

Retrieve Fast, Rerank Smart: Cooperative and Joint Approaches for Improved Cross-Modal Retrieval

Gregor Geigle^{*1}, Jonas Pfeiffer^{*1}, Nils Reimers¹,
Ivan Vulic², Iryna Gurevych¹

¹Ubiquitous Knowledge Processing Lab, Technical University of Darmstadt

²Language Technology Lab, University of Cambridge

www.ukp.tu-darmstadt.de

Abstract

Current state-of-the-art approaches to cross-modal retrieval process text and visual input jointly, relying on Transformer-based architectures with cross-attention mechanisms that attend over all words and objects in an image. While offering unmatched retrieval performance, such models: 1) are typically pretrained from scratch and thus less scalable, 2) suffer from huge retrieval latency and inefficiency issues, which makes them impractical in realistic applications. To address these crucial gaps towards both improved and efficient cross-modal retrieval, we propose a novel **fine-tuning framework** which **turns any pretrained text-image multi-modal model into an efficient retrieval model**. The framework is based on a **cooperative retrieve-and-rerank approach**, which combines: 1) **twin networks** to separately encode all items of a corpus, enabling efficient initial retrieval, and 2) a **cross-encoder component** for a more nuanced (i.e., smarter) ranking of the retrieved small set of items. We also propose to jointly fine-tune the two components with shared weights, yielding a more parameter-efficient model. Our experiments on a series of standard cross-modal retrieval benchmarks in monolingual, multilingual, and zero-shot setups, demonstrate improved accuracy and huge efficiency benefits over the state-of-the-art cross-encoders.

1. Introduction

Information-rich and efficient methods for dealing with large unstructured data in both computer vision and natural language processing (NLP) are required to process and understand huge amounts of user-created content and beyond. In multi-modal contexts, such methods enable fundamental applications such as *image retrieval*. A typical efficient **embedding-based approach** encodes images and text *separately* and then induces a shared high-dimensional multi-

modal feature space. This enables cross-modal retrieval, where standard distance metrics are used to identify the most similar examples for each query in the target data collection via standard nearest-neighbor search [3, 4, 29, 35, 2, 20].

While these embedding-based approaches have already been shown to achieve reasonable performance in search and retrieval applications, both monolingually for English [41, 15, 58, 52, 48] and in multilingual contexts [28, 54, 7], they cannot match performance of more recent **attention-based methods**. Here, a typical modus operandi is to apply a **cross-attention mechanism** between examples from the two modalities to compute their similarity score, relying on Transformer-based neural architectures [51]. Such so-called **multi-modal cross-encoders (CE)** [50, 38, 9, 31, 18, 32, 24] have to pass each text-image pair through the multi-modal encoder network to compute their similarity, as illustrated in Figure 1a. While the results accomplished by the CE methods look impressive [32, 6, 24], this comes at a prohibitive cost. In particular, they have extremely high search latency: processing a single text query with a target image collection of 1M items may take up to 36 minutes using a single NVIDIA V100 GPU (see Table 3). Due to this issue, they are evaluated only with extremely small benchmarks, i.e., the maximum size of typical target image collections for image retrieval tasks is 5k images, and evaluation still lasts ≈ 50 hours (see Table 4).¹ In sum, cross-encoders are impractical for deployment in realistic application scenarios, while the use of small benchmarks results in inflated and thus misleading evaluation performance.

In unimodal text-only setups, Transformed-based architectures have recently been integrated with embedding-based (EMB) methods [19, 45, 25, 21, 16, *inter alia*], yielding computationally more efficient sentence encoders. Instead of jointly encoding sentence pairs with cross-attention, a pre-

¹Consequently, it would be impossible to evaluate these CE approaches on newer larger benchmarks: e.g., the (extrapolated) evaluation time on a benchmark spanning 100,000 images exceeds 2 years with a single GPU.

* Both authors contributed equally to this work.

trained Transformer model (e.g., BERT [12]) is fine-tuned within a twin network with shared Transformer weights, as illustrated in Figure 1b. In a nutshell, each sentence is passed through the encoder separately, and a loss function is defined on top of the two respective *separately computed* encodings. However, despite their strong performance on sentence retrieval and similarity tasks [45, 16, 34], these encoders still cannot match the task performance of cross-encoders [25].

Motivated by these insights, in this work we aim to leverage *the best of both worlds* towards improved and more efficient *cross-modal search and retrieval*: **1)** efficiency and simplicity of EMB approaches based on twin networks, as well as **2)** expressiveness and cutting-edge performance of CE methods. We first provide a systematic comparative analysis on the effectiveness and efficiency of Transformer-based multi-modal EMB and CE methods across a range of image search evaluation benchmarks. We then propose two novel models which aim to blend the main strengths of CE and EMB. The idea behind the first model variant, termed **cooperative (SEP+CO)**, is **to retrieve and rerank with two separate, independently trained retrieval models**: **1)** an initial *top-k* list of potentially relevant items (i.e., texts or images) is retrieved by the more efficient EMB model, and then **2)** this *top-k* list is reranked “*smartly*” by the more accurate CE model, as illustrated in Figure 1c. Our second, **joint (JOIN+CO)** model variant also operates in the same retrieve-and-rerank setup, but it now trains a multi-modal cross-encoder and a multi-modal EMB model jointly with tied weights, as illustrated in Figure 1d. The retrieve step, where efficiency is paramount, is again executed by the EMB sub-model, and the precision-oriented rerank step is conducted via the CE sub-model.

We propose a general framework for cross-modal search and retrieval, where JOIN+CO and SEP+CO models are independent of the chosen pretrained vision-language representation architectures. The experiments are thus based on a state-of-the-art vision-language architecture OSCAR [32] (experiments in English) and M3P [24] (multilingual), and we demonstrate consistent improvements over the original OSCAR model on the standard benchmarks MSCOCO and Flickr30k and improvements over the original M3P in multiple languages on the Multi30k dataset. We also empirically validate huge efficiency benefits of the proposed framework.

Contributions. **1)** We construct and systematically evaluate twin-networks combined with multi-modal Transformers (EMB); they outperform all previous embedding-based approaches, but lag behind their CE counterparts. **2)** We evaluate EMB and CE approaches within a cooperative retrieve-and-rerank approach; their combination outperforms the individual models, while offering substantial efficiency boosts compared to CE methods. **3)** We propose a novel joint CE-EMB model (JOIN+CO), which is trained to simultaneously cross-encode and embed multi-modal input;

it achieves the highest scores overall, while maintaining retrieval efficiency. **4)** Finally, we propose a more realistic evaluation benchmark; we demonstrate harsh drops in overall cross-modal retrieval performance of all models in this more difficult scenario, calling for improved evaluation benchmarks and protocols in future work. Our code is available at: github.com/UKPLab/MMT-Retrieval.

2. Related Work

Efficient approaches to cross-modal image-text retrieval relied on the induction of shared multi-modal visual-semantic embedding spaces (VSEs) [17, 15, 48, 40]. In a multilingual setup, all languages share the same embedding space along with the visual data [28, 54, 7]. More recently, attention-based cross-encoder models, typically based on Transformer architectures [51] have considerably outperformed the VSE-based approaches. However, this comes at a severe cost of decreased retrieval efficiency and increased latency [30, 53]. The current state-of-the-art multi-modal models jointly encode and cross-attend over text tokens and image features [38, 50, 9, 31, 18, 32, 6, 24, *inter alia*]. These CE methods leverage image captioning datasets such as MSCOCO [33] and Flickr30k [43] and train a classification head which learns to identify whether or not an (*image, caption*) input pair constitutes an aligned pair. Each image-text combination must be passed through the network, which scales quadratically with the number of examples.

Our work is inspired by the work on EMB-based approaches in unimodal text-only setups. There, pretrained language models (LMs) such as BERT [12] or RoBERTa [36] are fine-tuned via twin-network architectures on auxiliary tasks such as semantic textual similarity [45, 25], paraphrasing [55], response retrieval [57, 22, 21, 25], or translation ranking [10, 16]. This process effectively turns the LMs into universal *sentence encoders* which are then be used off-the-shelf for efficient text-based monolingual and cross-lingual retrieval [34]. In this work, we first extend this idea to multi-modal setups, and then show that our cooperative and joint approach yields improved cross-modal retrieval models, while maintaining retrieval efficiency.

The work most closely related to ours includes contemporaneous models: CLIP [44], ALIGN [26], and VisualSparta [39]. CLIP and ALIGN use similar contrastive learning strategies as we do, but are cast as full-fledged *pretraining* architectures that learn from scratch and require magnitudes of more data than our approach. We show that it is possible to *fine-tune* pretrained models with fewer data and offer a general framework, applicable to a spectrum of pretrained models. Further, unlike prior work, we demonstrate the benefits of combining EMB-based (contrastive) learning with cross-encoders for improved and efficient retrieval.² Finally,

²As both CLIP [44] and ALIGN [26] dis-join the image and text com-

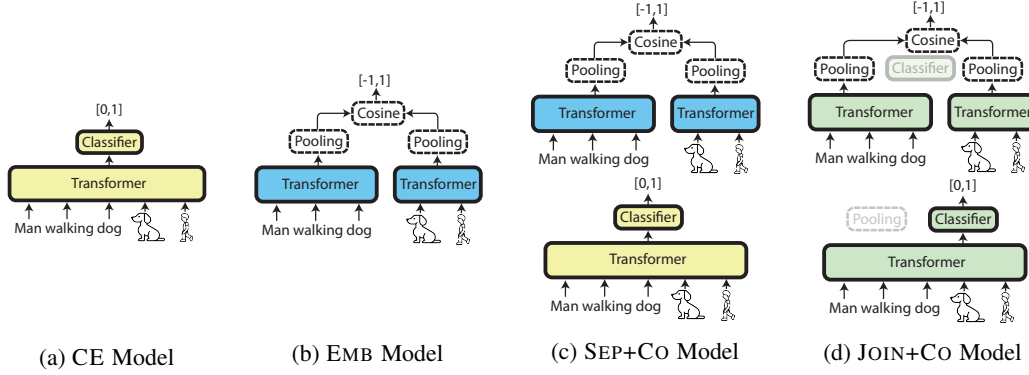


Figure 1: Different architectures for image and text retrieval. Equal colors and annotation indicate shared weights.

VisualSparta [39] fine-tunes OSCAR, but at the level of token (text) and image-region embeddings. This enables the use of extremely fast lookup tables for efficient retrieval. However, this comes with a major disadvantage: the model disposes of wider context information.³ Our cooperative methods do leverage the finer-grained information at retrieval.

3. Methodology

The predominant Transformer-based multi-modal text-vision architecture is a **single-stream encoder**; it shares the majority of weights between the two modalities, including the multi-head cross-attention [9, 31, 18, 32, 24]. The Transformer weights and text embeddings are typically initialized with weights of a pretrained LM (e.g., BERT [12] for English, XLM-R [11] for multilingual models), where the corresponding vocabulary and tokenizer is utilized. Images are preprocessed via object detection models such as Faster R-CNN [46] to extract feature representations for regions of interest [1]. The image features are passed through an **affine-transformation layer** which learns to align the vision input with the pre-trained Transformer. The position of the region of interest (or in some models also the region’s width and height) is used to generate positional embeddings. By combining these two representations, each object-region is passed into the Transformer separately.

The **cross-attention mechanism** of the Transformer attends over all text and image inputs at every layer, thus learning a joint representation of both modalities.

Similar to **masked language modeling** (MLM) in the text domain [12], multi-modal Transformer models are trained with **self-supervised objectives**. For pretraining, image-caption datasets (i.e., MSCOCO [33], Flickr30k [43], Conceptual Captions (CC) [47], and SBU [42]) are utilized. The pretrained multi-modal model is subsequently fine-tuned with task data for a particular downstream multi-modal task.

In this work, we focus on **different fine-tuning strategies** of the pretrained models for the downstream task of image and text retrieval. We illustrate these different approaches in Figure 1 and describe them in what follows.

3.1. Cross-Encoders

For image and text retrieval tasks, the prevailing approach with pretrained multi-modal Transformer models is to cross-encode each image-text combination (see Figure 1a).

Training. A pretrained model receives as input positive and negative pairs of images and captions. Negative pairs are also sampled from the training dataset (e.g., MSCOCO, Flickr30k). A **binary classification head** is placed on top of the Transformer model, where the contextualized embedding of the [CLS] token is passed into the classification head. The weights of the classifier together with the Transformer, word embeddings and image feature transformation matrices are fully fine-tuned using a cross-entropy loss:

$$\mathcal{L}_{CE}(i, c) = -\left(y \log p(i, c) + (1 - y) \log(1 - p(i, c))\right). \quad (1)$$

$p(i, c)$ indicates the probability of the input combination of image i and caption c to have the positive label (i.e., whether it is the correct image-caption combination); $y = 1$ if (i, c) is a positive pair and $y = 0$ if either the image or text has been replaced (i.e., a negative pair).⁴

Retrieval. At retrieval, all (i, c) combinations need to be processed, and are ranked by the probability $p(i, c)$. For instance, given a text query c , retrieving a single most relevant image i from a target image collection I proceeds as:

$$\arg \max(p(i, c), \forall i \in I) \quad (2)$$

Despite its typically high performance, this approach comes at high computational costs as each target instance needs to be passed through the entire network along with the query to obtain the score $p(i, c)$; that is, the approach does not leverage any precomputed representations during retrieval.

ponents in their methods, cross-attention over the instances is not possible.

³E.g., considering a query “two dogs and one cat”, the model is unable to match the numbers to the animals yielding likely worse retrieval results.

⁴Some cross-encoders such as UNITER [9] and VL-BERT [49] rely on another standard triplet loss function [8]; however, OSCAR [32] reports improved performance with cross-entropy.

3.2. Embedding-Based Retrieval Methods

Training. Each image and text caption is passed separately through the pretrained Transformer model, see Figure 1a. The contextualized representations are **mean-pooled** to represent the embedding of the respective image \mathbf{i} and text caption \mathbf{c} .⁵ The objective of the twin network is to place positive training instances (i, c) closely in the shared multi-modal space, while unrelated instances should be placed farther apart. This is formulated through a standard triplet loss function. It leverages (i, c, c') and (i, i', c) triplets, where (i, c) are positive image-caption pairs from the training corpus, while c' and i' are negative examples sampled from the training corpus such that image-caption pairs/instances (i, c') and (i', c) do not occur in the corpus. The triplet loss is then:

$$\mathcal{L}_{\text{EMB}}(i, c) = [\cos(\mathbf{i}, \mathbf{c}') - \cos(\mathbf{i}, \mathbf{c}) + \alpha]^+ + [\cos(\mathbf{i}', \mathbf{c}) - \cos(\mathbf{i}, \mathbf{c}) + \alpha]^+ \quad (3)$$

where $[\cdot]^+ = \max(0, \cdot)$, α defines a margin, and \mathbf{i}' and \mathbf{c}' are embeddings of respective image and caption negatives.

Sampling Negative Examples. The choice of negative examples may have a profound impact on training and performance, and it has been shown that selecting hard negative examples typically yields improved performance and faster learning curves [15]. However, detecting such hard negatives is only possible with EMB-based approaches, as cross-encoding all instances is computationally infeasible. We rely on the *In-Batch Hard Negatives (BHN)* method [23], a computationally efficient sampling of hard negative examples. In a nutshell, BHN randomly samples a set of N negative examples from the training corpus and then ranks them according to their distance to all positive examples; for each positive example, the closest negative example is selected as the *hardest* negative example. By scaling up N , the probability of sampling truly hard negatives increases.

Retrieval. The EMB approach enables **pre-encoding** of all the target collection items for efficient retrieval look-up.⁶ For instance, a text query q is encoded with the embedding model and the most similar pre-encoded instance from a target image collection I is retrieved: $\arg \max_{i \in I} \cos(\mathbf{i}, \mathbf{q})$.

This approach can easily scale to even billions of target images [27], but it cannot be guaranteed that the important idiosyncratic information necessary to distinguish truly relevant from related examples, is sufficiently encoded in the embedding. Further, the approach might not generalize well in low-resource scenarios as the model is not required to

learn finer-grained parts of the input if they are never demanded by the training data.

3.3. Separate Training, Cooperative Retrieval

We propose to combine the benefits of the two model types (CE and EMB) within a **cooperative retrieval approach** (SEP+CO), as illustrated in Figure 1c.

Training and Retrieval. Two models, one CE (see §3.1) and one EMB (see §3.2), are trained independently. Following that, the retrieval step is split into two stages. First, the efficient EMB model is used to retrieve the *top k* relevant items from the entire large collection instances, yielding a much smaller target collection I_k : $I_k = \text{top}_k(\{\cos(\mathbf{i}, \mathbf{q}) : \forall i \in I\})$, where $\text{top}_k(\cdot)$ retrieves a set of the top k most similar instances. Second, we rerank the instances from I_k with the more precise but computationally more expensive CE model: $\arg \max_{i \in I'} p(i, c)$. This cooperative approach thus combines the benefits of both approaches and is able to efficiently retrieve instances.⁷ However, given that this approach requires two models to be stored in memory, it is less parameter-efficient than the previous methods.

3.4. Joint Training, Cooperative Retrieval

Training and Retrieval. Instead of relying on two fully separated models, we propose to **train a single joint model**, able to both *cross-encode* and *embed*, see Figure 1d. The joint model with shared parameters trains by alternating between the respective sub-models and their input types. When cross-encoding, a dedicated prediction head is trained using a cross-entropy loss; Eq (1). In order to train the EMB-based sub-model, we again rely on a twin architecture with a triplet loss from Eq. (3).

Retrieval proceeds with the same two-step retrieve-and-rerank procedure from §3.3. We first obtain the set I_k with the much cheaper EMB-based submodel, and then rerank its items with the CE submodel. We again combine the best traits of the CE and EMB approaches, while maintaining parameter efficiency. By utilizing both learning objectives at training, the joint model is forced to observe the input from different viewpoints, thus improving its generalization capability while offering parameter efficiency.

4. Experimental Setup

Our fine-tuning framework from §3 can be applied to any pretrained multi-modal Transformer. In all the experiments, we opt for state-of-the-art pretrained multi-modal models for monolingual (English) and multilingual contexts: OSCAR [32] and M3P [24], respectively.

OSCAR is a single-stream multi-modal Transformer, with its weights initialized with those of the pretrained BERT-Base

⁵Following Reimers et al. [45] we opt for mean pooling as the final “aggregated” embedding; it also outperformed by a variant that uses the [CLS] token in our preliminary experiments.

⁶Note that precomputing such encodings does come with increased storage and memory demands; e.g., with a base Transformer architecture this requires an additional $\approx 3\text{KB}$ of memory for each embedding. A corpus of 1M image would amount to $\approx 3\text{GB}$ of required storage.

⁷Retrieval time from a pool of 1M images: 94ms (GPU), 13s (CPU).

model, and then subsequently fine-tuned on multi-modal data (see §3). Unlike prior work, OSCAR additionally uses object labels of detected regions: those labels serve as anchors for visual grounding, with large improvements achieved over its prior work. *M3P* is a single-stream multilingual multi-modal Transformer. Its weights are initialized with those of pretrained XLM-R Base, and then fine-tuned on multi-modal data (see §3) as well as multilingual text-only data.

Training and Test Data. We primarily experiment with the English image-text retrieval benchmarks MSCOCO and Flickr30k. They respectively comprise 123k and 31.8k image, with 5 captions describing each image. MSCOCO provides two test benchmarks of sizes 1k and 5k, where the smaller set is a subset of the 5k set. The standard Flickr30k test set consists of 1k images. In addition, we use the development set of Conceptual Captions (CC) [47] for zero-shot evaluation, and also to construct a larger and more difficult test set (see later in §6). The original CC dev set contained 15.8k images, but currently only 14k images are still available online.

For multilingual experiments, we use the standard Multi30k dataset [14, 13, 5], which extends Flickr30k with 5 German captions and one French and Czech caption per image. Its test set also comprises 1k images.

The evaluation metric is the standard *Recall-at-M* ($R@M$): it reports the proportion of queries for which the relevant target item is present within the top M retrieved items.

Training Setup and Hyperparameters. Our setup largely follows Li et al. [32] and Huang et al. [24] unless noted otherwise.⁸ We experiment with learning rates $[5e-5, 2e-5]$, and with the number of update steps between 25k and 125k. One batch contains 128 positive pairs plus 128 negative pairs with \mathcal{L}_{CE} . We use the AdamW optimizer [37] with a linear learning rate decay without warmup, and a weight decay of 0.05. We take model checkpoints every 5k steps and select the checkpoint with the best dev set performance.

4.1. Baselines and Model Variants

CE. The main baselines are state-of-the-art OSCAR and M3P models used in the standard CE setting, described in §3.1. We fully fine-tune the Transformer weights along with a randomly initialized classification head.⁹ At retrieval, we cross-encode each text-image combination and rank them according to the corresponding probability, see Eq. (2).

EMB. We rely on the BHE negative sampling strategy, finding that training for 30k steps, with a learning rate of $5e-5$, and with a margin $\alpha = 0.1$ works best.¹⁰

⁸Unlike Li et al. [32] we do not use object tags as additional input, as preliminary experiments suggested no improvement with object tags.

⁹We empirically verified that training for 100k steps and a learning rate of $2e-5$ (OSCAR) or $5e-5$ (M3P) performed best.

¹⁰We also experimented with *Approximate-nearest-neighbor Negative Contrastive Estimation (ANCE)* [56]; however, it did not yield performance benefits. See Appendix §A.3 for a detailed comparison.

SEP+CO. For the cooperative method without joint training (see §3.3), we retrieve the top $k = 20$ instances with EMB and re-rank each combination via CE.¹¹

JOIN+CO. We alternate between the two objective functions while training the joint model (see §3.4). We find that training for 60k update steps with a learning rate of $2e-5$ (OSCAR) or $5e-5$ (M3P) works best, the rest of the hparams are the same as with separately trained models. At SEP+CO retrieval, $k = 20$. To demonstrate the benefits of cooperative retrieval, we also evaluate two non-cooperative variants originating from the joint model: **JOIN+CE** uses the CE sub-model for a single-step CE-style retrieval, while **JOIN+EMB** operates in the fully EMB retrieval setup.

The underlying pretrained Transformer is denoted with a superscript: e.g., $\text{JOIN+CO}^{\text{OSCAR}}$ is read as: 1) pretrained OSCAR is 2) fine-tuned with the joint variant from §3.4, and 3) then used in the cooperative retrieval setup.

5. Results and Discussion

The main results on English-only monolingual datasets Flickr30k and MSCOCO are summarized in Table 1, while the scores on multilingual Multi30k are provided in Table 2.

As expected, all Transformer-based approaches (groups G2 and G3) substantially outperform the pre-Transformer models (G1). While this has already been established in prior work for CE methods, our findings confirm that the same holds also for the efficient EMB approach. This validates the effectiveness of Transformer architectures pretrained on large corpora for the retrieval task. $R@1$ scores with EMB lag slightly behind the CE scores, but the respective $R@10$ scores are mostly on-par. This suggests that the EMB approach is “coarser-grained”, and mostly relies on “global” interactions between the modalities. We investigate this conjecture further in §6. This is also illustrated by an example in Figure 2, with more examples available in the appendix. When dealing with related target items, CE’s cross-attention mechanism is able to explicitly attend over each token and image region, capturing additional (non-global) information relevant to the query.¹²

Most importantly, the relative comparison of $R@1$ versus $R@10$ scores empirically hints at the necessity of the retrieve-and-rerank cooperative approach: the EMB approach efficiently retrieves 10-20 relevant examples, but the increased

¹¹We provide an ablation study of different k values in §A.1. We have also experimented with training a CE model using hard negative ANCE samples from a pre-trained EMB model. However, the CE model is able to easily overfit on those negative examples, resulting in inferior performance.

¹²In the context of our example in Figure 2, while the high-level “global” concept of a *skiing person* is present in (almost) every example, the additional important information related to *what the person is wearing* is not adequately represented in the embeddings. Therefore, the EMB (sub)model does not rank this instance at the top position. The CE (sub)model then directly compares the instances, identifying that clothing is an important indicator and reranks the target examples accordingly.

Group	Model	Image Retrieval			Text Retrieval			Image Retrieval			Text Retrieval		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
		MSCOCO (5k)						Flickr30k (1k)					
G1. Pre-Transformer	VSE++ [15]	43.9	59.4	72.4	41.3	71.1	81.2	39.6	70.1	79.5	52.9	80.5	87.2
	SCAN [30]	38.6	69.3	80.4	50.4	82.2	90.0	48.6	77.7	85.2	67.9	90.3	95.8
	PFAN [53]	—	—	—	—	—	—	50.4	78.7	86.1	70.0	91.8	95.0
	SCG [48]	39.2	68.0	81.3	56.6	84.5	92.0	49.3	76.4	85.6	71.8	90.8	94.8
G2. Cross-Encoders (Inefficient for retrieval)	CE ^{UNITER} [9]	48.4	76.7	85.0	63.3	87.0	93.1	72.5	92.4	96.1	85.9	97.1	98.8
	CE ^{Unicoder-VL} [31]	46.7	76.0	85.3	62.3	87.1	92.8	71.5	90.9	94.9	86.2	96.3	99.0
	CE ^{VILLA} [18]	—	—	—	—	—	—	74.7	92.9	95.8	86.6	97.9	99.2
	CE ^{OSCAR} _† [32]	54.0	80.8	88.5	70.0	91.1	95.5	—	—	—	—	—	—
	CE ^{OSCAR} _‡	52.6	80.0	88.1	69.3	90.7	95.3	75.9	93.3	96.6	88.5	98.5	99.2
G3. Embedding-Based (Efficient for retrieval)	VisualSparta [39]	44.4	72.8	82.4	—	—	—	57.4	82.0	88.1	—	—	—
	EMB ^{OSCAR}	52.2	80.2	88.0	66.9	90.1	95.0	72.0	91.0	94.7	84.7	97.1	98.7
	SEP+Co ^{OSCAR}	52.8	80.5	88.5	70.2	91.6	95.0	76.0	93.0	95.0	88.7	98.3	99.2
	JOIN+Co ^{OSCAR}	54.7	81.3	88.9	70.8	91.0	95.2	76.4	93.6	96.2	89.4	97.7	99.0
	JOIN+CE ^{OSCAR}	54.6	81.1	88.8	70.6	91.0	95.1	76.5	93.4	96.3	89.0	97.9	99.1
	JOIN+EMB ^{OSCAR}	52.5	80.0	88.0	66.7	90.0	95.0	71.6	91.5	95.0	86.3	96.8	98.6

Table 1: Results on MSCOCO and Flickr30k (monolingual setups). The group G1 presents results from the literature with Pre-Transformer approaches. G2 denotes the results of recent cross-encoders with Transformers (CE*; §3.1). Here, † indicates the results taken directly from the literature [32], while ‡ indicates our own results achieved with the published model weights. G3 covers efficient retrieval methods that either retrieve images based only on distance metrics (EMB, §3.2), or rely on the SEP+Co (CO) approach (see §3.3 and §3.4). The last two lines present the results of the joint model without the cooperative retrieval step (see §4.1). Highest results per each group in **bold**, highest overall results are underlined.

Caption: A skier is skiing down the snow wearing a white shirt and black shorts.

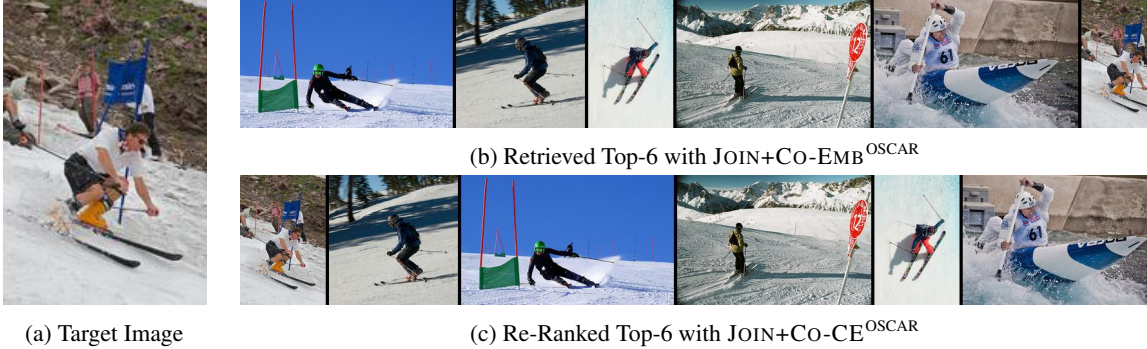


Figure 2: By first efficiently retrieving the top instances with the JOIN+Co-EMB^{OSCAR} submodel we identify (globally) most relevant target instances. The more precise, but less efficient JOIN+Co-CE^{OSCAR} submodel is then able to disentangle the specific intricacies of the target images. The ranking of the target images proceeds from left to right.

expressiveness of CE is required to refine the initially retrieved list. Moreover, the results in the cooperative setup even without joint training (SEP+Co^{OSCAR} and SEP+Co^{M3P}) demonstrate that the two models support each other: slight improvements are observed over the pure CE, while offering massive efficiency boosts over CE. Our speculation is that the EMB model filters out false positives, which in turn makes the CE model more robust.

The results of the JOIN+Co variant indicate that it is indeed possible to maintain retrieval efficiency with improved parameter efficiency: this approach performs on-par or even slightly outperforms the standard state-of-the-art CE models. The results verify that the two objective functions do not interfere with each other, and that a single model is able to both embed and cross-encode. We note that the JOIN+Co vari-

ant offers the best trade-off between parameter and retrieval efficiency, achieving the peak scores on the monolingual MSCOCO and Flickr30k benchmarks, and very competitive results on the multilingual Multi30k benchmark.

6. Further Analysis

We now discuss a series of additional experiments which further profile and analyze the proposed multi-modal retrieval approaches, focusing especially on the multiple efficiency aspects related to fine-tuning and retrieval stages.

Retrieval Efficiency and Larger Target Collections. The results in Table 1 indicate that current best-performing models achieve very high scores in absolute terms on the standard retrieval benchmarks. However, this is partially due

Type	Model	en	de	fr	cs	mean
G1. Pre-Transformer	MULE [28]	70.3	64.1	62.3	57.7	63.6
	S-LIWE [54]	76.3	72.1	63.4	59.4	67.8
	SMALR [7]	74.5	69.8	65.9	64.8	68.8
G2. Cross-Encoders	CE ^{M3P} _† [24]	86.7	82.2	73.5	70.2	78.2
	CE ^{M3P} _‡	83.7	79.4	76.5	74.6	78.6
G3. Emb-Based	EMB ^{M3P}	82.8	78.0	75.1	73.6	77.4
	SEP+Co ^{M3P}	84.8	80.5	77.5	75.6	79.6
	JOIN+Co ^{M3P}	83.0	79.2	75.9	74.0	78.0

Table 2: Results on Multi30k (multilingual setups). Following prior work [24], we report *mean Recall (mR)* scores: mR computes a total average score of Recall@1, Recall@5 and Recall@10 on image-to-text retrieval and text-to-image retrieval tasks. All methods in the comparison use text data from all four languages. We divide the models into groups G1-G3 as in Table 1. † indicates results taken directly from the literature [24] and ‡ indicates our own results.

Model	NVIDIA V100		CPU	
	50k	1M	50k	1M
EMB	16ms	37ms	0.2s	1.6s
SEP/JOIN+CO	74ms	94ms	6s	13s
CE	2min	36min	2.4h	47h

Table 3: Estimated retrieval latency for a single query with a target image collection spanning 50k or 1M images.

Model	1k	5k	100k
EMB	5s	30s	7min
SEP/JOIN+CO	5min	25min	8.5h*
CE	2h	50h	2.3a*

Table 4: Evaluation time for the full MSCOCO test sets spanning 1k, 5k, and 100k images. 1 NVIDIA V100 GPU with batch size 512. * denotes extrapolated values.

to the too small target collections; one undesired effect is that it becomes increasingly difficult to identify significant differences between model performances. Unfortunately, the inefficiency of CE models, as empirically validated in Table 3 and Table 4, has prevented the evaluation with larger target collections. However, more efficient fully EMB-based and SEP+CO methods now enable evaluation on larger target collections and in realistic scenarios.

We thus increase the target collection by merging test instances (i.e., target collections) from different available evaluation sets. In particular, we construct a target collection spanning 20k target items: it blends the test sets of MSCOCO (5k instances), Flickr30k (1k), and the development set of CC (14k). Note that we simply augment the target collection, but the query set with labels for each standardized evaluation task/set remains unchanged; in other words, the instances from other datasets are used as distractors that increase the search space and make the retrieval task more difficult. The results thus provide insights into the model performance in the target domain, as well as its robustness regarding out-of-

Model	Image Retrieval			Text Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
	<u>Flickr30k 1k+ CC 14k + MSCOCO 5k</u>					
EMB ^{OSCAR}	45.8	69.1	76.1	71.1	90.9	94.9
SEP+Co ^{OSCAR}	55.5	75.8	80.1	80.5	93.8	95.4
JOIN+Co ^{OSCAR}	55.9	77.5	82.9	81.0	92.9	94.9
	<u>MSCOCO 5k+ CC 14k + Flickr 1k</u>					
EMB ^{OSCAR}	40.6	68.5	78.1	62.5	87.7	93.3
SEP+Co ^{OSCAR}	43.7	72.1	81.2	68.2	90.4	94.3
JOIN+Co ^{OSCAR}	45.6	73.0	82.3	69.0	90.3	94.7

Table 5: Results with larger target collections. The dataset underlined indicates the actual standard task with the corresponding task data and labels used, while the instances from the datasets in *italic* are used as additional non-relevant test examples (i.e., distractors in the search space).

distribution data. We now observe more salient performance differences, which were less apparent or lacking with the smaller benchmarks. The pure EMB-based approach now substantially underperforms SEP/JOIN+CO variants. The JOIN+CO remains the best-scoring variant overall.

Zero-Shot Performance. Relying on multi-modal and multilingual representations fine-tuned for cross-modal retrieval, the proposed methods should also generalize to new unseen captions and images beyond the dataset used for fine-tuning. Therefore, we directly transfer the model fine-tuned on one dataset to the test data of another dataset (e.g., fine-tune on MSCOCO data, test on Flickr30k). As baselines, we use the reported zero-shot results of UNITER [9] for Flickr30k¹³ and we also evaluate the CLIP model.¹⁴

The zero-shot results, provided in Table 6, reveal that the CE variant slightly outperforms other approaches when transferring from Flickr30k to MSCOCO, while JOIN+Co^{OSCAR} remains competitive. However, for the opposite direction, we achieve considerable performance gains with the JOIN+Co^{OSCAR} variant. On CC, all variants considerably underperform CLIP; we speculate that it might be due to a more diverse set of images included in CC, including illustrations, which neither exist in MSCOCO nor Flickr30k. This means that CLIP has a considerable advantage on CC due to its exposure to massive amounts of data during pretraining.

Multilingual zero-shot results, where we fine-tune on the English Multi30k captions and test on the captions in other languages, are shown in Table 7. Cooperative approaches again excel; the highest scores are achieved by SEP+Co^{M3P}.

Sample Efficiency. We also analyze how the amount of image-text data for fine-tuning impacts the final retrieval performance; we therefore sample smaller datasets from the full MSCOCO training set, covering 1k, 10k, and 50k im-

¹³They do not report results for MSCOCO.

¹⁴CLIP has been trained on large amounts of multi-modal data specifically to learn strong representations for zero-shot use on various tasks. It is currently the best performing EMB-based model with published weights.

Loss	Image Retrieval			Text Retrieval			Image Retrieval			Text Retrieval			Image Retrieval			Text Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
	MSCOCO 5k						Flickr30k 1k						CC 14k					
JOIN+CO ^{OSCAR} _{In-Domain}	54.7	81.3	88.9	70.8	91.0	95.2	76.4	93.6	96.2	89.4	97.7	99.0	—	—	—	—	—	—
CE ^{UNITER}	—	—	—	—	—	—	66.2	88.4	92.9	80.7	95.7	98.0	—	—	—	—	—	—
CE ^{OSCAR}	47.8	75.7	84.6	61.8	86.2	92.0	67.2	88.5	92.7	81.0	95.5	97.8	—	—	—	—	—	—
CLIP	30.4	56.1	66.9	50.1	74.8	83.6	61.1	85.9	91.8	81.9	95.0	97.5	30.8	52.7	61.3	32.1	53.9	63.0
EMB ^{OSCAR}	37.6	64.4	75.0	52.0	78.1	86.3	63.3	86.4	91.6	78.2	94.0	97.3	13.8	29.4	37.9	14.4	29.6	37.6
SEP+CO ^{OSCAR}	47.6	73.9	81.2	62.8	83.8	88.7	67.6	89.0	93.1	82.4	96.3	98.2	16.8	34.3	41.9	17.0	33.5	41.5
JOIN+CO ^{OSCAR}	47.6	74.5	82.6	63.9	85.7	91.0	70.0	90.2	94.1	83.7	96.8	97.9	16.7	34.7	43.6	17.5	34.6	43.5

Table 6: Results for zero-shot evaluation on Flickr30k, MSCOCO, and CC. For Flickr30k and MSCOCO zero-shot results we train on the respective other datasets. For CC results we have trained on Flickr30k. JOIN+CO^{OSCAR}_{In-Domain} is the in-domain performance for the JOIN+CO approach and represents an upper-bound in experiment.

Model	en	de	fr	cs	Avg
CE ^{M3P} [24]	86.0	48.8	39.4	38.8	42.3
EMB ^{M3P}	81.3	52.4	49.7	39.6	47.2
CE ^{M3P}	84.2	52.6	49.6	33.4	45.2
SEP+CO ^{M3P}	84.4	55.6	52.2	39.8	49.2
JOIN+CO ^{M3P}	83.5	54.2	48.4	39.4	47.3

Table 7: Multilingual image-text retrieval results (in mR) on Multi30k. Models are trained on the English data only. Mean results are for the non-English languages.

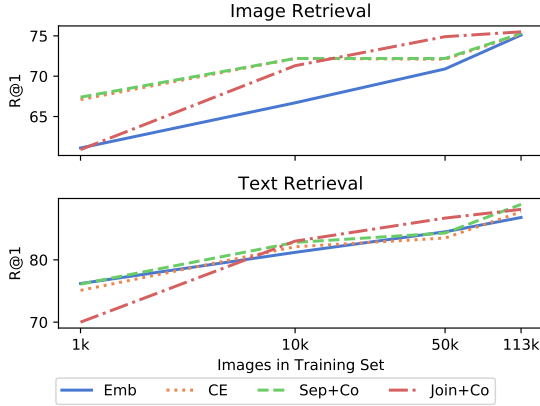


Figure 3: Impact of data size for fine-tuning on retrieval performance. MSCOCO training and test data; OSCAR as the underlying Transformer. Y-axis is in \log_{100} -scale.

ages with their captions (5 per image). The results, shown in Figure 3, reveal that embedding-based approaches in general are considerably less sample-efficient than cross-encoders. They particularly struggle in the lowest-data scenario with only 1k images available; this is also reflected in the lower performance of JOIN+CO in the 1k setup. A reason behind the more effective adaptation of CE to low-data regimes might be their richer “input consumption”: starting from 1k instances and 5k captions, CE runs a whole grid of $1k \times 5k$ items through its network, which might provide more learning signal with fewer data available. On the other hand, EMB-based approaches are expected to learn effective encoders of both modalities separately based solely on 1k images and 5k

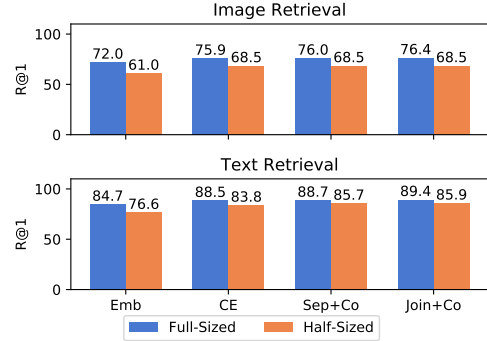


Figure 4: Half-sized vs. full-sized models on Flickr30k.

captions, without any cross-modal interaction.

Parameter Efficiency. We also provide a simple parameter efficiency analysis by initializing the models with pre-trained OSCAR weights, but only passing the representations through every second layer, effectively halving the total amount of Transformer parameters. The results are shown in Figure 4. The performance with all approaches using the “halved” model is around $\sim 90\%$ of the performance with the full Transformer. Overall, the JOIN+CO method again achieves the highest scores. This suggests that the proposed fine-tuning approaches are applicable also to smaller models, with similar relative trends in retrieval results.

7. Conclusion

We have proposed a novel framework that converts pre-trained multi-modal Transformers into effective *and* efficient cross-modal retrieval models. The framework is applicable to any pretrained model, and combines the efficiency of embedding-based (EMB) approaches with the accuracy of computationally more demanding cross-encoding (CE) approaches. Their synergistic effect at retrieval is achieved through a cooperative retrieve-and-rerank regime, where the initial retrieval from a large target collection is performed via efficient EMB approaches, followed by another accuracy-driven step via a CE model. Moreover, we have proposed a parameter-efficient joint fine-tuning regime which blends

EMB and CE into a single model with shared weights. Our results with state-of-the-art pretrained models across a range of standard monolingual and multilingual cross-modal retrieval tasks and setups have validated the strong performance of such cooperative and joint approaches; at the same time, we have demonstrated their retrieval efficiency, which makes them viable in realistic retrieval scenarios with large target collections. In future work, we will put more focus on zero-shot and few-shot retrieval scenarios, and expand the approach to more languages, modalities, and retrieval tasks.

Acknowledgments

Jonas Pfeiffer is supported by the LOEWE initiative (Hesse, Germany) within the emergenCITY center. The work of Nils Reimers has been supported by the German Research Foundation through the German-Israeli Project Cooperation (DIP, grant DA 1600/1-1 and grant GU 798/17-1) and has been funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE. The work of Ivan Vulić is supported by the ERC Consolidator Grant LEXICAL: Lexical Acquisition Across Languages (no 648909).

We thank Kevin Stowe and Christopher Klammer for insightful feedback and suggestions on a draft of this paper.

References

- [1] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018. 3
- [2] Alexandr Andoni and Piotr Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *2006 47th annual IEEE symposium on foundations of computer science (FOCS'06)*, 51(1):117–122, 2008. 1
- [3] Sunil Arya and David M. Mount. Approximate nearest neighbor queries in fixed dimensions. In *Proceedings of the Fourth Annual ACM/SIGACT-SIAM Symposium on Discrete Algorithms*, 25-27 January 1993, Austin, Texas, USA, pages 271–280, 1993. 1
- [4] Sunil Arya, David M. Mount, Nathan S. Netanyahu, Ruth Silverman, and Angela Y. Wu. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *Journal of the ACM (JACM)*, 45(6):891–923, 1998. 1
- [5] Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323, Belgium, Brussels, Oct. 2018. Association for Computational Linguistics. 5
- [6] Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. Multimodal pretraining unmasked: Unifying the vision and language berts. *arXiv preprint*, abs/2011.15124, 2020. 1, 2
- [7] Andrea Burns, Donghyun Kim, Derry Wijaya, Kate Saenko, and Bryan A. Plummer. Learning to scale multilingual representations for vision-language tasks. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IV*, volume 12349 of *Lecture Notes in Computer Science*, pages 197–213. Springer, 2020. 1, 2, 7
- [8] Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research*, 11:1109–1135, 2010. 3
- [9] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: universal image-text representation learning. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 104–120. Springer, 2020. 1, 2, 3, 6, 7
- [10] Muthuraman Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Learning cross-lingual sentence representations via a multi-task dual-encoder model. In *Proceedings of the 4th Workshop on Representation Learning for NLP*, pages 250–259, 2019. 2
- [11] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics, 2020. 3
- [12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. 2, 3
- [13] Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation*, pages 215–233, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. 5
- [14] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. 5
- [15] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: improving visual-semantic embeddings with hard negatives. In *British Machine Vision Conference 2018*,

- BMVC 2018, Newcastle, UK, September 3-6, 2018, page 12. BMVA Press, 2018. 1, 2, 4, 6
- [16] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding. *arXiv preprint*, abs/2007.01852, 2020. 1, 2
- [17] Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc’Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2121–2129, 2013. 2
- [18] 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 *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 1, 2, 3, 6
- [19] Mandy Guo, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Cer, Gustavo Hernandez Abrego, Keith Stevens, Noah Constant, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. Effective parallel corpus mining using bilingual sentence embeddings. In *Proceedings of WMT*, pages 165–176, 2018. 1
- [20] Kiana Hajebi, Yasin Abbasi-Yadkori, Hossein Shahbazi, and Hong Zhang. Fast approximate nearest-neighbor search with k-nearest neighbor graph. In *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, pages 1312–1317, 2011. 1
- [21] Matthew Henderson, Inigo Casanueva, Nikola Mrksic, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulic. ConveRT: Efficient and accurate conversational representations from transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020*, pages 2161–2174. Association for Computational Linguistics, 2020. 1, 2
- [22] Matthew Henderson, Ivan Vulic, Daniela Gerz, Inigo Casanueva, Pawel Budzianowski, Sam Coope, Georgios Spithourakis, Tsung-Hsien Wen, Nikola Mrksic, and Pei-Hao Su. Training neural response selection for task-oriented dialogue systems. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5392–5404. Association for Computational Linguistics, 2019. 2
- [23] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint*, abs/1703.07737, 2017. 4
- [24] Haoyang Huang, Lin Su, Di Qi, Nan Duan, Edward Cui, Taroon Bharti, Lei Zhang, Lijuan Wang, Jianfeng Gao, Bei Liu, Jianlong Fu, Dongdong Zhang, Xin Liu, and Ming Zhou. M3P: learning universal representations via multi-task multilingual multimodal pre-training. *arXiv preprint*, abs/2006.02635, 2020. 1, 2, 3, 4, 5, 7, 8
- [25] Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 1, 2
- [26] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint*, abs/2102.05918, 2021. 2
- [27] Jeff Johnson, Matthijs Douze, and Herve Jegou. Billion-scale similarity search with gpus. *arXiv preprint*, abs/1702.08734, 2017. 4
- [28] Donghyun Kim, Kuniaki Saito, Kate Saenko, Stan Sclaroff, and Bryan A. Plummer. MULE: multimodal universal language embedding. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 11254–11261. AAAI Press, 2020. 1, 2, 7
- [29] Eyal Kushilevitz, Rafail Ostrovsky, and Yuval Rabani. Efficient search for approximate nearest neighbor in high dimensional spaces. *SIAM Journal on Computing*, 30(2):457–474, 2000. 1
- [30] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IV*, volume 11208 of *Lecture Notes in Computer Science*, pages 212–228. Springer, 2018. 2, 6
- [31] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 11336–11344. AAAI Press, 2020. 1, 2, 3, 6
- [32] Xiujuan Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 121–137. Springer, 2020. 1, 2, 3, 4, 5, 6
- [33] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer, 2014. 2, 3
- [34] Robert Litschko, Ivan Vulic, Simone Paolo Ponzetto, and Goran Glavas. Evaluating multilingual text encoders for unsupervised cross-lingual retrieval. In *Advances in Information*

- Retrieval - 43rd European Conference on IR Research, ECIR 2020, Online, March 28 - April 1, 2021.* [2](#)
- [35] Ting Liu, Andrew W. Moore, Alexander G. Gray, and Ke Yang. An investigation of practical approximate nearest neighbor algorithms. In *Advances in Neural Information Processing Systems 17 [Neural Information Processing Systems, NIPS 2004, December 13-18, 2004, Vancouver, British Columbia, Canada]*, pages 825–832, 2004. [1](#)
- [36] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint*, abs/1907.11692, 2019. [2](#)
- [37] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. [5](#)
- [38] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23, 2019. [1](#), [2](#)
- [39] Xiaopeng Lu, Tiancheng Zhao, and Kyusong Lee. VisualSparta: Sparse transformer fragment-level matching for large-scale text-to-image search. *arXiv preprint*, abs/2101.00265, 2021. [2](#), [3](#), [6](#)
- [40] Shweta Mahajan, Teresa Botschen, Iryna Gurevych, and Stefan Roth. Joint wasserstein autoencoders for aligning multimodal embeddings. In *2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019*, pages 4561–4570, 2019. [2](#)
- [41] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2156–2164. IEEE Computer Society, 2017. [1](#)
- [42] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pages 1143–1151, 2011. [3](#)
- [43] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2641–2649, 2015. [2](#), [3](#)
- [44] 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. Learning transferable visual models from natural language supervision. *arXiv preprint*, abs/2103.00020, 2021. [2](#)
- [45] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics, 2019. [1](#), [2](#), [4](#)
- [46] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99, 2015. [3](#)
- [47] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 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 (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics. [3](#), [5](#)
- [48] Botian Shi, Lei Ji, Pan Lu, Zhendong Niu, and Nan Duan. Knowledge aware semantic concept expansion for image-text matching. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5182–5189. ijcai.org, 2019. [1](#), [2](#), [6](#)
- [49] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: pre-training of generic visual-linguistic representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. [3](#)
- [50] Hao Tan and Mohit Bansal. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 5099–5110. Association for Computational Linguistics, 2019. [1](#), [2](#)
- [51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. [1](#), [2](#)
- [52] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):394–407, Feb. 2019. [1](#)
- [53] Yaxiong Wang, Hao Yang, Xueming Qian, Lin Ma, Jing Lu, Biao Li, and Xin Fan. Position focused attention network for image-text matching. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 3792–3798. International Joint Conferences on Artificial Intelligence Organization, 7 2019. [2](#), [6](#)

- [54] Jonatas Wehrmann, Maurício Armani Lopes, Douglas M. Souza, and Rodrigo C. Barros. Language-agnostic visual-semantic embeddings. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 5803–5812. IEEE, 2019. [1](#), [2](#), [7](#)
- [55] John Wieting, Kevin Gimpel, Graham Neubig, and Taylor Berg-Kirkpatrick. Simple and effective paraphrastic similarity from parallel translations. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4602–4608. Association for Computational Linguistics, 2019. [2](#)
- [56] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint*, abs/2007.00808, 2020. [5](#), [12](#)
- [57] Yinfei Yang, Steve Yuan, Daniel Cer, Sheng-yi Kong, Noah Constant, Petr Pilar, Heming Ge, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Learning semantic textual similarity from conversations. In *Proceedings of the Third Workshop on Representation Learning for NLP*, pages 164–174, 2018. [2](#)
- [58] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions On Multimedia Computing, Communications And Applications*, 16(2):51:1–51:23, 2020. [1](#)

A. Appendix

A.1. Retrieving Top_k

We analyze different values for k for top_k retrieval of the embedding components in Table 8. Selecting small values for k significantly decreases the retrieval latency, as less instances need to be cross-encoded. However, selecting k values that are too small can come at a cost of precision, as the true positive instance might not be among the top_k retrieved instances of the embedding model (EMB). We find that with $k = 20$ we achieve the best trade-off between precision and retrieval latency.

A.2. Combining Ranking

We evaluate the ranking score combination of the two components CO-EMB and CO-CE in Table 11. We combine the ranking of the embedding model and the cross-encoder by summing over the scores using two different variations:

We directly add the scores in a weighted sum:

$$\text{ADD}_\lambda(e, c) = \lambda e + (1 - \lambda)c \quad (4)$$

where e and c are the embedding and cross-encoder similarity scores respectively and λ is a weighting parameter. The cross-encoder scores have been processed with a sigmoid

function so that both e and c are in the same value range. The final ranking is then defined by $\text{ADD}_\lambda(e, c)$.

We additionally experiment with 0-1 normalized ranking scores, with

$$\text{NORM}(e) = \frac{e - |\min(E)|}{\max(e) - \min(E)} \quad (5)$$

where $e \in E$. Amounting to:

$$\text{NORM_ADD}_\lambda(e, c) = \lambda \text{NORM}(e) + (1 - \lambda) \text{NORM}(c). \quad (6)$$

However, we find that relying solely on the cross-encoding achieves the best results and that adding embedding rankings does not improve performance.

A.3. Negative Sampling Strategies

An alternative negative sampling approach is *Approximate nearest neighbour Negative Contrastive Estimation (ANCE)* [56]. Here the cross-product distances of all examples are computed. With a list of the closest neighbours for each positive example, identifying truly hard negatives no longer relies on sufficiently large batch sizes as in BHN. However, with increasing number of update steps of the model the ANCE list becomes outdated and needs to be recomputed. We analyze the impact of the two negative sampling strategies ANCE and BHE combined with the objective functions triplet T and cross-entropy CE, on the downstream performance.

In Table 9 we present results for the embedding based model EMB^{OSCAR} and compare it to the OSCAR embeddings before training them using twin-networks (-). We find that the model cannot be used as an embedding model without training as it is not able to retrieve meaningful instances. When comparing the different negative sampling strategies we do not find that the more complex ANCE sampling approaches improves the performance in general.

In Table 10 we present results for the different negative sampling strategies for the JOIN+CO^{OSCAR} approach. Similar to the embedding based approach ANCE negative sampling does not have an advantage over simpler in-batch negative sampling (BHE) and normal cross-entropy loss (E) without negative sampling.

A.4. Additional Reranking Examples

We present additional examples for re-ranking in Figures 5-10. We select only examples where the display retrieved top- k images and re-ranked top- k images are identical to show-case how the cross-encoder changes the ranking.

Model	k	Image Retrieval			Text Retrieval			Image Retrieval			Text Retrieval			Image Retrieval			Text Retrieval		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
SEP+Co		MSCOCO 1k						MSCOCO 5k						Flickr30k					
	10	75.4	94.8	97.2	88.4	98.8	99.7	53.2	80.3	86.6	71.1	90.9	94.3	75.9	92.2	93.4	89.2	97.8	98.4
	20	75.3	95.2	98.1	87.9	98.9	99.8	52.8	80.5	88.5	70.2	91.6	95.0	76.0	93.0	95.0	88.7	98.3	99.2
	50	75.2	95.0	98.2	87.9	99.1	99.8	52.6	80.1	88.4	70.1	91.4	95.5	75.9	93.4	96.3	88.9	98.4	99.4
JOIN+Co	10	75.4	95.5	97.8	88.0	98.8	99.9	54.8	81.2	88.0	70.9	91.2	95.0	76.5	93.2	95.0	88.9	97.3	98.6
	20	75.5	95.4	98.2	88.1	98.6	99.5	54.7	81.3	88.9	70.8	91.0	95.2	76.4	93.6	96.2	89.4	97.7	99.0
	50	75.4	95.4	98.3	88.2	98.4	99.4	54.6	81.2	88.8	70.7	91.1	95.3	76.5	93.5	96.5	89.1	98.0	98.9

Table 8: Results with SEP+CO and JOIN+CO re-ranking the top- k candidates. **Bold** numbers indicate which k value resulted in the highest score for each separate model.

Loss	Image Retrieval			Text Retrieval			Image Retrieval			Text Retrieval			Image Retrieval			Text Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
	MSCOCO 1k						MSCOCO 5k						Flickr30k 1k					
-	1.0	4.2	7.1	0.9	6.8	10.6	0.4	1.4	2.4	0.5	2.1	3.7	0.8	3.7	5.7	0.8	2.7	6.0
T ^{BHE}	75.1	94.7	97.6	86.8	98.4	99.7	52.2	80.2	88.0	66.9	90.1	95.0	72.0	91.0	94.7	84.7	97.1	98.7
T ^{BHE} + T ^{ANCE}	73.2	93.9	97.2	86.5	98.4	99.8	51.3	78.9	87.4	65.6	89.1	94.4	72.4	91.4	95.2	85.7	97.5	99.1

Table 9: Results of the embedding models trained with triplet loss T^{BHE} and sampled triplet loss T^{ANCE} , alongside an embedding model initialized with the pre-training weights without any training (-), for both MSCOCO and Flickr30k.

Loss	Image Retrieval			Text Retrieval			Image Retrieval			Text Retrieval			Image Retrieval			Text Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
	MSCOCO 1k						MSCOCO 5k						Flickr30k 1k					
$T^{BHE} + E$	75.5	95.4	98.2	88.1	98.6	99.5	54.7	81.3	88.9	70.8	91.0	95.2	76.4	93.6	96.2	89.4	97.7	99.0
$T^{BHE} + T^{ANCE} + E$	75.5	95.2	98.1	87.6	98.3	99.9	54.0	80.7	88.6	70.2	90.7	95.1	76.3	93.5	96.4	88.6	97.6	99.0
$T^{BHE} + E^{ANCE}$	70.5	93.2	97.6	80.9	97.7	99.2	47.9	76.8	86.3	63.5	87.5	93.8	76.8	93.9	96.1	89.6	97.9	99.1
$T^{BHE} + T^{ANCE} + E^{ANCE}$	71.9	94.6	98.0	84.2	97.9	99.3	50.0	77.9	86.7	63.9	87.4	93.1	76.8	93.5	96.3	88.0	97.7	98.7

Table 10: Results of the JOIN+CO^{OSCAR} models trained with different negative sampling strategies. The triplet loss (T) is used for the embedding part whereas cross-entropy loss (E) is used for the cross-encoding part.

Model	Sum	λ	Image Retrieval			Text Retrieval		
			R@1	R@5	R@10	R@1	R@5	R@10
SEP+CO	-	0.0	76.0	93.0	95.0	88.7	98.3	99.2
	ADD	0.1	76.0	92.7	94.8	86.4	98.7	99.2
		0.5	75.7	92.6	94.7	85.9	98.5	99.2
		0.9	74.5	92.5	94.7	85.1	98.3	99.2
	NORM_ADD	0.1	70.8	90.2	93.8	86.2	98.5	99.2
		0.5	70.7	90.3	93.7	85.4	98.4	99.2
		0.9	70.3	90.1	93.7	83.8	97.6	98.8
	-	0.0	76.4	93.6	96.2	89.4	97.7	99.0
JOIN+CO	ADD	0.1	76.7	93.3	95.8	88.5	98.0	99.1
		0.5	75.6	93.1	95.5	87.2	97.8	99.1
		0.9	74.6	92.8	95.5	87.3	97.8	99.1
	NORM_ADD	0.1	72.8	92.0	95.2	87.6	97.9	99.2
		0.5	72.5	92.0	95.2	87.3	97.9	99.0
		0.9	72.3	91.8	95.2	86.4	97.0	99.0
	-	0.0	76.4	93.6	96.2	89.4	97.7	99.0
	-	0.0	76.0	93.0	95.0	88.7	98.3	99.2

Table 11: Results on Flickr30k for different combinations of the embedding and cross-encoder scores using the summing functions ADD_λ and $NORM_ADD_\lambda$ and different values for λ . - indicates the results for re-ranking using only the cross-encoder.

Caption: *This man is wearing a red helmet and flip-flops and driving a Spyder.*



(a) Target Image



(b) Retrieved images with JOIN+CO-EMB^{OSCAR}



(c) Re-Ranked images with JOIN+CO-CE^{OSCAR}

Figure 5

Caption: *Two people in a yellow kayak come along side a larger white kayak.*



(a) Target Image



(b) Retrieved images with JOIN+CO-EMB^{OSCAR}



(c) Re-Ranked images with JOIN+CO-CE^{OSCAR}

Figure 6

Caption: *A man with long hair and a beard is strimming a guitar.*



(a) Target Image



(b) Retrieved images with JOIN+CO-EMB^{OSCAR}



(c) Re-Ranked images with JOIN+CO-CE^{OSCAR}

Figure 7

Caption: *A man is standing center stage with a microphone as flames shoot upward behind him.*



(a) Target Image



(b) Retrieved images with JOIN+CO-EMB^{OSCAR}



(c) Re-Ranked images with JOIN+CO-CE^{OSCAR}

Figure 8

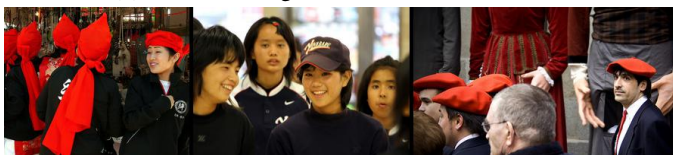
Caption: *An Asian girl wearing a red hat along side several other with red hats as well.*



(a) Target Image



(b) Retrieved images with JOIN+CO-EMB^{OSCAR}



(c) Re-Ranked images with JOIN+CO-CE^{OSCAR}

Figure 9

Caption: *The dog jumps up to catch the Frisbee.*



(a) Target Image



(b) Retrieved images with JOIN+CO-EMB^{OSCAR}



(c) Re-Ranked images with JOIN+CO-CE^{OSCAR}

Figure 10