Compositional Hard Negatives for Visual Semantic Embeddings via an Adversary

Avishek Joey Bose[1,3,*†] **Huan Ling**[1,2,*†] **Yanshuai Cao**[1]
[1]Borealis AI [2]University of Toronto [3]McGill University and MILA
joey.bose@mail.mcgill.ca huan.ling@mail.utoronto.ca yanshuai.cao@borealisai.com

Abstract

Learning high-quality representations for data from different modalities but with a shared underlying meaning has been a key building block for information retrieval. Further, hard negative mining has shown to be effective in forcing models to learn discriminaitve features. In this paper, we present a new technique for hard negative mining for learning visual-semantic embeddings for cross-modal retrieval. We focus on selecting hard negative pairs that are sampled by an adversarial generator. In settings with attention, our adversarial generator is able to compose harder negatives through novel combinations of image regions across different images for a given caption. We find that our approach leads to higher scores across the board for all R@K based metrics over the previous state of the art.

1 Introduction

Image-caption and caption-image retrieval [6, 5, 4, 7] require multimodal learning to find the underlying semantic association of data from different modalities while discarding irrelevant, superficial differences in statistics of the modalities. Such cross-modal association models typically have a contrastive learning objective that tries to separate observed similar cross-modal data pairs from dissimilar ones. The dissimilar pairs can be constructed in any way, for example by randomly sampling, i.e. noise contrastive estimation (NCE) [3, 9]. Simple heuristics used in NCE typically produce very dissimilar pairs, as learning progresses, the model learns to correctly separate similar from very dissimilar pairs, and useful learning signals become sparse. This suggests the need for hard negative examples as noted by others [4, 7].

This work explores adversarial negative mining for cross-modal retrieval between images and captions. We start by exploring a straightforward application of a generic exemplar-based adversarial search method [1] as a baseline. We then exploit the compositionality of images and build out-of-sample negatives that are not in the observed dataset. Our compositional adversary deconstructs some observed images into parts and selects some of the parts from potentially different images to reconstruct a new datum that does not necessarily exist in the training set, but fools the contrastive learning model as much as possible. This compositional approach yields harder negatives than previous heuristics and exemplar adversarial approaches and improves over the state-of-art image caption retrieval method, the Stacked Cross Attention Network (SCAN) model [7].

2 Background

Cross-Modal Retrieval. Given a query, the retrieval task is to rank a set of candidate answers according to relevancy and select the one or more top ones. Typically the measure of retrieval quality is recall at K ($R@K$), i.e. the fraction of queries for which the correct candidate is among the top K selections. For caption retrieval given an image, this amounts to picking the most accurate caption for a given image query from a set of captions (image-caption retrieval). Conversely, when the

[*] authors contributed equally
[†] Work done while author was an intern at Borealis AI

32nd Conference on Neural Information Processing Systems (NIPS 2018), ViGIL Workshop, Montréal, Canada.

query is a caption then the task is to retrieve the most relevant image(s) from a database of images (caption-image retrieval).

To formalize the problem, we assume that there is a joint distribution of matching image-caption pairs, the *positive distribution*, and our training set of image-caption pairs $D = \{i_k, c_k\}_{k=1}^{N}$ is a sample from this distribution. We refer to these pairs as positive sample points and other pairings not from the true positive joint distribution as negative samples —i.e. a random pairing of image and caption.

Given a similarity scoring function, $s(i, c)$, that accepts an image-caption pair as input, learning for cross-modal retrieval aims to increase similarity of positive pairs while decreasing similarity of negative ones. This can be posed as minimizing a contrastive objective that assigns higher scores to positive examples while assigning lower scores to negative ones. So a triplet contrastive loss can be formulated as: $L_{ic} = \sum_{k}^{N} \lfloor \alpha - s(i_k, c_k) + s(i_k, c'_k) \rfloor_+$. Here α is a margin, $\lfloor . \rfloor_+$ denotes clipping below at zero, c'_k is some other caption not corresponding to i_k, (forming a negative pair (i_k, c'_k)). A similar loss, L_{ci}, can be formulated by replacing images with a sampled one — i.e. i'_k. In fact, these two objectives could be combined together by taking the sum of L_{ic} and L_{ci} as the training objective.

Hard Negative Examples. The positive examples are readily available from the training set, but the choice of negative examples c'_k and i'_k are a design choice. Random sampling in NCE is unlikely to yield hard negatives, and as learning progresses it becomes even increasingly harder to pick a negative that contributes sufficient gradient, causing slow convergence.

If learning terminates prior to observing a sufficient number of hard negatives, the resulting model would do poorly in ranking hard examples. Because the positive and easy negatives do not sufficiently constraint the model decision function, i.e. there are too many features different between positives and easy negatives, and picking up any of the differences would satisfy the training objective. However, by presenting hard negatives, the model has to learn the most discriminative features for retrieval. Hence, the strategy for picking hard negatives is crucial here.

VSE++ and exemplar-based negatives. VSE++ [4] is an improvement of visual semantic embeddings by Kiros et al. [6], and was the previous SOTA for image-caption retrieval before SCAN [7] was proposed. VSE++ introduced hard negatives as an alternative to randomly sampling easy negatives. These hard negatives are taken to be the maximum violating datum over the mini-batch for a given training query c_k, i.e. an image from the current minibatch that is not i_k, but yields the largest L_{ci} when plugged in as negative i'_k. In this sense, VSE++ uses an exemplar based hard negative mining, with an adversary that is not parametrized by a model but explicitly enumerates over a candidate set and picks the hardest. This limits the candidate set to be the mini-batch and hence limits the hardness of the negatives. Contrary to the local candidate set in VSE++, ACE [1] amortizes the search for maximum violating datum by an inference network that learns a categorical distribution over any candidate set conditioned on a query. Although its search is no longer exact, it considers a much larger candidate set. In our case, this set can be taken to be all training data of the other modality, hence this global exemplar search in ACE can potentially find harder negatives than the local hard negatives in VSE++ and in Sec. 4 we demonstrate the benefit of ACE negatives over VSE++.

Stacked Cross Attention Networks (SCAN). We now briefly explain the key formulation details of SCAN [7], as we build our compositional adversary method on top of it.

VSE++ treats an image as a whole and extracts one feature vector for a single image, which is then used in computing similarity with any given sentence embedding. Stacked Cross Attention Networks (SCAN) [7] instead computes a similarity by taking into account the alignment and compatibility of parts in the image (objects) with parts in the caption (words).

To do so for a given image I and sentence T, SCAN uses Faster R-CNN [10] to extract multiple feature vectors, $V = \{v_k\}_{k=1}^{n}$, corresponding to objects in the image; and extracts (context-dependent) word vectors, $S = \{e_l\}_{l=1}^{m}$, using a GRU-RNN, where v_k and e_l have the same dimentionality. For caption-image retrieval, it then computes an caption-dependent representation of an object region k by dot-product attention between the object feature v_k and all word features e_l's in the caption, yielding $\tilde{v}_k$. This step allows SCAN to approximately align each object k with the corresponding word(s) for it. The dot product attention uses a modified cosine similarity between image region and word, i.e. $\bar{s}_{kl} = h(v_k \cdot e_l)$, where h consists of rectification at zero and normalization, details can be found in [7]. SCAN then computes the relevance of this object region k given the caption as the cosine similarity between the original object feature v_k and the caption-dependent object feature $\tilde{v}_k$, $R(v_k, \tilde{v}_k)$. If an object is mentioned by a word in the caption, then the cross attention softmax will

be peaked at one word that semantically describes the object, allowing the resulting caption-depdent object feature to be very close to v_k. However, if an object is not mentioned, then the cross attention is flatter across words, hence the resulting representation will be some average of many embedding vectors, resulting in smaller cosine with v_k. Finally, the similarity for image I and sentence T is either $S^{avg}(I,T) = 1/N \sum_{k=1}^{N} R(v_k, \tilde{v}_k)$ or when considering the minimum classification error formulation a log sum exponential loss is used as, $S^{LSE} = \log(\sum_{k=1}^{N} exp(\lambda_2(R(v_k, \tilde{v}_k)))^{1/\lambda_2}$ [7].

3 Method

SCAN still selects one observed datum (from the minibatch) as a negative sample, so it still uses an exemplar-based negative. Our approach, on the other hand, aims to compose a new sample by piecing together parts of different data items from the current minibatch. Note that we do not actually rebuild the image in pixels, but only find a set of image features that would correspond to a new image whose object detection regions correspond to this set of features. Because the set of possible negatives produced this way contains any original data items, the set of negative sample candidates that SCAN considers is a (small) subset of what our method searches over, hence our compositional adversary is potentially able to produce richer and harder set of negatives.

Formally, each image i in a minibatch of size N is composed of image features over n regions. Let $V_i = \{v_{i1}, ..., v_{ik}, ..., v_{in}\}$, be the set of image features for image i. Now, let $V = \bigcup_{i=1}^{N} V_i$, be the union over the sets of image features in the minibatch. Then given a caption j with m words, the goal of the compositional adversary is to sample one image region feature v_{ik} for each word l in caption j.

We do so by an amortized inference neural net which takes as inputs the caption word features, $\{e_l\}$, and produce a categorical distribution $P_l(.|\phi, \{e_l\})$ over $|V|$ regions in the minibatch for each word l, where ϕ denotes the parameter of the neural net. Sampling from those categorical distributions and collect the corresponding regions form a negative example $I' = \{v'_l\}$. Note that, because the S^{avg} just considers a set of object region features, we do not need to construct I' in pixels.

Using θ to denote the SCAN model parameters, then θ and the compositional adversarial miner's parameter ϕ are learned in a min-max game:

$$\min_{\phi} \max_{\theta} S_{\theta}^{avg}(I,T) - < S_{\theta}^{avg}(I',T) >_{P(I'|T,\phi)} \tag{1}$$

To update the adversary, due to the non differentiable sampling step, we use the REINFORCE gradient estimator for ϕ. Crafting a negative sample through this process allows the main embedding model to attend to every image region in a mini-batch for every word in the caption. Given sampled I', θ can be updated by gradient descent.

False negatives. When the hard negative is actually a legitimate candidate for retrieval, it is a *false negative*. Asking the objective to separate a positive example pair and a false negative pair is meaningless at best where gradient cancels, and is harmful at worst, as the model could overfit to irrelevant differences. Exactly which failure mode it leads to depends on the smoothness of the model, but neither is desirable. Hence to avoid false negatives, we mask out image features from the k-th image when using the k-th caption. Intuitively, without masking the generator may aggregate a set of image features to produce a negative sample that is very close if not exactly the same as k-th image, a.k.a a false negative.

4 Experiments

We perform experiments for both image and caption retrieval on the MS-COCO dataset [8, 2]. To train our VSE++ we adopt all settings from [4] while to train the SCAN model we adopt all settings from[7]. As shown in Table.1, for caption retrieval, ACE improves R@1 by 0.9 and R@10 by 1.7 compared to VSE++. On the SCAN model, we find that applying our compositional hard negative mining strategy is roughly as effective as the performance gain from going from VSE++ to SCAN on caption retrieval. In fact, the compositional hard negatives result in improvements over the SCAN model on every retrieval metric for *both* caption and image retrieval task, highlighting the importance of composing harder negatives rather than simply taking the hardest example in a minibatch. We report the sum over all recall scores, both image, and caption, and R@1 in particular in Fig.1 and observe that harder negatives as proposed through our approach achieves higher recall scores from fewer samples and is thus more sample efficient.

Table 1: MSCOCO results

Method	Caption Retrieval	Image Retrieval
	R@1 / R@5 / R@10	R@1 / R@5 / R@10
VSE	43.4 / **75.7** / 85.8	31.0 / 66.7 / 79.9
VSE++ (VGG-19)	43.6 / 74.8 / 84.6	33.7 / 68.8 /81.0
ACE VSE++ (VGG-19)	**44.5** / 75.4 /**86.3**	**34.6 / 69.1 /81.8**
VSE++ (ResNet + Attention)	64.7 / 93.0 / 97.2	50.0 / 83.8 / 92.4
SCAN (number from [7])	67.5 / 92.9 / 97.6	53.0 / 85.4/ 92.9
SCAN + Compositional ADV	**70.5 / 94.8 / 98.3**	**53.7 / 85.7/ 93.6**

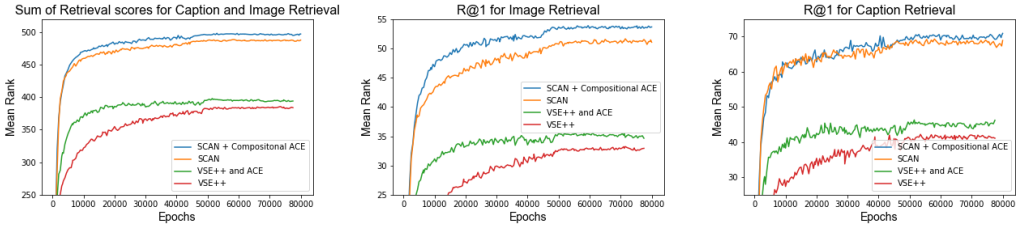

Figure 1: Performance of VSE++, SCAN models with and without ACE. **Left**: Sum of (R@1, R@5, R@10) for caption and images. **Centre**: R@1 Caption Retrieval. **Right**: R@1 Image Retrieval.

5 Conclusion

To summarize, we propose two schemes for hard negative mining for cross-modal embeddings. First we straightforwardly apply ACE for exemplar-based negative mining for before proposing a novel compositional adversary, which composes an unseen sample from features of observed data. We empirically validate the utility of our approach and observe improvements over the previous state of the art methods VSE++ and SCAN.

References

[1] Avishek Bose, Huan Ling, and Yanshuai Cao. Adversarial contrastive estimation. *arXiv preprint arXiv:1805.03642*, 2018.

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

[3] Chris Dyer. Notes on noise contrastive estimation and negative sampling. *arXiv preprint arXiv:1410.8251*, 2014.

[4] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improved visual-semantic embeddings. *arXiv preprint arXiv:1707.05612*, 2017.

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

[6] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models, 2014.

[7] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, , and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, 2018.

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[9] Andriy Mnih and Yee Whye Teh. A fast and simple algorithm for training neural probabilistic language models. *arXiv preprint arXiv:1206.6426*, 2012.

[10] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.