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

Gregor Geigle[*1], **Jonas Pfeiffer**[*1], **Nils Reimers**[1], **Ivan Vulić**[2], **Iryna Gurevych**[1]

[1]Ubiquitous Knowledge Processing Lab, Technical University of Darmstadt, Germany
[2]Language Technology Lab, University of Cambridge, United Kingdom
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 that turns any pretrained text-image multi-modal model into an efficient retrieval model. The framework is based on a cooperative retrieve-and-rerank approach that combines: **1)** twin networks (i.e., a bi-encoder) 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]

1 Introduction

Information-rich and efficient methods for dealing with large unstructured data in both computer vision and 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 *bi-encoder*[2] 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 identify the most similar examples for each query in the data collection via nearest-neighbor search (Arya et al., 1998; Kushilevitz et al., 2000; Liu et al., 2004; Andoni and Indyk, 2008; Hajebi et al., 2011).

These bi-encoder approaches have already been shown to achieve reasonable performance in search and retrieval applications, both monolingually for English (Nam et al., 2017; Faghri et al., 2018; Zheng et al., 2020; Wang et al., 2019a; Shi et al., 2019) and in multilingual contexts (Gella et al., 2017; Kádár et al., 2018; Kim et al., 2020; Wehrmann et al., 2019; Burns et al., 2020). However, 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 (Vaswani et al., 2017). Such so-called multi-modal *cross-encoders* (CE) (Tan and Bansal, 2019; Lu et al., 2019; Chen et al., 2020; Li et al., 2020a; Gan et al., 2020; Li et al., 2020b; Ni et al., 2021) pass each text-image pair through the multi-modal encoder to compute their similarity, see Figure 1a.

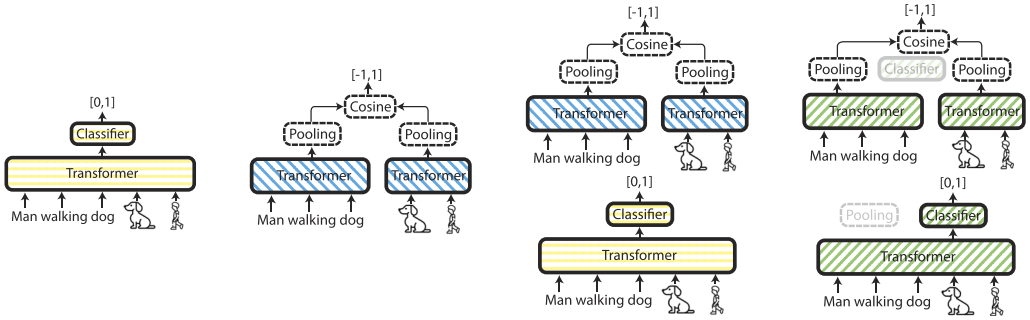
While the results accomplished by the CE methods look impressive (Li et al., 2020b; Bugliarello et al., 2021; Ni et al., 2021), this comes at a prohibitive cost. In particular, they have extremely high search latency: Processing a single text query with an image collection of 1M items may take up to 36 minutes using a single NVIDIA V100 GPU (see Table 3). Due to this issue, they are evaluated only with extremely small

[*]Both authors contributed equally to this work.

[1]We release the code and model weights at github.com/UKPLab/MMT-Retrieval.

[2]Also frequently referred to as *dual-encoder*.

Transactions of the Association for Computational Linguistics, vol. 10, pp. 503–521, 2022. https://doi.org/10.1162/tacl_a_00473
Action Editor: Jimmy Lin. Submission batch: 7/2021; Revision batch: 11/2021; Published 5/2022.



(a) CE Model (b) BE Model (c) SEP+COOP (d) JOINT+COOP

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

benchmarks, that is, the maximum size of typical image collections for image retrieval tasks is 5k images, and evaluation still lasts ≈50 hours (see Table 4).[3] 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, Transformer-based architectures have recently been integrated with bi-encoder (BE) methods (Guo et al., 2018; Reimers and Gurevych, 2019; Humeau et al., 2020; Henderson et al., 2020; Feng et al., 2020, *inter alia*), yielding computationally more efficient sentence encoders. Instead of jointly encoding sentence pairs with cross-attention, a pretrained Transformer model (e.g., BERT [Devlin et al., 2019]) 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 (Reimers and Gurevych, 2019; Litschko et al., 2021), these encoders cannot match the task performance of cross-encoders (Humeau et al., 2020).

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 BE 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 BE and CE methods across a range of image search evaluation benchmarks. We then propose two novel models that aim to blend the main strengths of CE and BE. The idea behind the first model variant, termed *cooperative* (SEP+COOP), 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 BE 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* (JOINT+COOP) model variant also operates in the same retrieve-and-rerank setup, but it now trains a multi-modal cross-encoder and a multi-modal BE model jointly with tied weights, as illustrated in Figure 1d. The retrieve step, where efficiency is paramount, is again executed by the BE 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 JOINT+COOP and SEP+COOP 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 (Li et al., 2020b) (experiments in English) and M3P (Ni et al., 2021) (multilingual), and we demonstrate consistent improvements over the original OSCAR model on the standard benchmarks MSCOCO and Flick30k and improvements over the original M3P in multiple languages on the Multi30k dataset. We empirically validate huge efficiency benefits of the proposed framework.

Contributions. **1)** We construct and systematically evaluate twin-networks combined with

[3]Consequently, 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.

multi-modal Transformers (BE); they outperform all previous bi-encoder approaches, but lag behind their CE counterparts. **2)** We evaluate BE 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-BE model (JOINT+ COOP), 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.

2 Related Work

Efficient approaches to cross-modal image-text retrieval relied on the induction of shared multi-modal visual-semantic embedding spaces (VSEs) (Frome et al., 2013; Faghri et al., 2018; Shi et al., 2019; Mahajan et al., 2019). In a multilingual setup, all languages share the same embedding space along with the visual data (Kim et al., 2020; Wehrmann et al., 2019; Burns et al., 2020). More recently, attention-based cross-encoder models, typically based on Transformer architectures (Vaswani et al., 2017) have considerably outperformed the VSE-based approaches. However, this comes at a severe cost of decreased retrieval efficiency and increased latency (Lee et al., 2018; Wang et al., 2019b). The current state-of-the-art multi-modal models jointly encode and cross-attend over text tokens and image features (Lu et al., 2019; Tan and Bansal, 2019; Chen et al., 2020; Li et al., 2020a; Gan et al., 2020; Li et al., 2020b; Bugliarello et al., 2021; Ni et al., 2021, *inter alia*). These CE methods leverage image captioning datasets such as MSCOCO (Lin et al., 2014) and Flick30k (Plummer et al., 2015) and train a classification head that 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.

To handle this quadratic increase, we use a cooperative retrieve-and-rerank approach. Although to the best of our knowledge this has not been proposed for cross-modal settings, it has a long history in NLP, where Yates et al. (2021) date it back to the 1960s (Simmons, 1965). Until recently, bag-of-words methods (BoW; e.g., BM25) were commonly used for the first retrieval step. For the second step, pretrained language models (LMs) were fine-tuned to either rerank candidates (Nogueira and Cho, 2019; Nogueira et al., 2019) or—for question-answering tasks—directly generated the answer span (Yang et al., 2019). More recent work on text-based retrieval and QA tasks has moved away from BoW methods towards learned (neural) models for the first retrieval step (Karpukhin et al., 2020; Qu et al., 2021; Xiong et al., 2021).

Our work is inspired by recent BE-based approaches in unimodal text-only setups. Here, LMs are fine-tuned via twin-network architectures on auxiliary tasks such as semantic textual similarity (Reimers and Gurevych, 2019; Humeau et al., 2020), paraphrasing (Wieting et al., 2019), response retrieval (Yang et al., 2018; Henderson et al., 2019; Henderson et al., 2020; Humeau et al., 2020), or translation ranking (Chidambaram et al., 2019; Feng et al., 2020). This effectively turns the LMs into universal *sentence encoders* which can then be used off-the-shelf for efficient text-based monolingual and cross-lingual retrieval (Litschko et al., 2021). In this work, we first extend this idea to multi-modal setups, and then show that our cooperative and joint approaches yields improved cross-modal retrieval models, maintaining retrieval efficiency.

Joint approaches like our JOINT+COOP model, which aim to align the retriever and reranker can be found in different forms: Boualili et al. (2020) ''mark'' exact matches from the bag-of-words retrieval for the reranker; Yan et al. (2021) share the parameters between a passage expander (which adds more relevant terms for a bag-of-words retriever) and the reranker; Hofstätter et al. (2020) distill knowledge from the reranker into the retriever model with soft labels generated by the teacher. Specifically for question-answering—where a two stage retriever-reader setup similar to the retrieve-and-rerank approach is common—research aims to synchronize the models through knowledge distillation from the reader to the retriever (Yang and Seo, 2020; Izacard and Grave, 2021) or by directly training both models end-to-end (Lee et al., 2019; Sachan et al., 2021a,b). The challenge here is that the reader and the retriever are coupled—the reader requires candidates from

the retriever that contain the solution. Our proposed reranker side-steps this problem as it uses no candidates from the retriever during training and only learns if a given input pair is (dis)similar. This way, we can train both components, the retriever and the reranker, side-by-side and align them by sharing their weights.

The work most closely related to ours includes contemporaneous models: ALBEF (Li et al., 2021), CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), and VisualSparta (Lu et al., 2021). ALBEF includes contrastive learning as one of its pretraining tasks but then uses a CE approach for downstream retrieval. 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 BE-based (contrastive) learning with cross-encoders for improved and efficient retrieval.[4] Finally, VisualSparta (Lu et al., 2021) 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.[5] 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 (Chen et al., 2020; Li et al., 2020a; Gan et al., 2020; Li et al., 2020b; Ni et al., 2021). The Transformer weights and text embeddings are typically initialized with weights of a pretrained LM (e.g., BERT [Devlin et al., 2019] for English, XLM-R [Conneau et al., 2020] for multilingual models), where the corresponding vocabulary and tokenizer are utilized. Images are preprocessed via object detection models such as Faster R-CNN (Ren et al., 2015) to extract feature representations for regions of interest (Anderson et al., 2018). The image features are passed through an affine-transformation layer which learns to align the vision input with the pretrained 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, multi-modal Transformer models are trained with self-supervised objectives. For pretraining, image-caption datasets (i.e., MSCOCO [Lin et al., 2014], Flickr30k [Plummer et al., 2015], Conceptual Captions (CC) [Sharma et al., 2018], and SBU [Ordonez et al., 2011]) are utilized. The pretrained multi-modal model is subsequently fine-tuned with multi-modal downstream task data.

We focus on different *fine-tuning strategies* of the pretrained models for the downstream task of image-text retrieval. We illustrate these 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 binary cross-entropy (BCE) loss:

$$\mathcal{L}_{\text{CE}}(i,c) = -\Big(y \log \text{p}(i,c) + (1-y)\log(1-\text{p}(i,c))\Big)$$

[4] As both CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) disjoin the image and text components in their methods, cross-attention over the instances is not possible.

[5] For example, considering a query ''two dogs and one cat'', the model is unable to match the numbers to the animals yielding likely worse retrieval results.

$\mathrm{p}(i, c)$ 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).[6]

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 the single most relevant image i from an image collection I proceeds as follows:

$$\arg\max(\mathrm{p}(i, c), \forall i \in I) \quad (1)$$

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 pre-computed representations during retrieval.

3.2 Bi-Encoders

Training. Each image and text caption is passed separately through the pretrained Transformer model (Figure 1b). The contextualized representations are mean-pooled to represent the embedding of the respective image $\mathbf{i}$ and text caption $\mathbf{c}$.[7] 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 same corpus such that image-caption pairs/instances (i, c') and (i', c) do not occur in the corpus. The triplet loss is then:

$$\mathcal{L}_{\mathrm{BE}}(i, c) = [\cos(\mathbf{i}, \mathbf{c}') - \cos(\mathbf{i}, \mathbf{c}) + \alpha]^+ + [\cos(\mathbf{i}', \mathbf{c}) - \cos(\mathbf{i}, \mathbf{c}) + \alpha]^+ \quad (2)$$

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

[6]Some cross-encoders such as UNITER (Chen et al., 2020) and VL-BERT (Su et al., 2020) rely on another triplet loss function (Chechik et al., 2010); however, OSCAR (Li et al., 2020b) reports improved performance with BCE.

[7]Following Reimers and Gurevych (2019), we opt for mean pooling as the final ''aggregated'' embedding; it outperformed another standard variant, which uses the `[CLS]` token, in our preliminary experiments.

Sampling Negative Examples. 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 (Faghri et al., 2018). However, detecting such hard negatives is only possible with BE-based approaches, as cross-encoding all instances is computationally infeasible. We rely on the *In-Batch Hard Negatives (BHN)* method (Hermans et al., 2017), 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 BE approach enables *pre-encoding* of all items for efficient retrieval look-up.[8] For instance, a text query q is encoded with the bi-encoder and the most similar pre-encoded instance from an image collection I is retrieved: $\arg\max_{i \in I} \cos(\mathbf{i}, \mathbf{q})$.

This approach can scale to even billions of images (Johnson et al., 2021), 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 combine the benefits of the two model types (CE and BE) within a *cooperative* retrieval approach (SEP+COOP), as illustrated in Figure 1c.

Training and Retrieval. Two models, one CE (§3.1) and one BE (§3.2), are trained independently. Following that, the retrieval step is split into two stages. First, the efficient BE model is used to retrieve the *top-k* relevant items from the entire large collection, yielding a much smaller collection I_k: $I_k = \mathrm{top}_k(\{\cos(\mathbf{i}, \mathbf{q}) : \forall i \in I\})$,

[8]Note that pre-computing the embedding does come with increased storage and memory demands; e.g., with a base Transformer architecture this requires an additional $\approx$ 3kB of memory for each embedding. A corpus of 1M images would amount to $\approx$ 3GB of required storage.

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: $\arg\max_{i \in I'} \text{p}(i, c)$. This cooperative approach thus combines the benefits of both approaches and is able to efficiently retrieve instances.[9] 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* (i.e., '*bi-encode*'), 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 BCE loss (§3.1). In order to train the BE-based sub-model, we again rely on a twin architecture with a triplet loss from Eq. (2).

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 BE-based submodel, and then rerank its items with the CE submodel. We combine the best traits of CE and BE, while maintaining parameter efficiency. Using 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 (Li et al., 2020b) and M3P (Ni et al., 2021), respectively.

OSCAR is a single-stream multi-modal Transformer (Bugliarello et al., 2021), with its weights initialized with those of the pretrained BERT Base 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 Flick30k. They comprise 123k and 31.8k images, respectively, 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 test set. The standard Flickr30k test set consists of 1k images. In addition, we use the development set of Conceptual Captions (CC) (Sharma et al., 2018) 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 (Elliott et al., 2016, 2017; Barrault et al., 2018), which extends Flickr30k with 5 German and one French and Czech caption per image. Its test s et 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. (2020b) and Ni et al. (2021) unless noted otherwise.[10] 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 $\mathcal{L}_{\text{CE}}$. We use the AdamW optimizer (Loshchilov and Hutter, 2019) 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 development set performance.

4.1 Baselines and Model Variants

CE. Our main baselines are OSCAR and M3P models used in the standard CE setting, described in §3.1. We fully fine-tune the Transformer

[9] Retrieval time for 1M images: 94ms (GPU), 13s (CPU).

[10] Unlike Li et al. (2020b), we do not use object tags as additional input, as preliminary experiments suggested no improvement with object tags.

weights along with a randomly initialized classification head.[11] At retrieval, we cross-encode each text-image combination and rank them according to the corresponding probability, see Eq. (1).

BE. We rely on BHN negative sampling, finding that training for 30k steps, with a learning rate of $5e-5$, and with a margin $\alpha = 0.1$ works best.[12]

SEP+COOP. For the cooperative method without joint training (§3.3), we retrieve the top-20 instances with BE and rerank them via CE.[13]

JOINT+COOP. 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 hyperparameters are the same as with separately trained models. For retrieval, we again set $k = 20$. To demonstrate the benefits of cooperative retrieval, we also evaluate two non-cooperative variants originating from the joint model: **JOINT+CE** uses the CE sub-model for a single-step CE-style retrieval, while **JOINT+BE** operates in the fully BE retrieval setup.

The underlying pretrained Transformer is denoted with a superscript: For example, JOINT+COOP$^{\text{OSCAR}}$ denotes that: 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, and 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 (PT) 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 BE approach. This validates the effectiveness of Transformer architectures pretrained on large corpora for the retrieval task. R@1 scores with BE lag slightly behind the CE scores, but the respective R@10 scores are mostly on-par. This suggests that the BE 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. 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. Although 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 BE (sub)model does not rank this instance at the top position. The CE (sub)model then directly compares the instances, identifying that clothing is important and reranks the target examples accordingly.

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 BE approach efficiently retrieves 20 relevant examples, but the increased expressiveness of CE is required to refine the initially retrieved list. Moreover, the results in the cooperative setup even without joint training (SEP+COOP$^{\text{OSCAR}}$ and SEP+COOP$^{\text{M3P}}$) 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 BE model filters out false positives, which in turn makes the CE model more robust.

The results of the JOINT+COOP 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 JOINT+COOP variant offers the best trade-off between parameter and retrieval efficiency, achieving the peak scores on

[11] Training for 100k steps and a learning rate of $2e-5$ (OSCAR) or $5e-5$ (M3P) performed best.

[12] We also experimented with *Approximate-nearest-neighbor Negative Contrastive Estimation (ANCE)* (Xiong et al., 2021); however, it did not yield performance benefits.

[13] We provide an ablation study of different k values in §6. We have also experimented with training a CE model using hard negative samples from a pretrained BE model. However, the CE model is able to easily overfit on those negative examples, resulting in inferior performance.

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++ (Faghri et al., 2018)	**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 (Lee et al., 2018)	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 (Wang et al., 2019b)	—	—	—	—	—	—	**50.4**	**78.7**	**86.1**	70.0	**91.8**	95.0
	SCG (Shi et al., 2019)	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} (Chen et al., 2020)	48.4	76.7	85.0	63.3	87.0	93.1	72.5	92.4	96.1	85.9	97.1	98.8
	$CE^{Unicoder\text{-}VL}$ (Li et al., 2020a)	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} (Gan et al., 2020)	—	—	—	—	—	—	74.7	92.9	95.8	86.6	97.9	**99.2**
	CE^{OSCAR}† (Li et al., 2020b)	**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. Bi-Encoders *(Efficient for retrieval)*	VisualSparta (Lu et al., 2021)	44.4	72.8	82.4	—	—	—	57.4	82.0	88.1	—	—	—
	BE^{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+$COOP^{OSCAR}$	52.8	80.5	88.5	70.2	**91.6**	95.0	76.0	93.0	95.0	88.7	**98.3**	**99.2**
	JOINT+$COOP^{OSCAR}$	**54.7**	**81.3**	**88.9**	**70.8**	91.0	**95.2**	**76.4**	**93.6**	**96.2**	**89.4**	97.7	99.0
	JOINT+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
	JOINT+BE^{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 (PT) approaches. G2 denotes the results of recent cross-encoders with Transformers (CE^*; §3.1). † indicates the results taken directly from the literature (Li et al., 2020b), and ‡ indicates our own results achieved with the model weights. G3 covers efficient retrieval methods that either retrieve images based only on distance metrics (BE, §3.2), or rely on the SEP+COOP 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.

Type	Model	en	de	fr	cs	mean
G1. PT	MULE	70.3	64.1	62.3	57.7	63.6
	S-LIWE	76.3	72.1	63.4	59.4	67.8
	SMALR	74.5	69.8	65.9	64.8	68.8
G2. CE	CE^{M3P}†	**86.7**	**82.2**	73.5	70.2	78.2
	CE^{M3P}‡	83.7	79.4	76.5	74.6	78.6
G3. BE	BE^{M3P}	82.8	78.0	75.1	73.6	77.4
	SEP+$COOP^{M3P}$	84.8	80.5	**77.5**	**75.6**	**79.6**
	JOINT+$COOP^{M3P}$	83.0	79.2	75.9	74.0	78.0

Table 2: Results on Multi30k (multilingual setups). Following prior work (Ni et al., 2021), we report *mean Recall (mR)* scores: mR computes an 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 (Ni et al., 2021) and ‡ indicates our own results. MULE (Kim et al., 2020); S-LIWE (Wehrmann et al., 2019); SMALR (Burns et al., 2020); CE^{M3P}† (Ni et al., 2021).

the monolingual MSCOCO and Flickr30k benchmarks, and very competitive results on the multilingual Multi30k benchmark.

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

Table 3: Retrieval latency for one query with an image collection of 50k or 1M images (with pre-encoded images) using a single GPU/ CPU. Batch size for cross-encoding of the query with the images is 512. CPU is an Intel Xeon Gold 6154.

6 Further Analysis

We now discuss a series of additional experiments that 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. We empirically validate the time efficiency of our cooperative approaches for retrieval in an image search scenario (Table 3) and for evaluation on huge datasets (Table 4). To allow for a fair comparison between the approaches, we implement the entire retrieval

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

(a) Target Image

(b) Retrieved Top-6 with JOINT+BE$^{\text{OSCAR}}$

(c) Reranked Top-6 with JOINT+CE$^{\text{OSCAR}}$

Figure 2: We efficiently retrieve the top instances with the JOINT+BE$^{\text{OSCAR}}$ submodel to identify the (globally) most relevant target instances. The more precise, but less efficient JOINT+CE$^{\text{OSCAR}}$ submodel then disentangles the specific intricacies of the images. Ranking proceeds from left to right.

Model	**1k**	**5k**	**100k**
BE	5s	30s	7min
SEP/JOINT+COOP	5min	25min	8.5h
CE	2h	50h	2.3a*

Table 4: Evaluation time for the MSCOCO test sets of 1k, 5k, and 100k images on an NVIDIA V100 with batch size 512. The time includes bi-encoding images and text, i.e., the embeddings are not pre-computed. * denotes extrapolated values.

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

Table 5: Results with larger benchmarks. 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).

pipeline—from model to nearest-neighbor search—in PyTorch without additional optimization such as multi-processing or optimized nearest-neighbor search libraries like FAISS (Johnson et al., 2021).

Our measurements confirm the efficiency of BEs in comparison to CEs. The cooperative approaches, which only have to cross-encode a constant number of items invariant of the collection size, are close in retrieval latency to BE for image search and remain feasible even for large datasets.

Larger Benchmarks. The results in Table 1 indicate that current top-performance models achieve very high scores in absolute terms on the standard retrieval benchmarks. However, this is partially due to too small image collections with only a few thousand instances; 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 Tables 3–4, has prevented evaluation with larger collections. However, more efficient fully BE-based and SEP+COOP methods now enable evaluation on larger collections and in realistic scenarios.

We thus increase the benchmark size by merging test instances from different available evaluation sets. In particular, we construct a collection spanning 20k images: It blends the test sets of MSCOCO (5k instances), Flickr30k (1k), and the development set of CC (14k). Note that we simply augment the benchmarks 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-distribution data. We now observe in Table 5 more salient performance

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					
JOINT+COOP$^{\text{OSCAR}}_{\text{In-Domain}}$	*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$^{\text{UNITER}}$	—	—	—	—	—	—	66.2	88.4	92.9	80.7	95.7	98.0	—	—	—	—	—	—
CE$^{\text{OSCAR}}$	**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**
BE$^{\text{OSCAR}}$	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+COOP$^{\text{OSCAR}}$	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
JOINT+COOP$^{\text{OSCAR}}$	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 results we train on the respective other datasets. For CC results we train on Flickr30k. JOINT+COOP$^{\text{OSCAR}}_{\text{In-Domain}}$ is the in-domain performance for the JOINT+COOP approach and here represents the upper-bound.

differences, which were lacking with the smaller benchmarks. The pure BE-based approach now substantially underperforms SEP/JOINT+COOP variants. The JOINT+COOP does remain the best-scoring variant overall.

Zero-Shot Performance. Relying on multimodal 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 (Chen et al., 2020) for Flickr30k[14] and we also evaluate the CLIP model.[15]

The zero-shot results in Table 6, reveal that the CE variant slightly outperforms other approaches when transferring from Flickr30k to MSCOCO, while JOINT+COOP$^{\text{OSCAR}}$ remains competitive. However, for the opposite direction, we achieve considerable performance gains with the JOINT+COOP$^{\text{OSCAR}}$ 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+COOP$^{\text{M3P}}$.

Model	en	de	fr	cs	Avg
CE$^{\text{M3P}}$ (Ni et al., 2021)	86.0	48.8	39.4	38.8	*42.3*
BE$^{\text{M3P}}$	81.3	52.4	49.7	39.6	*47.2*
CE$^{\text{M3P}}$	84.2	52.6	49.6	33.4	*45.2*
SEP+COOP$^{\text{M3P}}$	84.4	**55.6**	**52.2**	**39.8**	***49.2***
JOINT+COOP$^{\text{M3P}}$	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. Avg results of non-English languages.

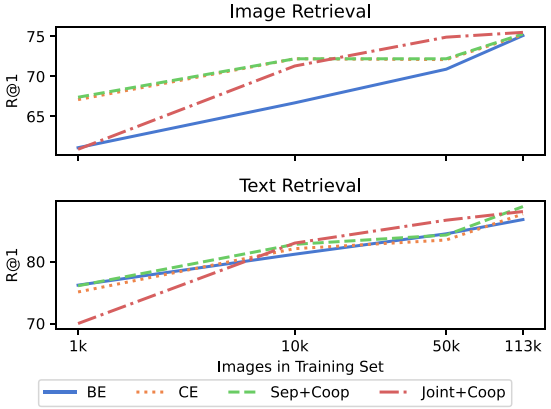


Figure 3: Impact of data size for fine-tuning on retrieval performance. MSCOCO training and test data; OSCAR as the underlying Transformer.

Sample Efficiency. We also analyze how the amount of image-text data for fine-tuning impacts the retrieval performance; we thus sample smaller datasets from the full MSCOCO training set, covering 1k, 10k, and 50k images with their captions (5 per image). The results in Figure 3 reveal that BE-based approaches in general are considerably less sample-efficient than cross-encoders. They

[14] They do not report results for MSCOCO.

[15] We use the ViT-B/32 model variant. Retrieval results from Radford et al. (2021) Table 13 use the (larger) ViT-L/14 variant that has not been released to the public.

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
		MSCOCO 1k						MSCOCO 5k						Flickr30k					
SEP+COOP	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**
JOINT+COOP	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+COOP and JOINT+COOP reranking the top-k candidates. **Bold** numbers indicate which k value resulted in the highest score for each separate model.

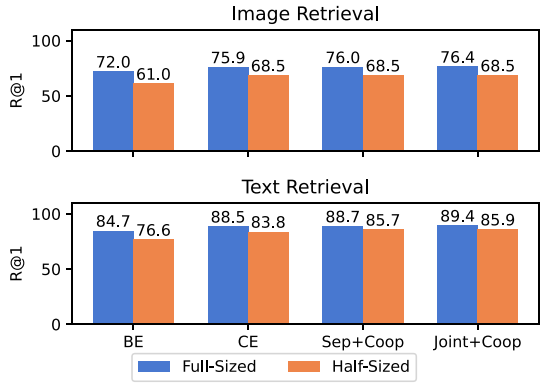


Figure 4: Half- vs. full-sized models on Flickr30k. With half-sized models, we skip every odd-numbered Transformer layer.

Model	Sum	λ	Image Retrieval			Text Retrieval		
			R@1	R@5	R@10	R@1	R@5	R@10
SEP+COOP	–		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
JOINT+COOP	–		76.4	93.6	96.2	89.4	97.7	99.0
	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

Table 9: Results on Flickr30k for different combinations of the embedding and cross-encoder scores using the functions ADD_λ and NORM_ADD_λ and different values for λ. - indicates the results for reranking using only the cross-encoder.

particularly struggle in the lowest-data scenario with only 1k images available; this is also reflected in the lower performance of JOINT+COOP in the 1k setup. A reason behind the more effective adaptation of CE to low-data regimes might be their richer ''input consumption'': With 1k images and 5k captions, CE runs a whole grid of 1k×5k items through its network, which provides more learning signal with fewer data available. On the other hand, BE-based approaches are expected to learn effective encoders of both modalities separately based solely on 1k images and 5k captions, without any cross-modal interaction.

Parameter Efficiency. We also provide a simple parameter efficiency analysis by initializing the models with pretrained 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 JOINT+COOP 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.

Retrieving Top-k. We analyze different values for k for top-k retrieval of the BE component in Table 8. Selecting small values for k significantly decreases the retrieval latency, as fewer 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 BE model. In our experiments, $k = 20$ achieves the best trade-off between precision and retrieval latency.

Combining Ranking. We evaluate the ranking score combination of the two components JOINT+BE and JOINT+CE in Table 9. We combine

the ranking of the bi-encoder submodel and the cross-encoder submodel by summing over the scores using two different variations:

(1) We directly add the scores in a weighted sum

$$\text{ADD}_\lambda(e, c) = \lambda e + (1 - \lambda)c \quad (3)$$

where e and c are the embedding cosine similarity 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 $\text{ADD}_\lambda(e, c)$.

(2) We separately 0-1-normalize the scores for the top-k candidates of the bi- and cross-encoder before combining them for $\text{NORM\_ADD}_\lambda(e, c)$, which is defined analog to $\text{ADD}_\lambda(e, c)$.

However, we find that relying solely on the cross-encoder achieves the best results. This suggests that the scores by the bi-encoder are useful in the ''global'' scope with all data to retrieve strong candidates but in the ''local'' scope of the top-k candidates, the cross-encoder is superior.

7 Conclusion

We proposed a novel framework that converts pretrained multi-modal Transformers into effective *and* efficient cross-modal retrieval models. The framework is applicable to any pretrained model and combines the efficiency of bi-encoder (BE) 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 collection is performed via efficient BE approaches, followed by another accuracy-driven step via a CE model. Moreover, we introduced a parameter-efficient joint fine-tuning regime that blends BE 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 validated the strong performance of such cooperative and joint approaches. At the same time, we demonstrated their retrieval efficiency, which makes them viable in realistic retrieval scenarios with large 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 tasks.

Acknowledgments

Ubiquitous Knowledge Processing Lab acknowledge the financial support of the German Federal Ministry of Education and Research (BMBF) under the promotional reference 13N15897 (MISRIK), the LOEWE initiative (Hesse, Germany) within the emergenCITY center, and the German Research Foundation (DFG) as part of the UKP-SQuARE project (grant GU 798/29-1). The work of Ivan Vulić has been supported by the ERC Consolidator grant LEXICAL (no 648909), ERC PoC grant MultiConvAI (no. 957356), and a research donation from Huawei.

We thank Kevin Stowe and Christopher Klamm for insightful feedback and suggestions on a draft of this paper and we thank the TACL reviewers and Action Editor for their valuable feedback and comments during the editing process.

References

P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6077–6086. IEEE Computer Society. https://doi.org/10.1109/CVPR.2018.00636

Alexandr Andoni and Piotr Indyk. 2008. 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. https://doi.org/10.1145/1327452.1327494

Sunil Arya, David M. Mount, Nathan S. Netanyahu, Ruth Silverman, and Angela Y. Wu. 1998. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *Journal of the ACM (JACM)*, 45(6):891–923. https://doi.org/10.1145/293347.293348

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. 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. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6402

Lila Boualili, Jose G. Moreno, and Mohand Boughanem. 2020. MarkedBERT: Integrating traditional IR cues in pre-trained language models for passage retrieval. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 1977–1980. ACM. https://doi.org/10.1145/3397271.3401194

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. 2021. Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language BERTs. *Transactions of the Association for Computational Linguistics*, 9:978–994. https://doi.org/10.1162/tacl_a_00408

Andrea Burns, Donghyun Kim, Derry Wijaya, Kate Saenko, and Bryan A. Plummer. 2020. 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. https://doi.org/10.1007/978-3-030-58548-8_12

Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. 2010. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research*, 11:1109–1135. https://doi.org/10.1007/978-3-642-02172-5_2

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: Universal image-text representation learning. In *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. https://doi.org/10.1007/978-3-030-58577-8_7

Muthuraman Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Learning cross-lingual sentence representations via a multi-task dual-encoder model. In *Proceedings of the 4th Workshop on Representation Learning for NLP, RepL4NLP@ACL 2019, Florence, Italy, August 2, 2019*, pages 250–259. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4330

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *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. https://doi.org/10.18653/v1/2020.acl-main.747

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. 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. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-4718

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-3210

Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. 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.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding. *arXiv preprint arXiv:2007.01852*.

Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc' Aurelio Ranzato, and Tomás Mikolov. 2013. DeViSE: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems 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. Curran Associates Inc.

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Spandana Gella, Rico Sennrich, Frank Keller, and Mirella Lapata. 2017. Image pivoting for learning multilingual multimodal representations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2839–2845. Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1303

Mandy Guo, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Cer, Gustavo Hernández Ábrego, Keith Stevens, Noah Constant, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Effective parallel corpus mining using bilingual sentence embeddings. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 165–176. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6317

Kiana Hajebi, Yasin Abbasi-Yadkori, Hossein Shahbazi, and Hong Zhang. 2011. Fast approximate nearest-neighbor search with k-nearest neighbor graph. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, pages 1312–1317. IJCAI.

Matthew Henderson, Iñigo Casanueva, Nikola Mrksic, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulic. 2020. 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. https://doi.org/10.18653/v1/2020.findings-emnlp.196

Matthew Henderson, Ivan Vulić, Daniela Gerz, Iñigo Casanueva, Pawel Budzianowski, Sam Coope, Georgios Spithourakis, Tsung-Hsien Wen, Nikola Mrkšić, and Pei-Hao Su. 2019. 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. https://doi.org/10.18653/v1/P19-1536

Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv: 1703.07737*.

Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. 2020. Improving efficient neural ranking models with cross-architecture knowledge distillation. *arXiv preprint arXiv:2010.02666*.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. 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.

Gautier Izacard and Edouard Grave. 2021. Distilling knowledge from reader to retriever for

question answering. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547. https://doi.org/10.1109/TBDATA.2019.2921572

Ákos Kádár, Desmond Elliott, Marc-Alexandre Côté, Grzegorz Chrupala, and Afra Alishahi. 2018. Lessons learned in multilingual grounded language learning. In *Proceedings of the 22nd Conference on Computational Natural Language Learning, CoNLL 2018, Brussels, Belgium, October 31 - November 1, 2018*, pages 402–412. Association for Computational Linguistics. https://doi.org/10.18653/v1/K18-1039

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6769–6781. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.550

Donghyun Kim, Kuniaki Saito, Kate Saenko, Stan Sclaroff, and Bryan A. Plummer. 2020. 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. https://doi.org/10.1609/aaai.v34i07.6785

Eyal Kushilevitz, Rafail Ostrovsky, and Yuval Rabani. 2000. Efficient search for approximate nearest neighbor in high dimensional spaces. *SIAM Journal on Computing*, 30(2):457–474. https://doi.org/10.1137/S0097539798347177

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. 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 6086–6096. Association for Computational Linguistics.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. 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. https://doi.org/10.1007/978-3-030-01225-0_13

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020a. 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. https://doi.org/10.1609/aaai.v34i07.6795

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven C. H. Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *arXiv preprint arXiv:2107.07651*.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020b. 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. https://doi.org/10.1007/978-3-030-58577-8_8

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *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. https://doi.org/10.1007/978-3-319-10602-1_48

Robert Litschko, Ivan Vulic, Simone Paolo Ponzetto, and Goran Glavas. 2021. Evaluating multilingual text encoders for unsupervised cross-lingual retrieval. In *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part I*, volume 12656 of *Lecture Notes in Computer Science*, pages 342–358. Springer. https://doi.org/10.1007/978-3-030-72113-8_23

Ting Liu, Andrew W. Moore, Alexander G. Gray, and Ke Yang. 2004. 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. MIT Press.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. VilBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23. Curran Associates Inc.

Xiaopeng Lu, Tiancheng Zhao, and Kyusong Lee. 2021. VisualSparta: An Embarrassingly Simple Approach to Large-scale Text-to-Image Search with Weighted Bag-of-words. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 5020–5029. Association for Computational Linguistics.

Shweta Mahajan, Teresa Botschen, Iryna Gurevych, and Stefan Roth. 2019. 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. https://doi.org/10.1109/ICCVW.2019.00557

Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. 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.

Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Dongdong Zhang, and Nan Duan. 2021. M3P: Learning universal representations via multitask multilingual multimodal pre-training. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 3977–3986. Computer Vision Foundation / IEEE.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*.

Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. 2019. Multi-stage document ranking with BERT. *arXiv preprint arXiv:1910.14424*.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. 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.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. 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. https://doi.org/10.1109/ICCV.2015.303

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 5835–5847. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Nils Reimers and Iryna Gurevych. 2019. 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. https://doi.org/10.18653/v1/D19-1410

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99. MIT Press.

Devendra Singh Sachan, Mostofa Patwary, Mohammad Shoeybi, Neel Kant, Wei Ping, William L. Hamilton, and Bryan Catanzaro. 2021a. End-to-end training of neural retrievers for open-domain question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6648–6662. Association for Computational Linguistics.

Devendra Singh Sachan, Siva Reddy, William L. Hamilton, Chris Dyer, and Dani Yogatama. 2021b. End-to-end training of multi-document reader and retriever for open-domain question answering. *arXiv preprint arXiv:2106.05346*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1238

Botian Shi, Lei Ji, Pan Lu, Zhendong Niu, and Nan Duan. 2019. 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.

Robert F. Simmons. 1965. Answering English questions by computer: A survey. *Communications of the ACM*, 8(1):53–70. https://doi.org/10.1145/363707.363732

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: Pre-training of generic visual-linguistic representations. In *8th International Conference*

on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5099–5110. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1514

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008. Curran Associates Inc.

Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. 2019a. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):394–407. https://doi.org/10.1109/TPAMI.2018.2797921

Yaxiong Wang, Hao Yang, Xueming Qian, Lin Ma, Jing Lu, Biao Li, and Xin Fan. 2019b. Position focused attention network 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 3792–3798. IJCAI. https://doi.org/10.24963/ijcai.2019/526

Jonatas Wehrmann, Maurício Armani Lopes, Douglas M. Souza, and Rodrigo C. Barros. 2019. 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. https://doi.org/10.1109/ICCV.2019.00590

John Wieting, Kevin Gimpel, Graham Neubig, and Taylor Berg-Kirkpatrick. 2019. 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. https://doi.org/10.18653/v1/P19-1453

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Ming Yan, Chenliang Li, Bin Bi, Wei Wang, and Songfang Huang. 2021. A unified pretraining framework for passage ranking and expansion. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 4555–4563. AAAI Press.

Sohee Yang and Minjoon Seo. 2020. Is retriever merely an approximator of reader? *arXiv preprint arXiv:2010.10999*.

Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with bertserini. 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, Demonstrations*, pages 72–77. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-4013

Yinfei Yang, Steve Yuan, Daniel Cer, Sheng yi Kong, Noah Constant, Petr Pilar, Heming Ge, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Learning semantic textual similarity from conversations. In *Proceedings of The Third Workshop on Representation Learning for NLP, Rep4NLP@ACL 2018, Melbourne, Australia, July 20, 2018*, pages 164–174.

Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-3022

Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. 2021. Pretrained transformers for text ranking: BERT and beyond. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 2666–2668. ACM. https://doi.org/10.1145/3404835.3462812

Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. 2020. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions On Multimedia Computing, Communications And Applications*, 16(2):51:1–51:23. https://doi.org/10.1145/3383184