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

Xiao Xu1, 2*, Chenfei Wu2, Shachar Rosenman3, Vasudev Lal3, Wanxiang Che1†, Nan Duan2†

1Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology
2Microsoft Research Asia
3Intel Labs, Cognitive Computing Research

{xu,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 use the last-layer uni-modal representations from deep pre-trained uni-modal encoders to the 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 checklists 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. 2021; 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
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 VQA v2 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 VQA v2 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 (VQA v2), 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.

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...
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 \(I \in \mathbb{R}^{H \times W \times C}\), 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 \(P \in \mathbb{R}^{N \times (P^2 C)}\), 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 \([\text{class}]\) token to the patch sequence and uses learnable 1D position embeddings \(V^\text{pos} \in \mathbb{R}^{(N+1) \times D_v}\), where \(D_v\) is the dimension of the visual encoder. The input visual representation can be calculated as follows:

\[
V_0 = [E_1^{\text{class}}; p_1 W_p; \ldots; p_N W_p] + V^\text{pos},
\]

where \(W_p \in \mathbb{R}^{(P^2 C) \times D_v}\) is the trainable linear projection layer and \(V_0 \in \mathbb{R}^{(N+1) \times D_v}\). Each layer of ViT consists of
a multi-head self-attention (MSA) block and a feed-forward network (FFN) block. We omit the computation details and simplify them as $\text{Encoder}^V$. The $\ell$-th layer representation can be denoted as $V_\ell = \text{Encoder}^V_\ell (V_{\ell-1}), \ell = 1 \ldots L_V$, 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\_BASE as our textual encoder. Each input sequence $w$ is tokenized by the byte-level Byte-Pair Encoding (BPE) (Sennrich, Haddow, and Birch 2016; Radford et al. 2019). $[\text{<s>}]$ token and $[\text{</s>}]$ token are added to the sequence as the start and end tokens, respectively. The input textual representation can be represented as:

$$T_0 = [E_{\text{<s>}}; E_w; \ldots; E_{w_m}; E_{\text{</s>}},] + T^\text{pos},$$

where $T_0 \in \mathbb{R}^{(M+2) \times D_t}$, $E$ is the word embedding matrix, $M$ is the number of tokens, $D_t$ is the dimension of the textual encoder, and $T^\text{pos}$ is the positional embedding matrix. Similarly, we denote the $\ell$-th layer of the textual encoder as $\text{Encoder}^T_\ell$, and the $\ell$-th layer representation can be denoted as $T_\ell = \text{Encoder}^T_\ell (T_{\ell-1}), \ell = 1 \ldots L_T$, 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, 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 $\ell$-th layer of the cross-modal encoder as $\text{Encoder}^Z_\ell$, 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:

$$Z^V_\ell = Z^V_{\ell-1},$$

$$Z^T_\ell = Z^T_{\ell-1},$$

$$Z^V_\ell, Z^T_\ell = \text{Encoder}^Z_\ell (Z^V_\ell, Z^T_\ell), \ell = 1 \ldots L_Z,$$

where $Z^V_\ell, Z^T_\ell$ is the output representation of the visual or textual part at the $\ell$-th layer. $Z^V_\ell, Z^T_\ell$ 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 $\text{Encoder}^Z_\ell$ (Equation 3 & 4). $Z^V_0, Z^T_0$ are initialized with the last-layer representations from pre-trained uni-modal encoders: $Z^V_0 = V_0, V_0, V^\text{type}, Z^T_0 = T_0, W_T + T^\text{type}$, where $W_V \in \mathbb{R}^{D_v \times D_z}$ and $W_T \in \mathbb{R}^{D_t \times D_z}$ are linear projections, $V^\text{type}$ and $T^\text{type}$ 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:

$$\tilde{Z}^V_\ell = \text{BridgeLayer}^V_\ell (Z^V_{\ell-1}, V_k W_V + V^\text{type}),$$

$$\tilde{Z}^T_\ell = \text{BridgeLayer}^T_\ell (Z^T_{\ell-1}, T_k W_T + T^\text{type}),$$

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

---

**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.
We pre-train BRIDGEOWER 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 [sep] 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, RoBERTaBASE 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 (VQA) (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 VQA, where we use 576 × 576 for a robust evaluation and fair comparison with METER. Standard settings and splits are used for all datasets. For VQA, where we follow the common practice (Goyal et al. 2017; Teney et al. 2018), convert VQA 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 VQA and Flickr30K datasets. Each model is initialized with CLIP-ViT-B-224/16 and RoBERTaBASE 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 VQA 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. Wₙ is initialized as zero so that row (i) is
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.

<table>
<thead>
<tr>
<th>BridgeLayer ($x, y$)</th>
<th># Params</th>
<th>Test-Dev</th>
<th>RSUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) $x + y$</td>
<td>18.4K</td>
<td>75.18</td>
<td>533.8</td>
</tr>
<tr>
<td>(b) $x \odot y$</td>
<td>18.4K</td>
<td>73.41</td>
<td>530.4</td>
</tr>
<tr>
<td>(c) $\alpha x + (1 - \alpha) y$, $\alpha \in \mathbb{R}^D$</td>
<td>26.0K</td>
<td>75.09</td>
<td>532.9</td>
</tr>
<tr>
<td>(d) $\alpha x + (1 - \alpha) y$, $\alpha = \sigma(W [x; y])$</td>
<td>11.8M</td>
<td>75.13</td>
<td>533.1</td>
</tr>
<tr>
<td>(e) $W_2 (\text{GeLU} (W_1 [x; y]))$</td>
<td>11.8M</td>
<td>74.55</td>
<td>532.2</td>
</tr>
<tr>
<td>(f) $\text{MCA} (x, y)$</td>
<td>23.6M</td>
<td>74.26</td>
<td>530.2</td>
</tr>
<tr>
<td>(g) $\text{FFN} (\text{MCA} (x, y))$</td>
<td>70.8M</td>
<td>73.54</td>
<td>511.1</td>
</tr>
<tr>
<td>(i) $x + y + W_z [x; y]$</td>
<td>11.8M</td>
<td>75.10</td>
<td>533.1</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Backbone</th>
<th>Visual</th>
<th>Textual</th>
<th>VQA2 Test-Dev</th>
<th>Flickr30K RSUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Det B-224/16 RoBERTa</td>
<td>69.98</td>
<td>70.83</td>
<td>448.0</td>
<td>455.7</td>
</tr>
<tr>
<td>VIT B-224/16 RoBERTa</td>
<td>70.26</td>
<td>72.24</td>
<td>472.7</td>
<td>476.9</td>
</tr>
<tr>
<td>VIT B-384/16 RoBERTa</td>
<td>70.52</td>
<td>72.38</td>
<td>472.8</td>
<td>477.1</td>
</tr>
<tr>
<td>CLIP-ViT-B/32 RoBERTa</td>
<td>72.19</td>
<td>72.91</td>
<td>508.8</td>
<td>512.0</td>
</tr>
<tr>
<td>CLIP-ViT-B/16 BERT</td>
<td>74.09</td>
<td>74.89</td>
<td>522.1</td>
<td>526.5</td>
</tr>
<tr>
<td>CLIP-ViT-B/6 BERT</td>
<td>74.06</td>
<td>74.85</td>
<td>530.7</td>
<td>533.8</td>
</tr>
</tbody>
</table>

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

### 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.

**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 $1 \times 10^{-5}$. 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 $288 \times 288$. 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 achieves state-of-the-art performance, in particular 78.73% accuracy on the VQA2 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 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 achieves the best performance, surpassing not only ALBEF with its specially designed pre-training objective, but also ALIGN with 1.8B pre-train images.

<table>
<thead>
<tr>
<th>$L_Z$</th>
<th># Params</th>
<th>VQA2 Test-Dev</th>
<th>Flickr30K RSUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meter</td>
<td>0.80</td>
<td>74.93</td>
<td>526.0</td>
</tr>
<tr>
<td>Ours</td>
<td>1.18</td>
<td>75.18</td>
<td>527.1</td>
</tr>
<tr>
<td>Meter</td>
<td>0.85</td>
<td>74.94</td>
<td>526.5</td>
</tr>
<tr>
<td>Ours</td>
<td>1.29</td>
<td>75.18</td>
<td>528.6</td>
</tr>
<tr>
<td>Meter</td>
<td>0.72</td>
<td>74.36</td>
<td>527.9</td>
</tr>
<tr>
<td>Ours</td>
<td>1.18</td>
<td>75.00</td>
<td>529.7</td>
</tr>
<tr>
<td>Meter</td>
<td>0.71</td>
<td>74.98</td>
<td>530.3</td>
</tr>
<tr>
<td>Ours</td>
<td>1.18</td>
<td>75.16</td>
<td>530.7</td>
</tr>
<tr>
<td>Meter</td>
<td>0.70</td>
<td>73.71</td>
<td>531.6</td>
</tr>
<tr>
<td>Ours</td>
<td>1.18</td>
<td>75.07</td>
<td>531.7</td>
</tr>
<tr>
<td>Meter</td>
<td>0.72</td>
<td>74.04</td>
<td>534.4</td>
</tr>
<tr>
<td>Ours</td>
<td>1.18</td>
<td>75.06</td>
<td>534.7</td>
</tr>
</tbody>
</table>
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. ** 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).

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\textsubscript{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 $14 \times 14$, then we set the image resolution to $294 \times 294$ in pre-training and to $574 \times 574$ in fine-tuning on the VQAv2. Other hyperparameters remain unchanged. As shown in Table 4, B\textsc{ridge}T\textsc{ower} 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, B\textsc{ridge}T\textsc{ower}_{\text{LARGE}} achieves 81.15% accuracy on the VQAv2 test-std set, surpassing the previous state-of-the-art OFA\textsubscript{LARGE} model by 0.48%. This further demonstrates the effectiveness and scalability of B\textsc{ridge}T\textsc{ower}. In addition, question-answer pairs from VG dataset are often used to extend the VQA2 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 M\textsc{eter} and B\textsc{ridge}T\textsc{ower} 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 B\textsc{ridge}T\textsc{ower}, but for M\textsc{eter}, 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 B\textsc{ridge}T\textsc{ower} is significantly larger than that of M\textsc{eter}, 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 M\textsc{eter}, B\textsc{ridge}T\textsc{ower} 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 $^3$.

$^2$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.


![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 M\textsc{eter} and B\textsc{ridge}T\textsc{ower} models.](image)

5 Conclusion and Future Work

We present B\textsc{ridge}T\textsc{ower}, 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 B\textsc{ridge}T\textsc{ower}, 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-
mannances 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


