BridgeTower: Building Bridges between Encoders in Vision-Language Representation Learning

Xiao Xu[1, 2*], **Chenfei Wu**[2], **Shachar Rosenman**[3], **Vasudev Lal**[3], **Wanxiang Che**[1†], **Nan Duan**[2†]

[1]Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology
[2]Microsoft Research Asia
[3]Intel Labs, Cognitive Computing Research
{xxu,car}@ir.hit.edu.cn, {chenfei.wu,nanduan}@microsoft.com, {shachar.rosenman,vasudev.lal}@intel.com

Abstract

Vision-Language (VL) models with the TWO-TOWER architecture have dominated visual-language representation learning in recent years. Current VL models either use lightweight uni-modal encoders and learn to extract, align and fuse both modalities simultaneously in a deep cross-modal encoder, or feed the last-layer uni-modal representations from the deep pre-trained uni-modal encoders into the top cross-modal encoder. Both approaches potentially restrict vision-language representation learning and limit model performance. In this paper, we propose BRIDGETOWER, which introduces multiple bridge layers that build a connection between the top layers of uni-modal encoders and each layer of the cross-modal encoder. This enables effective bottom-up cross-modal alignment and fusion between visual and textual representations of different semantic levels of pre-trained uni-modal encoders in the cross-modal encoder. Pre-trained with only 4M images, BRIDGETOWER achieves state-of-the-art performance on various downstream vision-language tasks. In particular, on the VQAv2 test-std set, BRIDGETOWER achieves an accuracy of 78.73%, outperforming the previous state-of-the-art model METER by 1.09% with the same pre-training data and almost negligible additional parameters and computational costs. Notably, when further scaling the model, BRIDGETOWER achieves an accuracy of 81.15%, surpassing models that are pre-trained on orders-of-magnitude larger datasets. Code and checkpoints are available at https://github.com/microsoft/BridgeTower.

1 Introduction

Vision-Language (VL) tasks aim to perceive, comprehend and fuse both visual and textual information in our complex multi-modal world and then produce cross-modal representations to address difficult cross-modal challenges, such as visual question answering, visual entailment, and image-text retrieval (Goyal et al. 2017; Xie et al. 2019; Young et al. 2014). Recently, by pre-training on large-scale image-text pairs, cross-modal representations have been improved considerably (Su et al. 2020; Lu et al. 2019; Chen et al. 2020; Zhang et al. 2021; Radford et al. 2021; Wang et al. 2021c; Li et al. 2021a; Dou et al. 2022; Wang et al. 2022b; Alayrac et al. 2022). Many elaborate Vision-Language Pre-training (VLP) objectives are proposed for mining cross-modal knowledge from image-text pairs, such as Masked Language Modeling (MLM) and Image-Text Matching (ITM).

Most existing VL models can be unified into the TWO-TOWER architecture, which consists of a visual encoder, a textual encoder, and a cross-modal encoder. The models differ in the design of the three encoders. Benefiting from the rapid progress and prominent performance of Vision Transformer (ViT) (Dosovitskiy et al. 2021) on various vision tasks, recent VL models can adopt ViT as a cross-modal or visual encoder without using region features from heavy and time-consuming pre-trained object detectors.

ViLT (Kim, Son, and Kim 2021) adopts linear projection and word embedding as lightweight uni-modal encoders, and uses ViT as the cross-modal encoder to extract, align and fuse the features of both modalities simultaneously. While parameter-efficient, it may be difficult for ViLT to learn intra- and cross-modal interactions concurrently, and thus its performance lags behind state-of-the-art performance on downstream VL tasks. METER (Dou et al. 2022) uses ViT and RoBERTa (Liu et al. 2019b) as pre-trained uni-modal encoders and feeds the last-layer uni-modal representations directly into the top cross-modal encoder. Although METER achieves performance competitive with the previous region-based state-of-the-art model VinVL (Zhang et al. 2021), it ignores and wastes different levels of semantic knowledge contained in different layers of pre-trained uni-modal encoders. Furthermore, the abstract representations from the last layer of pre-trained uni-modal encoders could be challenging for the top cross-modal encoder to learn cross-modal alignment and fusion (Lu et al. 2019; Tan and Bansal 2019).

It has been demonstrated that different layers encode different types of information in both vision (Zeiler and Fergus 2014; Dosovitskiy et al. 2021; Du et al. 2020; Raghu et al. 2021; Naseer et al. 2021) and language models (Peters et al. 2018b; Liu et al. 2019a; Jawahar, Sagot, and Seddah 2019). Dosovitskiy et al. (2021) and Raghu et al. (2021) find that lower layers of ViT attend both locally and globally, while higher layers primarily incorporate global information. Jawahar, Sagot, and Seddah (2019) find that the intermediate layers of BERT (Devlin et al. 2019) encode a rich hierarchy of linguistic information, starting with surface features at the bottom, syntactic features in the middle, and then semantic features at the top. Therefore, it makes perfect sense to utilize

[*] Contribution during internship at Microsoft. [†] Contact Person

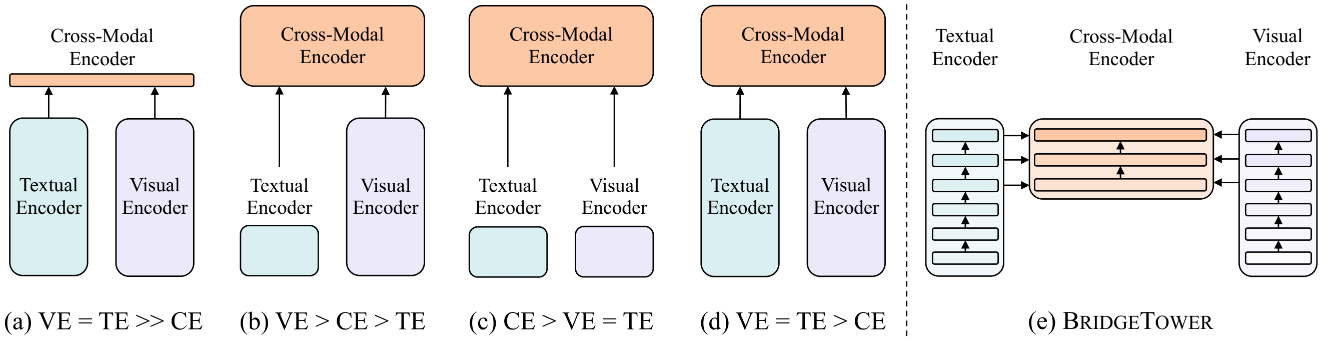


Figure 1: (a) – (d) are four categories of current TWO-TOWER vision-language models; (e) gives a brief illustration of the BRIDGETOWER architecture. VE, TE, and CE are short for the Visual Encoder, Textual Encoder, and Cross-modal Encoder, respectively. The height of each rectangle represents its relative computational cost. VE = TE indicates that the visual encoder and the textual encoder have the same or a similar number of parameters or computational costs. Illustration inspired by ViLT.

multi-layer uni-modal features to obtain effective improvements in both vision (Lin et al. 2017; Huang et al. 2017; Yu et al. 2018; Zheng et al. 2021; Xie et al. 2021) and language tasks (Peters et al. 2018a; Wang et al. 2018; Shen et al. 2018; Dou et al. 2018; Sun et al. 2019). The question, therefore, naturally arises: *can we build a bridge between different layers of pre-trained uni-modal encoders and the cross-modal encoder to utilize multi-layer uni-modal features?*

We propose BRIDGETOWER, a novel transformer-based VL model that takes full advantage of the features of different layers in pre-trained uni-modal encoders. By introducing multiple bridge layers, the top layers of uni-modal encoders can be connected with each layer of the cross-modal encoder. This enables effective bottom-up cross-modal alignment and fusion between visual and textual representations of different semantic levels of pre-trained uni-modal encoders in the cross-modal encoder. Moreover, in principle, the proposed BRIDGETOWER architecture is applicable to any visual, textual or cross-modal encoder.

We conduct extensive experiments on different design choices for BRIDGETOWER and fine-tune it on various downstream VL tasks. Experimental results show that with only 4M images for pre-training, our model achieves state-of-the-art performance on various downstream VL tasks, especially 78.73% accuracy on the VQAv2 test-std set, outperforming the previous state-of-the-art model METER by 1.09% with the same pre-training data and almost negligible additional parameters and computational costs. When further scaling the model, BRIDGETOWER achieves 81.15% accuracy on the VQAv2 test-std set, outperforming models that are pre-trained on orders-of-magnitude larger datasets.

Our contributions are threefold:

- We introduce BRIDGETOWER, a novel transformer-based VL model that achieves substantial improvements over previous state-of-the-art model METER both with and without pre-training.
- We propose using multiple bridge layers to connect the top layers of uni-modal encoders with each layer of the cross-modal encoder. Furthermore, we conduct extensive experiments on different design choices for BRIDGETOWER.
- We demonstrate the effectiveness of BRIDGETOWER on various VL downstream tasks, including visual question answering (VQAv2), visual entailment (SNLI-VE), and image-text retrieval (Flickr30K) tasks.

2 Related Work

2.1 TWO-TOWER Vision-Language Models

Following the taxonomy proposed by ViLT (Kim, Son, and Kim 2021), most VL models can be unified into the TWO-TOWER architecture shown in Figure 1(a) – (d). They feed last-layer representations of pre-trained uni-modal encoders into the top cross-modal encoder and can be differentiated by the depth of the textual, visual, and cross-modal encoders[1].

CLIP (Radford et al. 2021) and ALIGN (Jia et al. 2021) are representative models that directly perform a shallow fusion (*e.g.*, dot product) of last-layer representations of equally expressive pre-trained uni-modal encoders in the cross-modal encoder, as illustrated in Figure 1(a). The remaining models perform deep fusion in the multi-layer transformer-based cross-modal encoder but choose pre-trained uni-modal encoders with varying levels of expressiveness. Numerous works (Li et al. 2019; Su et al. 2020; Li et al. 2020a; Chen et al. 2020; Li et al. 2020b; Zhou et al. 2020; Zhang et al. 2021; Cho et al. 2021; Huang et al. 2020, 2021; Shen et al. 2021; Liu et al. 2022; Li et al. 2021b; Xia et al. 2021; Ni et al. 2021; Chen et al. 2022; Wang et al. 2022a; Alayrac et al. 2022) fall in the category of Figure 1(b) as they adopt various types of deep vision models (*e.g.*, Faster R-CNN (Ren et al. 2015), ResNet (He et al. 2016) or ViT (Dosovitskiy et al. 2021)) as their visual encoder to obtain region, grid, or patch features, and concatenate them with word embedding to feed into their top cross-modal encoder. The third category of models (Kim, Son, and Kim 2021; Wang et al. 2021c,b, 2022b), illustrated in Figure 1(c), utilizes lightweight visual and lightweight textual encoders and handles both modalities in a single transformer-based cross-modal encoder. In contrast, models (Lu et al. 2019; Tan and Bansal 2019; Kamath

[1] A cross-modal decoder can be placed on top of the cross-modal encoder, which is omitted since it is not the main study of this paper.

et al. 2021; Li et al. 2021a; Zeng, Zhang, and Li 2021; Dou et al. 2022; Wang et al. 2021a; Li et al. 2022b,c; Wang et al. 2022c; Yu et al. 2022; Li et al. 2022a), which belong to Figure 1(d) category, use expressive deep pre-trained uni-modal encoders and feed their last-layer representation into the top multi-layer cross-modal encoder.

Regardless of the visual, textual, or cross-modal encoders they utilize, most current models ignore the various levels of semantic information at the different layers of pre-trained uni-modal encoders, and simply utilize the last-layer uni-modal representations for cross-modal alignment and fusion. While the models belonging to Figure 1(c) appear to retain the possibility of utilizing different levels of uni-modal semantic information, it could be challenging for them to learn intra- and cross-modal interactions concurrently without modality-specific parameters. Their unconstrained cross-modal interaction could impede intra-modal interaction (Dou et al. 2022; Du et al. 2022). This may be the reason why the performance of ViLT lags behind models in the Figure 1(d) category, and why SimVLM (Wang et al. 2021c) and OFA (Wang et al. 2022b) need to use significantly more data to obtain competitive performance compared with METER.

Unlike current models, BRIDGETOWER, as shown in Figure 1(e), proposes using multiple bridge layers to connect the top layers of uni-modal encoders with each layer of the cross-modal encoder. This does not affect intra-modal interaction in the pre-trained uni-modal encoders, and enables different semantic levels of visual and textual representations to interact thoroughly and mildly in the bottom-up directions at each layer of the cross-modal encoder.

2.2 Multi-Layer Feature Utilization

Multi-layer feature utilization has been demonstrated to be an effective method of making full use of the information contained in different layers of neural networks to improve the representation and generalization capabilities of computer vision (Ronneberger, Fischer, and Brox 2015; Liu et al. 2016; Lin et al. 2017; Huang et al. 2017; Yu et al. 2018; Kirillov et al. 2019; Zheng et al. 2021; Xie et al. 2021; Naseer et al. 2021), natural language processing (Peters et al. 2018a; Wang et al. 2018; Shen et al. 2018; Dou et al. 2018; Jawahar, Sagot, and Seddah 2019; Sun et al. 2019; Dou et al. 2019) and multi-modal models (Dou et al. 2022; Nagrani et al. 2021).

Since Zeiler and Fergus (2014) introduce a visualization technique and find that different patterns are learned in different layers of CNN models, then many researchers exploit features of different layers to improve detection and semantic segmentation. U-Net (Ronneberger, Fischer, and Brox 2015) and FPN (Lin et al. 2017) propose to adopt lateral connections for associating feature maps from different layers across resolutions and semantic levels. The same idea is also applicable to ViT-based models. SETR (Zheng et al. 2021) and SegFormer (Xie et al. 2021) aggregate features from different layers to improve semantic segmentation performance. In natural language processing, researchers (Søgaard and Goldberg 2016; Hashimoto et al. 2017; Belinkov et al. 2017; Peters et al. 2018b; Jawahar, Sagot, and Seddah 2019; Liu et al. 2019a) find that Recurrent Neural Networks (RNN) (Hochreiter and Schmidhuber 1997) and BERT (Devlin et al. 2019) encode different types of semantic information in different layers. Hence, Peters et al. (2018a) and Sun et al. (2019) use the concatenation or weighted sum of representations from different layers of RNN or BERT as input for different task heads. Dou et al. (2018) explore layer aggregation with multi-layer attention mechanisms.

Recent multi-modal models exploit features from different layers. MBT (Nagrani et al. 2021) introduces simple bottleneck tokens at multiple layers to jointly model intra- and restricted cross-modal correlations. While MBT achieves good performance on audio-visual benchmarks, learning complex vision-language alignment and fusion via a limited number of bottleneck tokens instead of a cross-modal encoder maybe too difficult, which limits cross-modal alignment. METER feeds the weighted sum of representations from each layer of the bottom uni-modal encoder into the top cross-modal encoder; they find this can improve performance by a small margin without VLP but can degrade performance with VLP.

In a departure from existing models, BRIDGETOWER considers detailed interactions between the top layers of uni-modal encoders and each layer of the cross-modal encoder. It is intuitive to connect pre-trained uni-modal encoders and the cross-modal encoder via multiple bridge layers, in order to achieve comprehensive cross-modal alignment and fusion of the uni-modal representations of different semantic levels. Most importantly, unlike the simple multi-layer feature fusion method in METER, BRIDGETOWER can significantly improve performance both with and without vision-language pre-training on large-scale image-text data.

3 Approach

As shown in Figure 2, BRIDGETOWER consists of a visual encoder, a textual encoder and a cross-modal encoder with multiple lightweight bridge layers. Our goal is to build a bridge between each uni-modal encoder and the cross-modal encoder to enable comprehensive and detailed interaction at each layer of the cross-modal encoder. Our goal is not to develop new encoders; in principle, one can apply any visual, textual, or cross-modal encoder in the proposed architecture.

3.1 Visual Encoder

Recent works (Shen et al. 2021; Dou et al. 2022) show that CLIP's visual encoder has strong capabilities on VL tasks. We follow METER to adopt CLIP-ViT-B/16 as the pre-trained visual encoder. For each input 2D image $\mathbf{I} \in \mathbb{R}^{H \times W \times C}$, where (H, W) is the resolution of the input image and C is the number of channels, ViT reshape it into a sequence of flattened 2D patches $\mathbf{P} \in \mathbb{R}^{N \times (P^2 C)}$, where (P, P) is the image patch resolution and $N = \frac{HW}{P^2}$ is the number of patches. Similar to BERT, ViT also prepends the `[class]` token to the patch sequence and uses learnable 1D position embeddings $\mathbf{V}^{pos} \in \mathbb{R}^{(N+1) \times D_v}$, where D_v is the dimension of the visual encoder. The input visual representation can be calculated as follows:

$$\mathbf{V}_0 = [\mathbf{E}_{\texttt{[class]}}; \mathbf{p}_1 \mathbf{W_p}; \dots; \mathbf{p}_N \mathbf{W_p}] + \mathbf{V}^{pos}, \quad (1)$$

where $\mathbf{W_p} \in \mathbb{R}^{(P^2 C) \times D_v}$ is the trainable linear projection layer and $\mathbf{V}_0 \in \mathbb{R}^{(N+1) \times D_v}$. Each layer of ViT consists of

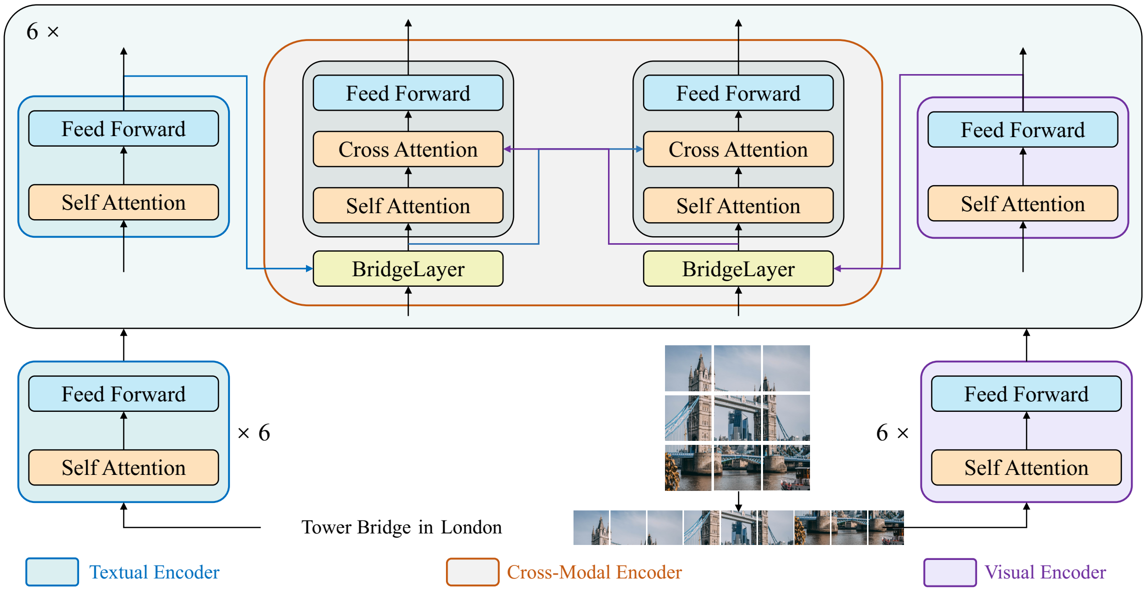


Figure 2: Illustration of BRIDGETOWER. BRIDGETOWER consists of a 12-layer textual encoder, a 12-layer visual encoder, and a 6-layer cross-modal encoder. Each of the top 6 layers of the visual and textual encoders is connected to each layer of the cross-modal encoder via bridge layers, which brings bottom-up alignment and fusion to the cross-modal encoder.

a multi-head self-attention (MSA) block and a feed-forward network (FFN) block. We omit the computation details and simplify them as $\mathrm{Encoder}^{\mathrm{V}}$. The ℓ-th layer representation can be denoted as $\mathbf{V}_\ell = \mathrm{Encoder}^{\mathrm{V}}_\ell(\mathbf{V}_{\ell-1}), \ell = 1 \dots L_V$, where L_V is the number of layers of the visual encoder.

3.2 Textual Encoder

Since RoBERTa achieves robust performance on a wide range of NLP tasks, we adopt RoBERTa$_{\mathrm{BASE}}$ as our textual encoder. Each input sequence $\mathbf{w}$ is tokenized by the byte-level Byte-Pair Encoding (BPE) (Sennrich, Haddow, and Birch 2016; Radford et al. 2019). `[<s>]` token and `[</s>]` token are added to the sequence as the start and end tokens, respectively. The input textual representation can be represented as:

$$\mathbf{T}_0 = [\mathbf{E}_{\texttt{[<s>]}}; \mathbf{E}_{w_1}; \dots; \mathbf{E}_{w_M}; \mathbf{E}_{\texttt{[</s>]}}] + \mathbf{T}^{pos}, \quad (2)$$

where $\mathbf{T}_0 \in \mathbb{R}^{(M+2)\times D_t}$, $\mathbf{E}$ is the word embedding matrix, M is the number of tokens, D_t is the dimension of the textual encoder, and $\mathbf{T}^{pos}$ is the positional embedding matrix. Similarly, we denote the ℓ-th layer of the textual encoder as $\mathrm{Encoder}^{\mathrm{T}}_\ell$, and the ℓ-th layer representation can be denoted as $\mathbf{T}_\ell = \mathrm{Encoder}^{\mathrm{T}}_\ell(\mathbf{T}_{\ell-1}), \ell = 1 \dots L_T$, where L_T is the number of layers of the textual encoder.

3.3 Cross-Modal Encoder with Bridge Layers

Hendricks et al. (2021) perform analysis on different types of attention mechanisms used in the existing transformer-based cross-modal encoders and demonstrate that the *co-attention* mechanism (Lu et al. 2019) performs best. This mechanism uses a different set of parameters for each modality. For example, for the visual part of the cross-modal encoder, the queries of each MSA block are from the visual modality. However, the keys and values are from the other modality (*i.e.*, the textual modality). We, therefore, adopt the *co-attention* mechanism. Formally, we define the ℓ-th layer of the cross-modal encoder as $\mathrm{Encoder}^{\mathrm{Z}}_\ell$, which consists of a visual part and a textual part. Each part consists of an MSA block, a multi-head cross-attention (MCA) block, and an FFN block. For brevity, the interactions at each layer are defined as:

$$\tilde{\mathbf{Z}}^{\mathrm{V}}_\ell = \mathbf{Z}^{\mathrm{V}}_{\ell-1}, \quad (3)$$

$$\tilde{\mathbf{Z}}^{\mathrm{T}}_\ell = \mathbf{Z}^{\mathrm{T}}_{\ell-1}, \quad (4)$$

$$\mathbf{Z}^{\mathrm{V}}_\ell, \mathbf{Z}^{\mathrm{T}}_\ell = \mathrm{Encoder}^{\mathrm{Z}}_\ell(\tilde{\mathbf{Z}}^{\mathrm{V}}_\ell, \tilde{\mathbf{Z}}^{\mathrm{T}}_\ell), \ell = 1 \dots L_Z, \quad (5)$$

where $\mathbf{Z}^{\{\mathrm{V,T}\}}_\ell$ is the output representation of the visual or textual part at the ℓ-th layer, $\tilde{\mathbf{Z}}^{\{\mathrm{V,T}\}}_\ell$ is the input of each part, and L_Z is the number of layers of the cross-modal encoder.

Generally, current VL models, such as METER, directly use the output representation of the previous layer as the input to $\mathrm{Encoder}^{\mathrm{Z}}_\ell$ (Equation 3 & 4). $\mathbf{Z}^{\mathrm{V}}_0, \mathbf{Z}^{\mathrm{T}}_0$ are initialized with the last-layer representations from pre-trained uni-modal encoders: $\mathbf{Z}^{\mathrm{V}}_0 = \mathbf{V}_{L_V}\mathbf{W}_V + \mathbf{V}^{type}, \mathbf{Z}^{\mathrm{T}}_0 = \mathbf{T}_{L_T}\mathbf{W}_T + \mathbf{T}^{type}$, where $\mathbf{W}_V \in \mathbb{R}^{D_V \times D_Z}$ and $\mathbf{W}_T \in \mathbb{R}^{D_T \times D_Z}$ are linear projections, $\mathbf{V}^{type}$ and $\mathbf{T}^{type}$ are the modality type embeddings.

However, in this paper, we propose using multiple bridge layers to connect the top layers of uni-modal encoders with each layer of the cross-modal encoder:

$$\tilde{\mathbf{Z}}^{\mathrm{V}}_\ell = \mathrm{BridgeLayer}^{\mathrm{V}}_\ell(\mathbf{Z}^{\mathrm{V}}_{\ell-1}, \mathbf{V}_k\mathbf{W}_V + \mathbf{V}^{type}), \quad (6)$$

$$\tilde{\mathbf{Z}}^{\mathrm{T}}_\ell = \mathrm{BridgeLayer}^{\mathrm{T}}_\ell(\mathbf{Z}^{\mathrm{T}}_{\ell-1}, \mathbf{T}_k\mathbf{W}_T + \mathbf{T}^{type}), \quad (7)$$

where k denotes the index of layer representations of uni-modal encoders. In this paper, $L_V = L_T = 12, L_Z = 6$ and we use the representations of the top 6 layers of uni-modal encoders, which means that $k = 7, \dots, 12$. Take the input of $\text{Encoder}_2^{\text{Z}}$ as an example:

$$\tilde{\mathbf{Z}}_2^{\text{V}} = \text{BridgeLayer}_2^{\text{V}}(\mathbf{Z}_1^{\text{V}}, \mathbf{V}_8\mathbf{W}_V + \mathbf{V}^{type}), \quad (8)$$

$$\tilde{\mathbf{Z}}_2^{\text{T}} = \text{BridgeLayer}_2^{\text{T}}(\mathbf{Z}_1^{\text{T}}, \mathbf{T}_8\mathbf{W}_T + \mathbf{T}^{type}). \quad (9)$$

Utilizing our proposed bridge layers, top-layer uni-modal representations can be bridged with each layer of the cross-modal encoder, thus incorporating different semantic levels of uni-modal representations into cross-modal interaction. In the spirit of ResNet and Transformer (Vaswani et al. 2017), the simplest formal definition of a bridge layer is:

$$\text{BridgeLayer}(x, y) = \text{LayerNorm}(x + y). \quad (10)$$

In Sec. 4.2, we describe the extensive experiments we conducted on different design choices for BRIDGETOWER, including the formal definition of bridge layers and the number of cross-modal layers.

3.4 Pre-training Objectives

We pre-train BRIDGETOWER with two commonly used vision-language pre-training objectives: MLM and ITM.

Masked Language Modeling. MLM is a common objective for language and vision-language pre-training. Given an image-text pair, following UNITER (Chen et al. 2020), we use conditional masking for MLM, which means we randomly mask 15% of tokens in the token sequence while keeping the input image patch sequence untainted. The model is trained to reconstruct the original tokens conditioned on incomplete input token sequence and complete observed image patch sequence. We adopt the same masking strategy and MLM task head as RoBERTa. The last-layer representation of the textual part of the cross-modal encoder is used as input to the MLM task head.

Image-Text Matching. ITM aims to predict whether the given image-text pair is positive (matched) or negative (mismatched). Matched and mismatched image-text pairs are fed into our model with the same probability. We pass the final representations of `[class]` and `[<s>]` token in the cross-modal encoder to the non-linear layer activated by `Tanh`, respectively. The concatenation of the outputs is fed into a linear classifier with cross-entropy loss for binary classification.

3.5 Fine-Tuning on Downstream Tasks

For visual question answering and visual entailment, we use the same strategy as for ITM. For image-text retrieval, following ALBEF (Li et al. 2021a), our model is jointly optimized with image-text contrastive (ITC) loss and ITM loss. Two linear projections are added on top of both uni-modal encoders to obtain uni-modal representations of image-text pairs, and then compute their contrastive similarity by dot product. Then, instead of randomly sampling negatives for the ITM task, for each image (text) in a mini-batch, we use the contrastive similarity distribution from the ITC task to sample one hard in-batch negative text (image). In inference, we first compute the contrastive similarity for all images and texts, and then take the top-k candidates and calculate their ITM scores for ranking.

4 Experiment

4.1 Implementation Details

BRIDGETOWER consists of a pre-trained textual encoder, RoBERTa$_{\text{BASE}}$ with 124M parameters, a pre-trained visual encoder, CLIP-ViT-B-224/16 with 86M parameters, and a random-initialized 6-layer cross-modal encoder with 113M parameters. For each layer of the cross-modal encoder, the hidden size is set to 768, the intermediate size of feed-forward networks is set to $3,072$, and the number of heads is set to 12. The maximum length of the text sequence is set to 50. The patch size is 16×16. Center-crop is used to resize each input image to the same resolution, and we also apply RandAugment (Cubuk et al. 2020) to the input images following previous works (Li et al. 2021a; Dou et al. 2022). We use the AdamW (Loshchilov and Hutter 2019) optimizer with a base learning rate of $2e^{-5}$ and weight decay of 0.01. The learning rate is warmed up for 10% of the total training steps and then decayed linearly to zero for the rest of the training steps. Following METER, the learning rate of the cross-modal encoder is five times higher than that of uni-modal encoders.

We evaluate BRIDGETOWER by fine-tuning the entire model on the visual question answering (VQAv2) (Goyal et al. 2017), visual entailment (SNLI-VE) (Xie et al. 2019), and image-text retrieval (Flickr30K) (Young et al. 2014) tasks. We use an image resolution of 384×384 for these downstream VL tasks, except for VQAv2, where we use 576×576 for a robust evaluation and fair comparison with METER. Standard settings and splits are used for all datasets. For VQAv2, where we follow the common practice (Goyal et al. 2017; Teney et al. 2018): convert VQAv2 to a classification task with $3,129$ answer classes; train the model with training data and validation data, and evaluate the model on the test-dev data.

4.2 Investigation and Analysis

In this section, we evaluate different design choices for BRIDGETOWER on the VQAv2 and Flickr30K datasets. Each model is initialized with CLIP-ViT-B-224/16 and RoBERTa$_{\text{BASE}}$ pre-trained weights, and then directly fine-tuned on the two downstream tasks without VLP. All experimental settings are the same as METER for fair comparisons. In our preliminary experiments, the uni-modal representations of the top layers perform much better than the middle and bottom layers. Thus, we use the top 6 layer representations of the uni-modal encoders as the corresponding inputs for each bridge layer in the bottom-up directions.

Design Choice I: Formal Definition of Bridge Layers Table 1 shows, perhaps unexpectedly but not very surprisingly, that row (a) provides the best results using the minimum number of parameters and achieves an accuracy of 75.18% on the test-dev set of VQAv2 and RSUM of 533.84 on the test set of Flickr30K. The additional parameters used for interpolation cause slight performance degradation in rows (c)&(d). Rows (e) – (h) try to incorporate generally used feature transformation forms into the bridge layer, but the additional computation and parameters instead lead to performance degradation. Inspired by ResNet and ViTDet (Li et al. 2022d), in row (i), we incorporate row (a) into row (e) as the residual connection. $\mathbf{W}_*$ is initialized as zero so that row (i) is

BridgeLayer(x, y)	# Params	Test-Dev	RSUM
(a) $x + y$	**18.4K**	**75.18**	**533.8**
(b) $x \odot y$	18.4K	73.41	530.4
(c) $\alpha x + (1 - \alpha)\, y, \alpha \in \mathbb{R}^{D_Z}$	26.0K	75.09	532.9
(d) $\alpha x + (1 - \alpha)\, y, \alpha = \sigma(\mathbf{W}\, [x; y])$	11.8M	75.13	533.1
(e) $\mathbf{W}\, [x; y]$	11.8M	74.55	532.2
(f) $\mathbf{W}_2\, (\mathrm{GeLU}\, (\mathbf{W}_1[x; y]))$	35.4M	74.26	530.2
(g) $\mathrm{MCA}\, (x, y)$	23.6M	73.67	514.3
(h) $\mathrm{FFN}\, (\mathrm{MCA}\, (x, y))$	70.8M	73.54	511.1
(i) $x + y + \mathbf{W}_*\, [x; y]$	11.8M	75.10	533.1

Table 1: Performance and number of parameters for different formal definitions of bridge layers. We omit the layer normalization used in each form. x denotes the output cross-modal representation of the previous layer and y denotes the corresponding input uni-modal representation. RSUM indicates the sum of recall metrics for image-text retrieval.

initially equivalent to row (a). Although the performance is significantly higher than row (e) (74.55 → 75.10), there is no significant gain compared with row (a). Hence, we choose row (a) (`Add&Norm`) as the default bridge layer.

Design Choice II: Number of Cross-Modal Layers In BRIDGETOWER, the cross-modal encoder is not located on top of the uni-modal encoders, but between them. Each cross-modal layer is connected to the corresponding layer of the uni-modal encoder by a bridge layer. Therefore, based on the two 12-layer uni-modal encoders we used, the number of cross-modal layers can be $[1, 12]$. Table 2 shows the performance of BRIDGETOWER with different numbers of cross-modal layers. It is illuminating to note that adding more cross-modal layers does not constantly improve performance, possibly because (i) more cross-modal layers are more difficult to train and are more data-hungry; (ii) uni-modal representations of top layers are beneficial to cross-modal alignment and fusion, while uni-modal representations of bottom layers may be less useful and even detrimental. We also evaluate METER and find that while the only difference between the two models is the bridge layers, BRIDGETOWER can achieve consistent performance gains for different numbers of cross-modal layers. It further illustrates that the bridge layers can facilitate effective cross-modal alignment and fusion with uni-modal representations of different semantic levels in the cross-modal encoder.

Apply Different Visual and Textual Backbones We apply different visual and textual backbones as pre-trained uni-modal encoders and directly fine-tune on downstream tasks to further investigate the impact brought by bridge layers. As shown in Table 3, no matter what visual and textual encoders we apply, the performances of BRIDGETOWER are consistently and significantly better than that of METER. This further demonstrates the effectiveness of our proposed BRIDGETOWER architecture and bridge layers for vision-language representation learning.

4.3 Comparison with Previous Arts

In this section, we describe how to pre-train BRIDGETOWER with the best-performing setting (Sec. 4.2) and compare its fine-tuning performance with previous works.

L_Z	# Params	VQAv2 Test-Dev		Flickr30K RSUM	
		METER	Ours	METER	Ours
2	37.8M	72.84	74.12 (↑ 1.28)	526.0	527.1 (↑ 1.1)
3	56.8M	73.47	74.36 (↑ 0.89)	526.5	528.6 (↑ 2.1)
4	75.6M	73.71	75.00 (↑ 1.29)	527.9	529.7 (↑ 1.8)
5	94.6M	73.80	74.98 (↑ 1.18)	528.8	531.8 (↑ 3.0)
6	113.4M	74.04	**75.18** (↑ 1.14)	530.7	**533.8** (↑ 3.1)
8	151.2M	73.97	75.07 (↑ 1.10)	530.0	531.6 (↑ 1.6)
10	189.0M	73.45	75.06 (↑ 1.61)	529.6	531.7 (↑ 2.1)
12	226.8M	71.88	74.94 (↑ 3.06)	528.7	531.4 (↑ 2.7)

Table 2: Performance of METER and BRIDGETOWER with different number of cross-modal layers. # Params denotes the number of parameters of the cross-modal encoder.

Visual Backbone	Textual Backbone	VQAv2 Test-Dev		Flickr30K RSUM	
		METER	Ours	METER	Ours
DeiT B-224/16	RoBERTa	69.98	70.83 (↑ 0.85)	448.0	455.7 (↑ 7.7)
ViT B-224/16	RoBERTa	70.26	72.24 (↑ 1.98)	472.7	476.9 (↑ 4.2)
ViT B-384/16	RoBERTa	70.52	72.38 (↑ 1.86)	472.8	477.1 (↑ 4.3)
CLIP-VIT-B/32	RoBERTa	72.19	72.91 (↑ 0.72)	508.8	512.0 (↑ 3.2)
CLIP-VIT-B/16	BERT	74.09	74.89 (↑ 0.80)	522.1	526.5 (↑ 4.4)
CLIP-VIT-B/16	RoBERTa	74.04	**75.18** (↑ 1.14)	530.7	**533.8** (↑ 3.1)

Table 3: Performance of METER and BRIDGETOWER with different visual and textual encoders. The image resolution of all CLIP visual backbones is 224×224.

Pre-training Setup. We use four public image-caption datasets for pre-training: Conceptual Captions (CC) (Sharma et al. 2018), SBU Captions (Ordonez, Kulkarni, and Berg 2011), MSCOCO Captions (Chen et al. 2015), and Visual Genome (VG) (Krishna et al. 2017). The total number of unique images in the combined data is 4M. We pre-train BRIDGETOWER for 100k steps on 8 NVIDIA A100 GPUs with a batch size of $4,096$. All the pre-training settings for BRIDGETOWER are the same as for METER for a fair comparison. The learning rate is set to $1e^{-5}$. No data augmentation is used except for center-crop (Radford et al. 2021; Dou et al. 2022). The image resolution in pre-training is set to 288×288. Other hyperparameters remain unchanged based on the experiments in Sec. 4.2.

Main Results. Table 4 and 5 show the performance of BRIDGETOWER compared with previous works on downstream VL tasks. With only 4M images for pre-training, BRIDGETOWER$_{\text{BASE}}$ achieves state-of-the-art performance, in particular 78.73% accuracy on the VQAv2 test-std set, outperforming the previous state-of-the-art model METER by 1.09% with the same pre-training setting and almost negligible additional parameters and computational costs. Remarkably, BRIDGETOWER$_{\text{BASE}}$ not only outperforms all base-size models that use the same or a larger number of pre-trained images, but it even outperforms some large-size models. A similar trend also occurs on the visual entailment and image-text retrieval tasks. On the Flickr30K dataset, BRIDGETOWER$_{\text{BASE}}$ achieves the best performance, surpassing not only ALBEF with its specially designed pre-training objective, but also ALIGN with 1.8B pre-train images.

Model	# Pre-train	Visual	Test-Dev	Test-Standard			
	Images	Backbone	Overall	Yes/No	Number	Other	Overall
Base-Size Models							
ViLT$_{\text{BASE}}$ (Kim, Son, and Kim 2021)	4M	ViT-B-384/32	71.26	-	-	-	-
UNITER$_{\text{BASE}}$ (Chen et al. 2020) $*$	4M	Faster R-CNN	72.70	-	-	-	72.91
VILLA$_{\text{BASE}}$ (Gan et al. 2020) $*$	4M	Faster R-CNN	73.59	-	-	-	73.67
UNIMO$_{\text{BASE}}$ (Li et al. 2021b)	4M	Faster R-CNN	73.79	-	-	-	74.02
ALBEF$_{\text{BASE}}$ (Li et al. 2021a) $*$	4M	DeiT-B-224/16	74.54	-	-	-	74.70
ALBEF$_{\text{BASE}}$ (Li et al. 2021a) $*$	14M	DeiT-B-224/16	75.84	-	-	-	76.04
VinVL$_{\text{BASE}}$ (Zhang et al. 2021)	5.7M	ResNeXt-152	75.95	-	-	-	76.12
VLMo$_{\text{BASE}}$ (Wang et al. 2021a)	4M	BEiT-B-224/16	76.64	-	-	-	76.89
BLIP$_{\text{BASE}}$ (Li et al. 2022b) $*$	14M	DeiT-B-224/16	77.54	-	-	-	77.62
Meter$_{\text{BASE}}$ (Dou et al. 2022)	4M	CLIP-ViT-B-224/16	77.68	92.49	58.07	69.20	77.64
mPLUG (Li et al. 2022a) $*$	4M	CLIP-ViT-B-224/16	77.94	-	-	-	77.96
OFA$_{\text{BASE}}$ (Wang et al. 2022b) $*\star$	54M	ResNet-101	77.98	-	-	-	78.07
SimVLM$_{\text{BASE}}$ (Wang et al. 2021c) $\star$	1.8B	ResNet-101	77.87	-	-	-	78.14
BLIP$_{\text{BASE}}$ (Li et al. 2022b) $*$	129M	DeiT-B-224/16	78.24	-	-	-	78.17
BridgeTower$_{\text{BASE}}$ (**Ours**)	**4M**	CLIP-ViT-B-224/16	**78.66**	**92.92**	**60.69**	**70.51**	**78.73**
BridgeTower$_{\text{BASE}}$ (**Ours**) $*$	**4M**	CLIP-ViT-B-224/16	**79.10**	**93.06**	**62.19**	**70.69**	**79.04**
Large-Size Models							
UNITER$_{\text{LARGE}}$ (Chen et al. 2020) $*$	4M	Faster R-CNN	73.82	-	-	-	74.02
VILLA$_{\text{LARGE}}$ (Gan et al. 2020) $*$	4M	Faster R-CNN	74.69	-	-	-	74.87
UNIMO$_{\text{LARGE}}$ (Li et al. 2021b)	4M	Faster R-CNN	75.06	-	-	-	75.27
VinVL$_{\text{LARGE}}$ (Zhang et al. 2021)	5.7M	ResNeXt-152	76.52	92.04	61.50	66.68	76.63
SimVLM$_{\text{LARGE}}$ (Wang et al. 2021c)	1.8B	ResNet-152	79.32	-	-	-	79.56
VLMo$_{\text{LARGE}}$ (Wang et al. 2021a)	4M	BEiT-L-224/16	79.94	-	-	-	79.98
OFA$_{\text{LARGE}}$ (Wang et al. 2022b) $*\star$	54M	ResNet-152	80.43	93.32	**67.31**	72.71	80.67
BridgeTower$_{\text{LARGE}}$ (**Ours**)	**4M**	CLIP-ViT-L-224/14	**81.25**	**94.69**	64.58	**73.16**	**81.15**
BridgeTower$_{\text{LARGE}}$ (**Ours**) $*$	**4M**	CLIP-ViT-L-224/14	**81.52**	**94.80**	66.01	**73.45**	**81.49**
Huge or even Larger Size Models							
SimVLM$_{\text{HUGE}}$ (Wang et al. 2021c)	1.8B	Larger ResNet-152	80.03	93.29	66.54	72.23	80.34
Meter$_{\text{HUGE}}$ (Dou et al. 2022)	14M	Florence-CoSwin-H	80.33	94.25	64.37	72.30	80.54
mPLUG (Li et al. 2022a) $*$	14M	CLIP-ViT-L-224/14	81.27	-	-	-	81.26
GIT2 (Wang et al. 2022a) $*$	10.5B	DaViT(4.8B)	81.74	92.90	67.06	75.77	81.92
OFA$_{\text{HUGE}}$ (Wang et al. 2022b) $*\star$	54M	ResNet-152	82.0	94.66	71.44	73.35	81.98
Flamingo (Alayrac et al. 2022) $\star$	2.3B	NFNet-F6	82.0	-	-	-	82.1
CoCa (Yu et al. 2022) $\star$	4.8B	ViT-G-288/18	82.3	94.55	70.25	74.46	82.33
BEiT-3 (Wang et al. 2022c)	28M	BEiT-3	84.19	**96.43**	**73.63**	75.92	84.18
PaLI (Chen et al. 2022)	1.6B	ViT-E-224	**84.3**	96.13	69.07	**77.58**	**84.34**

Table 4: Comparisons with previous models on visual question answering (VQAv2). The best score is bolded. The models are divided into base size and large/huge size. B, N and M in ViT-B-N/M denote the model size, image resolution and patch size, respectively. $*$ indicates that the model also uses VG-QA data to fine-tune on VQAv2. $\star$ denotes the model is trained from scratch. "# Pre-train Images" denotes the number of images in VLP (the images for pre-trained visual and textual backbones are not counted).

Model	# Pre-train	SNLI-VE		Flickr30K (1K test set)						
	Images	dev	test	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10	RSUM
Pre-trained on More Data										
ALIGN$_{\text{BASE}}$ (Jia et al. 2021)	1.8B	-	-	84.9	97.4	98.6	95.3	99.8	100.0	576.0
ALBEF$_{\text{BASE}}$ (Li et al. 2021a)	14M	80.80	80.91	85.6	97.5	98.9	95.9	99.8	100.0	577.7
Pre-trained on CC, SBU, MSCOCO and VG datasets										
ViLT$_{\text{BASE}}$ (Kim, Son, and Kim 2021)	4M	-	-	64.4	88.7	93.8	83.5	96.7	98.6	525.7
UNITER$_{\text{LARGE}}$ (Chen et al. 2020)	4M	79.30	79.38	75.6	94.1	96.8	87.3	98.0	99.2	550.9
VILLA$_{\text{LARGE}}$ (Gan et al. 2020)	4M	80.18	80.02	76.3	94.2	96.8	87.9	97.5	98.8	551.5
UNIMO$_{\text{LARGE}}$ (Li et al. 2021b)	4M	**81.11**	80.63	78.0	94.2	97.1	89.4	98.9	99.8	557.5
ALBEF$_{\text{BASE}}$ (Li et al. 2021a)	4M	80.14	80.30	82.8	96.7	98.4	94.3	99.4	99.8	571.4
Meter-CLIP-ViT$_{\text{BASE}}$ (Dou et al. 2022)	4M	80.86	**81.19**	82.2	96.3	98.4	94.3	**99.6**	99.9	570.7
BridgeTower$_{\text{BASE}}$ (**Ours**)	**4M**	**81.11**	**81.19**	**85.8**	**97.6**	**98.9**	**94.7**	**99.6**	**100.0**	**576.6**

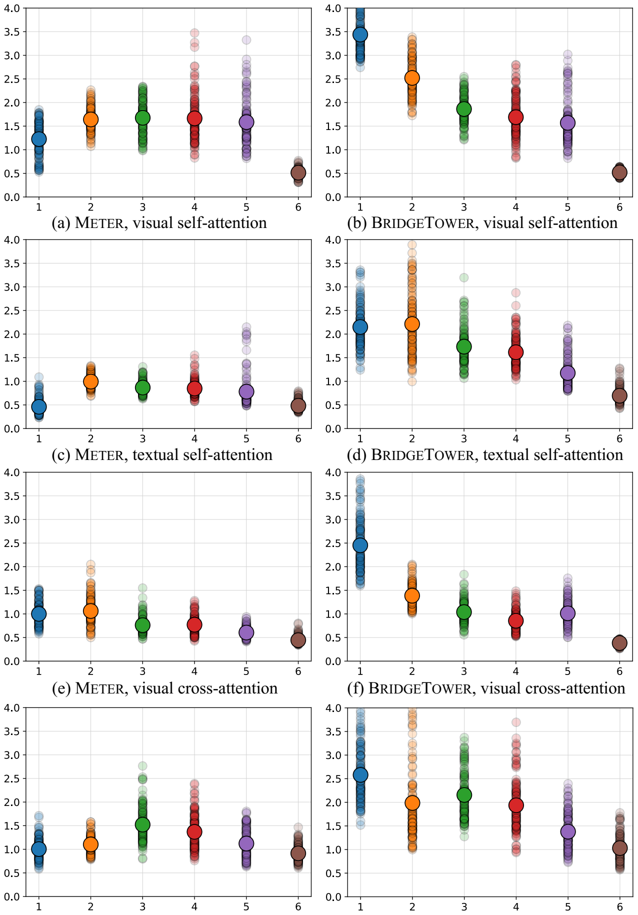
Table 5: Comparisons with previous models on visual entailment (SNLI-VE), image retrieval (IR) and text retrieval (TR) tasks (Flickr30K). The best score is bolded.

Scaling the Model. To investigate the effect of the scale of the model structure on performance, we replace our uni-modal encoders with the corresponding large version, *i.e.*, RoBERTa$_{\text{LARGE}}$ with 355M parameters and CLIP-ViT-L/14 with 304M parameters. For each layer of the cross-modal encoder, the hidden size is set to $1,024$, the intermediate size of feed-forward networks is set to $4,096$, and the number of heads is set to 16. The number of cross-modal encoder layers remains 6 so the number of parameters grows to 200M. The patch size is 14×14, then we set the image resolution to 294×294 in pre-training and to 574×574 in fine-tuning on the VQAv2. Other hyperparameters remain unchanged. As shown in Table 4, BridgeTower outperforms previous models trained with 10 times or even $1,000$ times more images, not only in the base size but also in the large size. Notably, BridgeTower$_{\text{LARGE}}$ achieves 81.15% accuracy on the VQAv2 test-std set, surpassing the previous state-of-the-art OFA$_{\text{LARGE}}$ model by 0.48%. This further demonstrates the effectiveness and scalability of BridgeTower. In addition, question-answer pairs from VG dataset are often used to extend the VQAv2 training data, thus further improving performance (Teney et al. 2018; Yu et al. 2019). Our performance of base and large size can be improved to 79.04% and 81.49% on the VQAv2 test-std set, respectively.

4.4 Visualization

Attention mechanism (Bahdanau, Cho, and Bengio 2015) is a critical and naturally interpretable component of transformer-based models. It is intuitive to analyze attention weights since it measures how much attention each token pays to the other tokens. Inspired by Xie et al. (2022), we compare the pre-trained Meter and BridgeTower models by analyzing the Kullback-Leibler (KL) divergence between attention weight distributions of different attention heads in each layer [2]. KL divergence can be seen as the diversity of attention heads. Higher/lower KL divergence means that different attention heads pay attention to different/similar tokens.

As shown in Figure 3, by comparing the KL divergence of the two models in each row, there are two distinct trends: (i) the diversity of attention heads becomes progressively smaller as the layer goes deeper for BridgeTower, but for Meter, the diversity of attention heads becomes progressively larger and then smaller as the layer goes deeper; (ii) the diversity of attention heads of each layer of BridgeTower is significantly larger than that of Meter, especially for the 1st to the 5th layer. Thus, for different attention heads of self-/cross-attention of the visual/textual part of the cross-modal encoder, compared with Meter, BridgeTower can aggregate more different tokens. We attribute this to our proposed bridge layer, which connects the top layers of uni-modal encoders with each cross-modal layer. Different semantic levels of visual and textual representations are introduced by bridge layers, facilitating more effective and informative cross-modal alignment and fusion at each cross-modal layer[3].



(a) Meter, visual self-attention (b) BridgeTower, visual self-attention

(c) Meter, textual self-attention (d) BridgeTower, textual self-attention

(e) Meter, visual cross-attention (f) BridgeTower, visual cross-attention

(g) Meter, textual cross-attention (h) BridgeTower, textual cross-attention

Figure 3: The KL divergence between attention distributions of different heads (small dots) and the averaged KL divergence (large dots) in each layer w.r.t. the layer number on the self-/cross-attention of the visual/textual part of the cross-modal encoder in the Meter and BridgeTower models.

5 Conclusion and Future Work

We present BridgeTower, a simple yet effective vision-language model that introduces multiple bridge layers to build a connection between the top layers of uni-modal encoders and each layer of the cross-modal encoder. This facilitates effective bottom-up cross-modal alignment and fusion between visual and textual representations of different semantic levels of the pre-trained uni-modal encoders in the cross-modal encoder. We experimentally prove the effectiveness of the proposed bridge layers and BridgeTower, which achieves remarkable performance in all downstream VL tasks with almost negligible additional parameters and computational costs. We hope that our work will draw more attention to the rich semantic knowledge latent in the different layers of uni-modal encoders. Incorporating such semantic knowledge into cross-modal alignment and fusion can yield more expressive and powerful vision-language representations. Furthermore, experiments with different visual and textual backbones as pre-trained uni-modal encoders demonstrate that the perfor-

[2] We also follow Xie et al. (2022) to analyze the averaged attention distance and entropy of attention weight distribution between the pre-trained two models, but no significant trends are found.

[3] Please check https://arxiv.org/abs/2206.08657 for more.

mances of our proposed BRIDGETOWER architecture are consistently and significantly better than that of METER.

In the future, we plan to improve BRIDGETOWER in the following aspects:

Different Pre-training Objectives. We followed METER to directly adopt the masked language modeling (MLM) and image-text matching (ITM) as pre-training objectives for a fair comparison. More pre-training objectives, such as image-text contrastive learning (ITC) and masked image modeling (MIM), could be incorporated to investigate their impact on BRIDGETOWER and further improve the performance.

Larger Scale Pre-training. We have pre-trained our model with 4M images both on the BASE and LARGE sizes. In both versions, BRIDGETOWER achieves lower accuracy on the "Number" type questions of VQAv2 than other models pre-trained with more data. We expect to investigate and further improve the performance of BRIDGETOWER after pre-training on larger-scale image-text data.

Generative Task. In this paper, we focus on discriminative tasks. It would be interesting to investigate the impact of the proposed bridge layer on the performance of a visual language generation task, such as image captioning.

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *ArXiv preprint*, abs/2204.14198.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proc. of ICLR*.

Belinkov, Y.; Durrani, N.; Dalvi, F.; Sajjad, H.; and Glass, J. 2017. What do Neural Machine Translation Models Learn about Morphology? In *Proc. of ACL*, 861–872.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft coco captions: Data collection and evaluation server. *ArXiv preprint*, abs/1504.00325.

Chen, X.; Wang, X.; Changpinyo, S.; Piergiovanni, A.; Padlewski, P.; Salz, D.; Goodman, S.; Grycner, A.; Mustafa, B.; Beyer, L.; et al. 2022. Pali: A jointly-scaled multilingual language-image model. *ArXiv preprint*, abs/2209.06794.

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. Uniter: Universal image-text representation learning. In *Proc. of ECCV*, 104–120.

Cho, J.; Lei, J.; Tan, H.; and Bansal, M. 2021. Unifying Vision-and-Language Tasks via Text Generation. In *Proc. of ICML*, volume 139, 1931–1942.

Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. 2020. RandAugment: Practical Automated Data Augmentation with a Reduced Search Space. In *Proc. of NeurIPS*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of NAACL*, 4171–4186.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proc. of ICLR*.

Dou, Z.; Tu, Z.; Wang, X.; Wang, L.; Shi, S.; and Zhang, T. 2019. Dynamic Layer Aggregation for Neural Machine Translation with Routing-by-Agreement. In *Proc. of AAAI*, 86–93.

Dou, Z.-Y.; Tu, Z.; Wang, X.; Shi, S.; and Zhang, T. 2018. Exploiting Deep Representations for Neural Machine Translation. In *Proc. of EMNLP*, 4253–4262.

Dou, Z.-Y.; Xu, Y.; Gan, Z.; Wang, J.; Wang, S.; Wang, L.; Zhu, C.; Zhang, P.; Yuan, L.; Peng, N.; Liu, Z.; and Zeng, M. 2022. An Empirical Study of Training End-to-End Vision-and-Language Transformers. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Du, C.; Wang, Y.; Wang, C.; Shi, C.; and Xiao, B. 2020. Selective feature connection mechanism: Concatenating multi-layer CNN features with a feature selector. *Pattern Recognition Letters*.

Du, Y.; Liu, Z.; Li, J.; and Zhao, W. X. 2022. A Survey of Vision-Language Pre-Trained Models. *ArXiv preprint*, abs/2202.10936.

Gan, Z.; Chen, Y.; Li, L.; Zhu, C.; Cheng, Y.; and Liu, J. 2020. Large-Scale Adversarial Training for Vision-and-Language Representation Learning. In *Proc. of NeurIPS*.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Proc. of CVPR*, 6325–6334.

Hashimoto, K.; Xiong, C.; Tsuruoka, Y.; and Socher, R. 2017. A Joint Many-Task Model: Growing a Neural Network for Multiple NLP Tasks. In *Proc. of EMNLP*, 1923–1933.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proc. of CVPR*, 770–778.

Hendricks, L. A.; Mellor, J.; Schneider, R.; Alayrac, J.-B.; and Nematzadeh, A. 2021. Decoupling the Role of Data, Attention, and Losses in Multimodal Transformers. *Transactions of the Association for Computational Linguistics*, 9: 570–585.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*.

Huang, G.; Liu, Z.; van der Maaten, L.; and Weinberger, K. Q. 2017. Densely Connected Convolutional Networks. In *Proc. of CVPR*, 2261–2269.

Huang, Z.; Zeng, Z.; Huang, Y.; Liu, B.; Fu, D.; and Fu, J. 2021. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Proc. of CVPR*, 12976–12985.

Huang, Z.; Zeng, Z.; Liu, B.; Fu, D.; and Fu, J. 2020. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *ArXiv preprint*, abs/2004.00849.

Jawahar, G.; Sagot, B.; and Seddah, D. 2019. What Does BERT Learn about the Structure of Language? In *Proc. of ACL*, 3651–3657.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.; Parekh, Z.; Pham, H.; Le, Q. V.; Sung, Y.; Li, Z.; and Duerig, T. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *Proc. of ICML*, volume 139, 4904–4916.

Kamath, A.; Singh, M.; LeCun, Y.; Synnaeve, G.; Misra, I.; and Carion, N. 2021. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proc. of ICCV*, 1780–1790.

Kim, W.; Son, B.; and Kim, I. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *Proc. of ICML*, volume 139, 5583–5594.

Kirillov, A.; Girshick, R. B.; He, K.; and Dollár, P. 2019. Panoptic Feature Pyramid Networks. In *Proc. of CVPR*, 6399–6408.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*.

Li, C.; Xu, H.; Tian, J.; Wang, W.; Yan, M.; Bi, B.; Ye, J.; Chen, H.; Xu, G.; Cao, Z.; et al. 2022a. mPLUG: Effective and Efficient Vision-Language Learning by Cross-modal Skip-connections. *ArXiv preprint*, abs/2205.12005.

Li, G.; Duan, N.; Fang, Y.; Gong, M.; and Jiang, D. 2020a. Unicoder-VL: A Universal Encoder for Vision and Language by Cross-Modal Pre-Training. In *Proc. of AAAI*, 11336–11344.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022b. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *ArXiv preprint*, abs/2201.12086.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021a. Align before fuse: Vision and language representation learning with momentum distillation. *Proc. of NeurIPS*, 34: 9694–9705.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019. VISUALBERT: ASimple AND PERFORMANT BASELINE FOR VISION AND LANGUAGE. *ArXiv preprint*, abs/1908.03557.

Li, W.; Gao, C.; Niu, G.; Xiao, X.; Liu, H.; Liu, J.; Wu, H.; and Wang, H. 2021b. UNIMO: Towards Unified-Modal Understanding and Generation via Cross-Modal Contrastive Learning. In *Proc. of ACL*, 2592–2607.

Li, W.; Gao, C.; Niu, G.; Xiao, X.; Liu, H.; Liu, J.; Wu, H.; and Wang, H. 2022c. UNIMO-2: End-to-End Unified Vision-Language Grounded Learning. In *Proc. of ACL Findings*, 3187–3201.

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. 2020b. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proc. of ECCV*, 121–137.

Li, Y.; Mao, H.; Girshick, R.; and He, K. 2022d. Exploring plain vision transformer backbones for object detection. *ArXiv preprint*, abs/2203.16527.

Lin, T.; Dollár, P.; Girshick, R. B.; He, K.; Hariharan, B.; and Belongie, S. J. 2017. Feature Pyramid Networks for Object Detection. In *Proc. of CVPR*, 936–944.

Liu, N. F.; Gardner, M.; Belinkov, Y.; Peters, M. E.; and Smith, N. A. 2019a. Linguistic Knowledge and Transferability of Contextual Representations. In *Proc. of NAACL*, 1073–1094.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. Ssd: Single shot multibox detector. In *Proc. of ECCV*, 21–37.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019b. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv preprint*, abs/1907.11692.

Liu, Y.; Wu, C.; Tseng, S.-Y.; Lal, V.; He, X.; and Duan, N. 2022. KD-VLP: Improving End-to-End Vision-and-Language Pretraining with Object Knowledge Distillation. In *Proc. of ACL Findings*, 1589–1600.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *Proc. of ICLR*.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Proc. of NeurIPS*, 13–23.

Nagrani, A.; Yang, S.; Arnab, A.; Jansen, A.; Schmid, C.; and Sun, C. 2021. Attention bottlenecks for multimodal fusion. *Proc. of NeurIPS*, 34: 14200–14213.

Naseer, M. M.; Ranasinghe, K.; Khan, S. H.; Hayat, M.; Shahbaz Khan, F.; and Yang, M.-H. 2021. Intriguing properties of vision transformers. *Proc. of NeurIPS*, 34: 23296–23308.

Ni, M.; Huang, H.; Su, L.; Cui, E.; Bharti, T.; Wang, L.; Zhang, D.; and Duan, N. 2021. M3p: Learning universal representations via multitask multilingual multimodal pre-training. In *Proc. of CVPR*, 3977–3986.

Ordonez, V.; Kulkarni, G.; and Berg, T. L. 2011. Im2Text: Describing Images Using 1 Million Captioned Photographs. In *Proc. of NeurIPS*, 1143–1151.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018a. Deep Contextualized Word Representations. In *Proc. of NAACL*, 2227–2237.

Peters, M. E.; Neumann, M.; Zettlemoyer, L.; and Yih, W.-t. 2018b. Dissecting Contextual Word Embeddings: Architecture and Representation. In *Proc. of EMNLP*, 1499–1509.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proc. of ICML*, volume 139, 8748–8763.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Raghu, M.; Unterthiner, T.; Kornblith, S.; Zhang, C.; and Dosovitskiy, A. 2021. Do vision transformers see like convolutional neural networks? *Proc. of NeurIPS*, 34: 12116–12128.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Proc. of NeurIPS*, 91–99.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*.

Sennrich, R.; Haddow, B.; and Birch, A. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proc. of ACL*, 1715–1725.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Proc. of ACL*, 2556–2565.

Shen, S.; Li, L. H.; Tan, H.; Bansal, M.; Rohrbach, A.; Chang, K.-W.; Yao, Z.; and Keutzer, K. 2021. How Much Can CLIP Benefit Vision-and-Language Tasks? *ArXiv preprint*, abs/2107.06383.

Shen, Y.; Tan, X.; He, D.; Qin, T.; and Liu, T.-Y. 2018. Dense Information Flow for Neural Machine Translation. In *Proc. of NAACL*, 1294–1303.

Søgaard, A.; and Goldberg, Y. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *Proc. of ACL*, 231–235.

Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; and Dai, J. 2020. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *Proc. of ICLR*.

Sun, C.; Qiu, X.; Xu, Y.; and Huang, X. 2019. How to fine-tune bert for text classification? In *Proc. of CCL*, 194–206.

Tan, H.; and Bansal, M. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *Proc. of EMNLP*, 5100–5111.

Teney, D.; Anderson, P.; He, X.; and van den Hengel, A. 2018. Tips and Tricks for Visual Question Answering: Learnings From the 2017 Challenge. In *Proc. of CVPR*, 4223–4232.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *Proc. of NeurIPS*, 5998–6008.

Wang, J.; Yang, Z.; Hu, X.; Li, L.; Lin, K.; Gan, Z.; Liu, Z.; Liu, C.; and Wang, L. 2022a. GIT: A Generative Image-to-text Transformer for Vision and Language. *ArXiv preprint*, abs/2205.14100.

Wang, P.; Yang, A.; Men, R.; Lin, J.; Bai, S.; Li, Z.; Ma, J.; Zhou, C.; Zhou, J.; and Yang, H. 2022b. Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. *ArXiv preprint*, abs/2202.03052.

Wang, Q.; Li, F.; Xiao, T.; Li, Y.; Li, Y.; and Zhu, J. 2018. Multi-layer Representation Fusion for Neural Machine Translation. In *Proc. of COLING*, 3015–3026.

Wang, W.; Bao, H.; Dong, L.; Bjorck, J.; Peng, Z.; Liu, Q.; Aggarwal, K.; Mohammed, O. K.; Singhal, S.; Som, S.; et al. 2022c. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *ArXiv preprint*, abs/2208.10442.

Wang, W.; Bao, H.; Dong, L.; and Wei, F. 2021a. VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts. *ArXiv preprint*, abs/2111.02358.

Wang, Z.; Wang, W.; Zhu, H.; Liu, M.; Qin, B.; and Wei, F. 2021b. Distilled Dual-Encoder Model for Vision-Language Understanding. *ArXiv preprint*, abs/2112.08723.

Wang, Z.; Yu, J.; Yu, A. W.; Dai, Z.; Tsvetkov, Y.; and Cao, Y. 2021c. Simvlm: Simple visual language model pretraining with weak supervision. *ArXiv preprint*, abs/2108.10904.

Xia, Q.; Huang, H.; Duan, N.; Zhang, D.; Ji, L.; Sui, Z.; Cui, E.; Bharti, T.; and Zhou, M. 2021. Xgpt: Cross-modal generative pre-training for image captioning. In *Proc. of NLPCC*, 786–797.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Proc. of NeurIPS*, 34: 12077–12090.

Xie, N.; Lai, F.; Doran, D.; and Kadav, A. 2019. Visual entailment: A novel task for fine-grained image understanding. *ArXiv preprint*, abs/1901.06706.

Xie, Z.; Geng, Z.; Hu, J.; Zhang, Z.; Hu, H.; and Cao, Y. 2022. Revealing the Dark Secrets of Masked Image Modeling. *ArXiv preprint*, abs/2205.13543.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2: 67–78.

Yu, F.; Wang, D.; Shelhamer, E.; and Darrell, T. 2018. Deep Layer Aggregation. In *Proc. of CVPR*, 2403–2412.

Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; and Wu, Y. 2022. Coca: Contrastive captioners are image-text foundation models. *ArXiv preprint*, abs/2205.01917.

Yu, Z.; Yu, J.; Cui, Y.; Tao, D.; and Tian, Q. 2019. Deep Modular Co-Attention Networks for Visual Question Answering. In *Proc. of CVPR*, 6281–6290.

Zeiler, M. D.; and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *Proc. of ECCV*, 818–833.

Zeng, Y.; Zhang, X.; and Li, H. 2021. Multi-Grained Vision Language Pre-Training: Aligning Texts with Visual Concepts. *ArXiv preprint*, abs/2111.08276.

Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; and Gao, J. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proc. of CVPR*, 5579–5588.

Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P. H.; et al. 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proc. of CVPR*, 6881–6890.

Zhou, L.; Palangi, H.; Zhang, L.; Hu, H.; Corso, J. J.; and Gao, J. 2020. Unified Vision-Language Pre-Training for Image Captioning and VQA. In *Proc. of AAAI*, 13041–13049.