# Improving Visual Semantic Embedding By Adversarial Contrastive Estimation

**Huan Ling***
Department of Computer Science
University of Toronto
huan.ling@mail.utoronto.ca

**Avishek Bose***
Department of Electrical and Computer Engineering
University of Toronto
joey.bose@mail.utoronto.ca

## Abstract

Learning by contrasting positive and negative samples is a general strategy adopted by many methods such as noise contrastive estimation, max margin estimation, contrastive divergence, etc. It has shown to be an effective way of solving difficult problems with intractable computational requirements. In all of these cases, positive examples are observed and hence losses and gradient of losses can be evaluated easily on them, but negative examples are not. Hard negative mining has shown to be effective in forcing models to learn discriminative features. In this project, we focus on selecting hard negative pairs that are sampled by an adversarial generator for the task learning visual semantic embeddings. We find that hard negative mining using an adversarial generator not only leads to higher scores across the board for all R@K based metrics but is also significantly more sample efficient and leads to faster convergence in fewer iterations.

## 1 Contribution

Equal contribution for code and report and slides. We cooperated to do all of those.

## 2 Introduction

Many models learn by contrasting losses on positive examples against those on negative examples; where the goal of learning involves up weighting scores on positive examples while simultaneously down weighting scores on the negative ones. Typically, positive pairs are directly sampled from training data and negative pairs are anything that are not directly observed but often sampled at random. In noise contrastive estimation (NCE) for word embeddings, a negative example is formed by replacing a component of a positive pair by randomly selecting a sampled word from the vocabulary, resulting in a fictitious word-context pair which would be unlikely to actually exist in the dataset. Fixed random sampling in approaches like NCE lead to easy negatives that are sub-optimal for learning discriminative representation as they do not force the model to find critical characteristics of observed positive data, which has been observed in applications outside NLP previously [13]. Even if hard negatives are occasionally reached, the infrequency means slow convergence. One approach to mine hard negatives instead of random sampling is to choose the hardest negatives in a mini-batch which results in a significant improvement over the previous state of the art for the image caption cross modal retrieval task [3].

In this work, we propose to augment the random sampling approaches via a mixture negative sampling distribution. Specifically, we add an adversarially learned conditional negative sampler that adapts to the underlying data and the embedding models throughout training. The adaptive conditional model engages in a minimax game with the primary embedding model, much like in Generative Adversarial Networks (GANs)[4], where a discriminator net (D), tries to distinguish samples

produced by a generator (G) from real data [5]. Unlike a typical GAN our generator does not generate data in the input space, but instead defines a categorial distribution over all data points in a candidate set from which we sample hard negatives. Concretely, our generator makes a discrete choice over a negative sample for the discriminator embedding model to train on. In this paper, we focus on visual-semantic embeddings for the generic task of cross-modal retrieval task. We show that our method yields harder negative examples for the discriminator which results in higher scores on all metrics while being more sample efficient than the previous state of the art.

## 3 Overview of Visual-Semantic Embeddings

For completeness, we first describe the problem formulation for learning visual-semantic embeddings for cross-model retrieval. We then analyze the specific hard negative mining strategy employed in VSE++, which we directly compare against our proposed adaptive adversarial sampler.

### 3.1 Joint Embeddings

Given an image $i$ and its corresponding caption $c$ taken from a set of image caption pairs $S = \{i_n, c_n\}_{n=1}^{N}$, along with a corresponding image encoder $\phi$ and caption encoder $\psi$. In particular, $\phi$ is a deep CNN (VGG-19 or Resnet-101) up to the final logits that encodes the feature representations of the image, while $\psi$ is a GRU based RNN encoder that maps the caption to its own embedding space. The transformation to a joint embedding space is then simply a linear projection of either of the encoded representations. A natural scoring function in embedded space is cosine similarity which amounts to a simple dot product between the embeddings. Formally, we define similarity function $S(i, c)$ over image $i$ and caption $c$, and it's formularized as following:

$$E_{image(i)} = \phi(i) \tag{1}$$
$$E_{caption(c)} = \psi(c) \tag{2}$$
$$f(i) = W_f^T E_{image(i)} \tag{3}$$
$$g(i) = W_g^T E_{caption(c)} \tag{4}$$
$$s(i, c) = f(i) \cdot g(c) \tag{5}$$

### 3.2 Cross-Modal Retrieval

Given a query, the retrieval task corresponds to retrieving the most relevant answer from a set of candidate answers. For caption retrieval given an image this amounts to picking the most accurate caption for the given query from a set of captions (image-caption retrieval). Conversely, when the query is a caption then the task is to retrieve the most relevant image(s) from a database of images (caption-image retrieval). We define the set of image-caption pairs $S$ to be the set of all positive examples and any other pairing of image captions that are not in $S$ to be negative examples. Following [3], we define a query as one of $i_k$ or $c_k$ and define its answer to be the remaining one. We used recall at K ($R@K$), which is the percentage of queries for which the positive answers is ranked among the top K answers, to evaluate the model.

### 3.3 Triplet Loss and Negative Mining

The goal of learning with contrastive objectives is to assign higher scores to positive examples while assigning lower scores to negative ones. In cases where we want to separate positive and negative examples by a margin $\alpha$ the objective function is then the same as triplet loss [11]. We formulate triplet loss for cross-modal retrieval to be:

$$L = \sum_{k}^{K} (\alpha - s(i, c) + s(i, c')) + (\alpha - s(i, c) + s(i', c)) \tag{6}$$

where K is number of data points. Minimizing triplet loss is equivalent to maximizing the margin between positive pairs and negative pairs. Positive examples are readily available but the choice of

negative examples are a design choice. As a standard baseline, for a random sampled mini-batch $B = \{(i, c)\}$, with image $i$, and caption $c$ we treat the set $N_1 = \{(i, c')\}$ where $c' \neq c$ and the set $N_2 = \{(i', c)\}$ where $i' \neq i$ as negative examples. As a result, we reformulate triplet loss as:

$$L = \sum_{k}^{K} ( \sum_{(i,c') \in N1} (\alpha - s(i, c) + s(i, c')) + \sum_{(i',c) \in N2} (\alpha - s(i, c) + s(i', c))) \tag{7}$$

In [3], authors introduce hard negative for visual semantic embedding task, which are negatives closest to each training query in embedding space. Then the triplet loss is rewritten as:

$$L = \sum_{k}^{K} ( \max_{(i,c') \in N1} (\alpha - s(i, c) + s(i, c')) + \max_{(i',c) \in N2} (\alpha - s(i, c) + s(i', c))) \tag{8}$$

# 4 Adversarial Contrastive Estimation

We now formally define our approach dubbed as Adversarial Contrastive Estimation (ACE). In its most general form, contrastive objectives of the following form are amenable to ACE:

$$L(\omega) = \mathbb{E}_{p(x^+, y^+, y^-)} \, l_\omega(x^+, y^+, y^-) \tag{9}$$

where $l_\omega(x^+, y^+, y^-)$ captures both the model with parameters $\omega$ and the loss that scores a positive tuple $(x^+, y^+)$ against a negative one $(x^+, y^-)$. $\mathbb{E}_{p(x^+, y^+, y^-)}(.)$ denotes expectation with respect to some joint distribution over positive and negative samples. Furthermore, by the law of total expectation, and the fact that given $x^+$, the negative sampling is not dependent on the positive label, i.e. $p(y^+, y^-|x^+) = p(y^+|x^+)p(y^-|x^+)$, Eq. 9 can be re-written as

$$\mathbb{E}_{p(x^+)} \left( \mathbb{E}_{p(y^+|x^+)p(y^-|x^+)} \, l_\omega(x^+, y^+, y^-) \right) \tag{10}$$

The loss function can either be separable in which case the expectation further decomposes into a difference in scores of positive and negative examples or a non separable where this decomposition is not possible. The general separable loss case is simply stated to be:

$$L(\omega) = E_{p^+(x)} \left( E_{p^+(y|x)} s_\omega \left( x, y \right) - E_{p^-(y|x)} \tilde{s}_\omega \left( x, y \right) \right) \tag{11}$$

where $s_\omega$ is score function over positive data pair and $\tilde{s}$ is score function over negative data pair , $p^+$ and $p^-$ are positive and negative sampling distributions respectively.

In our case, we define $p^-$ to be:

$$p^-(y|x) = \lambda p_{\text{noise}}(y) + (1 - \lambda) g_\theta(y|x) \tag{12}$$

where $p_{\text{noise}}$ is some fixed noise distribution, $g_\theta$ is a conditional distribution with learnable parameters $\theta$ and $\lambda$ is a hyperparameter. Learning $(\omega, \theta)$ then proceeds in a GAN-style min-max game:

$$\min_\omega \max_\theta V(\omega, \theta) = \min_\omega \max_\theta E_{p^+(x)} L(\omega, \theta; x) \tag{13}$$

where $s_\omega$ is similar to the discriminator in GAN. While $g_\theta(y|x)$ acts as the conditional generator. In ACE, the discriminator learns to distinguish positive examples sampled from real data from negative examples which are sampled from $p^-$ which is a mixture distribution consisting of a fixed noise distribution and an adversarially learned conditional distribution.

## 4.1 Discriminator

We now state our ACE model for the Visual Semantic Embedding task. We first define the pair score as the cosine similarity between the joint embeddings of an image and its corresponding caption. Thus, $s(i, c) = f(i) \cdot g(c)$ where $i$ is an image and $c$ is a caption belonging to the same paired sample, either positive or negative. We implement $f(i)$ as an input into Resnet101 to produce extracted features which are then projected to the joint embedding space via a linear transformation. Similarly, $g(c)$ takes a caption as input to a GRU to produce caption features that are then projected down to the joint embedding space via a linear transformation. We define our $L(\omega)$ to be the sum of hinges. Specifically, $L(\omega) = \sum_{c'}([\alpha - s(i, c) + s(i, c')]) + \sum_{i'}([\alpha - s(i, c) + s(i', c)])$, where $c'$ and $i'$ are sampled from $p^-$.

### 4.2 Learning the Generator

There is one important distinction to typical GAN: $g_\theta(y|x)$ defines a categorical distribution over possible $y$ values, and samples are drawn accordingly; in contrast to typical GAN over continuous data space such as images, where samples are generated by an implicit generative model that warps noise vectors into data points. Due to the discrete sampling step, $g_\theta$ cannot learn by receiving gradient through the discriminator. One possible solution is to use the Gumbel-softmax reparametrization trick [8], which gives a differentiable approximation. However, this differentiability comes at the cost of drawing $N$ Gumbel samples per each categorical sample, where $N$ is the number of categories. For cross-modal retrieval this scales with the candidate set of answers and is computationally intractable. Instead, we use the REINFORCE ([14]) gradient estimator for and the discriminator loss $l_\omega(x, y^+, y^-)$ acts as the reward. We design our generator to output a categorical distribution over a candidate set of answers. For caption-image retrieval this is the set of all images in the training set other than the actual correct label image and vice-versa for image caption retrieval. Concretely, the generator can be expressed by $t(i, c) = \text{softmax}(p(f(i), g(c)))$ where function $p$ takes image features and caption features as input and outputs a K dimension vector over the candidate set of answers.

## 5 Mixture negative training

Following section 2 and section 3, our final loss for visual-semantic embedding loss is then define to be:

$$\begin{aligned} L = \sum_{k}^{K} [(\max_{(i,c')\in N1} (\alpha - s(i,c) + s(i,c')) \quad + \max_{(i',c)\in N2} (\alpha - s(i,c) + s(i',c))) + \\ \sum_{c^-} ([\alpha - s(i,c) + s(i,c^-)]) + \sum_{i^-} ([\alpha - s(i,c) + s(i^-,c)])] \end{aligned} \tag{14}$$

Where $c^-$ and $i^-$ are sampled from generator $p^-$. The first term in Equation.14 finds the hardest negative from current batch (local hard negative). As described in [9] It also reduces easy negatives which will lead the model converge to a wrong local minimum. Furthermore, the second term introduces harder negative samples from the whole dataset (global hard negative).
In our model mixture negative training is essential.

### 5.1 Hard Negatives Analysis

To better understand the hard negative samples proposed in the original vse++ and our adversarial generator we take an active learning perspective. In active learning the prototypical setting is where there exists a small labeled set, $L$, which can readily be used by the model to learn from and a large unlabeled set, $U$, from which a sample(s) must be chosen to be labeled [12]. Thus the goal in active learning is to minimize the number of queries and the associated labeling cost while maximizing accuracy. Uncertainty sampling is a query strategy within the active learning framework where the model chooses queries that it is least certain about. In binary classification this is simply picking the sample whose posterior probability of being positive is closest to $0.5$ or in equations the optimal sample is $x^* = \arg\max(1 - P_\theta(\hat{y}|x))$. Hard negative mining in VSE++ and equation 8 is in fact a slightly modified multi-class version of uncertainty sampling. The only difference being the objective in VSE++ is combined over images and captions as the goal is to learn joint embeddings for cross-modal retrieval. Intuitively, hard negatives in vse++ are samples which are closest to the current point in embedded space but do not share the same label. In contrast, hard negatives through ACE need not necessarily be closest to the original point in embedded space but instead is the point that is most likely to fool the embedding model into thinking it that it was sampled from real data. Indeed, the adversarial objective in ACE optimizes for a different notion of "hardness" than uncertainty sampling and is directly correlated with the embedding models ability to distinguish these hard samples from real ones leading to quantitative performance gains.

## 6 Experiments

We inherit all settings from [3], we used VGG-19 as image encoder and the models are trained on $1C$(1 fold) dataset. We set the text encoder's hidden state size to be 300 and joint space dimensionality to be 1024. As shown in Table.1, for Caption Retrieval,ACE improves R@1 by 0.9 percentage and R@10 by 1.7 percentage. One interesting observation is, although our ACE is only applied to Caption Retrieval, Image retrieval scores are also improved by 0.9 percentage. We report sum over all recall scores and r1 in particular in Fig.1.
Furthermore, we report the mean and median ranking curves in Fig.2. As shown in figure, ACE does indeed make the model converge faster.

### 6.1 Limitations

The formulation of ACE as introduced in equation has a few notable drawbacks. Firstly, generator defines a categorical distribution over a set of possible actions which is computed via softmax over the number of data is points and is thus computationally expensive. Although ACE converges faster per iteration, it may converge more slowly on wall-clock time depending on the cost of the softmax. We believe that the computational cost could potentially be reduced via the Augment and Reduce variational inference trick of [10], or the application of adaptive softmax [6] but leave that as future work. Another limitation is that GAN training can suffer from instability and degeneracy where the generator probability mass collapses to a few modes or points. Much work has been done to stabilize GAN training in the continuous case [1, 7, 2]. In ACE, if the generator $g_\theta$ probability mass collapses to a few candidates, then after the discriminator successfully learns about these negatives, $g_\theta$ cannot adapt to select new hard negatives, because the REINFORCE gradient estimator relies on $g_\theta$ being able to explore other candidates during sampling. Therefore, if $g_\theta$ probability mass collapses, instead of leading to oscillation in typical GAN, the min-max game in ACE reaches an equilibrium where the discriminator wins and $g_\theta$ can no longer adapt. One potential fix that occasionally works is to add an entropy term to the generator loss that forces $g_\theta$ to have high entropy and thus not collapse but tuning this term is difficult and we leave it to future work.

## 7 Conclusion

In this report we propose Adversarial Contrastive Estimation as a general technique for improving vision semantic embedding problems that learn by contrasting observed and fictitious samples. Specifically, we use a generator network in a conditional GAN like setting to propose hard negative examples for our discriminator model. We find that a mixture distribution of local hard negative examples of vse++ along with an adaptive negative sampler that mines hard negatives leads to improved performance and faster convergence.

Table 1: ACE VSE++ results

| Method | Caption Retrieval | Image Retrieval |
|---|---|---|
| | R@1 / R@10 / Medr | R@1 / R@10 / Medr |
| VSE | 43.4 / 85.8 / 2 | 31.0 / 79.9 / 3 |
| VSE++ | 43.6 / 84.6 / 2.0 | 33.7 / 81.0 / 3.0 |
| ACE VSE++ | **44.5 / 86.3 / 2.0** | **34.6 / 81.8 / 3.0** |

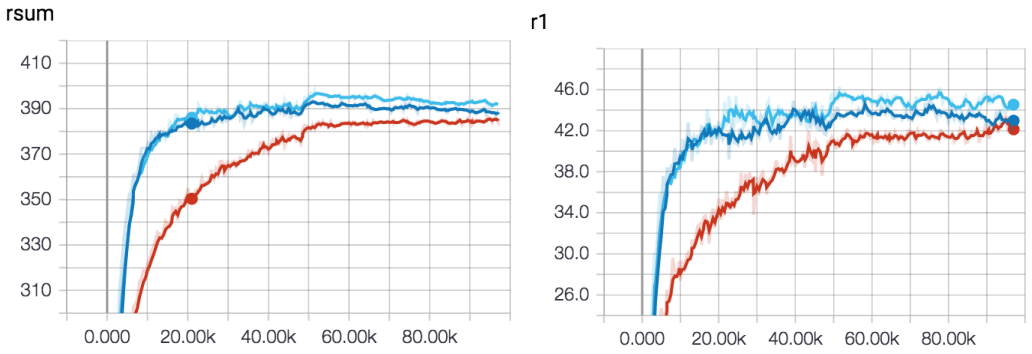


Figure 1: **Red**:VSE, **Blue**:VSE++,**LightBleu**:ACE. **Left**: sum of (r@1, r@5, r@10, r@1i, r@5i, r@10i) **Right**: R@1.

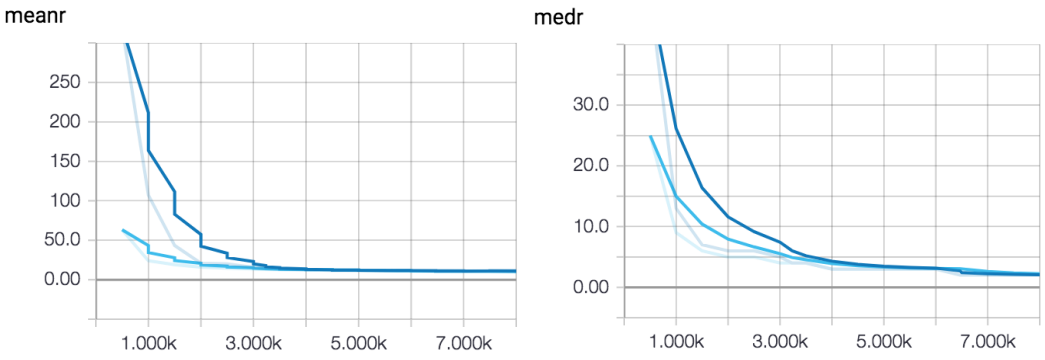


Figure 2: **Blue**:VSE++,**LightBleu**:ACE **Left**:Mean Rank **Right**: Median Rank

## References


[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[2] Yanshuai Cao, Gavin Weiguang Ding, Kry Yik-Chau Lui, and Ruitong Huang. Improving gan training via binarized representation entropy (bre) regularization. *International Conference on Learning Representations*, 2018. accepted as poster.

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

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[6] Edouard Grave, Armand Joulin, Moustapha Cissé, David Grangier, and Hervé Jégou. Efficient softmax approximation for gpus. *arXiv preprint arXiv:1609.04309*, 2016.

[7] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5769–5779, 2017.

[8] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[9] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollr. Focal loss for dense object detection. In *arXiv:1708.02002*, 201t.

[10] Francisco JR Ruiz, Michalis K Titsias, and David M Blei. Scalable large-scale classification with latent variable augmentation.

[11] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[12] Burr Settles. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114, 2012.

[13] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 761–769, 2016.

[14] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.