faster vision-language model. *arXiv preprint arXiv:2012.06946* (2020).
[80] Xin Wang, Wenhu Chen, Yuan-Fang Wang, and William Yang Wang. 2018. No metrics are perfect: Adversarial reward learning for visual storytelling. In *ACL*.
[81] Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony R. Dick, and Anton van den Hengel. 2016. What value do explicit high level concepts have in vision to language problems? In *CVPR*.
[82] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei

Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144* (2016).

[83] Qiaolin Xia, Haoyang Huang, Nan Duan, Dongdong Zhang, Lei Ji, Zhifang Sui, Edward Cui, Taroon Bharti, Xin Liu, and Ming Zhou. 2020. XGPT: Cross-modal generative pre-training for image captioning. *arXiv preprint arXiv:2003.01473* (2020).

[84] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *CVPR*.

[85] Yiran Xing, Zai Shi, Zhao Meng, Yunpu Ma, and Roger Wattenhofer. 2021. KM-BART: Knowledge enhanced multimodal BART for visual commonsense generation. *arXiv preprint arXiv:2101.00419* (2021).

[86] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.

[87] Pengcheng Yang, Fuli Luo, Peng Chen, Lei Li, Zhiyi Yin, Xiaodong He, and Xu Sun. 2019. Knowledgeable storyteller: A commonsense-driven generative model for visual storytelling. In *IJCAI*.

[88] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. 2019. Auto-encoding scene graphs for image captioning. In *CVPR*.

[89] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*.

[90] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring visual relationship for image captioning. In *ECCV*.

[91] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. 2017. Boosting image captioning with attributes. In *ICCV*.

[92] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL* 2 (2014), 67–78.

[93] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. ERNIE-ViL: Knowledge enhanced vision-language representations through scene graph. *arXiv preprint arXiv:2006.16934* (2020).

[94] 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 *CVPR*. IEEE Computer Society, 1307–1315.

[95] Cha Zhang, John C. Platt, and Paul A. Viola. 2006. Multiple instance boosting for object detection. In *NIPS*.

[96] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. VinVL: Making visual representations matter in vision-language models. *arXiv preprint arXiv:2101.00529* (2021).

[97] Guangxiang Zhao, Xu Sun, Jingjing Xu, Zhiyuan Zhang, and Liangchen Luo. 2019. MUSE: Parallel multi-scale attention for sequence to sequence learning. *arXiv preprint arxiv:1911.09483* (2019).

[98] Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J. Corso, and Marcus Rohrbach. 2019. Grounded video description. In *CVPR*.

[99] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and VQA. In *AAAI*.

[100] Linchao Zhu and Yi Yang. 2020. ActBERT: Learning global-local video-text representations. In *CVPR*.

Received August 2020; revised December 2020; accepted January 2021