U-BERT for Fast and Scalable Text-Image Retrieval

Tan Yu
Cognitive Computing Lab
Baidu Research
10900 NE 8th St
Bellevue, WA 98004, USA
tan.yu1503@gmail.com

Hongliang Fei
Cognitive Computing Lab
Baidu Research
10900 NE 8th St
Bellevue, WA 98004, USA
feihongliang0@gmail.com

Ping Li
Cognitive Computing Lab
Baidu Research
10900 NE 8th St
Bellevue, WA 98004, USA
pingli98@gmail.com

ABSTRACT

Exploiting cross-modal attention on image region features and text features, cross-modal BERT models have achieved higher accuracy than the embedding-based methods in cross-modal text-image retrieval. Nevertheless, cross-modal BERT models take image-text pairs as input, requiring a quadratic computational complexity. Thus, cross-modal BERT models are prohibitively slow and not scalable. A remedy is a two-stage strategy, wherein the first stage uses an embedding-based method to retrieve top K items and the second stage deploys the heavy cross-modal BERT to re-rank these K items. Nevertheless, to achieve a satisfying accuracy, K should be large, making the retrieval in the second phase still slow. In this paper, we propose a U-BERT model to achieve an effective and efficient cross-modal retrieval. Our model decomposes each image/text feature into an intra-modal component and an inter-modal component. In the first stage, U-BERT only uses the intra-modal component of the image/text features to obtain the text-image similarity scores based on two independent encoders, with a linear computation complexity. In the second stage, U-BERT reuses the intra-modal component and additionally use the inter-modal component as the complementary residual to enhance the intra-modal component's discriminating capability. By reusing the intra-modal component, we only need few Transformer layers to generate the inter-modal component, making the second phase efficient even facing a large K. Extensive experiments on public benchmarks demonstrate the efficiency and effectiveness of the proposed U-BERT.

CCS CONCEPTS

• **Computing methodologies → Visual content-based indexing and retrieval**; • **Information systems** → *Image search*; *Multilingual and cross-lingual retrieval*; **Multimedia and multimodal retrieval**.

KEYWORDS

retrieval, search, computer vision, cross-modal, deep learning, cross-lingual, natural language understanding

ACM Reference Format:
Tan Yu, Hongliang Fei, and Ping Li. 2022. U-BERT for Fast and Scalable Text-Image Retrieval. In *Proceedings of the 2022 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '22), July 11–12, 2022, Madrid, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3539813.3545148

1 INTRODUCTION

With the rapid growth of multimedia data, cross-modal retrieval has received significant attention from both academia [2, 44, 51, 52] and industry [19, 20, 35, 58]. Due to the modal gap between different data types, it is quite challenging to obtain a reliable similarity measurement. In this work, we focus on the retrieval task between two most common data types: text and image. Generally speaking, the existing solutions to image-text matching fall into two categories: embedding-based methods [7, 9, 12, 39, 44, 58] and attention-based methods [3, 8, 21, 24, 30–32, 36, 48, 53, 54, 57, 59, 60, 62, 63].

Embedding-based methods [7, 9, 12, 18, 39, 40, 44] map the images and texts into a joint feature space. They adopt two encoders to generate the image embedding and the text embedding separately. In the search phase, the similarity between a text and an image is calculated from the image embedding and the text embedding. In the training phase, the image encoder and the text encoder are optimized to increase the similarities between a text and its relevant images and decrease that between its counterparts. Decoupling the image features and sentence features, embedding-based methods enjoy high efficiency by precomputing and caching image embeddings. Owing to the high efficiency, embedding-based methods have been widely deployed in large-scale cross-modal retrieval systems.

In parallel, attention-based methods [3, 14, 16, 17, 21, 24, 32, 36, 48, 53, 54] represent an image by a set of region features and represent a text by a sequence of word features. When conducting the matching between an image and a sentence, attention-based methods [49] pay more attention to critical regions in the image and key words in the sentence. Benefited from exploiting fine-level matching, attention-based methods normally achieve higher retrieval accuracy than embedding-based methods. Recently, inspired by the great success of BERT [6] in natural language processing tasks, several cross-modal BERT models emerge [4, 25, 28, 32, 37, 46, 48, 61, 63] and achieve state-of-the-art performance.

Nevertheless, due to the dependency between the text features and image features, the attention-based methods are much slower than embedding-based methods. For instance, given N texts and N images, embedding-based methods only need to extract N text features and N image features separately with a complexity $O(N)$. In contrast, the cross-modal BERT consumes N^2 input text-image pairs to obtain text-image similarities, since the image features and

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ICTIR '22, July 11–12, 2022, Madrid, Spain
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9412-3/22/07…$15.00
https://doi.org/10.1145/3539813.3545148

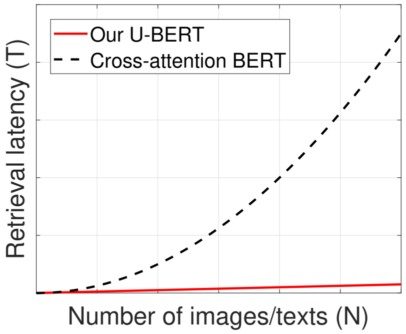

Figure 1: The scalability of cross-attention BERT and our U-BERT. Given N texts and N images, the retrieval latency (T) of the cross-attention BERT to obtain text-image similarities is a quadratic function of N. In contrast, the retrieval latency of our U-BERT is a linear function of N.

the text features are dependent in cross-modal BERT. Thus, the total computational complexity of the cross-modal BERT method is $O(N^2)$. The quadratic computational complexity of the cross-modal BERT method is significantly larger than the linear computational complexity of the embedding-based method. Despite that the cross-modal BERT methods have achieved state-of-the-art cross-modal retrieval accuracy, they are prohibitively slow and not scalable as visualized in Figure 1.

A trade-off between accuracy and efficiency is a two-stage strategy, which is recently investigated by LightningDOT [47] and Geigle *et al.* [11]. They adopt an efficient joint-embedding method for large-scale retrieval in the first stage to obtain a small-scale set of K candidate items. Then, it adopts a heavy cross-modal BERT model to re-rank the K candidate items for higher retrieval accuracy. Since K is smaller than the total number of reference items in the corpus, the efficiency is boosted. Nevertheless, to achieve a satisfactory retrieval accuracy, K cannot be too small. Thus the retrieval is still slow. When using re-ranking, LightningDOT [47] only achieves a 46× speed-up ratio over the cross-modal BERT UNITER-base [4].

To achieve an effective and efficient cross-modal retrieval, we propose a novel model as shown in Figure 2. It consists of two tall encoders BERT_{img} and BERT_{txt} as well as a low encoder $\text{BERT}_{\text{cross}}$. Since it is in a **U** shape, we term it as U-BERT. It decomposes the text feature $\mathbf{T}$ as well as the image feature $\mathbf{I}$ into an intra-modal component $\mathbf{I}_{\text{intra}}/\mathbf{T}_{\text{intra}}$ and an inter-modal component $\mathbf{I}_{\text{inter}}/\mathbf{T}_{\text{inter}}$. The intra-modal component of the text feature $\mathbf{T}_{\text{intra}}$ and that of the image feature $\mathbf{I}_{\text{intra}}$ are obtained from the text encoder BERT_{txt} and the image encoder BERT_{img}, separately. Meanwhile, the inter-modal component $\mathbf{T}_{\text{inter}}$ of the text feature and that of the image feature $\mathbf{I}_{\text{inter}}$ are obtained from $\text{BERT}_{\text{cross}}$. They exploit the cross-modal attention and serve as the complementary residues to enhance the intra-modal components. In the retrieval phase, we first only use the intra-modal components $\mathbf{I}_{\text{intra}}/\mathbf{T}_{\text{intra}}$ to retrieve top K reference items. Since the intra-modal components do not need cross-modal attention, they achieve fast retrieval as embedding-based methods. Then, in the second stage, we enhance the intra-modal components $\mathbf{I}_{\text{intra}}/\mathbf{T}_{\text{intra}}$ through the inter-modal components from $\text{BERT}_{\text{cross}}$. Thanks to reusing the intra-modal components, we only need a lightweight $\text{BERT}_{\text{cross}}$ to achieve an excellent retrieval accuracy. In fact, using a $\text{BERT}_{\text{cross}}$ with only 3 Transformer layers, the performance of our U-BERT is comparable to the cross-modal BERT with 12 Transformer layers. Benefited from the two-stage configuration, our U-BERT is efficient and scalable as shown in Figure 1.

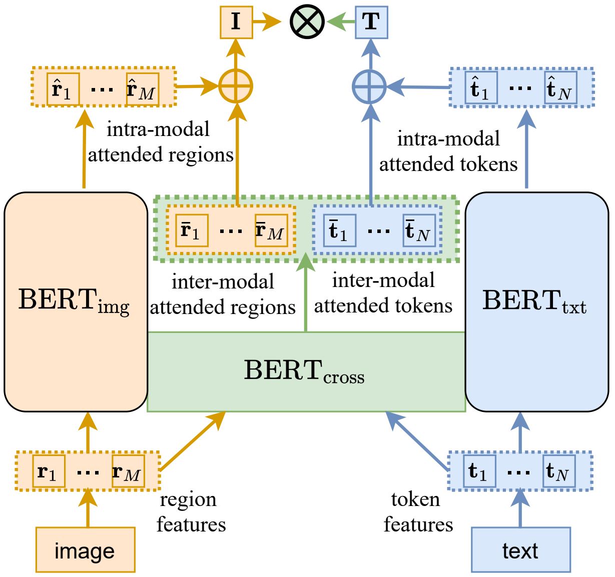
$\mathbf{I}$
$\otimes$
$\mathbf{T}$
$\hat{\mathbf{r}}_1 \cdots \hat{\mathbf{r}}_M$
$\hat{\mathbf{t}}_1 \cdots \hat{\mathbf{t}}_N$
intra-modal attended regions
intra-modal attended tokens
$\bar{\mathbf{r}}_1 \cdots \bar{\mathbf{r}}_M$
$\bar{\mathbf{t}}_1 \cdots \bar{\mathbf{t}}_N$
BERT_{img}
inter-modal attended regions
inter-modal attended tokens
BERT_{txt}
$\text{BERT}_{\text{cross}}$
$\mathbf{r}_1 \cdots \mathbf{r}_M$
$\mathbf{t}_1 \cdots \mathbf{t}_N$
region features
token features
image
text

Figure 2: The architecture of U-BERT. The region/token features are fed into a heavy $\text{BERT}_{\text{img}}/\text{BERT}_{\text{txt}}$ to generate an intra-modal component of the image/text features. Meanwhile, an inter-modal component of the text/image features is obtained from a lightweight $\text{BERT}_{\text{cross}}$. The final image/text features is obtained by summing up the intra-modal component and the inter-modal component.

We conduct comprehensive experiments on three benchmarks (MSCOCO1K, MSCOCO5K and Flickr30K) to show the efficiency and the effectiveness of U-BERT. We achieve a comparable retrieval accuracy with vanilla cross-modal BERT methods on these public benchmarks. Meanwhile, we obtain an around 160× speed-up ratio on Flickr30K as well as MSCOCO1K dataset, and an around 780× speed-up ratio on MSCOCO5K dataset.

2 RELATED WORK

Embedding-based methods. Early embedding-based methods [12, 39] are based on correlation canonical analysis (CCA) [13]. They project the text features and image features into a joint feature space. In this case, the text-image similarity can be determined by their distance in the feature space. Recently, deep neural networks [7, 9] are exploited to generate more effective text and image embeddings. For example, VSE++ [7] extracts the image embedding through a CNN and meanwhile generates the text embedding from an RNN. It optimizes the similarities between texts and images based on a triplet loss widely used in metric learning. To make the training more effective, VSE++ adopts hard negative mining to focus more on the hard negative samples. Xu *et al.* [53] propose a polynomial loss for more effective cross-modal matching through adaptively weighting the hard pairs. In parallel, MRAM [3] adopts adversarial learning to enhance the robustness of the learned metric for high effective cross-modal retrieval. The advantage of embedding-based methods is high efficiency. We only need to individually extract the global text/image embedding for each text/image. Then the similarity between a text and an image is efficiently obtained by computing the similarity between their embeddings. Recently, CLIP [40] and Align [18] exploit the embedding-based method for self-supervised learning based on text-image pairs. They crawl huge-scale text-image corpus from the website and train an image encoder as well a text encoder based on the text-image pairs. After that, the trained image encoder will be fine-tuned in the downstream tasks such as image recognition, segmentation and object detection.

Attention-based methods exploit the interactions between local features of images and texts. They represent an image by a set of region features and a text by a set of words. They conduct fine-level matching between region features and word features. A pioneering attention-based method, SCAN [24] pays more attention to the regions and words with high relevance and suppresses the bad effects from the distracting background. Note that SCAN only exploits the attention in the late stage when the region features and word features have already been obtained. To exploit attention more thoroughly, recent attention-based methods adopt a graph-convolution network [26] or Transformer layers [50] to exploit the cross-modal attention in the early stage. Recently, inspired by the success achieved by BERT [6] in natural language processing tasks, many cross-modal BERT methods are proposed. Based on the structure, they can be grouped into two categories: 1) single-stream methods [4, 10, 25, 28, 45] and 2) two-stream methods [5, 32, 33, 48]. The single-stream model simply concatenates the word features and region features. It adopts a single BERT model to exploit the cross-modal attention between region features and word features. In contrast, the two-stream model uses two BERT models. The image-stream BERT model uses region features to attend word features, and the text-stream BERT model utilizes word features to attend region features. Both the single-stream model and the two-stream model design several pre-training tasks to enhance their cross-modal understanding capability. Benefited from exploiting cross-modal attention and pre-training on large-scale datasets, cross-modal BERT methods have achieved excellent performance in many cross-modal understanding tasks, such as image captioning, visual question answering, and cross-modal retrieval. Recently, VILLA [10] improves the robustness of the cross-modal BERT model through adversarial learning. OSCAR [28] improves the cross-modal BERT model by exploiting additional tags from object detectors. ERNIE-ViL [56] boosts the cross-modal understanding capability through exploiting the external knowledge. UNIMO [27] proposes a unified learning framework for learning the text and the visual representation jointly. VinVL [65] further improves OSCAR by using more effective vision features. They are prohibitively slow for large-scale cross-modal retrieval in real applications. ViLT [22] builds a pure Transformer architecture for fast cross-modal understanding. It does not need the expensive computational cost from pre-trained object detector and thus the efficiency is boosted. Nevertheless, in the cross-modal retrieval task, ViLT also suffers from quadratic computational cost, which is too slow in large-scale retrieval applications. Recently, Nie *et al.* [38] investigates the effectiveness of pure MLP-based architecture for cross-modal understanding. Nevertheless, the cross-modal BERT methods take text-image pairs as input, leading to quadratic computational complexity.

Hybrid methods. To improve the efficiency, VisualSparta [33] adopts a lightweight cross-modal attention operation only on local features from the output of the BERT model. Inflate & Shrink [30] exploits the late-stage cross-modal attention and additionally adopts the knowledge distilling to further reduce the computation cost. In parallel, LightningDOT [47] and Gregor *et al.* [11] adopt a two-stage strategy. In the first stage, they adopt an efficient embedding-based method to rank reference items in the corpus and obtains top K relevant items. Then it deploys the heavy cross-modal BERT to re-rank the top K relevant items in the initial list. Since the number of items for re-ranking (K) is much smaller than the total number of items in the corpus, the retrieval efficiency is significantly boosted. To save the memory, Gregor *et al.* [11] share the weights between the embedding-based model in the first stage and the cross-attention model in the second stage. Nevertheless, their speed-up ratio over the attention based methods is still relatively low. The U-BERT in this work also adopts the two-stage scheme. We decouple the text/image feature into an intra-modal component and an inter-modal component. The intra-modal component is generated from an embedding-based method and is used in the first-stage ranking to obtain the top-K items. Then, in the second stage, we reuse the intra-modal component and combine it with the inter-modal component obtained from a cross-attention model for re-ranking. Thanks to reusing the intra-modal component in the second stage, we only need a lightweight cross-modal BERT in the re-ranking stage and thus our U-BERT is significantly faster than the aforementioned two-stage methods including Sun et al. [47] and Geigle et al. [11].

3 PRELIMINARY

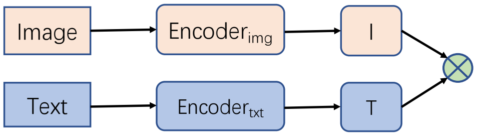

Figure 3: The architecture of embedding-based methods. The image encoder extracts the image feature I and the text encoder generates the text feature T. The text-image similarity is computed from T and I.

Embedding-based methods extract the image embedding for the image I and the text embedding for the text T through two encoders. As visualized in Figure 3, the image encoder extracts the image embedding $\mathbf{I}$ and the text encoder extracts the text embedding $\mathbf{T}$. The similarity between I and T is determined by the similarity between the image embedding $\mathbf{I}$ and the text embedding $\mathbf{T}$:

$$s(T, I) = \frac{\langle \mathbf{T}, \mathbf{I} \rangle}{\|\mathbf{T}\|_2 \|\mathbf{I}\|_2}. \tag{1}$$

Since the embedding-based methods individually encode the texts and images, they are efficient for large-scale cross-modal retrieval. Given Q texts and P images, joint-embedding methods only need to

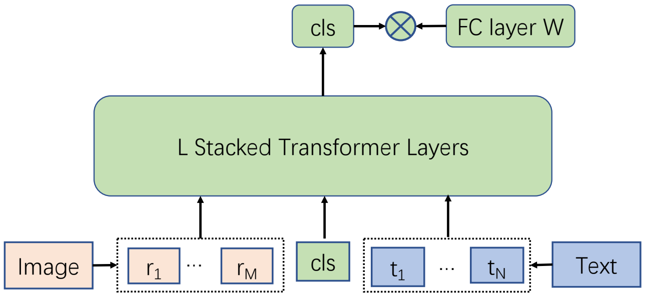

Figure 4: The architecture the cross-modal BERT. Region features, token features and the special token feature are concatenated into a long sequence, which is fed into a stack of L Transformer layers to exploit the cross-modal attention. The attended special token feature is further fed into a fully-connected (FC) layer to obtain the text-image similarity.

extract P image embeddings and Q text embeddings with computation cost of $O(P+Q)$. Nevertheless, due to the lack of interactions between text and image features, the embedding based methods are normally less accurate than attention-based methods.

Cross-modal BERT can be categorized into one-stream architecture [4, 10, 25, 28, 45] and two-stream architecture [5, 32, 33, 48]. Both of them exploit the cross-modal attention to enhance the discriminating capability of the image's region features and text's token features. For the easiness of illustration, we mainly introduce the one-stream architecture due to its simplicity in the formulation. Given an image I, a set of regions $\{r_i\}_{i=1}^M$ are detected by a pretrained object detector, e.g. faster R-CNN [41]. Each region r_i is represented by a region feature $\mathbf{r}_i$, which is obtained by summing up its visual feature $\mathbf{v}_i$ from faster R-CNN, its positional feature obtained based on the coordinates of the region, and an image type embedding $\mathbf{e}_i$:

$$\mathbf{r}_i = \mathbf{v}_i + \mathbf{p}_i + \mathbf{e}_i, \tag{2}$$

Each region feature $\mathbf{r}_i$ represents a candidate object in the image.

In parallel, a text T in the input is converted into a sequence of tokens $[t_1, \cdots, t_N]$ by a tokenizer. Each token t_i ($i \in [1, N]$) is represented by a token feature $\mathbf{t}_i$, which is obtained by summing up its word embedding, positional embedding, and type embedding. One-stream cross-modal BERT concatenates the region features of the image $\{\mathbf{r}_i\}_{i=1}^M$, token features from the text $\{\mathbf{t}_i\}_{i=1}^N$, and a special token feature $\mathbf{t}_{\text{cls}}$ into a long sequence $\mathcal{S}$:

$$\mathcal{S} = [\mathbf{r}_1, \cdots, \mathbf{r}_M, \mathbf{t}_1, \cdots, \mathbf{t}_N, \mathbf{t}_{\text{cls}}]. \tag{3}$$

After that, as visualized in Figure 4, cross-modal BERT feeds $\mathcal{S}$ into a stack of Transformer layers to exploit the cross-modal attention:

$$\bar{\mathcal{S}} = \text{Transformer}_{\times L}(\mathcal{S}) = [\bar{\mathbf{r}}_1, \cdots, \bar{\mathbf{r}}_M, \bar{\mathbf{t}}_1, \cdots, \bar{\mathbf{t}}_N, \bar{\mathbf{t}}_{\text{cls}}]. \tag{4}$$

Then the attended feature of the special token $\bar{\mathbf{t}}_{\text{cls}}$ is fed into a fully-connected (FC) layer to obtain the text-image similarity:

$$s(T, I) = \text{FC}(\bar{\mathbf{t}}_{\text{cls}}). \tag{5}$$

Benefited from exploiting cross-modal attention, cross-modal BERT methods achieve higher retrieval recall than joint-embedding methods. Nevertheless, the cross-modal BERT takes the text-image pair as input and is slow in large-scale cross-modal retrieval. Given Q texts and P images, the cross-modal BERT needs to compute PQ text-image pairs with $O(PQ)$ computational complexity to obtain their similarities, which is prohibitively slow for large-scale cross-modal retrieval in real applications.

Re-ranking. A trade-off between retrieval efficiency and effectiveness is a two-stage retrieval strategy. In the first stage, given a query, we can efficiently obtain the relevance between the query and all reference items through an embedding-based method. We select the most relevant K reference items based on the initial matching in the first stage. In the second stage, we use a cross-modal BERT to re-rank the obtained K reference items from the first stage. By exploiting the cross-modal attention in the re-ranking stage, it can achieve higher retrieval accuracy than embedding-based methods. Meanwhile, the number of images for re-ranking, K, is much less than the total number of reference items. Thus, the re-ranking phase is much faster than directly using cross-modal BERT to rank all reference items in the corpus. The computational cost in the re-ranking stage is $O(KC)$, where C is the computation cost for computing a text-image pair in the cross-modal BERT. To achieve a fast retrieval, KC should not be large. On the other hand, to achieve an effective re-ranking, K should be large enough. To maintain a small KC and a large K, the only choice is to decrease C. We will introduce our U-BERT taking small C for fast re-ranking.

4 U-BERT

The computation cost of cross-modal BERT, C, is linear with the number of Transformer layers. We define the computation cost of a single Transformer layer as c. Straightforwardly, we have $C = Lc$ where L is the number of Transformer layers. A naive solution to reduce C is using less Transformer layers. Nevertheless, using less Transformer layers in cross-modal BERT leads to significant performance drop as shown in Unicoder-VL [25]. Below we introduce our U-BERT for fast and accuracy cross-modal retrieval.

4.1 Architecture

To reduce the computation cost and meanwhile maintains the high retrieval accuracy, U-BERT decomposes the image feature as well as the sentence feature into two components: the intra-modal component and the inter-modal component. Suppose the image feature $\mathbf{I} = \alpha\mathbf{I}_{\text{intra}} + \mathbf{I}_{\text{inter}}$ where $\mathbf{I}_{\text{intra}}$ denotes the intra-modal component, $\mathbf{I}_{\text{inter}}$ denotes the inter-modal component, and α is a positive constant to balance the contributions from each component. Without cherry-picking, we set $\alpha = 0.5$ on all experiments, by default. In parallel, the text feature $\mathbf{T} = \alpha\mathbf{T}_{\text{intra}} + \mathbf{T}_{\text{inter}}$. The merit of this form of decomposition is that the intra-modal components $\mathbf{T}_{\text{intra}}$ and $\mathbf{I}_{\text{intra}}$ are only depended on the features from the single modal but independent from the other modal. If we use only the intra-modal components $\mathbf{T}_{\text{intra}}$ and $\mathbf{I}_{\text{intra}}$, it naturally degenerates to an embedding-based method. In parallel, the inter-modal components $\mathbf{T}_{\text{inter}}$ and $\mathbf{I}_{\text{inter}}$ exploit the cross-modal attention, which are the complementary residual components to further boost the discriminating capability of $\mathbf{T}_{\text{intra}}$ and $\mathbf{I}_{\text{intra}}$.

The intra-modal component of the image feature, $\mathbf{I}_{\text{intra}}$, is obtained from BERT_{img}. To be specific, given an image with a set of region features $\mathcal{R} = \{\mathbf{r}_1, \cdots, \mathbf{r}_M\}$, we feed the regions features $\mathcal{R}$ into BERT_{I} to obtain the intra-modal attended region features:

$$\hat{\mathcal{R}} = \text{BERT}_{\text{img}}(\mathcal{R}) = \{\hat{\mathbf{r}}_1, \cdots, \hat{\mathbf{r}}_M\}. \tag{6}$$

The image feature's intra-modal component is obtained by mean-pooling the intra-modal attended region features:

$$\mathbf{I}_{\text{intra}} = \frac{1}{M}\sum_{i=1}^{M} \hat{\mathbf{r}}_i. \tag{7}$$

In parallel, the intra-modal component of the text feature, $\mathbf{T}_{\text{intra}}$, is generated by BERT_{T}. To be specific, given a text with a sequence of token features $\mathcal{T} = [\mathbf{t}_1, \cdots, \mathbf{t}_N]$, we feed them into BERT_{txt} to obtain the intra-modal attended token features:

$$\hat{\mathcal{T}} = \text{BERT}_{\text{txt}}(\mathcal{T}) = [\hat{\mathbf{t}}_1, \cdots, \hat{\mathbf{t}}_N]. \tag{8}$$

The text feature's intra-modal component is obtained by mean-pooling the intra-modal attended token features:

$$\mathbf{T}_{\text{intra}} = \frac{1}{N}\sum_{i=1}^{N} \hat{\mathbf{t}}_i. \tag{9}$$

Meanwhile, the inter-modal component of the image feature $\mathbf{I}_{\text{inter}}$ as well as that of the sentence feature $\mathbf{T}_{\text{inter}}$ is obtained from a lightweight cross-attention encoder $\text{BERT}_{\text{cross}}$. Compared with BERT_{img} and BERT_{txt}, $\text{BERT}_{\text{cross}}$ contains much fewer Transformer layers. In practical, using only 3 Transformer layers, $\text{BERT}_{\text{cross}}$ has achieved competitive performance as the vanilla cross-modal BERT with 12 Transformer layers. The region features from the image and the token features from the text are concatenated together:

$$\mathcal{S} = [\mathcal{R}, \mathcal{T}] = [\mathbf{r}_1, \cdots, \mathbf{r}_M, \mathbf{t}_1, \cdots, \mathbf{t}_N]. \tag{10}$$

$\text{BERT}_{\text{cross}}$ takes $\mathcal{S}$ as input and generates the inter-modal attended region features and token features:

$$\bar{\mathcal{S}} = \text{BERT}_{\text{cross}}(\mathcal{S}) = [\bar{\mathbf{r}}_1, \cdots, \bar{\mathbf{r}}_M, \bar{\mathbf{t}}_1, \cdots, \bar{\mathbf{t}}_N]. \tag{11}$$

The inter-modal component of the image feature is obtained by mean-pooling the inter-modal attended region features:

$$\mathbf{I}_{\text{inter}} = \frac{1}{M}\sum_{i=1}^{M} \mathbf{r}_i, \tag{12}$$

Similarly, the inter-modal component of the text feature is:

$$\mathbf{T}_{\text{inter}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{t}_i. \tag{13}$$

To make the magnitude of $\mathbf{I}_{\text{inter}}$, $\mathbf{I}_{\text{intra}}$, $\mathbf{T}_{\text{inter}}$, $\mathbf{T}_{\text{intra}}$ in a comparable scale, we conduct ℓ_2 normalization on them.

4.2 Two-stage Retrieval

First stage. The proposed U-BERT naturally supports a two-stage retrieval process. To be specific, in the first stage, only the image's intra-modal component $\mathbf{I}_{\text{intra}}$ and the text's intra-modal component $\mathbf{T}_{\text{intra}}$ are used. The similarity between the image and the text is determined by the cosine similarity between $\mathbf{I}_{\text{intra}}$ and $\mathbf{T}_{\text{intra}}$. The first stage follows the spirit of the embedding-based method. Since $\mathbf{I}_{\text{intra}}$ and $\mathbf{T}_{\text{intra}}$ are only depended on the features from the single modal, the first stage is very efficient. Given a text query, the most K relevant images, $\{I_k\}_{k=1}^{K}$, are obtained based on the similarities between their intra-modal features with the query's intra-modal feature. They will be re-ranked in the second stage.

Second stage. We conduct re-ranking on the retrieved K images $\{I_k\}_{k=1}^{K}$ in the second stage for higher retrieval accuracy. We reuse the image's intra-modal component $\mathbf{I}_{\text{intra}}$ and the text's intra-modal component $\mathbf{T}_{\text{intra}}$. In addition, we obtain the inter-modal components $\mathbf{T}_{\text{inter}}$ and $\mathbf{I}_{\text{inter}}$ as the complementary information to enhance the discriminating capability of the intra-modal components. That is, we sum $\mathbf{T}_{\text{intra}}$ and $\mathbf{T}_{\text{inter}}$ to obtain the enhanced text feature $\mathbf{T}$ and sum up $\mathbf{I}_{\text{intra}}$ and $\mathbf{I}_{\text{inter}}$ to attain the enhanced image feature $\mathbf{I}$. Then the enhanced similarity between each image I_k and the text query T is determined by the cosine similarity between their enhanced features. After that, the retrieved images $\{I_k\}_{k=1}^{K}$ are re-ranked based on the similarities based on the enhanced features.

4.3 Training

Recall that, in this first stage, the similarity between an image I and a text is determined by the cosine similarity between their intra-modal features:

$$s_1(I, T) = \frac{\langle \mathbf{I}_{\text{intra}}, \mathbf{T}_{\text{intra}} \rangle}{\|\mathbf{I}_{\text{intra}}\| \|\mathbf{T}_{\text{intra}}\|}. \tag{14}$$

We optimize the text-image similarity in the first stage through a batch-wise triplet loss. To be specific, each batch consists of B ground-truth text-image pairs $\{(I_i, T_i)\}_{i=1}^{B}$. Among them, an image I_i is only relevant with the text T_i and is irrelevant with other texts T_j $(j \neq i)$. We obtain a $B \times B$ similarity matrices between B images and B texts based on their intra-modal features as Eq. (14). Then we construct a batch-wise triplet loss defined as

$$\mathcal{L}_1 = \sum_{i=1}^{B}\sum_{j\neq i}\Big\{[s_1(I_i, T_j) - s_1(I_i, T_i) + m]_+ + [s_1(I_j, T_i) - s_1(I_i, T_i) + m]_+\Big\},$$

where m is a predefined margin. In parallel, in the second stage, the image-text similarity is determined by the cosine similarity between their enhanced features:

$$s_2(I, T) = \frac{\langle \alpha\mathbf{I}_{\text{intra}} + \mathbf{I}_{\text{inter}}, \alpha\mathbf{T}_{\text{intra}} + \mathbf{T}_{\text{inter}} \rangle}{\|\alpha\mathbf{I}_{\text{intra}} + \mathbf{I}_{\text{inter}}\| \|\alpha\mathbf{T}_{\text{intra}} + \mathbf{T}_{\text{inter}}\|}. \tag{15}$$

To optimize the text-image similarity in the second stage, we devise another triplet loss $\mathcal{L}_2$ in the same manner as $\mathcal{L}_1$:

$$\mathcal{L}_2 = \sum_{i=1}^{B}\sum_{j\neq i}\Big\{[s_2(I_i, T_j) - s_2(I_i, T_i) + m]_+ + [s_2(I_j, T_i) - s_2(I_i, T_i) + m]_+\Big\}.$$

The whole training process consists of two steps. In the first step, we only use $\mathcal{L}_1$ to optimize BERT_{img} and BERT_{txt}. In the second step, BERT_{img} and BERT_{txt} are fixed. We only use $\mathcal{L}_2$ to update $\text{BERT}_{\text{cross}}$. An alternative solution is using a combined loss $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2$ to updates BERT_{img}, BERT_{txt} and $\text{BERT}_{\text{cross}}$ simultaneously. Nevertheless, the performance of this alternative solution is not as competitive as the two-step counterpart. We also conduct hard-negative mining as VSE++ [7] in $\mathcal{L}_1$ and $\mathcal{L}_2$.

5 EXPERIMENTS

Datasets. We conduct experiments on two public datasets, including MSCOCO [29] and Flickr30K [55]. MS COCO contains 123, 287 images. Each image has five textual descriptions. It was divided into 82, 783 training, 5, 000 validation, and 5, 000 testing samples. Following [21], we move 30, 504 images in the validation split to the training set. Experiments are conducted on both 1K and 5K testing settings. The 1K testing settings contain 1, 000 images and 5, 000 texts. In the meanwhile, 5K testing settings contain 5, 000 images and 25, 000 texts. The 1K settings are termed as MSCOCO1K, and 5K settings are termed as MSCOCO5K. Flickr30K consists of 31, 783 images from the Flickr website. Following [21], we split the dataset into 29, 783 training samples, 1, 000 validation samples and 1, 000 testing samples. The cross-modal retrieval accuracy is evaluated by image-to-text and text-to-image recall@K, which is the percentage of ground-truth matchings appearing in the top K-ranked results.

Settings. For each image, we extract 32 bounding boxes using an object detector, faster-RCNN [41], pretrained on Visual Genome dataset [23] provided by [1]. Following Unicoder-VL [25], we set the maximal token length as 44. We use the BERT-base model [6] as the backbone for both BERT_I and BERT_T in our U-BERT model. BERT-base model consists of 12 Transformer layers with 12 heads, and the hidden size is 768. In contrast, BERT_C in our U-BERT model only contains two Transformer layers. Each Transformer layer is with 12 heads, and the hidden size is 768 as well. The dropout ratio in each Transformer layer is set as 0.1. The training is conducted on a Linux server equipped with 4 NVIDIA V100 GPU cards. We adopt the ADAM optimizer and utilize float16 operations supported by the apex package to speed up the training and reduce the GPU memory consumption. By default, we set the batch size as 216. Besides, we will also report the measured retrieval latency in the testing phase based on a single NVIDIA V100 GPU.

5.1 Text-to-Image Retrieval

In this section, we evaluate the performance of the proposed U-BERT in text-to-image retrieval. For both MSCOCO1K and Flickr30K datasets, in the testing phase, they use 5, 000 text queries to retrieve the relevant images from a corpus of 1, 000 sentences.

Comparisons with baselines. To demonstrate the effectiveness and efficiency of the proposed U-BERT, we compare it with several baselines. We first compare with the embedding-based baseline. It is implemented by simply removing the BERT_cross in U-BERT and only retain the image encoder BERT_img and the text encoder BERT_txt. Since the embedding-based baseline encodes the text and image features individually, it is very efficient for retrieval. However, its retrieval accuracy is not competitive with cross-modal BERT due to a lack of cross-modal attention. As shown in Table 1, the embedding-based baseline only achieves a 59.9 text-to-image (T2I) recall@1 on MSCOCO1K and a 51.6 text-to-image T2I recall@1 on Flickr30K. Then we compare U-BERT with the cross-modal BERT, which adopts the same architecture as Unicode-VL with 12 Transformer layers. Benefited from exploiting the cross-modal attention, the cross-model baseline achieves high retrieval accuracy at the cost of significant time consumption. As shown in Table 1, the cross-modal BERT baseline achieves a 67.5 text-to-image T2I recall@1 on MSCOCO1K dataset and a 59.9 T2I recall@1 on Flickr30K dataset, which are higher than the accuracy achieved by the joint-embedding baseline. But it takes a higher latency than the embedding-based baseline in the retrieval.

Table 1: Comparisons with baseline methods in the text-to-image retrieval task on MSCOCO1K and Flickr30K datasets.

Method	MSCOCO1K R@			Flickr30K R@			Latency
	1	5	10	1	5	10	
Embed	59.9	89.1	94.6	51.6	80.2	87.9	6s
Cross	67.5	91.9	96.9	59.9	85.2	90.8	4832s
Two-stage	67.3	91.8	96.6	59.8	85.1	90.6	102s
U-BERT	67.1	91.5	96.2	59.5	85.0	90.6	31s

Then we compare U-BERT with the two-stage baseline. In the first stage, it adopts the embedding-based method to get an initial ranking list. Then in the second stage, it re-ranks the top 20 retrieved images using the cross-modal BERT with 12 Transformer layers. As shown in Table 1, by exploiting the trade-off between the efficiency and the accuracy, the two-stage baseline achieves higher accuracy than the embedding-based baseline and higher efficiency than the cross-modal BERT baseline. It achieves comparable accuracy with the cross-modal BERT baseline and meanwhile takes much less latency. Our U-BERT also takes the two-stage retrieval strategy. In the first stage, U-BERT only uses the intra-modal components in the image and text features, equivalent to the embedding-based baseline. In the second stage, U-BERT re-ranks the retrieved top 20 retrieved images from the first stage by considering the multi-modal components of the image features and the text features. Note that BERT_C used in U-BERT only contains 3 Transformer layers, whereas the cross-modal BERT used in the two-stage baseline contains 12 Transformer layers. Thus, our U-BERT is more efficient than the two-stage baseline. As shown in Table 1, our U-BERT only takes 31 seconds latency, which achieves an around 3.3× speed-up ratio over the two-stage baseline.

Table 2: Influence of the number of retrieved items for re-ranking (K) on text-to-image retrieval.

K	MSCOCO1K R@			Flickr30K R@			Latency
	1	5	10	1	5	10	
5	66.9	89.1	94.6	59.1	80.2	87.9	12s
10	67.1	91.0	94.6	59.7	84.2	87.9	19s
20	67.1	91.5	96.2	59.5	85.0	90.6	31s
1000	67.0	91.8	96.7	59.6	84.9	90.6	1219s

Influence of K. As shown in Table 2, we evaluate the influence of the number of retrieved items for retrieval, K. We change K among $\{5, 10, 20, 1000\}$. Since MSCOCO1K and Flickr30K datasets contain only 1000 images in the testing split, when $K = 1000$, it is equivalent to re-ranking the whole testing split. Intuitively, as K increases, more images will be re-ranked, bringing higher retrieval accuracy. Meanwhile, more images involved in the re-ranking phase will bring more computational cost as well. As shown in Table 2, when $K = 5$, the proposed U-BERT has achieved satisfied R@1 on both MSCOCO1K and Flickr30K datasets. But it does not influence the recall@5 and recall@10 when $K = 5$ since the re-ranking is only conducted on the top 5 retrieved images. To boost recall@10, K

should be larger than 10. As shown in Table 2, when $K = 20$, it achieves competitive R@5 and R@10 as that when $K = 1000$. By default, we set $K = 20$.

Influence of L_c. We evaluate the influence of L_c, the number of Transformer layers in $\mathrm{BERT}_{\mathrm{cross}}$, on the retrieval accuracy. Straightforwardly, more Transformer layers will lead to higher capability of modeling the text-image relevance. Note that $\mathrm{BERT}_{\mathrm{cross}}$ only serves to generate the complementary residues, $\mathbf{T}_{\mathrm{inter}}$ and $\mathbf{I}_{\mathrm{inter}}$ to enhance the intra-modal image and text features $\mathbf{T}_{\mathrm{intra}}$ and $\mathbf{I}_{\mathrm{intra}}$. Since $\mathbf{T}_{\mathrm{intra}}$ and $\mathbf{I}_{\mathrm{intra}}$ have already been empowered the capability to model the text-image relevance, we do not need to devise a heavy $\mathrm{BERT}_{\mathrm{cross}}$. In fact, a lightweight $\mathrm{BERT}_{\mathrm{cross}}$ is enough to achieve an excellent accuracy. As shown in Table 3, when L_c increases from 1 to 3, the retrieval accuracy increases considerably. But the accuracy saturates as L_c surpasses 3. Taking both accuracy and efficiency into consideration, we set $L_c = 3$ by default.

Table 3: Influence of the number of Transformer layers in $\mathrm{BERT}_{\mathrm{C}}$ (L_{C}) on text-to-image retrieval.

L_{C}	MSCOCO1K R@			Flickr30K R@			Latency
	1	5	10	1	5	10	
1	64.0	90.8	95.9	56.3	82.9	89.3	14s
2	66.4	91.5	96.2	57.0	84.3	89.9	23s
3	67.1	91.5	96.2	59.5	85.0	90.6	31s
12	67.0	91.7	96.1	59.7	85.1	90.6	108s

5.2 Image-to-Text Retrieval

We evaluate the retrieval recall and latency time of our U-BERT in the image-to-text retrieval. For both MSCOCO1K and Flickr30K datasets, in the testing phase, they use $1,000$ image queries to retrieve the relevant texts from a corpus of $5,000$ sentences.

Comparisons with baselines. We first compare U-BERT with the joint-embedding baseline. Due to a lack of cross-modal attention, the joint-embedding baseline might not achieve a competitive retrieval accuracy.

Table 4: Comparisons with baseline methods in the image-to-text retrieval task on MSCOCO1K and Flickr30K datasets.

Method	MSCOCO1K R@			Flickr30K R@			Latency
	1	5	10	1	5	10	
Embed	74.1	94.0	97.5	66.2	87.7	92.9	6s
Cross	79.7	97.2	98.7	76.0	94.4	96.6	4832s
Two-stage	79.6	97.0	98.6	75.5	94.6	97.1	105s
U-BERT	79.1	97.0	98.6	75.6	94.1	96.7	33s

As shown in Table 4, the embedding-based baseline (E) only achieves a 74.1 recall@1 on MSCOCO1K dataset and a 66.2 recall@1 on Flickr30K dataset. Then we compare U-BERT with the cross-modal BERT baseline with 12 Transformer layers. By exploiting the cross-modal attention, the cross-modal BERT baseline achieves high retrieval accuracy. However, it takes quadratic computational complexity, which is prohibitively slow for real applications. As shown in Table 4, the cross-modal BERT baseline achieves a 79.7 recall@1 on MSCOCO1K dataset and a 76.0 recall@1 on Flickr30K dataset. Nevertheless, the cross-modal BERT baseline takes 4832 seconds latency, much slower than the embedding-based baseline with only 6 seconds. Then we compare with the two-stage baseline. As shown in Table 4, by exploiting the trade-off between accuracy and efficiency, the two-stage baseline achieves a comparable accuracy with the cross-modal BERT baseline with only 105 seconds latency. Meanwhile, our U-BERT also takes the two-stage retrieval strategy. Due to using a light-weight $\mathrm{BERT}_{\mathrm{cross}}$, the proposed U-BERT is considerably faster than the two-stage baseline and maintains a high retrieval accuracy as shown in Table 4.

Influence of K. We vary K among $\{25, 50, 100, 5000\}$. Since the testing split only contains 5000 texts in MSCOCO-1K and Flickr30K datasets, $K = 5000$ is equivalent to re-ranking all texts in the corpus. Intuitively, as K increases, more texts will be involved in re-ranking, which tends to lead to higher retrieval recall but brings more computational cost in the re-ranking stage.

Table 5: Influence of the number of retrieved items for re-ranking (K) on image-to-text retrieval.

K	MSCOCO1K R@			Flickr30K R@			Latency
	1	5	10	1	5	10	
25	79.1	97.0	98.4	75.0	93.0	95.4	13s
50	79.1	97.1	98.5	75.5	93.3	96.4	20s
100	79.1	97.0	98.6	75.6	94.1	96.7	33s
5000	79.1	96.9	98.6	75.2	94.2	96.8	1219s

As shown in Table 5, when $K = 25$, on MSCOCO1K dataset, it has achieved competitive retrieval accuracy as that when $K = 5000$. In contrast, on Flickr30K dataset, only when K increases to 100, it achieves a comparable retrieval as that when $K = 5000$. By default, we set $K = 100$ on both MSCOCO1K and Flickr30K datasets in the image-to-text retrieval task. Note that, in the text-to-image retrieval task, we only set $k = 20$. That is, the value of K in the image-to-text retrieval task is 5× as that used in the text-to-image task. This is in accord with the fact that, in the testing split, the total number of texts is 5× as the total number of images.

Table 6: Influence of the number of Transformer layers in $\mathrm{BERT}_{\mathrm{cross}}$ (L_{c}) in the image-to-text retrieval task.

L_{c}	MSCOCO1K R@			Flickr30K R@			Latency
	1	5	10	1	5	10	
1	76.4	96.1	98.3	70.4	91.6	96.0	15s
2	78.5	96.8	98.5	72.3	92.6	95.3	24s
3	79.1	97.0	98.6	75.6	94.1	96.7	33s
12	79.3	97.1	98.7	75.9	94.3	96.9	114s

Influence of L_c. As shown in Table 6, when L_c increases from 1 to 3, the retrieval accuracy increases considerably. For instance, using a single layer, $L_c = 1$, it only achieves a 76.4 recall@1 on MSCOCO1K benchmark and a 70.4 recall@1 on Flickr30K benchmark. When L_c increases to 3, it achieves a 79.1 recall@1 on MSCOCO1K benchmark and a 75.6 recall@1 on Flickr30K benchmark. Meanwhile, the retrieval latency also increases from 15 seconds to 33 seconds. Moreover, the accuracy saturates as L_c surpasses 3. Considering both efficiency and effectiveness, we set $L_c = 3$ by default. It is the same as that used in the T2I retrieval task.

5.3 Pretraining

Pretraining on large-scale datasets can enhance the performance of cross-modal BERT methods [4, 25, 28]. Cross-modal BERT methods devise multiple pretraining tasks, including masked language modeling (MLM), masked region modeling (MRM), and text-image matching. We also investigate the influence of pretraining on the performance of the proposed U-BERT in cross-modal retrieval. Note that the above-mentioned retrieval recall is from the proposed U-BERT without pretraining. Following Unicoder-VL, we utilize two public datasets, SBU Captions [39] and Conceptual Captions [43] to pretrain the proposed U-BERT. Since U-BERT focuses on cross-modal retrieval, we only adopt the triplet loss $\mathcal{L}_1$ and $\mathcal{L}_2$ (in Section 4.3) for pretraining. Afterwards, we finetune the pretrained U-BERT on target datasets, i.e., MSCOCO1K and Flickr30K.

Table 7: Influence of pre-training on the text-to-image (T2I) retrieval task and the image-to-text (I2T) retrieval task.

Task	Pretrain	MSCOCO1K R@			Flickr30K R@		
		1	5	10	1	5	10
T2I		67.1	91.5	96.2	59.5	85.0	90.6
T2I	✓	69.8	92.6	97.0	66.4	89.0	93.3
I2T		79.1	97.0	98.6	75.6	94.1	96.7
I2T	✓	83.8	97.5	98.9	81.2	96.0	98.1

As shown in Table 7, pretraining considerably improves the performance of our U-BERT on both T2I and I2T retrieval tasks. Specifically, for the T2I retrieval task, pretraining improves the recall@1 from 67.1 to 69.8 on MSCOCO1K dataset and from 59.5 to 66.4 on the Flickr30K dataset. Meanwhile, on I2T retrieval task, pretraining improves the recall@1 from 79.1 to 83.8 on MSCOCO1K dataset and from 75.6 to 81.2 on Flickr30K dataset.

5.4 MSCOCO5K

To further demonstrate the effectiveness and efficiency of the proposed U-BERT, we test it on a larger testing split, MSCOCO5K. It has the same training split as the MSCOCO1K benchmark but contains 5, 000 images and 25, 000 texts for testing. The scale of the MSCOCO5K testing split is as 5 times as that of the MSCOCO1K testing split. As shown in Table 8, on the MSCOCO5K testing split, the embedding-based baseline cannot achieve a competitive retrieval recall as the cross-modal BERT baseline and our U-BERT. To be specific, in the text-to-image retrieval task, the embedding-based baseline only achieves a 39.4 recall@1, whereas our U-BERT achieves a 46.2 recall@1. In the image-to-text retrieval task, the embedding-based baseline obtains a 52.6 recall@1 but our U-BERT achieves a 62.2 recall@1. In the meanwhile, our U-BERT achieves a comparable retrieval recall as cross-modal BERT but takes much less latency in the retrieval. To be specific, cross-modal BERT takes around 34 hours (h), whereas our U-BERT only takes approximately 2.5 minutes (m). That is, our U-BERT achieves an around 780× speed-up ratio over the cross-modal BERT baseline.

Table 8: Comparisons among the embedding-based baseline, cross-modal BERT and U-BERT on MSCOCO5K.

Method	T2I R@			I2T R@			Latency
	1	5	10	1	5	10	T2I/I2T
Embed	39.4	70.1	80.8	52.6	81.6	89.0	29s/29s
Cross	46.0	75.0	84.8	62.4	86.6	92.4	34h/34h
U-BERT	46.2	75.1	84.8	62.2	86.6	92.8	2.5m/2.5m

5.5 Comparisons with State-of-the-art Methods

We compare our U-BERT with three types of methods, including embedding-based methods, cross-attention methods, and fast cross-attention methods. Embedding-based methods we compare include VSE++ [7], PSVE [44], SCO [17] and TIMAM [42] and ALIGN [18]. They are very fast in retrieval. Except for ALIGN, they cannot achieve competitive retrieval accuracy as cross-attention methods as shown in Table 9. To be specific, the best embedding-based method, TIMAM only achieves a 43.6 T2I R@1 on Flickr30K dataset, which is outperformed by one of the earliest cross-attention methods, SCAN [24]. It is worth noting that the excellence of ALIGN is owing to being pretrained on a huge dataset of 1.8 billion text-image pairs. Thus, it is unfair to directly compare it with other methods which are only pre-trained on SBU Captions and Conceptual Captions datasets with only 3 million text-image pairs.

Then we compare U-BERT with several cross-attention methods. We divide them into two groups. The first group of methods simply adopt the soft-attention operation, self-attention operation, or graph convolution to exploit the cross-modal attention including SCAN [24], ACMM [15], VSRN [26], IMRAM [3], CSVE [50], and MMnas [64]. Benefited from exploiting the cross-modal attention, they achieve higher retrieval accuracy than embedding-based methods at the cost of quadratic computational cost. The second group of methods include Uni-VL [25], UNITER [25], OSCAR [28], VILLA [10], UNIMO [27] and ERNIE-ViL [56]. Inspired by the success of BERT, they exploit not only cross-modal attention but also devise several pre-training tasks for improving accuracy. As shown in Table 9, these methods significantly outperforms the embedding-based methods and achieve state-of-the-art cross-modal retrieval accuracy. Nevertheless, as we mentioned previously, in the real large-scale retrieval scenario, they are prohibitively slow due to quadratic computational complexity.

At last, we compare our U-BERT with two-stage methods including LigDOT [47] and CO$^{\text{oscar}}$ [11]. By exploiting the trade-off between efficiency and accuracy, they achieve higher retrieval speed and comparable retrieval accuracy as cross-modal BERT methods. As shown in Table 9, our U-BERT cannot achieve as high accuracy as LigDOT [47] and CO$^{\text{oscar}}$ [11]. The higher accuracies of LigDOT [47] and CO$^{\text{oscar}}$ [11] are mainly owing to the fact that they additionally exploit the tags as OSCAR, whereas we only use the bounding box features from the object detector. But our U-BERT achieves higher efficiency than LightningDOT [47] and CO$^{\text{oscar}}$ [11]. To be specific, LightningDOT [47] only achieves a 48× speed-up ratio over the cross-modal BERT on Flickr30K dataset. In contrast, our U-BERT achieves a 130× speed-up ratio. Note that, both LightningDOT [47] and CO$^{\text{oscar}}$ [11] outperform our U-BERT. This is due to the fact that they use a much better cross-modal BERT in the re-ranking phase than ours. Specifically, the re-ranking model used in LightningDOT is UNITER or OSCAR pretrained on several large-scale datasets with multiple devised pretraining tasks. In

Table 9: Comparisons with state-of-the-art methods. We compare the proposed U-BERT model with embedding-based (Embed) methods, cross-modal BERT (Cross) methods and two-stage (Two) methods.

Method	Type	MSCOCO1K						Flickr30K						MSCOCO5K					
		T2I R@			I2T R@			T2I R@			I2T R@			T2I R@			I2T R@		
		1	5	10	1	5	10	1	5	10	1	5	10	1	5	10	1	5	10
VSE++ [7]	Embed	52.0	83.1	92.0	64.6	89.1	95.7	39.6	70.1	79.5	52.9	80.5	87.2	30.3	59.1	72.4	41.3	69.2	81.2
PSVE [44]	Embed	55.2	86.5	93.7	69.2	91.6	96.6	–	–	–	–	–	–	32.4	63.0	75.0	45.2	74.3	84.5
SCO [17]	Embed	56.7	87.5	94.8	69.9	92.9	97.5	41.1	70.5	81.1	55.5	82.0	89.3	33.1	62.9	75.5	42.8	72.3	83.0
TIMAM [42]	Embed	–	–	–	–	–	–	42.6	71.6	81.9	53.1	78.8	87.6	–	–	–	–	–	–
ALIGN [18]	Embed	–	–	–	–	–	–	84.9	97.4	98.6	95.3	99.8	100.0	59.9	83.3	89.8	77.0	93.5	96.9
SCAN [24]	Cross	58.8	89.0	93.1	72.7	94.8	98.4	48.6	77.7	85.2	67.4	90.3	95.8	34.4	63.7	75.7	46.4	77.4	87.2
ACMM [15]	Cross	58.2	87.3	93.9	81.9	98.0	99.3	50.2	76.8	84.7	80.0	95.5	98.2	36.7	65.1	76.7	63.5	88.0	93.6
VSRN [26]	Cross	62.8	89.7	95.1	76.2	94.8	98.2	54.7	81.8	88.2	71.3	90.6	96.0	40.5	70.6	81.1	53.0	81.1	89.4
MMCA [53]	Cross	61.6	89.8	95.2	74.8	95.6	97.7	54.8	81.4	87.8	74.2	92.8	96.4	38.7	69.7	80.8	54.0	82.5	90.7
IMRAM [3]	Cross	61.7	89.1	95.0	76.7	95.6	98.5	53.9	79.4	87.2	74.1	93.0	96.6	39.7	69.1	79.8	53.7	83.2	91.0
CSVE [50]	Cross	59.9	89.4	95.2	74.8	95.1	98.3	52.9	80.4	87.8	73.5	92.1	95.8	–	–	–	–	–	–
MMnas [64]	Cross	–	–	–	–	–	–	60.7	85.1	90.5	78.3	94.6	97.4	–	–	–	–	–	–
VSparta [34]	Cross	68.2	91.8	96.3	–	–	–	57.4	82.0	88.1	–	–	–	44.4	72.8	82.4	–	–	–
Uni-VL [25]	Cross	69.7	93.5	97.2	84.3	97.3	99.3	71.5	90.9	94.9	86.2	96.3	99.0	46.7	76.0	85.3	62.3	87.1	92.8
ERNIE-ViL [56]	Cross	–	–	–	–	–	–	74.4	92.7	95.9	86.7	97.8	99.0	–	–	–	–	–	–
UNIMO [27]	Cross	–	–	–	–	–	–	74.6	93.4	96.0	89.7	98.4	99.1	–	–	–	–	–	–
ViLT [22]	Cross	–	–	–	–	–	–	61.9	86.8	92.8	81.4	95.6	97.6	41.3	72.0	82.5	61.8	86.2	92.6
UNITER [25]	Cross	–	–	–	–	–	–	72.5	92.3	95.9	85.9	97.1	98.8	50.3	78.5	87.2	64.4	87.4	93.1
OSCAR [28]	Cross	**75.7**	**95.2**	**98.3**	**88.4**	**99.1**	**99.8**	–	–	–	–	–	–	**57.5**	**82.8**	89.8	73.5	92.2	**96.0**
VILLA [10]	Cross	–	–	–	–	–	–	74.7	92.9	95.8	86.6	**97.9**	**99.2**	–	–	–	–	–	–
LigDOT [47]	Two	–	–	–	–	–	–	72.6	93.1	96.1	86.5	97.5	98.9	57.4	82.7	**89.9**	**74.2**	**92.4**	**96.0**
CO$^{\text{oscar}}$ [11]	Two	–	–	–	–	–	–	**76.4**	**93.6**	**96.2**	**89.4**	97.7	99.0	54.7	81.3	88.9	70.8	91.0	95.2
U-BERT	Ours	69.8	92.6	97.0	83.8	97.5	98.9	66.4	89.0	93.3	81.2	96.0	98.1	46.2	75.1	84.8	62.2	86.6	92.8

Table 10: Comparisons with LightningDOT [47] and CO$^{\text{oscar}}$ [11] on latency time.

Method	MSCOCO1K	MSCOCO5K
LightningDOT [47]/CO$^{\text{oscar}}$ [11]	21s	16m
U-BERT (ours)	3.3s	2.5m

the meanwhile, the re-ranking model used in CO$^{\text{oscar}}$ [11] is OSCAR [28] which exploits the additional tags besides the visual features. In contrast, our U-BERT is only pretrained on SBU Captions and Conceptual Captions datasets. Meanwhile, we do not exploit the tags used in OSCAR. We also compare latency time with LightningDOT [47] and CO$^{\text{oscar}}$ [11] on MSCOCO1K and MSCOCO5K datasets and the results are in Table 10. Since LightningDOT and CO$^{\text{oscar}}$ adopt the same architecture, we merge them in the table. As shown in the table, LightningDOT/CO$^{\text{oscar}}$ takes 21 seconds on MSCOCO1K and 16 minutes on MSCOCO5K. In contrast, our U-BERT only takes 3.3 seconds on MSCOCO1K and 2.5 minutes on MSCOCO5K. That is, our U-BERT achieves an around 6× speed-up ratio over the compared LightningDOT/CO$^{\text{oscar}}$ model. The faster inference time of U-BERT is because our U-BERT only stacks 3 Transformer layers for re-ranking whereas LightningDOT and CO$^{\text{oscar}}$ adopt 12 Transformer layers for re-ranking. Meanwhile, the OSCAR model used in LightningDOT and CO$^{\text{oscar}}$ additionally takes bounding box labels as input, leading to a longer input sequence than our U-BERT without exploiting bounding box labels. The significant efficiency improvement can considerably improve the user experience in practical applications.

6 CONCLUSION

In this work, we propose a U-BERT for fast and scalable text-image retrieval in practical applications. It decomposes the text as well as image feature into an intra-modal component and an inter-modal component. The intra-modal component of the text feature and the image feature is obtained from the text/image encoder exploiting single-modality features. They follow the spirit of the global feature used in the embedding-based method and thus support efficient cross-modal retrieval. Meanwhile, the inter-modal component of the text and image features is obtained by exploiting the cross-modal attention on the features from both image and text modalities. They serve as the complementary residue to enhance the discriminating power of the intra-modal components. U-BERT is deployed in a two-stage retrieval pipeline. In the first stage, only the intra-modal components are utilized to retrieve a small set of relevant candidates for re-ranking efficiently. In the second stage, we compute the inter-modal components for the retrieved candidates for a higher retrieval accuracy. Benefited from the two-stage configuration, our U-BERT is faster and more scalable than the mainstream cross-modal BERT methods. Systematic experiments conducted on three public benchmarks including MSCOCO1K, MSCOCO5K, and Flickr30K demonstrate the high effectiveness and efficiency of the proposed U-BERT model.

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Salt Lake City, UT, 6077–6086.

[2] Yue Cao, Mingsheng Long, Jianmin Wang, Qiang Yang, and Philip S. Yu. 2016. Deep Visual-Semantic Hashing for Cross-Modal Retrieval. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. San Francisco, CA, 1445–1454.

[3] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. 2020. IMRAM: Iterative Matching With Recurrent Attention Memory for Cross-Modal Image-Text Retrieval. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, 12652–12660.

[4] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: UNiversal Image-TExt Representation Learning. In *Proceedings of the 16th European Conference on Computer Vision (ECCV), Part XXX*. Glasgow, UK, 104–120.

[5] Jaemin Cho, Jiasen Lu, Dustin Schwenk, Hannaneh Hajishirzi, and Aniruddha Kembhavi. 2020. X-LXMERT: Paint, Caption and Answer Questions with Multi-Modal Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online, 8785–8805.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. Minneapolis, MN, 4171–4186.

[7] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. In *Proceedings of the British Machine Vision Conference (BMVC)*. Newcastle, UK.

[8] Hongliang Fei, Tan Yu, and Ping Li. 2021. Cross-lingual Cross-modal Pretraining for Multimodal Retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. Online, 3644–3650.

[9] Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomás Mikolov. 2013. DeViSE: A Deep Visual-Semantic Embedding Model. In *Advances in Neural Information Processing Systems (NIPS)*. Lake Tahoe, NV, 2121–2129.

[10] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-Scale Adversarial Training for Vision-and-Language Representation Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*. virtual.

[11] Gregor Geigle, Jonas Pfeiffer, Nils Reimers, Ivan Vulić, and Iryna Gurevych. 2021. Retrieve Fast, Rerank Smart: Cooperative and Joint Approaches for Improved Cross-Modal Retrieval. *arXiv preprint arXiv:2103.11920* (2021).

[12] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. 2014. A Multi-View Embedding Space for Modeling Internet Images, Tags, and Their Semantics. *Int. J. Comput. Vis.* 106, 2 (2014), 210–233.

[13] David R. Hardoon, Sándor Szedmák, and John Shawe-Taylor. 2004. Canonical Correlation Analysis: An Overview with Application to Learning Methods. *Neural Comput.* 16, 12 (2004), 2639–2664.

[14] Yan Huang, Yang Long, and Liang Wang. 2019. Few-Shot Image and Sentence Matching via Gated Visual-Semantic Embedding. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI)*. Honolulu, HI, 8489–8496.

[15] Yan Huang and Liang Wang. 2019. ACMM: Aligned Cross-Modal Memory for Few-Shot Image and Sentence Matching. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul, Korea, 5773–5782.

[16] Yan Huang, Wei Wang, and Liang Wang. 2017. Instance-Aware Image and Sentence Matching with Selective Multimodal LSTM. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI, 7254–7262.

[17] Yan Huang, Qi Wu, Chunfeng Song, and Liang Wang. 2018. Learning Semantic Concepts and Order for Image and Sentence Matching. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Salt Lake City, UT, 6163–6171.

[18] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*. Virtual Event, 4904–4916.

[19] Xiaowei Jia, Handong Zhao, Zhe Lin, Ajinkya Kale, and Vipin Kumar. 2020. Personalized Image Retrieval with Sparse Graph Representation Learning. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*. Virtual Event, CA, 2735–2743.

[20] Rong Jin. 2020. Large-scale Multi-modal Search and QA at Alibaba. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval (SIGIR)*. Virtual Event, China, 8.

[21] Andrej Karpathy and Fei-Fei Li. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA, 3128–3137.

[22] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*. Virtual Event, 5583–5594.

[23] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *Int. J. Comput. Vis.* 123, 1 (2017), 32–73.

[24] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked Cross Attention for Image-Text Matching. In *Proceedings of the 15th European Conference on Computer Vision (ECCV), Part IV*. Munich, Germany, 212–228.

[25] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-VL: A Universal Encoder for Vision and Language by Cross-Modal Pre-Training. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*. New York, NY, 11336–11344.

[26] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2019. Visual Semantic Reasoning for Image-Text Matching. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul, Korea, 4653–4661.

[27] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021. UNIMO: Towards Unified-Modal Understanding and Generation via Cross-Modal Contrastive Learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics.

[28] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. OSCAR: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In *Proceedings of the 16th European Conference on Computer Vision (ECCV), Part XXX*. Glasgow, UK, 121–137.

[29] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Proceedings of the 13th European Conference on Computer Vision (ECCV), Part V*. Zurich, Switzerland, 740–755.

[30] Haoliang Liu, Tan Yu, and Ping Li. 2021. Inflate and Shrink:Enriching and Reducing Interactions for Fast Text-Image Retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[31] Jiaheng Liu, Tan Yu, Hanyu Peng, Mingming Sun, and Ping Li. 2022. Cross-Lingual Cross-Modal Consolidation for Effective Multilingual Video Corpus Moment Retrieval. In *Findings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

[32] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*. Vancouver, Canada, 13–23.

[33] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-Task Vision and Language Representation Learning. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition CVPR)*. Seattle, WA, 10434–10443.

[34] Xiaopeng Lu, Tiancheng Zhao, and Kyusong Lee. 2021. VisualSparta: An Embarrassingly Simple Approach to Large-scale Text-to-Image Search with Weighted Bag-of-words. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL/IJCNLP)*. Virtual Event, 5020–5029.

[35] Shaunak Mishra, Mikhail Kuznetsov, Gaurav Srivastava, and Maxim Sviridenko. 2021. VisualTextRank: Unsupervised Graph-based Content Extraction for Automating Ad Text to Image Search. In *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, Feida Zhu, Beng Chin Ooi, and Chunyan Miao (Eds.). ACM, 3404–3413.

[36] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual Attention Networks for Multimodal Reasoning and Matching. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI, 2156–2164.

[37] Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Dongdong Zhang, and Nan Duan. 2021. M3P: Learning Universal Representations via Multitask Multilingual Multimodal Pre-Training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. virtual, 3977–3986.

[38] Yixin Nie, Linjie Li, Zhe Gan, Shuohang Wang, Chenguang Zhu, Michael Zeng, Zicheng Liu, Mohit Bansal, and Lijuan Wang. 2021. MLP Architectures for Vision-and-Language Modeling: An Empirical Study. *arXiv preprint arXiv:2112.04453* (2021).

[39] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2Text: Describing Images Using 1 Million Captioned Photographs. In *Advances in Neural Information*

Processing Systems (NIPS). Granada, Spain, 1143–1151.

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*. Virtual Event, 8748–8763.

[41] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems (NIPS)*. Montreal, Canada, 91–99.

[42] Nikolaos Sarafianos, Xiang Xu, and Ioannis A. Kakadiaris. 2019. Adversarial Representation Learning for Text-to-Image Matching. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul, Korea, 5813–5823.

[43] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*. Melbourne, Australia, 2556–2565.

[44] Yale Song and Mohammad Soleymani. 2019. Polysemous Visual-Semantic Embedding for Cross-Modal Retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, 1979–1988.

[45] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*. Addis Ababa, Ethiopia.

[46] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. VideoBERT: A Joint Model for Video and Language Representation Learning. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul, Korea, 7463–7472.

[47] Siqi Sun, Yen-Chun Chen, Linjie Li, Shuohang Wang, Yuwei Fang, and Jingjing Liu. 2021. LightningDOT: Pre-training Visual-Semantic Embeddings for Real-Time Image-Text Retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. Online.

[48] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China, 5099–5110.

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems (NIPS)*. Long Beach, CA, 5998–6008.

[50] Haoran Wang, Ying Zhang, Zhong Ji, Yanwei Pang, and Lin Ma. 2020. Consensus-Aware Visual-Semantic Embedding for Image-Text Matching. In *Proceedings of the 6th European Conference on Computer Vision (ECCV), Part XXIV*. 18–34.

[51] Wei Wang, Beng Chin Ooi, Xiaoyan Yang, Dongxiang Zhang, and Yueting Zhuang. 2014. Effective Multi-Modal Retrieval based on Stacked Auto-Encoders. *Proc. VLDB Endow.* 7, 8 (2014), 649–660.

[52] Wei Wang, Xiaoyan Yang, Beng Chin Ooi, Dongxiang Zhang, and Yueting Zhuang. 2016. Effective deep learning-based multi-modal retrieval. *VLDB J.* 25, 1 (2016), 79–101.

[53] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. 2020. Multi-Modality Cross Attention Network for Image and Sentence Matching. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, 10938–10947.

[54] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*. Lille, France, 2048–2057.

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

[56] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE-ViL: Knowledge Enhanced Vision-Language Representations through Scene Graphs. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI)*. Virtual Event, 3208–3216.

[57] Tan Yu, Hongliang Fei, and Ping Li. 2022. Cross-Probe BERT for Fast Cross-Modal Search. In *Proceedings of the 45th International ACM SIGIR conference on research and development in Information Retrieval (SIGIR)*.

[58] Tan Yu, Xuemeng Yang, Yan Jiang, Hongfang Zhang, Weijie Zhao, and Ping Li. 2021. TIRA in Baidu Image Advertising. In *Proceedings of the 37th IEEE International Conference on Data Engineering (ICDE)*. Chania, Greece, 2207–2212.

[59] Tan Yu, Yi Yang, Hongliang Fei, Yi Li, Xiaodong Chen, and Ping Li. 2021. Assorted Attention Network for Cross-Lingual Language-to-Vision Retrieval. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM)*. Virtual Event.

[60] Tan Yu, Yi Yang, Yi Li, Xiaodong Chen, Mingming Sun, and Ping Li. 2020. Combo-Attention Network for Baidu Video Advertising. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*. Virtual Event, CA, 2474–2482.

[61] Tan Yu, Yi Yang, Yi Li, Xiaodong Chen, Mingming Sun, and Ping Li. 2020. Combo-Attention Network for Baidu Video Advertising. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*. Virtual Event, CA, USA, 2474–2482.

[62] Tan Yu, Yi Yang, Yi Li, Lin Liu, Hongliang Fei, and Ping Li. 2021. Heterogeneous Attention Network for Effective and Efficient Cross-modal Retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. Virtual Event, Canada, 1146–1156.

[63] Tan Yu, Yi Yang, Yi Li, Lin Liu, Mingming Sun, and Ping Li. 2021. Multi-modal Dictionary BERT for Cross-modal Video Search in Baidu Advertising. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM)*. Virtual Event.

[64] Zhou Yu, Yuhao Cui, Jun Yu, Meng Wang, Dacheng Tao, and Qi Tian. 2020. Deep Multimodal Neural Architecture Search. In *Proceedings of the 28th ACM International Conference on Multimedia (MM)*. Virtual Event / Seattle, WA, 3743–3752.

[65] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. VinVL: Revisiting Visual Representations in Vision-Language Models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. virtual, 5579–5588.