Cross-Modal Retrieval Augmentation for Multi-Modal Classification

Shir Gur[†*], Natalia Neverova[‡], Chris Stauffer[‡], Ser-Nam Lim[‡], Douwe Kiela[‡], Austin Reiter[‡]
[†]Tel Aviv University; [‡]Facebook AI
shir.gur@cs.tau.ac.il
{nneverova,cstauffer,sernamlim,dkiela,areiter}@fb.com

Abstract

Recent advances in using retrieval components over external knowledge sources have shown impressive results for a variety of downstream tasks in natural language processing. Here, we explore the use of unstructured external knowledge sources of images and their corresponding captions for improving visual question answering (VQA). First, we train a novel alignment model for embedding images and captions in the same space, which achieves substantial improvements in performance on image-caption retrieval w.r.t. similar methods. Second, we show that retrieval-augmented multi-modal transformers using the trained alignment model improve results on VQA over strong baselines. We further conduct extensive experiments to establish the promise of this approach, and examine novel applications for inference time such as hot-swapping indices.

1 Introduction

Neural networks augmented with non-parametric retrieval components have recently shown impressive results in NLP (Khandelwal et al., 2019; Guu et al., 2020; Lewis et al., 2020; Izacard and Grave, 2020). In this work, we introduce a novel image-caption alignment model architecture and utilize it in various retrieval-augmented multi-modal transformer models, achieving substantial improvements over strong baselines.

Retrieval components are promising because they allow for easy revision and expansion of their memory, as compared to their parametric counterparts. They provide more interpretability, as well as better factual consistency with trusted knowledge sources (Shuster et al., 2021). In the multi-modal setting, retrieval augmentation allows for leveraging the strengths of text-based models—as evidenced by the strong performance of BERT-based models in vision-and-language (Lu et al., 2019; Li et al., 2019b; Kiela et al., 2019)—via cross-modal translation from images to text. Being able to seamlessly "hot swap" knowledge sources without the need for re-training the model affords a unique scalability not typically seen in the traditional deep learning literature. Nearest neighbor methods are known to be strong baselines in the vision and language domain (Devlin et al., 2015).

Our contributions are as follows. We introduce a simple, yet effective, novel cross-modal alignment architecture called DXR (Dense X-modal Retriever). DXR achieves a substantial increase in performance on both COCO (Chen et al., 2015) and Flickr30k (Young et al., 2014) image-caption retrieval, with respect to similar methods. We subsequently use DXR as a retrieval component augmenting several multi-modal transformer architectures. We show that retrieval augmentation yields impressive results irrespective of the exact input strategy, with good performs on VQA for retrieval-augmented versions of well-known multi-modal transformer architectures, from VisualBERT (Li et al., 2019b) and ViLBERT (Lu et al., 2019)—which use bounding-box features—to Movie+MCAN (Nguyen et al., 2020)—which uses grid features. We name our overall method XTRA, for X-modal Transformer Retrieval Augmentation. We conduct extensive experiments on various datasets to shed light on XTRA's performance and explore the effect of in-domain versus out-of-domain retrieval, index size and inference time applications. Our experiments show that XTRA outperforms parametric-only pre-training techniques that have access to the same data. To our knowledge, this is the first work to showcase the promise of hybrid parametric and non-parametric models for the vision and language domain.

2 Related Work

Cross-Modal Retrieval Prior work in cross-modal retrieval can be divided into two primary

[1]This work was done when Shir Gur was an intern at FAIR.

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 111–123
November 7–11, 2021.

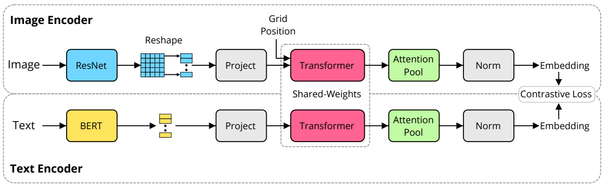
categories: (i) methods that use grid-features and/or vector representations of the embedding space, and (ii) methods that use detection features, sequence representations, or share information between the two modalities for computing the similarity metric.

The first category consists of methods such as RRF (Liu et al., 2017) and DPC (Zheng et al., 2017) which use two network branches, for image and text. CMPM (Zhang and Lu, 2018) introduced a Bi-directional LSTM to learn image and text embeddings. The most relevant work in this category is VSE++ (Faghri et al., 2017), which focuses on hard negative mining and a ranking loss. Recently, two methods that use substantial amounts of data were proposed, CLIP (Radford et al., 2021) which uses 0.4 Billion image-text pairs, and ALIGN (Jia et al., 2021) which uses 1.8 Billion noisy image-text pairs. Both methods use a dual encoder that produced and embedding vector for each modality. For fair comparison and reproducibility, we train on and compare against methods that use the open source COCO and Flickr30K datasets.

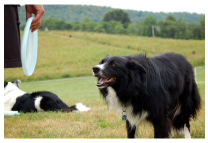
The second category generally exploits the use of detection features, which enforces an additional complexity. Methods such as TERN (Messina et al., 2020b), TERAN (Messina et al., 2020a), SAEM (Wu et al., 2019) and MMCA (Wei et al., 2020), use transformer modules to obtain modality-specific embeddings. TERAN, as well as SCAN (Lee et al., 2018), utilize sequence similarities. SCO (Huang et al., 2018) and VSRN (Li et al., 2019a) learn, in addition to image-text alignment, to generate the caption from the image embedding. MMCA, as well as CAMP (Wang et al., 2019), fuses image and text information to obtain the final embeddings. VisualSparta (Lu et al., 2021) uses fragment-level interaction to compute similarity scores. Other methods, such as Unicoder-VL (Li et al., 2020a), Oscar (Li et al., 2020b) and UNITER (Chen et al., 2020) are trained for multi-modal alignment as a pre-training task. While these models perform well, they suffer from high computational complexity as we discuss in Sec. 3.4.

External Knowledge Source Methods The use of an external knowledge source (KS) has gained much attention in the field of natural language processing (NLP), such as the work of Verga et al. (2020). Our work is inspired by that of Lewis et al. (2020), which introduced RAG, a generic approach for a variety of downstream NLP tasks using a learned retriever (DPR; Karpukhin et al., 2020) to augment the inputs by marginalizing across passages retrieved from Wikipedia. In the multi-modal domain, previous efforts have focused on building different types of KS, such as the work of Zhu et al. (2014); Chen et al. (2013); Divvala et al. (2014); Sadeghi et al. (2015) and Zhu et al. (2015), which use web information for the construction of the KS. Methods that use an external KS for a downstream task use a structured KS, such as the work of Narasimhan et al. (2018); Narasimhan and Schwing (2018); Wang et al. (2015, 2018) and Zhu et al. (2017). Zhu et al. (2017) introduced an iterative method for VQA tasks. Marino et al. (2019) introduced OK-VQA, a novel VQA dataset that requires the use of an external KS. Fan et al. (2020) applied a KS to multi-modal dialogue. In our work, we focus on a more naturally aligned KS, in the form of images and captions, which better reflects the data generated in newspapers and social media.

Multi-modal Classification In this work, we investigate the potential advantages of using an external KS for the popular and challenging VQA domain, a multi-modal classification task. Current methods for VQA use pre-training on different datasets in order to gain better performance. In our experiments, we show performance for three different methods, (i) VisualBERT (Li et al., 2019b), which is based on the BERT model by Devlin et al. (2018), (ii) ViLBERT (Lu et al., 2019), which fuses text and image modalities using co-attentional transformer layers, and (iii) MoVie+MCAN (Nguyen et al., 2020) (A similar method was introduced by Jiang et al. (2020)), which uses a modulated convolutional bottleneck for the image backbone. Other methods such as Pythia (Jiang et al., 2018), VLBERT (Su et al., 2019) and MMBT (Kiela et al., 2019) can benefit from our method, as well as more recent work such as UNITER (Chen et al., 2020), which use the alignment task for pre-training their models. Oscar (Li et al., 2020b), while using extensive data for pre-training, also introduces the use of objects' tags as additional inputs. Because the architecture of UNITER and Oscar is close to the ones we experiment with, we focus our work on our three selected models. We further note that MoVie+MCAN uses grid features instead of detection features, *i.e.*, no detector is needed (as opposed to most methods), which adds to our approach's broad applicability.



(a)



A black and white dog lays next to a frisbee

The black and white dog stands near a person holding a frisbee

A black and white dog with a frisbee by its feet

A black and white dog sitting next to a frisbee

A dog looking at a man holding a frisbee with another dog laying down

(b)

Figure 1: (a) Cross-modal alignment architecture. We use a pre-trained ResNet-152 and BERT as feature extractors with an in-batch hinge loss. (b) Sample query image and retrieved captions from the COCO dataset. Ground truth captions are colored in blue (best viewed in color).

3 Method

Our methodology is composed of two disjoint parts: (i) for a given external knowledge source $\mathcal{K}$, consisting of m modalities, we train a model (the *Retriever)* to align between the different modalities. (ii) Given a knowledge source $\mathcal{K}$ and an alignment model, we train a downstream model (the *Reader)* by augmenting its inputs with extra data from $\mathcal{K}$.

3.1 Cross-modal Alignment

Let $\mathcal{K}$ consist of m modalities, where each sample $s_i = (s_i^0, \ldots, s_i^m) \in \mathcal{K}$ is a tuple of m elements, corresponding to different modalities. Our alignment model encompasses m encoders E_m, each composed of a feature-extraction module F_m, projection P_m, shared Transformer layer T with attention pooling, and optional normalization $\mathcal{N}$:

$$E_m(x) = \mathcal{N}(T(P_m(F_m(x)))) \quad (1)$$

From this point, we will consider the two-modality case of images and captions, as illustrated in Fig. 1. For text and image feature extractors, F_1 and F_2, we use a pre-trained BERT model, and a pre-trained ResNet152 CNN backbone on ImageNet, respectively. Images are represented with convolutional grid features, chosen for robustness and speed, which are flattened across the spatial dimension. The projection layers P_m project each modality to a constant dimension d. Projected sequences are then forwarded to a shared Transformer-encoding layer, and aggregated by an attention pooling layer, resulting in a vector representation for each modality. Following Faghri et al. (2017), we normalize the text embeddings using $L2$ normalization, projecting all embeddings to the unit-sphere, due to image-caption imbalance (see Sec. 4.1).

We train our dense cross-modal retriever (DXR) using a contrastive loss, specifically using an in-batch hinge penalty with hard negatives (Faghri et al., 2017). Given a batch, consisting of b samples, $s_1 \ldots s_b$, for each sample s_i, let s_i^1 and s_i^2 be the positive pairs and s_i^1 and $s_{j \neq i}^2$ the negative pairs. We compute the pair-wise similarity between the two modalities, using a dot product:

$$s_i^{2\prime} = \max_{j \neq i} \langle s_i^1, s_j^2 \rangle, \quad s_i^{1\prime} = \max_{j \neq i} \langle s_j^1, s_i^2 \rangle \quad (2)$$

$$\mathcal{L}_{hard} = \sum_i [\alpha + \langle s_i^1, s_i^{2\prime} \rangle - \langle s_i^1, s_i^2 \rangle] + \sum_i [\alpha + \langle s_i^{1\prime}, s_i^2 \rangle - \langle s_i^1, s_i^2 \rangle] \quad (3)$$

where $s_i^{1\prime}$ and $s_i^{2\prime}$ are the hardest samples inside the batch, and α is the margin constant.

3.2 Indexing and Retrieving

Given a knowledge source $\mathcal{K}$, we construct an index by computing the embeddings of each sample in $\mathcal{K}$ using some alignment model (the *Retriever*), which can be trained on any arbitrary knowledge source. Following Lewis et al. (2020), we use FAISS (Johnson et al., 2017) as our indexer platform for fast KNN queries. We introduce two variants: we either construct separate indices $I_{\mathcal{K}}^m$ for each of the modalities; or we construct one joint index $I_{\mathcal{K}}$ that encompasses all modalities and where a KNN query will return a mixed modality result. Fig. 2 illustrates the two independent features of the alignment model and external knowledge source.

The retrieval process then consists of input query q, encoder E_m and indexer $I_{\mathcal{K}}$ (or $I_{\mathcal{K}}^m$). $I_{\mathcal{K}}$ takes as an input an embedding query $e_q = E_m(q)$ and k, and returns the k-nearest indices $i_1 \ldots i_k$, corresponding to the k-nearest embeddings. We then index data from $\mathcal{K}$, resulting in m retrieval sets $r^m = (r_1^m \ldots r_{n_m}^m)$, one for each modality, each consisting of varying number of samples n_m, where $\sum_{i=1}^m n_m = k$. When using $I_{\mathcal{K}}^m$, a single modality m is returned, resulting in

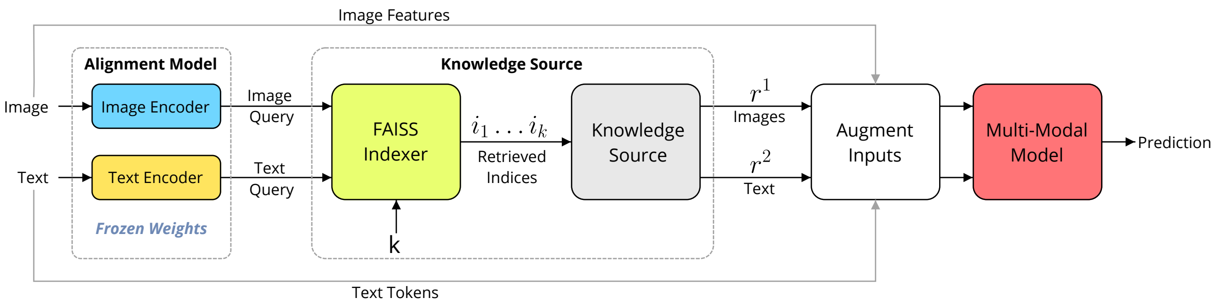


Figure 2: Illustration of our end-to-end framework. The trained cross-modal alignment is used to extract features as queries to a FAISS indexer. The k retrieved indices are used to access data from the external knowledge source, and augment the input by appending each of the k retrievals to the relative modality. For VQA, we only query the input image and retrieve k captions.

$r^m = (r_1^m \dots r_k^m)$: For simplicity, we define the retriever by $R(q, E_m, I_\mathcal{K}, k) := \{r^1, \dots, r^m\}$.

3.3 End-to-End Fusion

Let M be any multi-modal reader model, applied to a specific downstream task that takes as an input $x = (x^1, \dots, x^m)$ consisting of m modalities and outputs prediction y. The method augments the input x by concatenating the retrieved samples to their corresponding input modalities, resulting in the augmented input x':

$$x' = (x^1 \circ r_1^1 \circ \dots \circ r_{n_1}^1, \dots, x^m \circ r_1^m \circ \dots \circ r_{n_m}^m) \tag{4}$$

The resulting end-to-end training of model M minimizes a loss $\mathcal{L}(M(x'), y)$, with the same hyperparameters as in the non-retrieval augmented case. Fig. 2 illustrates the complete model.

3.4 Time-Complexity

As introduced in Sec. 2, we consider two types of retrievers, (i) methods such as ours, that use Maximum Inner Product Search (MIPS), where each modality is computed independently, and (ii) methods that have entangled computation of similarity between the different modalities, *i.e.*, that cannot compute an independent embedding. Assuming a KS of size N, and a forward-pass with $O(1)$ time-complexity, in type (i), the embeddings of the entire knowledge source need to be computed only once, with queries embedded independently. In our experiments, we use FAISS with "Hierarchical Navigable Small World" search, which as shown by Johnson et al. (2017) is $O(AD[\log N]v)$, where A and v are constants, and D is the degree of the graph. Therefore, the total time complexity of retrieving is $O(AD[\log N]v)$. On the other hand, methods of type (ii) must compute pairwise similarities between a query sample, and all samples in the dataset, resulting in a much less efficient $O(N)$.

4 Experiments

In this section, we describe the two experimental settings of the alignment model and the end-to-end downstream task training and evaluation. All models and experiments are implemented and performed with the MMF library (Singh et al., 2020a).

4.1 Datasets

We use three common datasets for training and evaluating retrieval and VQA tasks. Flickr-30K (Young et al., 2014) is composed of 30,000 images, with 5 captions each. Following Karpathy and Fei-Fei (2015), we use 1000 images for validation and 1000 images for testing. COCO (Chen et al., 2015) is a well-known dataset that contains 120,000 images, with 5 captions each. We use the splits from Karpathy and Fei-Fei (2015) as well, resulting in 80K images for training, 5K images for validation and 5K images for testing. Following Faghri et al. (2017), we add an additional 30K images for training, and uses the same 1K and 5K splits. Conceptual Captions (Sharma et al., 2018) is a dataset that contains image-caption pairs, composed of 3M samples for training and 100K for validation, which we use to test our retrieval model.

The proposed datasets differ in two major axes: (i) size, with CC at 3M image-caption pairs much larger than the smaller COCO and Flickr30K datasets; and (ii) domain gap, with e.g. CC datasets being very different in both the visual and textual domain from COCO, as shown in Singh et al. (2020b). Flickr30K is similar to COCO, but has even fewer examples.

4.2 Cross-Modal Retrieval

In the cross-modal retrieval task, we deal with two modalities: images and captions. We evaluate retrieval in both directions, denoted as Text $\rightarrow$ Image and Image $\rightarrow$ Text, where the left-hand-side indicates the query and the other indicates the retrieved domain. To ensure an apples-to-apples comparison, we here report results for methods that also use grid-features and vector representations. For a full comparison with other prior work, see Appendix A. Models are trained for 100K iterations with a warm-up of 2k iterations, batch size of 256, using the Adam optimizer with a learning rate of 0.0001 where the (pre-trained unimodal) feature encoder's learning rate is multiplied by 0.1. The hinge loss margin hyperparameter m is set to 0.2.

4.3 Downstream Tasks

After training the alignment models for each dataset—Flickr30K, COCO and CC—we build indices for each, as defined in Sec 3.2. Note that for COCO, we only use the training set for indexing, while for Flickr30K and CC, we use the entire set of train/val/test. This is done for fair comparison on the VQA task, which relies on COCO training-set images. Our experiments focus on VQA as the downstream task, however we note that extension to other multi-modal tasks is straightforward. The inputs of the VQA task are image and text tuples, and it is cast as a classification problem over a set of answers. In VQA, information regarding the content of the image, such as the amount, color and location of objects is often very correlated with the question and answer. Therefore, captions serve as good auxiliary information, while similar/retrieved images are less informative in that sense. Hence, we use the *separate indices* variant using cross-modal image to text translation, *i.e.*, we retrieve text captions of similar images to serve as augmentation data. We experiment with all three datasets, evaluating different training and inference variants.

5 Results

5.1 Cross-Modal Retrieval

Tab. 1 and 2 show retrieval results on COCO and Flickr30K, respectively, comparing similar methods that use grid-features and vector representations for the embedding space. Reported numbers correspond to Recall-at-1/5/10 on the `test-sets`. As can be seen, our method significantly outperforms previous work when trained on the same datasets. We also added the results for the state-of-the-art CLIP and ALIGN, which both use significantly larger amounts of external training data (0.4 and 1.8 Billion resp.). Appendix A compares to a wide range of additional methods.

While CC is not commonly used in the retrieval literature, we use it for our downstream task. Using DXR, we obtain the following results for CC: R@1: 25.1 R@5: 50.1 and R@10: 61.9 for Text $\rightarrow$ Image, and R@1: 25.4 R@5: 50.9 and R@10: 61.8 for Image $\rightarrow$ Text. The alignment model trained on CC is used for training in the downstream VQA task. We notice that performance degrades as the dataset size increases, which could affect the downstream task since we query from the entire dataset.

5.2 Visual Question Answering

We experiment with three common multimodal models: VisualBERT (Li et al., 2019b), ViLBERT (Lu et al., 2019), and the current winner of the VQA 2.0 challenge, Movie+MCAN (Nguyen et al., 2020), each along with three different knowledge sources (COCO, CC and Flickr30K). Following Jiang et al. (2020), we use the `val-set` split for ablations. We also report results on the VQA `test-dev` and `test-std` splits.

Tab. 3 and 4 summarize four different training settings: (i) **vanilla** - models using pre-trained BERT; (ii) **PT** - task agnostic masked language model pre-training on the knowledge source dataset; (iii) **5-GT** - training with the 5 ground truth captions from COCO; and (iv) **XTRA-10C** - training via our method, using the knowledge source indicated and alignment model trained on that source, using 10 retrieved captions. We see that using the five ground truth (GT) COCO captions as additional data (bottom row of Tab. 3), sets a soft upper bound for our approach. On the one hand, GT captions contain relevant information about the content of the image; on the other hand, other captions from the knowledge source may additionally serve as rich, useful descriptions. We also see that our method increases performance across all baselines, even with respect to pre-training. This suggests that our non-parametric hybrid method serves as a good alternative for parametric-only pre-training.

For the MoVie+MCAN model, we also report results for `test-dev` and `test-std` for COCO as our KS, setting our best model to be Movie+MCAN+XTRA-10C, obtaining a score of

	COCO 1K						COCO 5K					
	Text → Image			Image → Text			Text → Image			Image → Text		
Method	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
DPC	47.1	79.9	90.0	65.6	89.8	95.5	25.3	53.4	66.4	41.2	70.5	81.1
VSE++	52.0	83.1	92.0	64.6	89.1	95.7	30.3	59.1	72.4	41.3	69.2	81.2
CMPM	44.6	78.8	89.0	56.1	86.3	92.9	22.9	50.2	63.8	31.1	60.7	73.9
DXR	**56.8**	**88.2**	**94.9**	**67.0**	**93.0**	**97.6**	**33.9**	**64.9**	**77.4**	**44.9**	**75.2**	**84.7**
CLIP†	-	-	-	-	-	-	37.8	62.4	72.2	58.4	81.5	88.1
ALIGN†	-	-	-	-	-	-	45.6	69.8	78.6	58.6	83.0	89.7

Table 1: Retrieval results for COCO, comparing only methods that use raw images as input, and vector representations for the embedding space. We denote by † methods that train on substantial amount of novel data. Additional methods can be found in Appendix A.

	Text → Image			Image → Text		
Method	R@1	R@5	R@10	R@1	R@5	R@10
RRF	35.4	68.3	79.9	47.6	77.4	87.1
CMPM	37.3	65.7	75.5	49.6	76.8	86.1
DPC	39.1	69.2	69.2	55.6	81.9	89.5
VSE++	39.6	69.6	79.5	52.9	79.1	87.2
DXR	**50.6**	**78.8**	**86.7**	**65.1**	**87.3**	**92.6**
CLIP†	68.7	90.6	95.2	88.0	98.7	99.4
ALIGN†	75.7	93.8	96.8	88.6	98.7	99.7

Table 2: Retrieval results for Flickr30K, comparing only methods that use raw images as input, and vector representations for the embedding space. We denote by † methods that train on substantial amount of novel data. Additional methods can be found in Appendix A.

Knowledge Source	Training Type	Visual BERT	ViLBERT
Flickr30K	XTRA 10-C	66.77	67.32
CC	PT	64.34	68.14
	XTRA-10C	67.49	67.37
	PT + XTRA-10C	67.53	69.17
COCO	PT	64.54	67.58
	XTRA-10C	**68.98**	69.07
	PT + XTRA-10C	67.71	**69.90**
	Vanilla	63.54	67.56
	5-GT	**69.61**	**71.50**

Table 3: VQA Results for Visual-BERT and ViLBERT models on COCO `val-set`. **Vanilla** - models use pre-trained BERT model. **PT** - Pre-Training with the knowledge source. **XTRA-10C** - training via our method using the knowledge source indicated and alignment model trained on that knowledge source, using 10 retrieved captions.

Flickr30K	CC	COCO `val`	COCO `test` `dev`	COCO `test` `std`	Vanilla	5-GT
69.70	69.02	**71.52**	**72.80**	**73.12**	71.16	**71.80**

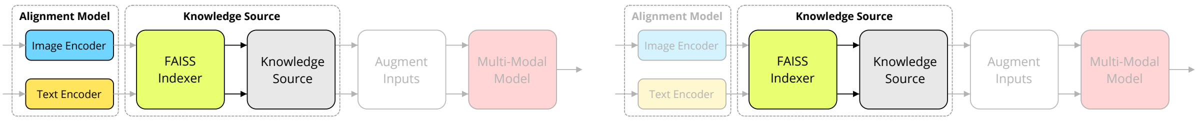
Table 4: VQA Results for MoVie+MCAN model, using XTRA-10C training type.

73.12 for `test-std` (with single model performance). Jiang et al. (2020) reported 72.71 on `test-dev` while training on the same data as our method (COCO *train+val*), while our approach achieves **72.8**. Nguyen et al. (2020) on the other hand, train with a larger VQA dataset using COCO and Visual Genome (VG) (Krishna et al., 2017), reporting 72.91 on `test-dev`.

5.3 Hot Swap

Our method is devised such that querying and retrieving from the knowledge source is independent of the downstream model, enabling the swap of the alignment model and/or knowledge source during inference. This affords interesting explorations. We describe two forms of "hot swapping": (i) the entire knowledge source and its trained alignment model are replaced with a new one and corresponding alignment model – we refer to this as "out-of-domain"; (ii) the knowledge source used for retrieving is swapped, but the alignment model remains the same as was originally trained with the downstream model. In this case, we build a new retriever for the new knowledge source, using the original alignment model – we call this "in-domain". Fig. 3 illustrates the two cases.

In Fig. 4 we show different inference results for hot swapping. All models in this experiment are trained using 10 retrieved captions. The title of each graph represents the trained model, followed by the trained knowledge source and the knowledge source to which we swap. In addition, we show inference results for training with the swapped knowledge source, *e.g.* training with CC knowledge source and alignment model from scratch, using 10 retrievals. As can be seen, "in-



(a) (b)

Figure 3: Two Hot-Swap configurations of the knowledge source during inference. **(a)** both the alignment model and the knowledge source are replaced with new ones built using a new dataset. **(b)** only the knowledge source is replaced, and the indexer is built using the old alignment model.

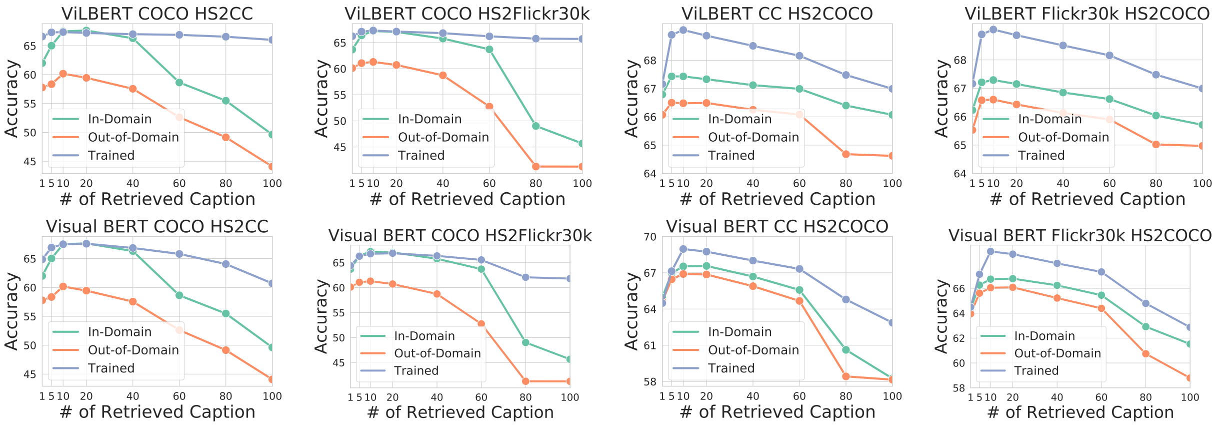


Figure 4: Hot-Swap results. Each row corresponds to a different reader model. Each graph shows **(a)** Training with different amount of retrieved captions. **(b)** Using the trained model with 10-cap, we inference with different amount of captions. **(c)** Hot swapping between knowledge sources.

Knowledge Source	Visual BERT	ViLBERT
COCO	58.77 (68.98) ↓ 10.21	45.60 (69.07) ↓ 23.47
CC	63.15 (67.49) ↓ 4.34	63.50 (67.37) ↓ 3.87
Flickr30K	61.86 (66.77) ↓ 4.91	59.34 (67.32) ↓ 7.98

Table 5: VQA performance using "unplugged" retrieval-less models trained with 10 retrieved captions, showing the highest drop for the in-domain COCO, where retrieved examples are the most informative.

domain" hot swapping performance is significantly higher than "out-of-domain". We hypothesize that the reader model has learned an implicit structure of the alignment space. Surprisingly, when training with COCO as the knowledge source, "in-domain" hot swapping performs similarly, for the same amount of trained retrievals (10), as training with an alternative knowledge source and alignment model. On the other hand, we observe a decrease in generalization due to different amounts of retrieval during inference-time. Conversely, hot swapping to COCO from CC or Flickr30K does not result in the same performance as training with COCO as the knowledge source and alignment model, yet, performance and generalization do not degrade. Qualitative results of "in-domain" hot swapping are presented in Fig 5. Novel useful information such as "cobblestone street" is retrieved from CC without having to train the alignment model on that particular source.

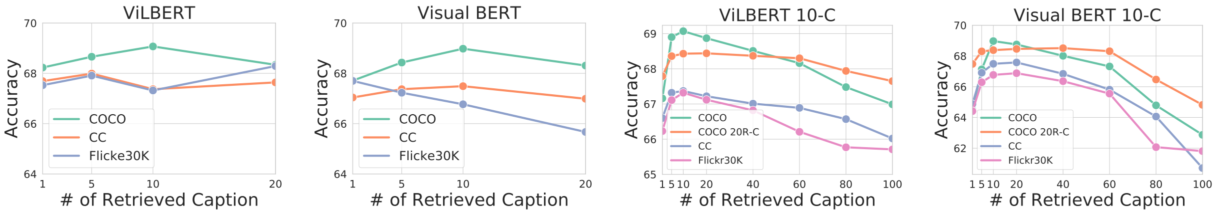
5.4 Ablation Study

In this study, we explore the use of different amounts of retrieval during training and inference, as well as doing inference without retrieving - which we call *unplugged*. We further explore the relationship between pre-training and XTRA.

Number of Retrievals We experiment with different amounts of retrieved captions during training and inference. In Fig 6 (a), we show the performance of our method when training with different amounts of retrieval, and different knowledge sources. As can be observed, training with 10 captions and COCO as the knowledge source results in the best performance. In Fig 6 (b), we show the inference performance for models trained using 10 retrievals. In addition, we show the inference performance of the same model, trained with random amounts of retrieval, between 1 and 20, on the COCO dataset (COCO 20R-C). With this, the best performance is given when we do inference with the same amount of trained retrievals, and

Query Image COCO `val-set`	Retrieved Captions No Hotswap COCO `train-set`	Flickr30K Hotswap `train+val+test sets`	CC Hotswap `train+val sets`
	A dog that is lying down on a sidewalk	A dog asleep on the streets	A dog lies down on a cobblestone street
	A dog with a muzzle on is lying on the sidewalk	A tan male bulldog sleeping on a sidewalk	The dog is lying on the cobblestone street
	A happy stray puppy lies in the street	Cute dog sleeping on the sidewalk	A dog laying on the side of the street
	A dog is laying and resting on a walkway	A dog lying on the sidewalk	A dog with a collar on lying on the street

Figure 5: Sample top-4 result for "in-domain" Hot-Swap. The model was trained using COCO as the knowledge source, and 10 retrieved captions. Left - Query image from VQA `val-set`. Columns refer to the different hot-swaps, showing retrieved captions.



(a) (b)

Figure 6: Ablation study of our method. **(a)** - Training with different amount of retrieved captions. **(b)** - Using the trained model with 10-cap, we inference with different amount of captions.

this then degrades as the number of retrievals differs from how the model was trained. We also see that training with a varying number of retrievals achieves better generalization to different amounts of retrievals during inference, as can be seen in Fig 6 (b), where COCO 20R-C performance is maintained for up to 60 retrieved examples.

Unplugged Performance One interesting observation we make is the ability to "unplug" the knowledge source by not retrieving during inference-time. Tab. 5 shows a noticeable decrease in performance, indicating the dependency of the reader on the retrieved data during training. When training with COCO as the knowledge source, introducing captions that are very related to the input images is biasing the model to depend on the retrieved captions. For CC and Flickr30K, the domain gap between the downstream task and the knowledge source lessens this gap in unplugged performance. Surprisingly, while ViLBERT performance is generally better than Visual BERT, using our method, the opposite is true when *unplugging* the knowledge source.

External Knowledge Source & Pre-training The use of a retrieval mechanism over external knowledge sources raises intriguing questions, e.g.: 1) is augmentation better than pre-training?; and 2) can pre-training help the external knowledge source? Tab. 3 shows results on COCO and CC. We find that our method is significantly better than pre-training alone, while using pre-training followed by XTRA causes the performance to vary with respect to the reader architecture (e.g., pre-training helps XTRA with ViLBERT, but not with VisualBERT). Tab. 3 also shows that fine-tuning our method after pre-training on the same knowledge source yields better performance over pre-training across all knowledge sources and architectures.

6 Conclusion

In this work, we presented a novel approach that proposes the use of external non-parametric knowledge sources in multi-modal transformer models. We trained a powerful alignment model, DXR, for performing retrieval over external knowledge sources. We showed that our method XTRA yields gains in performance when using an in-domain knowledge source on VQA. We conducted a variety of experiments to show the sensitivity and effects of the knowledge source with various choices of hyperparameters. Future research and applications of our method include improved interpretability via retrieved data and predictions for verification processes, the demonstration of increased safety and information security by hot-swapping, and unplugged versions of models and new architectures that take advantage of out-of-domain knowledge source. We hope that our approach inspires further work in the direction of hybrid parametric non-parametric models for multi-modal problems.

References

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. 2013. Neil: Extracting visual knowledge from web data. In *Proceedings of the IEEE international conference on computer vision*, pages 1409–1416.

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 *European Conference on Computer Vision*, pages 104–120. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Saurabh Gupta, Ross Girshick, Margaret Mitchell, and C Lawrence Zitnick. 2015. Exploring nearest neighbor approaches for image captioning. *arXiv preprint arXiv:1505.04467*.

Santosh K Divvala, Ali Farhadi, and Carlos Guestrin. 2014. Learning everything about anything: Webly-supervised visual concept learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3270–3277.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*.

Angela Fan, Claire Gardent, Chloe Braud, and Antoine Bordes. 2020. Augmenting transformers with knn-based composite memory for dialogue. *arXiv preprint arXiv:2004.12744*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.

Yan Huang, Qi Wu, Chunfeng Song, and Liang Wang. 2018. Learning semantic concepts and order for image and sentence matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6163–6171.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*.

Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. 2020. In defense of grid features for visual question answering. *arXiv preprint arXiv:2001.03615*.

Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. 2018. Pythia v0. 1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2019. Generalization through memorization: Nearest neighbor language models. *arXiv preprint arXiv:1911.00172*.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401*.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. 2020a. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, pages 11336–11344.

Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2019a. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4654–4662.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019b. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020b. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Yu Liu, Yanming Guo, Erwin M Bakker, and Michael S Lew. 2017. Learning a recurrent residual fusion network for multimodal matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4107–4116.

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*, pages 13–23.

Xiaopeng Lu, Tiancheng Zhao, and Kyusong Lee. 2021. Visualsparta: Sparse transformer fragment-level matching for large-scale text-to-image search. *arXiv preprint arXiv:2101.00265*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3195–3204.

Nicola Messina, Giuseppe Amato, Andrea Esuli, Fabrizio Falchi, Claudio Gennaro, and Stéphane Marchand-Maillet. 2020a. Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. *arXiv preprint arXiv:2008.05231*.

Nicola Messina, Fabrizio Falchi, Andrea Esuli, and Giuseppe Amato. 2020b. Transformer reasoning network for image-text matching and retrieval. *arXiv preprint arXiv:2004.09144*.

Medhini Narasimhan, Svetlana Lazebnik, and Alexander Schwing. 2018. Out of the box: Reasoning with graph convolution nets for factual visual question answering. In *Advances in neural information processing systems*, pages 2654–2665.

Medhini Narasimhan and Alexander G Schwing. 2018. Straight to the facts: Learning knowledge base retrieval for factual visual question answering. In *Proceedings of the European conference on computer vision (ECCV)*, pages 451–468.

Duy-Kien Nguyen, Vedanuj Goswami, and Xinlei Chen. 2020. Revisiting modulated convolutions for visual counting and beyond. *arXiv preprint arXiv:2004.11883*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Fereshteh Sadeghi, Santosh K Kumar Divvala, and Ali Farhadi. 2015. Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1456–1464.

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.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*.

Amanpreet Singh, Vedanuj Goswami, Vivek Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. 2020a. Mmf: A multimodal framework for vision and language research. `https://github.com/facebookresearch/mmf`.

Amanpreet Singh, Vedanuj Goswami, and Devi Parikh. 2020b. Are we pretraining it right? digging deeper into visio-linguistic pretraining. *arXiv preprint arXiv:2004.08744*.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Pat Verga, Haitian Sun, Livio Baldini Soares, and William W Cohen. 2020. Facts as experts: Adaptable and interpretable neural memory over symbolic knowledge. *arXiv preprint arXiv:2007.00849*.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 2018. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427.

Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony Dick. 2015. Explicit knowledge-based reasoning for visual question answering. *arXiv preprint arXiv:1511.02570*.

Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. 2019. Camp: Cross-modal adaptive message passing for text-image retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5764–5773.

Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. 2020. Multi-modality cross attention network for image and sentence matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10941–10950.

Yiling Wu, Shuhui Wang, Guoli Song, and Qingming Huang. 2019. Learning fragment self-attention embeddings for image-text matching. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2088–2096.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Ying Zhang and Huchuan Lu. 2018. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 686–701.

Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, and Yi-Dong Shen. 2017. Dual-path convolutional image-text embedding. corr abs/1711.05535 (2017). *arXiv preprint arXiv:1711.05535*.

Yuke Zhu, Alireza Fathi, and Li Fei-Fei. 2014. Reasoning about object affordances in a knowledge base representation. In *European conference on computer vision*, pages 408–424. Springer.

Yuke Zhu, Joseph J Lim, and Li Fei-Fei. 2017. Knowledge acquisition for visual question answering via iterative querying. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1154–1163.

Yuke Zhu, Ce Zhang, Christopher Ré, and Li Fei-Fei. 2015. Building a large-scale multimodal knowledge base system for answering visual queries. *arXiv preprint arXiv:1507.05670*.

A Retrieval

Tab. 6, 7 show a complete comparison of the different alignment methods in the cross-modal alignment literature. The top part corresponds to methods which use vector representations, grid-features, and do not share information between the modality branches. The bottom part shows the rest of the methods.

	Text → Image			Image → Text		
Method	R@1	R@5	R@10	R@1	R@5	R@10
RRF	35.4	68.3	79.9	47.6	77.4	87.1
CMPM	37.3	65.7	75.5	49.6	76.8	86.1
DPC	39.1	69.2	69.2	55.6	81.9	89.5
VSE++	39.6	69.6	79.5	52.9	79.1	87.2
DXR	**50.6**	**78.8**	**86.7**	**65.1**	**87.3**	**92.6**
CLIP†	68.7	90.6	95.2	88.0	98.7	99.4
ALIGN†	75.7	93.8	96.8	88.6	98.7	99.7
TERN	41.1	71.9	81.2	53.2	79.4	86.0
SCO	41.1	70.5	80.1	55.5	82.0	89.3
SAEM	52.4	81.1	88.1	69.1	91.0	95.1
SCAN	48.6	77.7	85.2	67.4	90.3	95.8
CAMP	51.5	77.1	85.3	68.1	89.7	95.2
VSRN	54.7	81.8	88.2	71.3	90.6	96.0
TERAN	56.5	81.2	88.2	70.8	90.9	95.5
MMCA	54.8	81.4	87.8	74.2	92.8	96.4
Unicoder-VL	71.5	90.9	94.9	86.2	96.3	99.0
UNITER	73.6	93.0	95.9	88.2	98.4	99.0

Table 6: Retrieval results for Flickr30K. **Top** - methods that use raw images as input, and vector representations for the embedding space. **Bottom** Methods that use detection features or sequence similarity measures. We denote by † methods that train on substantial amount of novel data.

	COCO 1K						COCO 5K					
	Text → Image			Image → Text			Text → Image			Image → Text		
Method	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
DPC	47.1	79.9	90.0	65.6	89.8	95.5	25.3	53.4	66.4	41.2	70.5	81.1
VSE++	52.0	83.1	92.0	64.6	89.1	95.7	30.3	59.1	72.4	41.3	69.2	81.2
CMPM	44.6	78.8	89.0	56.1	86.3	92.9	22.9	50.2	63.8	31.1	60.7	73.9
DXR	**56.8**	**88.2**	**94.9**	**67.0**	**93.0**	**97.6**	**33.9**	**64.9**	**77.4**	**44.9**	**75.2**	**84.7**
CLIP†	-	-	-	-	-	-	37.8	62.4	72.2	58.4	81.5	88.1
ALIGN†	-	-	-	-	-	-	45.6	69.8	78.6	58.6	83.0	89.7
TERN	51.9	85.6	93.6	63.7	90.5	96.2	28.7	59.7	72.7	38.4	69.5	81.3
SCO	56.7	87.5	94.8	69.9	92.9	97.5	33.1	62.9	75.5	42.8	72.3	83.0
SAEM	57.8	88.6	94.9	71.2	94.1	97.7	-	-	-	-	-	-
SCAN	58.8	88.4	94.8	72.7	94.8	98.4	38.6	69.3	80.4	50.4	82.2	90.0
CAMP	58.5	87.9	95.0	72.3	94.8	98.3	39.0	68.9	80.2	50.1	82.1	89.7
VSRN	62.8	89.7	95.1	76.2	94.8	98.2	40.5	70.6	81.1	53.0	81.1	89.4
TERAN	65.0	91.2	96.4	77.7	95.9	98.6	42.6	72.5	82.9	55.6	83.9	91.6
MMCA	61.6	89.8	95.2	74.8	95.6	97.7	38.7	69.7	80.8	54.0	82.5	90.7
Unicoder-VL	69.7	93.5	97.2	84.3	97.3	99.3	46.7	76.0	85.3	62.3	87.1	92.8
UNITER	-	-	-	-	-	-	51.7	78.4	86.9	66.6	89.4	94.2
Oscar	78.2	95.8	98.3	89.8	98.8	99.7	57.5	82.8	89.8	73.5	92.2	96.0

Table 7: Retrieval results for COCO. **Top** - methods that use raw images as input, and vector representations for the embedding space. **Bottom** Methods that use detection features or sequence similarity measures. We denote by † methods that train on substantial amount of novel data.