UNITER: UNiversal Image-TExt Representation Learning

Yen-Chun Chen(✉), Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu

Microsoft Dynamics 365 AI Research, Redmond, USA
{yen-chun.chen,lindsey.li,licheng.yu,ahmed.elkholy,fiahmed,
zhe.gan,yu.cheng,jingjl}@microsoft.com

Abstract. Joint image-text embedding is the bedrock for most Vision-and-Language (V+L) tasks, where multimodality inputs are simultaneously processed for joint visual and textual understanding. In this paper, we introduce UNITER, a UNiversal Image-TExt Representation, learned through large-scale pre-training over four image-text datasets (COCO, Visual Genome, Conceptual Captions, and SBU Captions), which can power heterogeneous downstream V+L tasks with joint multimodal embeddings. We design four pre-training tasks: Masked Language Modeling (MLM), Masked Region Modeling (MRM, with three variants), Image-Text Matching (ITM), and Word-Region Alignment (WRA). Different from previous work that applies joint random masking to both modalities, we use conditional masking on pre-training tasks (*i.e.*, masked language/region modeling is conditioned on full observation of image/text). In addition to ITM for global image-text alignment, we also propose WRA via the use of Optimal Transport (OT) to *explicitly* encourage fine-grained alignment between words and image regions during pre-training. Comprehensive analysis shows that both conditional masking and OT-based WRA contribute to better pre-training. We also conduct a thorough ablation study to find an optimal combination of pre-training tasks. Extensive experiments show that UNITER achieves new state of the art across six V+L tasks (over nine datasets), including Visual Question Answering, Image-Text Retrieval, Referring Expression Comprehension, Visual Commonsense Reasoning, Visual Entailment, and NLVR^2 (Code is available at https://github.com/ChenRocks/UNITER.).

1 Introduction

Most Vision-and-Language (V+L) tasks rely on joint multimodal embeddings to bridge the semantic gap between visual and textual clues in images and text,

Y.-C. Chen, L. Li and L. Yu—Equal contribution.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-58577-8_7) contains supplementary material, which is available to authorized users.

A. Vedaldi et al. (Eds.): ECCV 2020, LNCS 12375, pp. 104–120, 2020.
https://doi.org/10.1007/978-3-030-58577-8_7

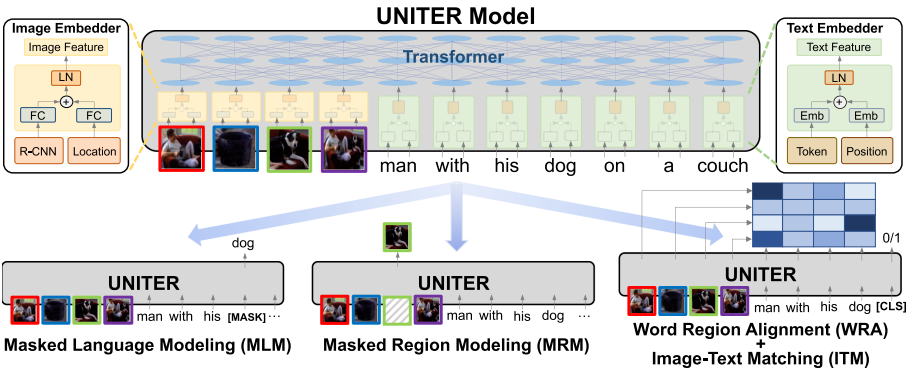


Fig. 1. Overview of the proposed UNITER model (best viewed in color), consisting of an Image Embedder, a Text Embedder and a multi-layer Transformer, learned through four pre-training tasks (Color figure online)

although such representations are usually tailored for specific tasks. For example, MCB [8], BAN [14] and DFAF [10] proposed advanced multimodal fusion methods for Visual Question Answering (VQA) [3]. SCAN [18] and MAttNet [45] studied learning latent alignment between words and image regions for Image-Text Retrieval [40] and Referring Expression Comprehension [13]. While each of these models has pushed the state of the art on respective benchmarks, their architectures are diverse and the learned representations are highly task-specific, preventing them from being generalizable to other tasks. This raises a million-dollar question: can we learn a universal image-text representation for all V+L tasks?

In this spirit, we introduce **UN**iversal **I**mage-**TE**xt **R**epresentation (**UNITER**), a large-scale pre-trained model for joint multimodal embedding. We adopt Transformer [39] as the core of our model, to leverage its elegant self-attention mechanism designed for learning contextualized representations. Inspired by BERT [6], which has successfully applied Transformer to NLP tasks through large-scale language modeling, we pre-train UNITER through four pre-training tasks: (*i*) Masked Language Modeling (MLM) *conditioned on image*; (*ii*) Masked Region Modeling (MRM) *conditioned on text*; (*iii*) Image-Text Matching (ITM); and (*iv*) Word-Region Alignment (WRA). To further investigate the effectiveness of MRM, we propose three MRM variants: (*i*) Masked Region Classification (MRC); (*ii*) Masked Region Feature Regression (MRFR); and (*iii*) Masked Region Classification with KL-divergence (MRC-kl).

As shown in Fig. 1, UNITER first encodes image regions (visual features and bounding box features) and textual words (tokens and positions) into a common embedding space with Image Embedder and Text Embedder. Then, a Transformer module is applied to learn generalizable contextualized embeddings for each region and each word through well-designed pre-training tasks. Compared with previous work on multimodal pre-training [1,19,20,23,33,37,50]:

(*i*) our masked language/region modeling is conditioned on full observation of image/text, rather than applying joint random masking to both modalities; (*ii*) we introduce a novel WRA pre-training task via the use of Optimal Transport (OT) [5,29] to *explicitly* encourage fine-grained alignment between words and image regions. Intuitively, OT-based learning aims to optimize for distribution matching via minimizing the cost of transporting one distribution to another. In our context, we aim to minimize the cost of transporting the embeddings from image regions to words in a sentence (and vice versa), thus optimizing towards better cross-modal alignment. We show that both conditional masking and OT-based WRA can successfully ease the misalignment between images and text, leading to better joint embeddings for downstream tasks.

To demonstrate the generalizable power of UNITER, we evaluate on six V+L tasks across nine datasets, including: (*i*) VQA; (*ii*) Visual Commonsense Reasoning (VCR) [48]; (*iii*) NLVR^2 [34]; (*iv*) Visual Entailment [42]; (*v*) Image-Text Retrieval (including zero-shot setting) [18]; and (*vi*) Referring Expression Comprehension [46]. Our UNITER model is trained on a large-scale V+L dataset composed of four subsets: (*i*) COCO [21]; (*ii*) Visual Genome (VG) [16]; (*iii*) Conceptual Captions (CC) [32]; and (*iv*) SBU Captions [26]. Experiments show that UNITER achieves new state of the art with significant performance boost across all nine downstream datasets. Moreover, training on additional CC and SBU data (containing unseen images/text in downstream tasks) further boosts model performance over training on COCO and VG only.

Our contributions are summarized as follows: (*i*) We introduce UNITER, a powerful UNiversal Image-TExt Representation for V+L tasks. (*ii*) We present Conditional Masking for masked language/region modeling, and propose a novel Optimal-Transport-based Word-Region Alignment task for pre-training. (*iii*) We achieve new state of the art on a wide range of V+L benchmarks, outperforming existing multimodal pre-training methods by a large margin. We also present extensive experiments and analysis to provide useful insights on the effectiveness of each pre-training task/dataset for multimodal encoder training.

2 Related Work

Self-supervised learning utilizes original data as its own source of supervision, which has been applied to many Computer Vision tasks, such as image colorization [49], solving jigsaw puzzles [25,38], inpainting [27], rotation prediction [11], and relative location prediction [7]. Recently, pre-trained language models, such as ELMo [28], BERT [6], GPT2 [31], XLNet [44], RoBERTa [22] and ALBERT [17], have pushed great advances for NLP tasks. There are two keys to their success: effective pre-training tasks over large language corpus, and the use of Transformer [39] for learning contextualized text representations.

More recently, there has been a surging interest in self-supervised learning for multimodal tasks, by pre-training on large-scale image/video and text pairs, then finetuning on downstream tasks. For example, VideoBERT [36] and CBT [35] applied BERT to learn a joint distribution over video frame features and linguistic tokens from video-text pairs. ViLBERT [23] and LXMERT [37] introduced

the two-stream architecture, where two Transformers are applied to images and text independently, which is fused by a third Transformer in a later stage. On the other hand, B2T2 [1], VisualBERT [20], Unicoder-VL [19] and VL-BERT [33] proposed the single-stream architecture, where a single Transformer is applied to both images and text. VLP [50] applied pre-trained models to both image captioning and VQA. More recently, multi-task learning [24] and adversarial training [9] were used to further boost the performance. VALUE [4] developed a set of probing tasks to understand pre-trained models.

Our Contributions. The key differences between our UNITER model and the other methods are two-fold: (*i*) UNITER uses conditional masking on MLM and MRM, *i.e.*, masking only one modality while keeping the other untainted; and (*ii*) a novel Word-Region Alignment pre-training task via the use of Optimal Transport, while in previous work such alignment is only implicitly enforced by task-specific losses. In addition, we examine the best combination of pre-training tasks through a thorough ablation study, and achieve new state of the art on multiple V+L datasets, often outperforming prior work by a large margin.

3 UNiversal Image-TExt Representation

In this section, we first introduce the model architecture of UNITER (Sect. 3.1), then describe the designed pre-training tasks and V+L datasets used for pre-training (Sect. 3.2 and 3.3).

3.1 Model Overview

The model architecture of UNITER is illustrated in Fig. 1. Given a pair of image and sentence, UNITER takes the visual regions of the image and textual tokens of the sentence as inputs. We design an Image Embedder and a Text Embedder to extract their respective embeddings. These embeddings are then fed into a multi-layer Transformer to learn a cross-modality contextualized embedding across visual regions and textual tokens. Note that the self-attention mechanism in Transformer is order-less, thus it is necessary to explicitly encode the positions of tokens and the locations of regions as additional inputs.

Specifically, in *Image Embedder*, we first use Faster R-CNN[1] to extract the visual features (pooled ROI features) for each region. We also encode the location features for each region via a 7-dimensional vector.[2] Both visual and location features are then fed through a fully-connected (FC) layer, to be projected into the same embedding space. The final visual embedding for each region is obtained by summing up the two FC outputs and then passing through a layer normalization (LN) layer. For *Text Embedder*, we follow BERT [6] and tokenize the input sentence into WordPieces [41]. The final representation for each sub-word

[1] Our Faster R-CNN was pre-trained on Visual Genome object+attribute data [2].

[2] $[x_1, y_1, x_2, y_2, w, h, w * h]$ (normalized top/left/bottom/right coordinates, width, height, and area.).

token[3] is obtained via summing up its word embedding and position embedding, followed by another LN layer.[4]

We introduce four main tasks to pre-train our model: Masked Language Modeling *conditioned on image regions* (MLM), Masked Region Modeling *conditioned on input text* (with three variants) (MRM), Image-Text Matching (ITM), and Word-Region Alignment (WRA). As shown in Fig. 1, our MRM and MLM are in analogy to BERT, where we randomly mask some words or regions from the input and learn to recover the words or regions as the output of Transformer. Specifically, word masking is realized by replacing the token with a special token `[MASK]`, and region masking is implemented by replacing the visual feature vector with all zeros. Note that each time we only mask one modality while keeping the other modality intact, instead of randomly masking both modalities as used in other pre-training methods. This prevents potential misalignment when a masked region happens to be described by a masked word (detailed in Sect. 4.2).

We also learn an instance-level alignment between the whole image and the sentence via ITM. During training, we sample both positive and negative image-sentence pairs and learn their matching scores. Furthermore, in order to provide a more fine-grained alignment between word tokens and image regions, we propose WRA via the use of Optimal Transport, which effectively calculates the minimum cost of transporting the contextualized image embeddings to word embeddings (and vice versa). The inferred transport plan thus serves as a propeller for better cross-modal alignment. Empirically, we show that both conditional masking and WRA contributes to performance improvement (in Sect. 4.2). To pre-train UNITER with these tasks, we randomly sample one task for each mini-batch, and train on only one objective per SGD update.

3.2 Pre-training Tasks

Masked Language Modeling (MLM). We denote the image regions as $\mathbf{v} = \{\mathbf{v}_1, ..., \mathbf{v}_K\}$, the input words as $\mathbf{w} = \{\mathbf{w}_1, ..., \mathbf{w}_T\}$, and the mask indices as $\mathbf{m} \in \mathbb{N}^M$.[5] In MLM, we randomly mask out the input words with probability of 15%, and replace the masked ones $\mathbf{w_m}$ with special token `[MASK]`.[6] The goal is to predict these masked words based on the observation of their surrounding words $\mathbf{w}_{\backslash\mathbf{m}}$ and all image regions $\mathbf{v}$, by minimizing the negative log-likelihood:

$$\mathcal{L}_{\text{MLM}}(\theta) = -\mathbb{E}_{(\mathbf{w},\mathbf{v})\sim D} \log P_\theta(\mathbf{w_m}|\mathbf{w}_{\backslash\mathbf{m}}, \mathbf{v}), \tag{1}$$

[3] We use word/sub-word and token interchangeably throughout the rest of the paper.

[4] We also use a special modality embedding to help the model distinguish between textual and visual input, which is similar to the 'segment embedding' in BERT. This embedding is also summed before the LN layer in each embedder. For simplicity, this modality embedding is omitted in Fig. 1.

[5] $\mathbb{N}$ is the natural numbers, M is the number of masked tokens, and $\mathbf{m}$ is the set of masked indices.

[6] Following BERT, we decompose this 15% into 10% random words, 10% unchanged, and 80% `[MASK]`.

where θ is the trainable parameters. Each pair $(\mathbf{w}, \mathbf{v})$ is sampled from the whole training set D.

Image-Text Matching (ITM). In ITM, an additional special token `[CLS]` is fed into our model, which indicates the fused representation of both modalities. The inputs to ITM are a sentence and a set of image regions, and the output is a binary label $y \in \{0, 1\}$, indicating if the sampled pair is a match. We extract the representation of `[CLS]` token as the joint representation of the input image-text pair, then feed it into an FC layer and a sigmoid function to predict a score between 0 and 1. We denote the output score as $s_\theta(\mathbf{w}, \mathbf{v})$. The ITM supervision is over the `[CLS]` token.[7] During training, we sample a positive or negative pair $(\mathbf{w}, \mathbf{v})$ from the dataset D at each step. The negative pair is created by replacing the image or text in a paired sample with a randomly-selected one from other samples. We apply the binary cross-entropy loss for optimization:

$$\mathcal{L}_{\mathrm{ITM}}(\theta) = -\mathbb{E}_{(\mathbf{w},\mathbf{v})\sim D}[y \log s_\theta(\mathbf{w}, \mathbf{v}) + (1 - y) \log(1 - s_\theta(\mathbf{w}, \mathbf{v}))]). \tag{2}$$

Word-Region Alignment (WRA). We use Optimal Transport (OT) for WRA, where a transport plan $\mathbf{T} \in \mathbb{R}^{T\times K}$ is learned to optimize the alignment between $\mathbf{w}$ and $\mathbf{v}$. OT possesses several idiosyncratic characteristics that make it a good choice for WRA: (*i*) *Self-normalization*: all the elements of $\mathbf{T}$ sum to 1 [29]. (*ii*) *Sparsity*: when solved exactly, OT yields a sparse solution $\mathbf{T}$ containing $(2r - 1)$ non-zero elements at most, where $r = \max(K, T)$, leading to a more interpretable and robust alignment [29]. (*iii*) *Efficiency*: compared with conventional linear programming solvers, our solution can be readily obtained using iterative procedures that only require matrix-vector products [43], hence readily applicable to large-scale model pre-training.

Specifically, $(\mathbf{w}, \mathbf{v})$ can be considered as two discrete distributions $\boldsymbol{\mu}, \boldsymbol{\nu}$, formulated as $\boldsymbol{\mu} = \sum_{i=1}^{T} \mathbf{a}_i \delta_{\mathbf{w}_i}$ and $\boldsymbol{\nu} = \sum_{j=1}^{K} \mathbf{b}_j \delta_{\mathbf{v}_j}$, with $\delta_{\mathbf{w}_i}$ as the Dirac function centered on $\mathbf{w}_i$. The weight vectors $\mathbf{a} = \{\mathbf{a}_i\}_{i=1}^{T} \in \Delta_T$ and $\mathbf{b} = \{\mathbf{b}_j\}_{j=1}^{K} \in \Delta_K$ belong to the T- and K-dimensional simplex, respectively (*i.e.*, $\sum_{i=1}^{T} \mathbf{a}_i = \sum_{j=1}^{K} \mathbf{b}_j = 1$), as both $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are probability distributions. The OT distance between $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ (thus also the alignment loss for the $(\mathbf{w}, \mathbf{v})$ pair) is defined as:

$$\mathcal{L}_{\mathrm{WRA}}(\theta) = \mathcal{D}_{ot}(\boldsymbol{\mu}, \boldsymbol{\nu}) = \min_{\mathbf{T}\in\Pi(\mathbf{a},\mathbf{b})} \sum_{i=1}^{T}\sum_{j=1}^{K} \mathbf{T}_{ij} \cdot c(\mathbf{w}_i, \mathbf{v}_j)\,, \tag{3}$$

where $\Pi(\mathbf{a}, \mathbf{b}) = \{\mathbf{T} \in \mathbb{R}_{+}^{T\times K} | \mathbf{T}\mathbf{1}_m = \mathbf{a}, \mathbf{T}^\top \mathbf{1}_n = \mathbf{b}\}$, $\mathbf{1}_n$ denotes an n-dimensional all-one vector, and $c(\mathbf{w}_i, \mathbf{v}_j)$ is the cost function evaluating the distance between $\mathbf{w}_i$ and $\mathbf{v}_j$. In experiments, the cosine distance $c(\mathbf{w}_i, \mathbf{v}_j) = 1 - \frac{\mathbf{w}_i^\top \mathbf{v}_j}{||\mathbf{w}_i||_2 ||\mathbf{v}_j||_2}$ is used. The matrix $\mathbf{T}$ is denoted as the transport plan, interpreting the alignment between two modalities. Unfortunately, the exact minimization

[7] Performing this during pre-training also alleviates the mismatch problem between pre-training and downstream finetuning tasks, since most of the downstream tasks take the representation of the `[CLS]` token as the joint representation.

over $\mathbf{T}$ is computational intractable, and we consider the IPOT algorithm [43] to approximate the OT distance (details are provided in the supplementary file). After solving $\mathbf{T}$, the OT distance serves as the WRA loss that can be used to update the parameters θ.

Masked Region Modeling (MRM). Similar to MLM, we also sample image regions and mask their visual features with a probability of 15%. The model is trained to reconstruct the masked regions $\mathbf{v_m}$ given the remaining regions $\mathbf{v}_{\backslash\mathbf{m}}$ and all the words $\mathbf{w}$. The visual features of the masked region are replaced by zeros. Unlike textual tokens that are represented as discrete labels, visual features are high-dimensional and continuous, thus cannot be supervised via class likelihood. Instead, we propose three variants for MRM, which share the same objective base:

$$\mathcal{L}_{\mathrm{MRM}}(\theta) = \mathbb{E}_{(\mathbf{w},\mathbf{v})\sim D} f_\theta(\mathbf{v_m}|\mathbf{v}_{\backslash\mathbf{m}}, \mathbf{w}). \tag{4}$$

1) **Masked Region Feature Regression (MRFR)** MRFR learns to regress the Transformer output of each masked region $\mathbf{v}_\mathbf{m}^{(i)}$ to its visual features. Specifically, we apply an FC layer to convert its Transformer output into a vector $h_\theta(\mathbf{v}_\mathbf{m}^{(i)})$ of same dimension as the input ROI pooled feature $r(\mathbf{v}_\mathbf{m}^{(i)})$. Then we apply L2 regression between the two: $f_\theta(\mathbf{v_m}|\mathbf{v}_{\backslash\mathbf{m}}, \mathbf{w}) = \sum_{i=1}^{M} \|h_\theta(\mathbf{v}_\mathbf{m}^{(i)}) - r(\mathbf{v}_\mathbf{m}^{(i)})\|_2^2$.

2) **Masked Region Classification (MRC)** MRC learns to predict the object semantic class for each masked region. We first feed the Transformer output of the masked region $\mathbf{v}_\mathbf{m}^{(i)}$ into an FC layer to predict the scores of K object classes, which further goes through a softmax function to be transformed into a normalized distribution $g_\theta(\mathbf{v}_\mathbf{m}^{(i)}) \in \mathbb{R}^K$. Note that there is no ground-truth label, as the object categories are not provided. Thus, we use the object detection output from Faster R-CNN, and take the detected object category (with the highest confidence score) as the label of the masked region, which will be converted into a one-hot vector $c(\mathbf{v}_\mathbf{m}^{(i)}) \in \mathbb{R}^K$. The final objective minimizes the cross-entropy (CE) loss: $f_\theta(\mathbf{v_m}|\mathbf{v}_{\backslash\mathbf{m}}, \mathbf{w}) = \sum_{i=1}^{M} \mathrm{CE}(c(\mathbf{v}_\mathbf{m}^{(i)}), g_\theta(\mathbf{v}_\mathbf{m}^{(i)}))$.

3) **Masked Region Classification with KL-Divergence (MRC-kl)** MRC takes the most likely object class from the object detection model as the hard label (w.p. 0 or 1), assuming the detected object class is the ground-truth label for the region. However, this may not be true, as no ground-truth label is available. Thus, in MRC-kl, we avoid this assumption by using soft label as supervision signal, which is the raw output from the detector (*i.e.*, a distribution of object classes $\tilde{c}(\mathbf{v}_m^{(i)})$). MRC-kl aims to distill such knowledge into UNITER as [12], by minimizing the KL divergence between two distributions: $f_\theta(\mathbf{v_m}|\mathbf{v}_{\backslash\mathbf{m}}, \mathbf{w}) = \sum_{i=1}^{M} D_{KL}(\tilde{c}(\mathbf{v}_\mathbf{m}^{(i)})||g_\theta(\mathbf{v}_\mathbf{m}^{(i)}))$.

3.3 Pre-training Datasets

We construct our pre-training dataset based on four existing V+L datasets: COCO [21], Visual Genome (VG) [16], Conceptual Captions (CC) [32], and SBU

Table 1. Statistics on the datasets used for pre-training. Each cell shows #image-text pairs (#images)

Split	In-domain		Out-of-domain	
	COCO captions	VG dense captions	Conceptual captions	SBU captions
train	533K (106K)	5.06M (101K)	3.0M (3.0M)	990K (990K)
val	25K (5K)	106K (2.1K)	14K (14K)	10K (10K)

Captions [26]. Only image and sentence pairs are used for pre-training, which makes the model framework more scalable, as additional image-sentence pairs are easy to harvest for further pre-training.

To study the effects of different datasets on pre-training, we divide the four datasets into two categories. The first one consists of image captioning data from COCO and dense captioning data from VG. We call it "In-domain" data, as most V+L tasks are built on top of these two datasets. To obtain a "fair" data split, we merge the raw training and validation splits from COCO, and exclude all validation and test images that appear in downstream tasks. We also exclude all co-occurring Flickr30K [30] images via URL matching, as both COCO and Flickr30K images were crawled from Flickr and may have overlaps.[8] The same rule was applied to Visual Genome as well. In this way, we obtain 5.6M image-text pairs for training and 131K image-text pairs for our internal validation, which is half the size of the dataset used in LXMERT [37], due to the filtering of overlapping images and the use of image-text pairs only. We also use additional Out-of-domain data from Conceptual Captions [32] and SBU Captions [26] for model training.[9] The statistics on the cleaned splits are provided in Table 1.

4 Experiments

We evaluate UNITER on six V+L tasks[10] by transferring the pre-trained model to each target task and finetuning through end-to-end training. We report experimental results on two model sizes: UNITER-base with 12 layers and UNITER-large with 24 layers.[11]

[8] A total of 222 images were eliminated through this process.

[9] We apply the same URL matching method, excluding 109 images from training.

[10] VQA, VCR, NLVR^2, Visual Entailment, Image-Text Retrieval, and Referring Expression Comprehension. Details about the tasks are listed in the supplementary.

[11] UNITER-base: L = 12, H = 768, A = 12, Total Parameters = 86M. UNITER-large: L = 24, H = 1024, A = 16, Total Parameters = 303M (L: number of stacked Transformer blocks; H: hidden activation dimension; A: number of attention heads). 882 and 3645 V100 GPU hours were used for pre-training UNITER-base and UNITER-large.

Table 2. Evaluation on pre-training tasks and datasets using VQA, Image-Text Retrieval on Flickr30K, NLVR^2, and RefCOCO+ as benchmarks. All results are obtained from UNITER-base. Averages of R@1, R@5 and R@10 on Flickr30K for Image Retrieval (IR) and Text Retrieval (TR) are reported. Dark and light grey colors highlight the top and second best results across all the tasks trained with In-domain data

Pre-training Data		Pre-training Tasks	Meta-Sum	VQA	IR (Flickr)	TR (Flickr)	NLVR^2	Ref-COCO+
				test-dev	val	val	dev	val^d
None	1	None	314.34	67.03	61.74	65.55	51.02	68.73
Wikipedia + BookCorpus	2	MLM (text only)	346.24	69.39	73.92	83.27	50.86	68.80
In-domain (COCO+VG)	3	MRFR	344.66	69.02	72.10	82.91	52.16	68.47
	4	ITM	385.29	70.04	78.93	89.91	74.08	72.33
	5	MLM	386.10	71.29	77.88	89.25	74.79	72.89
	6	MLM + ITM	393.04	71.55	81.64	91.12	75.98	72.75
	7	MLM + ITM + MRC	393.97	71.46	81.39	91.45	76.18	73.49
	8	MLM + ITM + MRFR	396.24	71.73	81.76	92.31	76.21	74.23
	9	MLM + ITM + MRC-kl	397.09	71.63	82.10	92.57	76.28	74.51
	10	MLM + ITM + MRC-kl + MRFR	399.97	71.92	83.73	92.87	76.93	74.52
	11	MLM + ITM + MRC-kl + MRFR + WRA	400.93	72.47	83.72	93.03	76.91	74.80
	12	MLM + ITM + MRC-kl + MRFR (w/o cond. mask)	396.51	71.68	82.31	92.08	76.15	74.29
Out-of-domain (SBU+CC)	13	MLM + ITM + MRC-kl + MRFR + WRA	396.91	71.56	84.34	92.57	75.66	72.78
In-domain + Out-of-domain	14	MLM + ITM + MRC-kl + MRFR + WRA	**405.24**	**72.70**	**85.77**	**94.28**	**77.18**	**75.31**

4.1 Downstream Tasks

In VQA, VCR and NLVR^2 tasks, given an input image (or a pair of images) and a natural language question (or description), the model predicts an answer (or judges the correctness of the description) based on the visual content in the image. For Visual Entailment, we evaluate on the SNLI-VE dataset. The goal is to predict whether a given image semantically entails an input sentence. Classification accuracy over three classes ("Entailment", "Neutral" and "Contradiction") is used to measure model performance. For Image-Text Retrieval, we consider two datasets (COCO and Flickr30K) and evaluate the model in two settings: Image Retrieval (IR) and Text Retrieval (TR). Referring Expression (RE) Comprehension requires the model to select the target from a set of image region proposals given the query description. Models are evaluated on both ground-truth objects and detected proposals[12] (MAttNet [45]).

For VQA, VCR, NLVR^2, Visual Entailment and Image-Text Retrieval, we extract the joint embedding of the input image-text pairs via a multi-layer perceptron (MLP) from the representation of the `[CLS]` token. For RE Comprehension, we use the MLP to compute the region-wise alignment scores. These MLP layers are learned during the finetuning stage. Specifically, we formulate VQA, VCR, NLVR^2, Visual Entailment and RE Comprehension as classification prob-

[12] The evaluation splits of RE comprehension using detected proposals are denoted as val^d, test^d, etc.

lems and minimize the cross-entropy over the ground-truth answers/responses. For Image-Text Retrieval, we formulate it as a ranking problem. During finetuning, we sample three pairs of image and text, one positive pair from the dataset and two negative pairs by randomly replacing its sentence/image with others. We compute the similarity scores (based on the joint embedding) for both positive and negative pairs, and maximize the margin between them through triplet loss.

4.2 Evaluation on Pre-training Tasks

We analyze the effectiveness of different pre-training settings through ablation studies over VQA, NLVR^2, Flickr30K and RefCOCO+ as representative V+L benchmarks. In addition to standard metrics[13] for each benchmark , we also use Meta-Sum (sum of all the scores across all the benchmarks) as a global metric.

Firstly, we establish two baselines: Line 1 (L1) in Table 2 indicates no pre-training is involved, and L2 shows the results from MLM initialized with pre-trained weights from [6]. Although MLM trained on text only did not absorb any image information during pre-training, we see a gain of approximately +30 on Meta-Sum over L1. Hence, we use the pre-trained weights in L2 to initialize our model for the following experiments.

Secondly, we validate the effectiveness of each pre-training task through a thorough ablation study. Comparing L2 and L3, MRFR (L3) achieves better results than MLM (L2) only on NLVR^2. On the other hand, when pre-trained on ITM (L4) or MLM (L5) only, we observe a significant improvement across all the tasks over L1 and L2 baselines. When combining different pre-training tasks, MLM + ITM (L6) improves over single ITM (L4) or MLM (L5). When MLM, ITM and MRM are jointly trained (L7–L10), we observe consistent performance gain across all the benchmarks. Among the three variants of MRM (L7–L9), we observe that MRC-kl (L9) achieves the best performance (397.09) when combined with MLM + ITM, while MRC (L7) the worst (393.97). When combining MRC-kl and MRFR together with MLM and ITM (L10), we find that they are complimentary to each other, which leads to the second highest Meta-Sum score. The highest Meta-Sum Score is achieved by MLM + ITM + MRC-kl + MRFR + WRA (L11). We observe significant performance improvements from adding WRA, especially on VQA and RefCOCO+. It indicates the fine-grained alignment between words and regions learned through WRA during pre-training benefits the downstream tasks involving region-level recognition or reasoning. We use this optimal pre-training setting for the further experiments.

Additionally, we validate the contributions of conditional masking through a comparison study. When we perform random masking on both modalities simultaneously during pre-training, *i.e.*, w/o conditional masking (L12), we observe a decrease in Meta-Sum score (396.51) compared to that with conditional masking (399.97). This indicates that the conditional masking strategy enables the model to learn better joint image-text representations effectively.

[13] Details about the metrics are listed in the supplementary.

Lastly, we study the effects of pre-training datasets. Our experiments so far have been focused on In-domain data. In this study, we pre-train our model on Out-of-domain data (Conceptual Captions + SBU Captions). A performance drop (396.91 in L13) from the model trained on In-domain data (COCO + Visual Genome) (400.93 in L11) shows that although Out-of-domain data contain more images, the model still benefits more from being exposed to similar downstream images during pre-training. We further pre-train our model on both In-domain and Out-of-domain data. With doubled data size, the model continues to improve (405.24 in L14).

4.3 Results on Downstream Tasks

Table 3 presents the results of UNITER on all downstream tasks. Both our base and large models are pre-trained on In-domain+Out-of-domain datasets, with the optimal pre-training setting: MLM+ITM+MRC-kl+MRFR+WRA. The implementation details of each task are provided in the supplementary file. We compare with both task-specific models and other pre-trained models on each downstream task. SOTA task-specific models include: MCAN [47] for VQA, MaxEnt [34] for NLVR^2, B2T2 [1] for VCR, SCAN [18] for Image-Text Retrieval, EVE-Image [42] for SNLI-VE, and MAttNet for RE Comprehension (RefCOCO, RefCOCO+ and RefCOCOg).[14] Other pre-trained models include: ViLBERT [23], LXMERT [37], Unicoder-VL [19], VisualBERT [20] and VLBERT [33].

Results show that our UNITER-large model achieves new state of the art across all the benchmarks. UNITER-base model also outperforms the others by a large margin across all tasks except VQA. Specifically, our UNITER-base model outperforms SOTA by approximately +2.8% for VCR on Q→AR, +2.5% for NLVR^2, +7% for SNLI-VE, +4% on R@1 for Image-Text Retrieval (+15% for zero-shot setting), and +2% for RE Comprehension.

Note that LXMERT pre-trains with downstream VQA (+VG+GQA) data, which may help adapt the model to VQA task. However, when evaluated on unseen tasks such as NLVR^2, UNITER-base achieves 3% gain over LXMERT. In addition, among all the models pre-trained on image-text pairs only, our UNITER-base outperforms the others by >1.5% on VQA.

It is also worth mentioning that both VilBERT and LXMERT observed two-stream model outperforms single-stream model, while our results show empirically that with our pre-training setting, single-stream model can achieve new state-of-the-art results, with much fewer parameters (UNITER-base: 86M, LXMERT: 183M, VilBERT: 221M).[15]

For VCR, we propose a two-stage pre-training approach: (*i*) pre-train on standard pre-training datasets; and then (*ii*) pre-train on downstream VCR dataset. Interestingly, while VLBERT and B2T2 observed that pre-training is not very

[14] MAttNet results are updated using the same features as the others. More details are provided in the supplementary file.

[15] The word embedding layer contains excessive rare words, thus excluded from the parameter counts.

Table 3. Results on downstream V+L tasks from UNITER model, compared with task-specific state-of-the-art (SOTA) and previous pre-trained models. ZS: Zero-Shot, IR: Image Retrieval and TR: Text Retrieval

Tasks		SOTA	ViLBERT	VLBERT (Large)	Unicoder -VL	VisualBERT	LXMERT	UNITER Base	UNITER Large
VQA	test-dev	70.63	70.55	71.79	-	70.80	72.42	72.70	**73.82**
	test-std	70.90	70.92	72.22	-	71.00	72.54	72.91	**74.02**
VCR	Q→A	72.60	73.30	75.80	-	71.60	-	75.00	**77.30**
	QA→R	75.70	74.60	78.40	-	73.20	-	77.20	**80.80**
	Q→AR	55.00	54.80	59.70	-	52.40	-	58.20	**62.80**
NLVR^2	dev	54.80	-	-	-	67.40	74.90	77.18	**79.12**
	test-P	53.50	-	-	-	67.00	74.50	77.85	**79.98**
SNLI-VE	val	71.56	-	-	-	-	-	78.59	**79.39**
	test	71.16	-	-	-	-	-	78.28	**79.38**
ZS IR (Flickr)	R@1	-	31.86	-	48.40	-	-	66.16	**68.74**
	R@5	-	61.12	-	76.00	-	-	88.40	**89.20**
	R@10	-	72.80	-	85.20	-	-	92.94	**93.86**
IR (Flickr)	R@1	48.60	58.20	-	71.50	-	-	72.52	**75.56**
	R@5	77.70	84.90	-	91.20	-	-	92.36	**94.08**
	R@10	85.20	91.52	-	95.20	-	-	96.08	**96.76**
IR (COCO)	R@1	38.60	-	-	48.40	-	-	50.33	**52.93**
	R@5	69.30	-	-	76.70	-	-	78.52	**79.93**
	R@10	80.40	-	-	85.90	-	-	87.16	**87.95**
ZS TR (Flickr)	R@1	-	-	-	64.30	-	-	80.70	**83.60**
	R@5	-	-	-	85.80	-	-	**95.70**	**95.70**
	R@10	-	-	-	92.30	-	-	**98.00**	97.70
TR (Flickr)	R@1	67.90	-	-	86.20	-	-	85.90	**87.30**
	R@5	90.30	-	-	96.30	-	-	97.10	**98.00**
	R@10	95.80	-	-	99.00	-	-	98.80	**99.20**
TR (COCO)	R@1	50.40	-	-	62.30	-	-	64.40	**65.68**
	R@5	82.20	-	-	87.10	-	-	87.40	**88.56**
	R@10	90.00	-	-	92.80	-	-	93.08	**93.76**
Ref-COCO	val	87.51		-	-	-	-	91.64	**91.84**
	testA	89.02	-	-	-	-	-	92.26	**92.65**
	testB	87.05	-	-	-	-	-	90.46	**91.19**
	val^d	77.48	-	-	-	-	-	81.24	**81.41**
	testA^d	83.37	-	-	-	-	-	86.48	**87.04**
	testB^d	70.32	-	-	-	-	-	73.94	**74.17**
Ref-COCO+	val	75.38	-	80.31	-	-	-	83.66	**84.25**
	testA	80.04	-	83.62	-	-	-	86.19	**86.34**
	testB	69.30	-	75.45	-	-	-	78.89	**79.75**
	val^d	68.19	72.34	72.59	-	-	-	75.31	**75.90**
	testA^d	75.97	78.52	78.57	-	-	-	81.30	**81.45**
	testB^d	57.52	62.61	62.30	-	-	-	65.58	**66.70**
Ref-COCOg	val	81.76	-	-	-	-	-	86.52	**87.85**
	test	81.75	-	-	-	-	-	86.52	**87.73**
	val^d	68.22	-	-	-	-	-	74.31	**74.86**
	test^d	69.46	-	-	-	-	-	74.51	**75.77**

Table 4. Experiments on two-stage pre-training for VCR. Results are from UNITER-base on VCR val split. Stage I and Stage II denote first-stage and second-stage pre-training

Stage I	Stage II	Q→A	QA→ R	Q → AR
N	N	72.44	73.71	53.52
N	Y	73.52	75.34	55.6
Y	N	72.83	75.25	54.94
Y	Y	**74.56**	**77.03**	**57.76**

Table 5. Experiments on three modified settings for NLVR^2. All models use pre-trained UNITER-base

Setting	dev	test-P
Triplet	73.03	73.89
Pair	75.85	75.80
Pair-biattn	**77.18**	**77.85**

helpful on VCR, we find that the second-stage pre-training can significantly boost model performance, while the first-stage pre-training still helps but with limited effects (results shown in Table 4). This indicates that the proposed two-stage approach is highly effective in our pre-trained model over new data that are unseen in pre-training datasets.

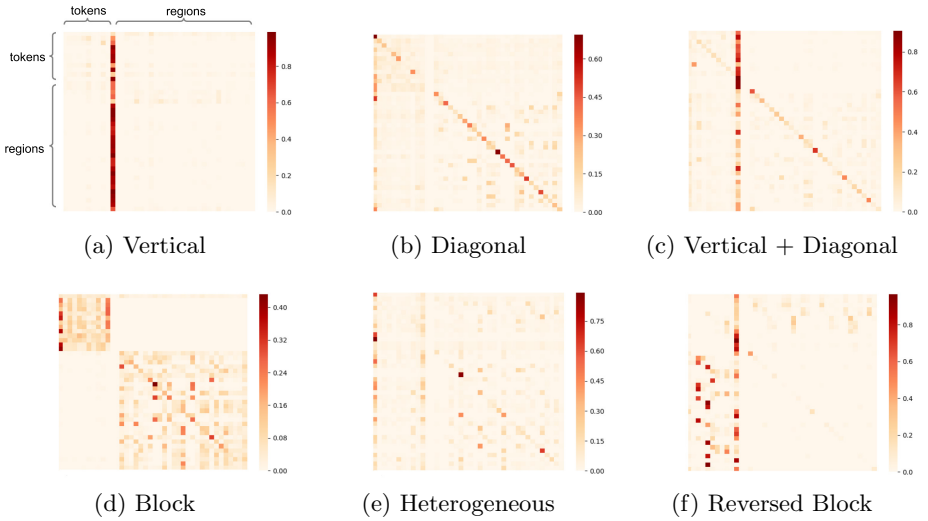
Different from other tasks, NLVR^2 takes two images as input. Thus, directly finetuning UNITER pre-trained with image-sentence pairs might not lead to optimal performance, as the interactions between paired images are not learned during the pre-training stage. Thus, we experimented with three modified settings on NLVR^2: (*i*) *Triplet*: joint embedding of images pairs and query captions; (*ii*) *Pair*: individual embedding of each image and each query caption; and (*iii*) *Pair-biattn*: a bidirectional attention is added to the *Pair* model to learn the interactions between the paired images.

Comparison results are presented in Table 5. The *Pair* setting achieves better performance than the *Triplet* setting even without cross-attention between the image pairs. We hypothesize that it is due to the fact that our UNITER is pre-trained with image-text pairs. Thus, it is difficult to finetune a pair-based pre-trained model on triplet input. The bidirectional attention mechanism in the *Pair-biattn* setting, however, compensates the lack of cross-attention between images, hence yielding the best performance with a large margin. This show that with minimal surgery on the top layer of UNITER, our pre-trained model can adapt to new tasks that are very different from pre-training tasks.

4.4 Visualization

Similar to [15], we observe several patterns in the attention maps of the UNITER model, as shown in Fig. 2. Note that different from [15], our attention mechanism operates in both inter- and intra-modality manners. For completeness, we briefly discuss each pattern here:

- *Vertical:* attention to special tokens `[CLS]` or `[SEP]`;
- *Diagonal:* attention to the token/region itself or preceding/following tokens/regions;
- *Vertical + Diagonal:* mixture of vertical and diagonal;



(a) Vertical (b) Diagonal (c) Vertical + Diagonal

(d) Block (e) Heterogeneous (f) Reversed Block

Fig. 2. Visualization of the attention maps learned by the UNITER-base model

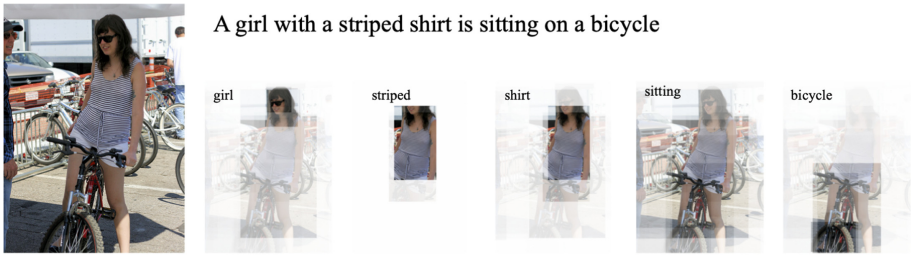


Fig. 3. Text-to-image attention visualization example

- *Block:* intra-modality attention, *i.e.*, textual self-attention and visual self-attention;
- *Heterogeneous:* diverse attentions that cannot be categorized and is highly dependent on actual input;
- *Reversed Block:* inter-modality attention, *i.e.*, text-to-image and image-to-text attention.

Note that *Reversed Block* (Fig. 2f) shows cross-modality alignment between tokens and regions. In Fig. 3, we visualize several examples of text-to-image attention to demonstrate the local cross-modality alignment between regions and tokens.

5 Conclusion

In this paper, we present UNITER, a large-scale pre-trained model providing UNiversal Image-TExt Representations for Vision-and-Language tasks. Four

main pre-training tasks are proposed and evaluated through extensive ablation studies. Trained with both in-domain and out-of-domain datasets, UNITER outperforms state-of-the-art models over multiple V+L tasks by a significant margin. Future work includes studying early interaction between raw image pixels and sentence tokens, as well as developing more effective pre-training tasks.

References

1. Alberti, C., Ling, J., Collins, M., Reitter, D.: Fusion of detected objects in text for visual question answering. In: EMNLP (2019)
2. Anderson, P., et al.: Bottom-up and top-down attention for image captioning and visual question answering. In: CVPR (2018)
3. Antol, S., et al.: VQA: visual question answering. In: ICCV (2015)
4. Cao, J., Gan, Z., Cheng, Y., Yu, L., Chen, Y.C., Liu, J.: Behind the scene: revealing the secrets of pre-trained vision-and-language models. arXiv preprint arXiv:2005.07310 (2020)
5. Chen, L., Gan, Z., Cheng, Y., Li, L., Carin, L., Liu, J.: Graph optimal transport for cross-domain alignment. In: ICML (2020)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. In: NAACL (2019)
7. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: ICCV (2015)
8. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding. In: EMNLP (2017)
9. Gan, Z., Chen, Y.C., Li, L., Zhu, C., Cheng, Y., Liu, J.: Large-scale adversarial training for vision-and-language representation learning. arXiv preprint arXiv:2006.06195 (2020)
10. Gao, P., et al.: Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In: CVPR (2019)
11. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. In: ICLR (2018)
12. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
13. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: ReferItGame: referring to objects in photographs of natural scenes. In: EMNLP (2014)
14. Kim, J.H., Jun, J., Zhang, B.T.: Bilinear attention networks. In: NeurIPS (2018)
15. Kovaleva, O., Romanov, A., Rogers, A., Rumshisky, A.: Revealing the dark secrets of BERT. In: EMNLP (2019)
16. Krishna, R., et al.: Visual genome: connecting language and vision using crowdsourced dense image annotations. Int. J. Comput. Vis. **123**, 32–73 (2017). https://doi.org/10.1007/s11263-016-0981-7
17. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: Albert: a lite BERT for self-supervised learning of language representations. In: ICLR (2020)
18. Lee, K.-H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11208, pp. 212–228. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01225-0_13

19. Li, G., Duan, N., Fang, Y., Jiang, D., Zhou, M.: Unicoder-VL: a universal encoder for vision and language by cross-modal pre-training. In: AAAI (2020)
20. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: VisualBERT: a simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557 (2019)
21. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
22. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
23. Lu, J., Batra, D., Parikh, D., Lee, S.: ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: NeurIPS (2019)
24. Lu, J., Goswami, V., Rohrbach, M., Parikh, D., Lee, S.: 12-in-1: multi-task vision and language representation learning. In: CVPR (2020)
25. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving Jigsaw puzzles. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 69–84. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_5
26. Ordonez, V., Kulkarni, G., Berg, T.L.: Im2Text: describing images using 1 million captioned photographs. In: NeurIPS (2011)
27. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: feature learning by inpainting. In: CVPR (2016)
28. Peters, M.E., et al.: Deep contextualized word representations. In: NAACL (2018)
29. Peyré, G., Cuturi, M., et al.: Computational optimal transport. Found. Trends® Mach. Learn. **11**(5–6), 355–607 (2019)
30. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models. In: ICCV (2015)
31. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners (2019)
32. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: a cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: ACL (2018)
33. Su, W., et al.: VL-BERT: pre-training of generic visual-linguistic representations. In: ICLR (2020)
34. Suhr, A., Zhou, S., Zhang, I., Bai, H., Artzi, Y.: A corpus for reasoning about natural language grounded in photographs. In: ACL (2019)
35. Sun, C., Baradel, F., Murphy, K., Schmid, C.: Contrastive bidirectional transformer for temporal representation learning. arXiv preprint arXiv:1906.05743 (2019)
36. Sun, C., Myers, A., Vondrick, C., Murphy, K., Schmid, C.: VideoBERT: a joint model for video and language representation learning. In: ICCV (2019)
37. Tan, H., Bansal, M.: LXMERT: learning cross-modality encoder representations from transformers. In: EMNLP (2019)
38. Trinh, T.H., Luong, M.T., Le, Q.V.: Selfie: self-supervised pretraining for image embedding. arXiv preprint arXiv:1906.02940 (2019)
39. Vaswani, A., et al.: Attention is all you need. In: NeurIPS (2017)
40. Wang, L., Li, Y., Lazebnik, S.: Learning deep structure-preserving image-text embeddings. In: CVPR (2016)
41. Wu, Y., et al.: Google's neural machine translation system: bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016)

42. Xie, N., Lai, F., Doran, D., Kadav, A.: Visual entailment: a novel task for fine-grained image understanding. arXiv preprint arXiv:1901.06706 (2019)
43. Xie, Y., Wang, X., Wang, R., Zha, H.: A fast proximal point method for Wasserstein distance. arXiv:1802.04307 (2018)
44. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V.: XLNet: generalized autoregressive pretraining for language understanding. In: NeurIPS (2019)
45. Yu, L., et al.: MAttNet: modular attention network for referring expression comprehension. In: CVPR (2018)
46. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 69–85. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_5
47. Yu, Z., Yu, J., Cui, Y., Tao, D., Tian, Q.: Deep modular co-attention networks for visual question answering. In: CVPR (2019)
48. Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: visual commonsense reasoning. In: CVPR (2019)
49. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 649–666. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_40
50. Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J.J., Gao, J.: Unified vision-language pre-training for image captioning and VQA. In: AAAI (2020)