

LOOPITR: Combining Dual and Cross Encoder Architectures for Image-Text Retrieval

Jie Lei¹, Xinlei Chen², Ning Zhang², Mengjiao Wang²,
Mohit Bansal¹, Tamara L. Berg², Licheng Yu²

¹UNC Chapel Hill ²Meta AI

{jielei, mbansal}@cs.unc.edu

{xinleic, ningzhang, mengjiaow, tlberg, lichengyu}@fb.com

Abstract

Dual encoders and cross encoders have been widely used for image-text retrieval. Between the two, the dual encoder encodes the image and text independently followed by a dot product, while the cross encoder jointly feeds image and text as the input and performs dense multi-modal fusion. These two architectures are typically modeled separately without interaction. In this work, we propose **LOOPITR**, which combines them in the same network for joint learning. Specifically, we let the **dual encoder** provide hard negatives to the **cross encoder**, and use the more discriminative cross encoder to distill its predictions back to the dual encoder. Both steps are efficiently performed together in the same model. Our work centers on empirical analyses of this combined architecture, putting the main focus on the design of the distillation objective. Our experimental results highlight the benefits of training the two encoders in the same network, and demonstrate that distillation can be quite effective with just a few hard negative examples. Experiments on two standard datasets (Flickr30K and COCO) show our approach achieves state-of-the-art dual encoder performance when compared with approaches using a similar amount of data.

1. Introduction

There are two widely used architectures for image-text retrieval: (i) **dual encoder**, with two separate streams (weights might be shared [22, 23, 43]) that encodes the image and text modalities independently, and the matching score is obtained via a simple dot product [5, 20, 29, 54]; and (ii) **cross encoder**, which takes the joint sequence of image and text as input, and performs dense cross-modal interactions (e.g., cross-attention [62]) in a single stream. The final image-text matching score is often predicted with a classifier [10, 42, 44]. These two types of encoders are typically learned separately without interaction.

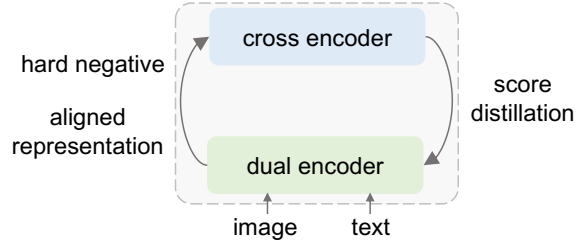


Figure 1. LOOPITR architecture for image-text retrieval. Dual encoder and cross encoder architectures are typically learned separately, whereas in LOOPITR, we combine them together in the same network via a loop interaction: (i) *bottom-up*, dual encoder provides aligned representations and hard negatives for cross encoder learning; (ii) *up-down*, cross encoder’s scores are used as dual encoder’s soft label for additional supervision. With this loop structure, we observe a performance gain for both encoders.

In this work, we propose **LOOPITR** (pronounced similarly to *Jupiter*), a new model that integrates the dual and cross encoder architectures into a single network via a loop-style structure for image-text retrieval, as in Figure 1. Given a pair of image and text, we first feed both inputs into the dual encoder, independently encoding them into their respective embedding sequences, and then apply a cross encoder on top of the two encoded sequences. Rather than only injecting the supervision on the very top of the model as [40, 49, 60], we make both encoders learn to align the image and text modalities. The dual encoder is trained with a contrastive loss, which has shown to be good at learning representations [6, 9, 24, 54]. Meanwhile, we also use the dual encoder to mine in-batch hard negatives for cross encoder training as [40]. Different from previous work [10] which requires additional offline computation to mine the hard negatives at every epoch, our approach does not introduce any extra training cost, as the hard negatives are readily available from the dual encoder’s contrastive scores. On the other hand, the fine-grained alignment information from the cross encoder also benefits the learning of the dual

encoder – more informative signals flow back to the dual encoder via back-propagation. Empirically, we show the two encoders mutually benefit from joint training.

Cross encoders tend to capture richer information than dual encoders due to their deeper cross-modal interaction. Thus, we further add a **distillation loss** to the architecture by distilling the image-text matching scores from the cross encoder to the dual encoder. As distillation for image-text retrieval has rarely been explored, in this work, we put our focus on the designs of this distillation objective and present comprehensive ablations around it. We first examine ~~how to construct the score distributions for the distillation between the two architectures~~. The standard distillation objective [25] is designed for classification tasks and requires a teacher score distribution as input. To extend the distillation for image-text retrieval, we follow [50] to construct a set of negatives for each positive pair. In [50], all in-batch negatives are used, and only a single distribution for each image-text pair is constructed (from a text to its image negatives). In our study, we find that distillation can be effective with *a few hard negatives*, and that both image negatives and text negatives are useful. These findings lead to a more efficient distillation objective for image-text retrieval.

Our second investigation is on ~~the type of teacher~~. By default, we use an **online cross encoder teacher** that learns along with the **dual encoder student** in the same network. We compare it with two other designs, including using an **offline trained model** [50] and a **momentum model** [24]. We observe all of these teachers work reasonably well, but the simplest online teacher performs the best. Besides, using an online teacher saves both memory (as no extra model storage is needed) and computation cost (by reusing online computation results from cross encoder training). This also simplifies the distillation process into a single run, eliminating the need to train a teacher model separately [25, 50].

Most existing image-text retrieval methods [10, 44] adopt a two-stage pipeline, pre-training followed by fine-tuning. Thus, we also study when to perform the distillation in these two stages. Experimental results suggest that distillation in pre-training gives more performance gain than that in fine-tuning, and the best performance is achieved by applying distillation in both stages.

Overall, our contributions are: (i) LOOPITR, a unified architecture that combines dual and cross encoders in the same model, with an efficient online distillation objective that improves the dual encoder for image-text retrieval; (ii) We present a comprehensive ablation of LOOPITR, with a focus on the cross-to-dual distillation objective design.

2. Related Work

Language-based Retrieval has been widely studied in various scenarios, e.g., document or passage retrieval [32, 51, 52], image retrieval [5, 10, 17, 29, 35, 54], video re-

trieval [37, 46, 64], and moment retrieval [1, 38, 41]. Particularly, for image-text retrieval, there are two main-stream methods: (i) *dual encoder*, where the input text queries and the images are encoded separately into dense vectors of the same visual semantic embedding (VSE) space using two encoders [5, 17, 22, 23, 29, 35, 50, 54]. This family of models runs fast, as the image and text similarity is simply computed via a dot product between their embeddings; and (ii) *cross encoder* [10, 21, 39, 42, 44, 49, 67], where the input text and image pair are jointly encoded by a cross encoder architecture, e.g., a transformer encoder [62]. While they are often treated as separate architectures, in this work we study a unified architecture that jointly learns both together. Our work shares a similar model architecture as [40], where both have the dual encoders as the base and a cross encoder on top. However, [40] primarily aims to improve the cross encoder, while we lay more focus on utilizing the more powerful cross encoder to improve the fast dual encoder. We provide a comprehensive analyses around this design.

Distillation [4, 25] is a technique for transferring knowledge from a relatively more powerful but complex model (or an ensemble of models) to a simpler one. This has been widely studied in vision [3, 19, 26, 61], language [28, 30, 34, 55, 59, 66], and multi-modal [18, 50, 63] domains. The students in most of these work learns from an offline trained teacher model. As comparison, we explore using an online teacher [57, 68] that is trained in the same network together with the student. Notably most previous works share the same prediction format between the teacher and student (e.g., both are classification scores). However, in our unified model the cross encoder teacher produces a two-way classification score while the dual encoder student produces a cosine similarity score, causing discrepancy for distillation. Thus, we propose an additional simple but important pre-processing step before distillation.

3. Method

In this section, we describe our proposed unified architecture, LOOPITR. Figure 2 shows a conceptual comparison of it to dual and cross encoders.

Dual Encoder typically consists of two encoders, an image encoder \mathcal{F} that transforms the input image x into a visual feature sequence $\mathcal{F}(x)$, and a text encoder \mathcal{G} that transforms the text sentence y into a textual feature sequence $\mathcal{G}(y)$. Two projection heads (with pooling) $\phi_{\mathcal{F}}$ and $\phi_{\mathcal{G}}$ are applied to get a vector representation for image and text respectively: $\phi_{\mathcal{F}}(\mathcal{F}(x))$ and $\phi_{\mathcal{G}}(\mathcal{G}(y))$. The similarity score of image x and text y is measured by dot product:

$$s_{x,y} = \phi_{\mathcal{F}}(\mathcal{F}(x))^T \phi_{\mathcal{G}}(\mathcal{G}(y)). \quad (1)$$

For brevity, we omit the notations of $\phi_{\mathcal{F}}$ and $\phi_{\mathcal{G}}$, and denote the similarity as $s_{x,y} = \mathcal{F}(x)^T \mathcal{G}(y)$. The goal of the

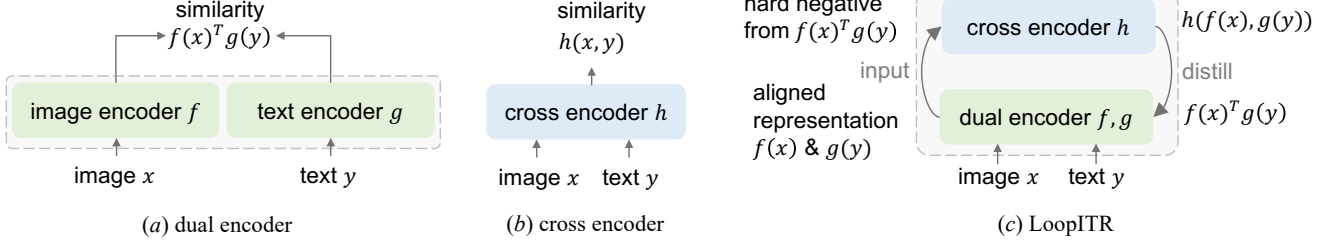


Figure 2. Comparison of LOOPITR to cross encoder and dual encoder architectures. (a): dual encoder separately encodes image x and text y into their respective sequence representations $f(x)$ and $g(y)$, and then produces a similarity score as the dot product their pooled representations $\phi(f(x))^T \phi(g(y))$ (ϕ is a pooling function, we omit it in the figure for brevity). (b): cross encoder h takes image x and text y jointly as inputs, and perform cross-modal interaction (e.g., cross-attention) and then produces a similarity score $h(x, y)$. Cross encoders are typically more powerful than dual encoders as they can perform more fine-grained image-text matching, while dual encoders often run much faster than cross encoders as the image/text representations can be pre-computed, and the similarity scores are computed with a simple dot product. (c): LOOPITR that combines these two, in which the bottom dual encoder provides aligned representations and hard negatives as inputs the cross encoder, and the cross encoder gives additional supervision to dual encoder via distillation.

dual encoder is to bring the matched image-text pairs in the embedding space closer than the unmatched ones. This is often achieved by using a **image-text contrastive (ITC) loss**:

$$p_i^{i2t} = \frac{\exp(s_{x_i, y_i} / \tau)}{\sum_j \exp(s_{x_i, y_j} / \tau)}, \quad p_i^{t2i} = \frac{\exp(s_{x_i, y_i} / \tau)}{\sum_j \exp(s_{x_j, y_i} / \tau)}, \quad (2)$$

$$\mathcal{L}_{itc} = - \sum_{i=1}^n (\log p_i^{i2t} + \log p_i^{t2i}), \quad (3)$$

where τ is a learned temperature parameter.

Cross Encoder. While the dual encoder runs fast in retrieval, it only applies an extremely shallow interaction between image and text (dot product), which in turn limits its performance. As a comparison, the cross encoder tackles the issue by using multiple layers of cross attention [62] for more fine-grained alignment between the input image patches and text tokens. Specifically, given a cross encoder \mathcal{H} , it takes the image x and text y as inputs, and output **classification scores** (two classes, positive and negative) $\mathbf{h}(x, y) = [h_{x,y}^{pos}, h_{x,y}^{neg}] = \mathcal{H}(x, y) \in \mathbb{R}^2$, where $h_{x,y}^{pos}$ denotes the classification score of the image-text pair (x, y) being the positive pair (in the positive class), and $h_{x,y}^{neg}$ being the negative. A softmax normalization is then applied $\mathbf{p}_{x,y}^{itm} = \text{softmax}(\mathbf{h}(x, y))$. The model is trained using an **image-text matching (ITM) objective**:

$$\mathcal{L}_{itm} = \sum_{i=1}^n (-\log \mathbf{p}_{x_i, y_i}^{itm}[0] + \log \mathbf{p}_{x_i, y_i}^{itm}[1] + \log \mathbf{p}_{\hat{x}_i, y_i}^{itm}[1]), \quad (4)$$

where $\mathbf{p}_{x_i, y_i}^{itm}$ is the two-class probabilities for the i -th image-text pair (x_i, y_i) . \hat{y}_i is a negative text w.r.t. image x_i and \hat{x}_i is a negative image w.r.t. text y_i . The negatives can be randomly sampled [33, 44, 49] or mined from hard negatives based on similarity scores [10, 40, 49]. In addition, during pre-training, cross encoders are often trained with a **Masked Language Model (MLM) loss** [14] \mathcal{L}_{mlm} .

LOOPITR: Combined Dual Encoder and Cross Encoder. Dual encoders and cross encoders are often seen as separate architectures. In LOOPITR (shown in Figure 3), we combine them together. Specifically, we use a dual encoder to produce aligned representations as the inputs to the cross encoder, i.e., the cross encoder’s matching scores are now computed as $h(x, y) = \mathcal{H}(\mathcal{F}(x), \mathcal{G}(y))$. It has been shown in [40] that such aligned representations help improve the cross encoder. The negatives in the cross encoder’s ITM objective \mathcal{L}^{itm} in Equation 4 are mined on-line [40] based on the dual encoder’s image-text contrastive scores.

Due to deeper interaction, the cross encoder contains more accurate matching signals than the dual encoder. We propose to distill the knowledge from the cross encoder teacher to the dual encoder student. The standard distillation objective [25] is defined for classification tasks, where a teacher probability distribution $\mathbf{q} \in \mathbb{R}^c$ over the c classes are used as soft labels for learning student probability distribution $\mathbf{p} \in \mathbb{R}^c$, via cross entropy $H(\mathbf{p}, \mathbf{q})$. However, such distribution is not directly available for image-text retrieval.

To address this issue, we construct an **image negative set** and a **text negative set**. Specifically, given a matched image-text pair (x_i, y_i) , we take m in-batch text hard negatives and pair it with the image $\{(x_i, \hat{y}_i)\}$, and denote $\mathcal{N}_i^{txt} = \{(x_i, y_i)\} \cup \{(x_i, \hat{y}_i)\}$ with $|\mathcal{N}_i^{txt}| = m + 1$ as the set of image-text pairs with the text hard negatives. Thus we can define the predicted probability distribution $\mathbf{p}_i^{txt} \in \mathbb{R}^{m+1}$ over \mathcal{N}_i^{txt} from the dual encoder, with $p_{i,k}^{txt}$ being its k -th element, denoting the probability of the k -th pair in \mathcal{N}_i^{txt} being the matched pair. $p_{i,k}^{txt}$ is computed as:

$$p_{i,k}^{txt} = \frac{\exp(s_{x_i, y_k} / \tau)}{\sum_{(x,y) \in \mathcal{N}_i^{txt}} \exp(s_{x,y} / \tau)}. \quad (5)$$

Note that $s_{x,y}$ is the matching score for image-text pair

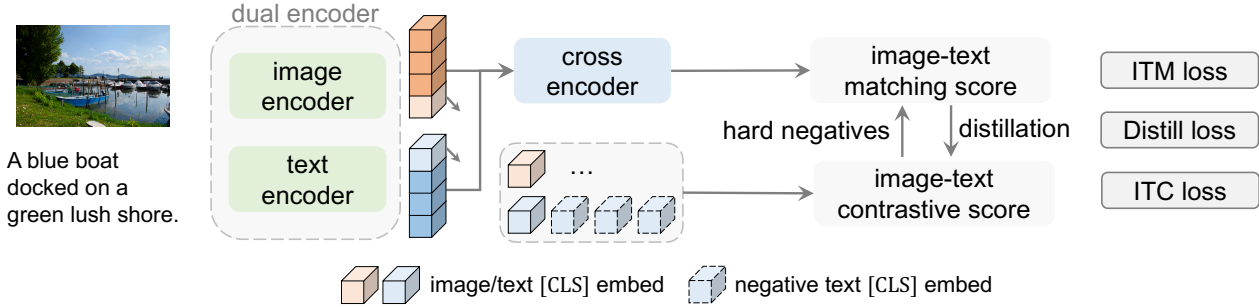


Figure 3. Our implementation of LOOPITR. It consists of an *image encoder* and a *text encoder* as its *dual encoder*, and a *cross encoder* on the top. All the encoders are implemented as transformers. The dual encoder encodes input image and text into their respective embedding sequences, and computes Image-Text Contrastive (ITC) scores via dot product of the [CLS] embeddings. The embedding sequences are also forwarded into the cross encoder to compute Image-Text Matching (ITM) scores of positive and in-batch hard negative (from ITC) image-text pairs. As cross encoder is typically more powerful than the dual encoder, we distill the knowledge from the ITM scores to the ITC scores via an online distillation loss. For brevity, MLM loss is not shown in this figure. See details in Section 3.

(x, y) from the dual encoder, and τ is the same temperature parameter used in Equation 2. For the cross encoder teacher, we replace $s(x, y)$ with $h_{x,y}^{pos}$:

$$q_{i,k}^{txt} = \frac{\exp(h_{x_k, y_k}^{pos}/\tau)}{\sum_{(x,y) \in \mathcal{N}_i^{txt}} \exp(h_{x,y}^{pos}/\tau)}, \quad (6)$$

where $q_{i,k}^{txt}$ is the k -th element of the cross encoder score distribution $\mathbf{q}_i^{txt} \in \mathbb{R}^{m+1}$. The loss is calculated as:

$$\mathcal{L}_{distill}^{txt} = \sum_{i=1}^n H(\mathbf{p}_i^{txt}, \text{stop-grad}(\mathbf{q}_i^{txt})). \quad (7)$$

Since our teacher and student network are trained together within the same model, we use stop-grad to indicate stopping gradients from back-propagating to \mathbf{q}_i^{txt} , and thus the cross encoder. This is crucial as it prevents the outputs of the network (i.e., the inputs to the loss) from “collapsing”. A similar idea has been exploited in the siamese network for representation learning [8].

The final distillation loss is a simple summation of these two terms $\mathcal{L}_{distill} = \mathcal{L}_{distill}^{txt} + \mathcal{L}_{distill}^{img}$. Besides, we also use Masked Language Modeling [14] loss during pretraining. Overall, the model is trained with the sum of losses:

$$\mathcal{L} = \mathcal{L}_{itc} + \mathcal{L}_{itm} + \mathcal{L}_{mlm} + \mathcal{L}_{distill}. \quad (8)$$

Prior work [50] also proposed a distillation objective for cross encoder to dual encoder knowledge transfer. Our formulation differs from it in the following aspects: (i) **negative types**: we construct the negative image-text pairs for both image and text, instead of only constructing for text. In Section 4.2 we empirically demonstrate that our symmetric formulation gives the best overall performance. (ii) **negative sampling method**: [50] samples all the other examples within the batch as negatives for distillation, while we use only the top- m hard negatives ranked by the dual encoder.

In Section 4.2 we show that the performance saturates at as few as 4 hard negatives, indicating that only the hard negatives are informative in the distillation process. This has an important implication – that we need only compute the cross encoder scores for the top- m hard negatives, making the computation more efficient. In the extreme case of using only a single hard negative, we can directly take the inputs to \mathcal{L}_{itm} as teacher scores *without extra cross encoder computation*. (iii) **online vs. offline teacher**: we use an online teacher that is trained together with the student in the same model, while [50] follows the standard distillation [25] to use a separate offline teacher. In Section 4.2 we show that distillation from an online teacher works well. Moreover, our training can be performed in a single run without an extra run for learning the teacher model.

Model Architecture. We show an overview of our implementation of the proposed LOOPITR architecture in Figure 3. Our dual encoder consists of a separate image encoder and a separate text encoder. The image encoder is a **12-layer ViT model** [15], initialized using weights from a DeiT [61] model trained on ImageNet [13] 1K dataset. The text encoder is a **6-layer transformer encoder** [62] initialized from the first 6 layers’ weights of BERT-base [14]. Each transformer layer in the image and text encoders mainly consists of a self-attention layer and a feedforward net (FFN). After the input image and text are encoded, we take their [CLS] token presentations to compute dual encoder’s ITC scores.

The cross encoder is a **6-layer transformer encoder**, each layer mainly consisting of a self-attention layer, a cross-attention layer and an FFN. It is initialized from the last 6-layer of BERT-base. The text output from the text encoder is forwarded as input to the cross encoder with the image output injected via the cross-attention layer as the key and value. The ITM score is computed based on the [CLS]

token representation from the last layer.

4. Experiments

4.1. Dataset and Implementation Details

Pre-Training. We pre-train our model on four popular image-text datasets, COCO Captions [7, 45], Visual Genome (VG) Captions [36], SBU Captions [53] and Conceptual Captions (CC) [56], with in total around 4M images. This is the same set of datasets used in [10, 40, 58] for pre-training. Our model is pre-trained for 30 epochs with a learning rate of $1e-4$, weight decay 0.02. We use AdamW [48] optimizer and warm up the learning rate in the first 1,000 iterations followed by cosine decay to $1e-5$. We also warm up the weight of the distillation loss in the first epoch from 0 to 1, and keep it to be 1 for the rest of the training. The number of negatives m is set to 4, the initial value of trainable temperature parameter τ is set to 0.07. We use input image resolution of 256×256 and apply RandAugment [12] (without color augmentations). The batch size is set to 64 per GPU, and we use 8 NVIDIA A100 GPUs for training. The training takes around 2.5 days.

Downstream. After pre-training, we fine-tune our model on COCO [7, 45] and Flickr30K [65] for image-text retrieval. For COCO, we use the Karpathy split [31], which contains 113K training images, 5k validation images, and 5k test images. For Flickr30K, we use the standard split with 29K training images, 1K validation images, and 1K test images. Each image in these two datasets is paired with 5 captions. During fine-tuning, for both datasets, we use a batch size of 512 with input image resolution 512×512 as in [5], and a peak learning rate of $1e-5$. We train the model for 10 epochs for COCO and 25 epochs for Flickr30K.

4.2. Architecture and Distillation Ablations

In this section, we study the design of this unified architecture, putting our focus on the distillation objective. If not otherwise stated, we fine-tune the models on COCO karpathy train split [31] from the same pre-trained checkpoint trained on COCO+VG+SBU+CC without distillation, and report R@1 on the 5K val split. The default input image resolution is set to 384×384 . We use a batch size of 256 and train the models for 5 epochs.

How many negatives should we use? During distillation, for each text caption, we construct a bag of m hard negative images, together with its matched image to form a distribution. Similarly, we construct another distribution with m hard negative text captions and a positive text caption for each image. To understand how the number of negatives m affects the model performance, we evaluate model variants that use $m \in \{1, 4, 9, 14, 31\}$ negative examples. The results are shown in Table 1. Compared to the base model

#negatives m	cross encoder		dual encoder		#extra fwd pairs
	TR	IR	TR	IR	
- (w/o distillation)	74.98	57.41	62.64	46.78	0
1	75.02	57.44	63.10	47.67	0
4	75.00	57.76	63.64	47.99	3
9	74.92	57.63	63.82	47.74	8
14	74.96	57.63	63.20	47.95	13
31 (all negatives)	74.70	57.31	63.60	47.92	30

Table 1. Effect of number of negatives in distillation (R@1 on COCO 5k val split). *TR*=Text Retrieval, *IR*=Image Retrieval. *#extra fwd pairs* denotes the number of extra negative image-text pairs that are used in the cross encoder’s forward pass for computing distillation teacher score. As we can reuse one negative pair score computed for \mathcal{L}_{itm} , *#extra fwd pairs* equals to $m-1$.

negative sampling method	cross encoder		dual encoder	
	TR	IR	TR	IR
-	74.98	57.41	62.64	46.78
random	74.96	57.55	62.86	46.83
hard	75.00	57.76	63.64	47.99

Table 2. Effect of negative sampling method in distillation. We use 4 negatives (i.e., $m=4$) for all the models with distillation.

that is fine-tuned without distillation, we notice the performance of models with distillation improve significantly for dual encoder in both text retrieval and image retrieval, while for cross encoder, the results stay similar. Note that in this experiment, distillation is only applied during fine-tuning (to save computation), we expect more significant improvement when distillation is used in both pre-training and fine-tuning, as we discuss later in this section.

It is also worth noting that, while prior work [50] uses all in-batch examples ($m=31$) as negatives, here we show that using a single hard negative example ($m=1$) already provides a notable performance gain for dual encoder (without introducing extra cross encoder forward cost as it reuses the negative score computed for \mathcal{L}_{itm}), and the model performance saturates when we use 4 negatives.¹ Since cross encoder computation is quite expensive, this makes our approach more efficient than [50].

Hard negatives are more informative. Our distillation strategy only considers top- m negatives to construct the distribution. In Table 2 we compare with an alternative approach that uses m random negatives for distillation. We notice that while using random negatives gives slightly better performance than not using distillation, it performs significantly worse than the model that uses hard negatives. This makes sense in that the dual encoder is already powerful enough to clearly distinguish positive examples w.r.t. most of the negative examples, and it only struggles with a few hard negatives. Therefore hard negatives are more in-

¹Using $m=4$ is $1.4\times$ faster in training than using all negatives ($m=31$).

negative type	cross encoder		dual encoder	
	TR	IR	TR	IR
-	74.98	57.41	62.64	46.78
image ($\mathcal{L}_{distill}^{img}$)	75.04	57.76	61.94	47.84
text ($\mathcal{L}_{distill}^{txt}$)	74.68	57.52	63.86	47.26
image + text	75.00	57.76	63.64	47.99

Table 3. Effect of negative types in distillation.

when to distill		cross encoder		dual encoder	
pre-training	fine-tuning	TR	IR	TR	IR
-	-	74.98	57.41	62.64	46.78
-	✓	75.00	57.76	63.64	47.99
✓	-	75.02	57.67	64.70	49.96
✓	✓	74.96	57.82	65.00	50.53

Table 4. Effect of using distillation at different stages.

formative than random negatives in the distillation process.

Both image and text negatives are important for distillation. Our distillation involves distillation from two directions, one from the image’s perspective and another from the text’s perspective. For example, for each image, we construct a set of negative text captions along with its matched positive text caption to form a distribution for distillation, the corresponding loss is written as $\mathcal{L}_{distill}^{txt}$ in Section 3. Similarly, we have $\mathcal{L}_{distill}^{img}$ for each text with its negative images. In [50], only image negatives are considered for distillation. Here, we study how these two types of negatives affect the model performance. The results are shown in Table 3. From the table, we notice that adding image negatives significantly improves the image retrieval performance of the dual encoder, while adding text negatives boosts the performance of text retrieval. The best overall performance is achieved by combining both image and text negatives.

Should we use distillation in both pre-training and fine-tuning? Current image-text retrieval models, and vision-and-language models [10, 40, 44] in general, typically follow a two-stage training paradigm: a pre-training stage that trains the model on a large corpus of image-text data (possibly noisy), followed by a fine-tuning stage that tunes the model on a specific downstream task dataset (e.g., COCO Captions [7]). Here we study the effect of using the proposed distillation objective either only in pre-training or fine-tuning, or both. The results are shown in Table 4.

When using distillation in even only one of the stages, we notice a notable performance gain in the dual encoder compared to the model that does not use distillation. Another observation is that using distillation in pre-training provides more performance boost than using it only in fine-tuning. This is not surprising as previous work [61] shows that distillation typically benefits from training with more data and a longer training schedule. Overall, the best per-

teacher type	cross encoder		dual encoder	
	TR	IR	TR	IR
(a) -	74.98	57.41	62.64	46.78
(b) offline from (a)	74.68	57.64	63.20	47.65
(c) momentum	74.88	57.61	62.98	47.84
(d) online (step=10)	74.90	57.66	63.02	47.80
(e) online (step=100)	74.92	57.67	63.12	47.79
(f) online	75.00	57.76	63.64	47.99

Table 5. Effect of using online and offline teacher.

	cross encoder		dual encoder	
	TR	IR	TR	IR
w/o stop-grad	56.5	38.19	63.12	47.7
w/ stop-grad	75.0	57.76	63.64	47.99

Table 6. Distillation with and without stop-grad.

formance is achieved by the model that uses distillation in both pre-training and fine-tuning.

Online vs. offline teacher. Standard distillation [25] trains a student model with supervision from an offline learnt teacher model which is kept constant during distillation. In LOOPITR, we use an online teacher instead, which is dynamically evolving along with the student model. We examine the effect of using an online teacher compared to that using a standard offline teacher. The results are shown in Table 5. The offline teacher model is the trained model from Table 5a. Except for the offline teacher, we also compare with a momentum teacher that updates from the online model every step with a momentum of 0.995 (Table 5c), and two online model variants that update the teacher model by copying the online model every 10 or 100 steps (Table 5c,d, they can also be seen as momentum models with momentum 0, and update every 10 or 100 steps). From the table we see that all distillation variants improve the performance of the baseline model in Table 5a, showing that distillation is useful in all the cases. The model with offline teacher performs similarly to the model with a momentum teacher, while both being worse than the model with online teacher.

The role of stop-grad in Equation 7. In visual representation learning, SimSiam [8] suggests that stop-grad is important for siamese network from collapsing to a trivial solution. Since predictions and targets in our distillation formulation are also from the same network, it may suffer the same collapsing as siamese networks. Therefore, in Equation 7 we also use stop-grad to disable the gradients flow from the distillation loss to the cross encoder. To study the effect of using stop-grad, we compare a variant that does not use stop-grad in Table 6. We notice that, though the model performance does not collapse to a random number as in SimSiam, we do observe performance drop for both encoders, and especially for the cross encoder,

Method	#PT images	COCO (5K test set)						Flickr30K (1K test set)					
		TR			IR			TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
ALIGN [29]	1.2B	77.0	93.5	96.9	59.9	83.3	89.8	95.3	99.8	100.0	84.9	97.4	98.6
ThinkingFastSlow [50]	3.1M	-	-	-	-	-	-	-	-	-	72.1	91.5	95.2
UNITER [10]	4M	65.7	88.6	93.8	52.9	79.9	88.0	87.3	98.0	99.2	75.6	94.1	96.8
VILLA [21]	4M	-	-	-	-	-	-	87.9	97.5	98.8	76.3	94.2	96.8
LightningDot [58] + UNITER	4M	64.6	87.6	93.5	50.3	78.7	87.5	86.5	97.5	98.9	72.6	93.1	96.1
ViLT [33]	4M	61.5	86.3	92.7	42.7	72.9	83.1	83.5	96.7	98.6	64.4	88.7	93.8
OSCAR [44]	4M	70.0	91.1	95.5	54.0	80.8	88.5	-	-	-	-	-	-
RerankSmart [22] + OSCAR	4M	70.8	91.0	95.2	54.7	81.3	88.0	89.4	97.7	99.0	76.4	93.6	96.2
ALBEF [40]	4M	73.1	91.4	96.0	56.8	81.5	89.2	94.3	99.4	99.8	82.8	96.7	98.4
LOOPITR (cross-enc)	4M	75.1	92.4	96.7	58.0	82.8	89.7	94.4	99.4	99.9	83.4	96.4	98.2

Table 7. Comparison to state-of-the-art image-text retrieval methods on COCO and Flickr30k datasets. Gray indicates models trained with significantly larger amount of data. Here we show cross encoder performance for LOOPITR.

	cross encoder		dual encoder	
	TR	IR	TR	IR
(a) base	74.98	57.41	62.64	46.78
(b) w/o cross encoder (\mathcal{L}_{itm})	-	-	60.24	43.56
(c) w/o dual encoder (\mathcal{L}_{itc})	65.26	50.98	-	-
(d) w/o hard negative for cross encoder	67.82	51.63	60.70	46.09

Table 8. Combining cross encoder and dual encoder. Line b-d each removes a component from the base model in Line a.

where the performance drops almost 20 points. Here we do not see a severe collapse as in SimSiam, probably because our model is also trained with other objectives (e.g., \mathcal{L}_{itc} and \mathcal{L}_{itm}) instead of a single one in SimSiam.

The effect of combining dual encoder and cross encoder. LOOPITR combines dual and cross encoder in the same architecture, optimized jointly. To examine how they affect each other, we compare the base model (without distillation) with the variants that removed the losses associated with one of them. The results are shown in Table 8. Table 8b removes the cross encoder and only trains the dual encoder part, its dual encoder performance drops significantly compared to the base model in Table 8a, suggesting that cross encoder objective is also useful in learning a better dual encoder. We hypothesize that the dual encoder benefits from the fine-grained alignment learned in the cross encoder.

Similarly, when removing the dual encoder objective \mathcal{L}_{itc} in Table 8c, we observe a notable performance drop on the cross encoder. While cross encoder is able to perform dense and fine-grained interactions and alignment between the two modalities, it does so implicitly, which can be harder to learn. In contrast, the dual encoder performs explicit alignment between the input image and text, and using better-aligned representation as inputs helps the cross encoder to focus more on fine-grained details.

In addition, we also study the effect of using hard negatives for cross encoder’s \mathcal{L}_{itm} objective. We compare our

base model (Table 8a which uses hard negatives provided by the dual encoder) with a variant (Table 8d) that uses random negatives. We notice that when switching from hard negatives to random negatives, the performance of both cross and dual encoders drops, and the drop is especially significant for cross encoder. This suggests that using hard negatives is crucial for cross encoder’s performance, and a better cross encoder has a positive impact on the dual encoder.

Summary: (i) Distillation from the cross encoder teacher to the dual encoder student is useful in improving the dual encoder’s performance; (ii) The distillation objective is effective using as few as a single negative (i.e., $m=1$); (iii) Using hard negatives as the anchor for distillation is crucial as they are more informative than random negatives; (iv) Image negatives are more important in improving image retrieval performance while text negatives are more useful in improving text retrieval performance. Combining both types of negatives gives the best overall performance; (v) Distillation is helpful either used during pre-training or fine-tuning, and it is more helpful if used in both stages; (vi) Applying stop-grad on cross encoder teacher’s score is important to prevent the model from collapsing; (vii) On-line teacher gives better performance than an offline teacher or a momentum teacher; (viii) Dual encoder and cross encoder help improve each other when trained together, and hard negatives are important for cross encoder learning.

4.3. Comparison to State-of-the-art

In Table 7, we compare LOOPITR’s cross encoder to state-of-the-art approaches on COCO and Flickr30K for image-text retrieval. We notice our model achieves strong performance on both text retrieval (TR) and image retrieval (IR). The improvement over prior state-of-the-art is more significant on COCO compared to that of Flickr30K, possibly due to the performance of the latter is almost saturated.

In Table 9, we compare LOOPITR’s dual encoder with the other dual encoder methods. Our model shows bet-

Method	#PT images	COCO (5K test set)						Flickr30K (1K test set)					
		TR			IR			TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
ALIGN [29]	1.2B	77.0	93.5	96.9	59.9	83.3	89.8	95.3	99.8	100.0	84.9	97.4	98.6
VSE ∞ [5]	1B	66.4	89.3	-	51.6	79.3	-	88.4	98.3	99.5	74.2	93.7	96.8
LightningDot [58]	4M	60.1	85.1	91.8	45.8	74.6	83.8	83.9	97.2	98.6	69.9	91.1	95.2
RerankSmart [22] *	4M	66.9	90.1	95.0	52.2	80.2	88.0	86.3	96.8	98.6	71.6	91.5	95.0
LOOPITR (dual-enc)	4M	67.6	90.5	95.4	51.7	79.2	87.5	89.6	98.6	99.5	77.2	94.3	97.6

Table 9. Comparison to dual encoder image-text retrieval methods on COCO and Flickr30k datasets. * the COCO performance for RerankSmart may benefit from pre-training on extra datasets (VQA [2] and GQA [27]) annotated on COCO images. For VSE ∞ , we specify #images used in image-only pre-training, while for other models, we specify #images used in image-text pre-training.

Method	#PT images	Flickr30K (1K test set)					
		TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10
ALIGN [29]	1.2B	88.6	98.7	99.7	75.7	93.8	96.8
CLIP [54]	400M	88.0	98.7	99.4	68.7	90.6	95.2
UNITER [10]	4M	80.7	95.7	98.0	66.2	88.4	92.9
OSCAR [44]	4M	81.0	95.5	97.8	67.2	88.5	92.7
RerankSmart [22]	4M	78.2	94.0	97.3	63.3	86.4	91.6
ALBEF [40]	4M	90.5	98.8	99.7	76.8	93.7	96.7
LOOPITR (dual)	4M	82.3	96.7	98.8	70.3	91.2	95.6
LOOPITR (cross)	4M	91.8	99.0	100.0	79.2	94.4	97.1

Table 10. Zero-shot retrieval results on Flickr30K.

ter overall performance on Flickr30K and comparable performance on COCO. The slightly lower performance on COCO image retrieval compared to RerankSmart [22] might be because its pre-training uses extra datasets (VQA [2] and GQA [27]) annotated on COCO images.

In Table 10, we show zero-shot results on Flickr30K, for both dual and cross encoder (from the same trained LOOPITR). LOOPITR’s cross encoder achieves the best overall performance, and is even better than ALIGN [29] and CLIP [54] which are trained on a significantly larger amount of data. Meanwhile, thanks to distillation, LOOPITR’s dual encoder is also better than multiple strong cross encoder methods, e.g., OSCAR [44] and UNITER [10], and achieves comparable image retrieval performance to CLIP, showing the benefit of leveraging a strong cross encoder to improve dual encoder. On the other hand, the performance of the dual encoder of our model still lags behind the cross encoder, especially on R@1. We hypothesize this difference mainly comes from the dual encoder’s inability of performing fine-grained matching required for accurate retrieval.

VQA results. In addition to image-text retrieval, we also validate the model’s effectiveness on VQA [2]. To adapt our model for VQA, we add an answer decoder to generate the answers as natural language text [11, 40]. We use an input image resolution of 512x512 with a batch size of 256, and we fine-tune the model for 8 epochs with a peak learning rate of 2e-5. In Table 11, we show our model demonstrates strong results compared to previous approaches.

Method	test-dev	test-std
VL-BART [11]	-	71.3
LXMERT [60]	72.42	72.54
UNITER [10]	72.70	72.91
OSCAR [44]	73.16	73.44
VILLA [21]	73.59	73.67
UNIMO [43]	73.79	74.02
ViLT [33]	70.94	-
ALBEF [40]	74.54	74.70
LOOPITR	75.18	75.20

Table 11. Results on VQA dataset [2].

5. Discussion and Conclusion

Previous approaches often view dual encoders [5, 17, 54] and cross encoders [10, 40, 44] as separate architectures for modeling. In this work, we propose a unified approach that combines the two types of architectures in the same network for image-text retrieval. We conduct a comprehensive ablation study around this unified architecture and especially lay our focus on the design of the distillation objective – which is employed to efficiently transfer the knowledge from cross encoder to dual encoder in an online manner. We empirically demonstrate the effectiveness of joint training of these two architectures, and show that distillation is quite effective in improving dual encoder’s performance. We hope our work will encourage future explorations on combining dual and cross encoders for image-text retrieval, and vision-and-language learning in general.

Limitations: While our distillation objective is more efficient than [50], it still induces an additional 6% computation cost at training with $m=4$ (it does not hurt inference speed). This cost is neglectable when we set $m=1$, but it underperforms the setting of $m=4$. One interesting future work could be exploring the extreme case of $m=1$.

Societal Impact: The predictions from the developed system reflect the distribution of the data used for training, and they can be inaccurate and biased by the data. Therefore, users should not completely rely on our system for making real-world decisions.

Method	I \rightarrow T	T \rightarrow I	T \rightarrow T	I \rightarrow I
ALIGN (1.2B images) [29]	78.1	61.8	45.4	49.4
VSE ∞ (1B images) [5]	67.9	53.6	46.7	51.3
DE [16]	55.9	41.7	42.6	38.5
LOOPITR (dual-encoder)	69.2	53.5	44.7	46.7

Table 12. R@1 results on COCO 5k test, with CxC [16].

Method	encode type	TR	IR	time
(a) ViLT [33]	cross	83.5	64.4	6613s
(b) VSE ∞ (1B images) [5]	dual	88.4	74.2	45s
(c) LOOPITR (dual)	dual	89.6	77.2	11s
(d) LOOPITR (cross) rerank top 16	dual+cross	94.5	83.4	76s
(e) LOOPITR (cross) no rerank	cross	94.4	83.1	2342s

Table 13. Query time comparison on the full Flickr 1K test split. Time cost is calculated with a single A100 GPU.

Acknowledgement. This research was partially done when Jie was an intern with Meta AI, and was later supported at UNC by ARO Award W911NF2110220 and DARPA MCS Grant N66001-19-2-4031. The views contained in this article are those of the authors and not of the funding agency.

A. Additional Experiments

Comparison on COCO 5K test set using CxC annotations. In Table 12 we compare LOOPITR with baseline methods on COCO 5k test split with CxC annotations [16]. CxC augments the original COCO dataset with human semantic similarity judgments for 267,095 intra- and inter-modality pairs. We note from the table that our method achieves competitive performance even compared to strong method like VSE ∞ that leverages visual encoder trained on 1B images.

Query time comparison. Comparing to the efficient cross encoder ViLT and strong dual encoder VSE ∞ in Table 13, LOOPITR (dual) gives better results while runs faster. Our full cross encoder (Table 13e) greatly improves the dual encoder results, though with increased time cost (still faster than ViLT). With reranking (Table 13d),² we notice a large speedup without performance degradation.

B. Additional Implementation Details

Pre-Training. For pre-training, we train the model with distillation $m=4$, and we use both image and text negatives. The image resolution is set to 256×256 . The training follows the default setting in Table 14.

²In this setting, we use cross encoder to rerank the top k retrieved candidates from the dual encoder, as in [40]. As reranking speeds up inference without hurting performance, we report ‘cross encoder’ results in this setting if not otherwise stated.

config	value
optimizer	AdamW
base learning rate	1e-4
min learning rate	1e-5
weight decay	0.02
optimizer momentum	$\beta_1, \beta_2=0.9, 0.999$
batch size	512
learning rate schedule	cosine decay [47]
warmup iterations	2,000
image augmentation	RandAug(N=2, M=7) [12]
training epochs	30

Table 14. Pre-Training settings.

config	value
optimizer	AdamW
base learning rate	1e-5
min learning rate	1e-6
weight decay	0.02
optimizer momentum	$\beta_1, \beta_2=0.9, 0.999$
batch size	512
learning rate schedule	cosine decay [47]
image augmentation	RandAug (N=2, M=7) [12]
training epochs	10

Table 15. COCO retrieval fine-tuning settings.

config	value
optimizer	AdamW
base learning rate	1e-5
min learning rate	1e-6
weight decay	0.02
optimizer momentum	$\beta_1, \beta_2=0.9, 0.999$
batch size	512
learning rate schedule	cosine decay [47]
image augmentation	RandAug (N=2, M=7) [12]
training epochs	25

Table 16. Flickr30K retrieval fine-tuning settings.

config	value
optimizer	AdamW
base learning rate	2e-5
min learning rate	1e-6
weight decay	0.02
optimizer momentum	$\beta_1, \beta_2=0.9, 0.999$
batch size	512
learning rate schedule	cosine decay [47]
image augmentation	RandAug (N=2, M=7) [12]
training epochs	8

Table 17. VQA fine-tuning settings.

Fine-Tuning. After pre-training, we fine-tune the model on image-text retrieval task on COCO and Flickr30K, and

visual question answering on VQA [2]. The image resolution is set to 512×512 as in [5], the positional embeddings are interpolated following [61]. For COCO and Flickr30K retrieval, we train the model with distillation $m=4$, and we use both image and text negatives, and we follow the settings in Table 15 for COCO, Table 16 for Flickr30K. For VQA, we follow the settings in Table 17.

References

- [1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, 2017. 2
- [2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015. 8, 10
- [3] Lei Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In *NeurIPS*, 2013. 2
- [4] Cristian Bucilua, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *KDD*, 2006. 2
- [5] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *CVPR*, 2021. 1, 2, 5, 8, 9, 10
- [6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 1
- [7] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv*, 2015. 5, 6
- [8] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021. 4, 6
- [9] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, 2021. 1
- [10] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. In *ECCV*, 2020. 1, 2, 3, 5, 6, 7, 8
- [11] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *ICML*, 2021. 8
- [12] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPR Workshops*, 2020. 5, 9
- [13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 4
- [14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 3, 4
- [15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 4
- [16] Parekh et al. Crisscrossed captions: Extended intra inter-modal semantic similarity judgments for ms-coco. In *EACL*, 2021. 9
- [17] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *BMVC*, 2018. 2, 8
- [18] Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lijuan Wang, Yezhou Yang, and Zicheng Liu. Compressing visual-linguistic model via knowledge distillation. *ICCV*, 2021. 2
- [19] Zhiyuan Fang, Jianfeng Wang, Lijuan Wang, Lei Zhang, Yezhou Yang, and Zicheng Liu. Seed: Self-supervised distillation for visual representation. In *ICCV*, 2021. 2
- [20] Andrea Frome, Greg Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc’Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *NeurIPS*, 2013. 1
- [21] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. In *NeurIPS*, 2020. 2, 7, 8
- [22] Gregor Geigle, Jonas Pfeiffer, Nils Reimers, Ivan Vulic, and Iryna Gurevych. Retrieve fast, rerank smart: Cooperative and joint approaches for improved cross-modal retrieval. *arXiv*, 2021. 1, 2, 7, 8
- [23] Shir Gur, Natalia Neverova, Chris Stauffer, Ser-Nam Lim, Douwe Kiela, and Austin Reiter. Cross-modal retrieval augmentation for multi-modal classification. In *Findings of EMNLP*, 2021. 1, 2
- [24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 1, 2
- [25] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NeurIPS Deep Learning Workshop*, 2014. 2, 3, 4, 6
- [26] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv*, 2017. 2
- [27] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 8
- [28] Gautier Izacard and Edouard Grave. Distilling knowledge from reader to retriever for question answering. In *ICLR*, 2021. 2
- [29] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv*, 2021. 1, 2, 7, 8, 9
- [30] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding. In *EMNLP*, 2020. 2
- [31] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 5

- [32] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wentaoh Yih. Dense passage retrieval for open-domain question answering. *arXiv*, 2020. 2
- [33] Wonjae Kim, Bokyoung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021. 3, 7, 8, 9
- [34] Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *EMNLP*, 2016. 2
- [35] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv*, 2014. 2
- [36] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017. 5
- [37] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *CVPR*, 2021. 2
- [38] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *ECCV*, 2020. 2
- [39] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, 2020. 2
- [40] Junnan Li, Ramprasaath R Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021. 1, 2, 3, 5, 6, 7, 8, 9
- [41] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. In *EMNLP*, 2020. 2
- [42] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv*, 2019. 1, 2
- [43] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. In *ACL*, 2021. 1, 8
- [44] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020. 1, 2, 3, 6, 7, 8
- [45] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5
- [46] Yang Liu, Samuel Albanie, Arsha Nagrai, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In *BMVC*, 2020. 2
- [47] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2017. 9
- [48] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 5
- [49] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 1, 2, 3
- [50] Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In *CVPR*, 2021. 2, 4, 5, 6, 7, 8
- [51] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. Ms marco: A human generated machine reading comprehension dataset. In *CoCo@ NIPS*, 2016. 2
- [52] Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with bert. *arXiv*, 2019. 2
- [53] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *NeurIPS*, 2011. 5
- [54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv*, 2021. 1, 2, 8
- [55] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing, NeurIPS*, 2019. 2
- [56] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 5
- [57] Dawei Sun, Anbang Yao, Aojun Zhou, and Hao Zhao. Deeply-supervised knowledge synergy. In *CVPR*, 2019. 2
- [58] Siqi Sun, Yen-Chun Chen, Linjie Li, Shuohang Wang, Yuwei Fang, and Jingjing Liu. Lightningdot: Pre-training visual-semantic embeddings for real-time image-text retrieval. In *NACCL*, 2021. 5, 7, 8
- [59] Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. Mobilebert: a compact task-agnostic bert for resource-limited devices. In *ACL*, 2020. 2
- [60] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP*, 2019. 1, 8
- [61] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 2, 4, 6, 10
- [62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 1, 2, 3, 4
- [63] Jianfeng Wang, Xiaowei Hu, Pengchuan Zhang, Xiujun Li, Lijuan Wang, Lei Zhang, Jianfeng Gao, and Zicheng Liu. Minivlm: A smaller and faster vision-language model. *arXiv*, 2020. 2
- [64] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 2

- [65] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2014. 5
- [66] Hang Zhang, Yeyun Gong, Yelong Shen, Jiancheng Lv, Nan Duan, and Weizhu Chen. Adversarial retriever-ranker for dense text retrieval. *arXiv*, 2021. 2
- [67] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*, 2021. 2
- [68] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *CVPR*, 2018. 2