Self-Supervised Visual Representations for Cross-Modal Retrieval

Yash Patel
The Robotics Institute, Carnegie Mellon University, PA, USA
yashp@andrew.cmu.edu

Lluis Gomez
Computer Vision Center, Universitat Autonoma de Barcelona, Spain
gomez@cvc.uab.es

Marçal Rusiñol
Computer Vision Center, Universitat Autonoma de Barcelona, Spain
marcal@cvc.uab.es

Dimosthenis Karatzas
Computer Vision Center, Universitat Autonoma de Barcelona, Spain
dimos@cvc.uab.es

C.V. Jawahar
CVIT, KCIS, IIIT Hyderabad, India
jawahar@iiit.ac.in

ABSTRACT

Cross-modal retrieval methods have been significantly improved in last years with the use of deep neural networks and large-scale annotated datasets such as ImageNet and Places. However, collecting and annotating such datasets requires a tremendous amount of human effort and, besides, their annotations are usually limited to discrete sets of popular visual classes that may not be representative of the richer semantics found on large-scale cross-modal retrieval datasets. In this paper, we present a self-supervised cross-modal retrieval framework that leverages as training data the correlations between images and text on the entire set of Wikipedia articles. Our method consists in training a CNN to predict: (1) the semantic context of the article in which an image is more probable to appear as an illustration (global context), and (2) the semantic context of its caption (local context). Our experiments demonstrate that the proposed method is not only capable of learning discriminative visual representations for solving vision tasks like image classification and object detection, but that the learned representations are better for cross-modal retrieval when compared to supervised pre-training of the network on the ImageNet dataset.

KEYWORDS

Self-Supervised Learning, Visual Representations, Cross-Modal Retrieval

ACM Reference Format:
Yash Patel, Lluis Gomez, Marçal Rusiñol, Dimosthenis Karatzas, and C.V. Jawahar. 2019. Self-Supervised Visual Representations for Cross-Modal Retrieval. In . , 9 pages.

1 INTRODUCTION

The emergence of large-scale annotated datasets such as ImageNet [6], Places [47] and MS-COCO [19] has undoubtedly been one of the key ingredients for the tremendous impact of deep learning on almost every computer vision task. However, there is a major issue with the supervised learning setup in large scale datasets; collecting and manually annotating those datasets requires a great amount of human effort. On the other hand, the fact that the annotations on such datasets are usually limited to discrete sets of popular visual classes, may not necessarily be an optimal training setup for cross-modal retrieval datasets that usually cover a set of broader and richer semantic concepts.

As an alternative to the fully supervised setup, self-supervised learning methods aim at learning discriminative visual features by designing auxiliary tasks for which the target labels are free to obtain. These labels provide supervision for the training of computer vision models the same way as in supervised learning, but the supervisory signal can be directly obtained from the training data, either from the images themselves [7, 25] (uni-modal training) or from a complementary modality that is found naturally correlated with them [1, 12, 22] (multi-modal training). Unlike supervised learning, where visual features are learnt from human generated labels, in self-supervised learning labels are automatically obtained from the training data.

In this paper we present a self-supervised cross-modal retrieval framework that leverages as supervisory signal the correlations found between images and text on a large collection of illustrated articles in order to learn discriminative visual features that could potentially transfer well to any general computer vision task, such as image classification or object detection. We hypothesise that the learned representations by using such approach can be used more naturally in a cross-modal retrieval framework than the representations learned from annotated datasets for image classification.

Our intuition follows from the observation that illustrated encyclopedic articles, like Wikipedia's ones, are well organized and contain a detailed textual description of their subject while certain aspects of the subject are illustrated by images. Those images complement the text and at the same time provide context to our imagination. Furthermore, the captions associated with these images specifically describe their contents. These observations, and the large-scale availability of such articles, lead us to treat representation learning for cross-modal retrieval as a self-supervised visual representation learning task. We demonstrate that rich visual representations can be learned by training a network to predict the global (article-level) and local (caption-level) semantic contexts in which an image appears, and at the same time, the learned representations can be used to perform cross-modal retrieval with promising results.

Gomez and Patel *et al.* [12, 23] have proposed in the past self-supervised representation learning using Wikipedia articles. Their method consists in learning a Latent Dirichlet Allocation (LDA) model from the entire corpus of text articles, and then training a

, ,
2019.

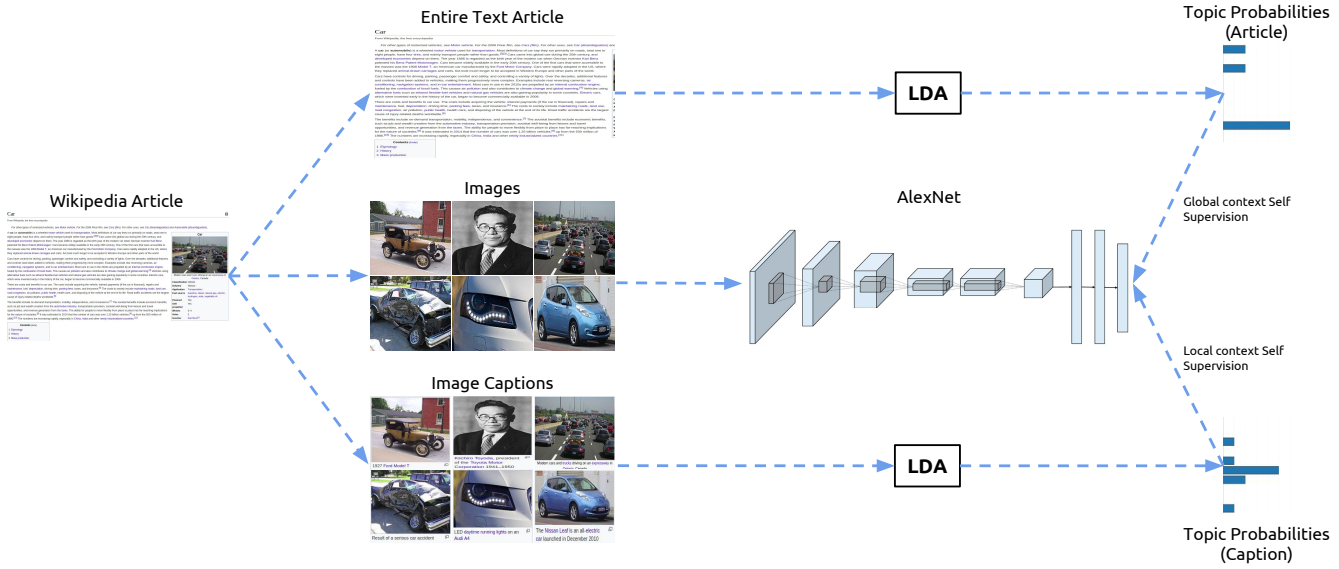

Figure 1: Method overview: Wikipedia articles contain textual description of a subject, these articles are also accompanied with illustrative images supporting the text. These images are often accompanied by captions. A Latent Dirichlet Allocation (LDA) [3] topic modeling framework generates a global contextual representation of the textual information from entire text article. The same LDA model generates a local contextual representation from the per-image caption. These two text representations are jointly used to supervised the training of deep CNN.

CNN to predict the semantic context of images by using as training labels the semantic level representations (the probability distribution over semantic topics) of the articles in which they appear, as provided by the LDA model. An assumption made in their method is that all the images within a given text article have the same target semantic representation, which is obtained from the LDA model. However, images within a Wikipedia article can be drastically different in terms of appearance and semantic content. To overcome this, we create a new *Wikipedia dataset with captions* which is similar to the one used in the TextTopicNet [12, 23] method, but also containing the image captions from Wikipedia. Thus, as illustrated in Figure 1, the training data in our method comes in a triplet form (image, text article, image caption).

Our intuition is that adding another target representation based on image captions could provide more image specific training self-supervision. Furthermore, we experimentally show that our training procedure leads to significantly better results for both cross-modal retrieval and image classification.

Following are the major contributions made in this paper:

- We propose a multi-task learning framework to train a CNN that predicts text representations obtained from text articles (global context) and per-image captions (local context).
- We experimentally demonstrate that the self-supervisedly learned visual features are generic enough for other computer vision tasks and outperform other self-supervised and naturally supervised approaches on standard benchmarks.
- Without using any form of semantic information, our method outperforms both unsupervised and supervised approaches on cross-modal retrieval (image-to-text and text-to-image) benchmarks on Wikipedia [27] and Pascal sentences datasets [10].
- The Wikipedia image-article dataset [23] consist of only images and text articles, and as an auxiliary contribution, we release a large scale dataset obtained from English Wikipedia consisting of images, per-image captions and co-occurring text articles.

The rest of the paper is structured as follows. In Section 2, previous work is reviewed. In Section 3, details of training dataset are elaborated. In Section 4 the proposed method is described and in Section 5 evaluated. The paper is concluded in Section 6.

2 RELATED WORK

2.1 Self-Supervised Visual Representations

As an alternative to fully-supervised algorithms, there has recently been a growing interest in self-supervised or naturally-supervised approaches that make use of non-visual signals, intrinsically correlated to the image, as a form to supervise visual feature learning. The objective of those methods is to learn visual representations (without human annotations) that are generic to work well across a wide range of object classes and at the same time are discriminating enough to be useful for classical computer vision tasks such at image classification, object detection, semantic segmentation etc.

2.1.1 Unsupervised Visual Representation Learning. Work in unsupervised data-dependent methods for learning visual features has been mainly focused on algorithms that learn filters one layer at a time. A number of unsupervised algorithms have been proposed to that effect, such as sparse-coding, restricted Boltzmann machines

Figure 2: Samples from Wikipedia dataset with captions. In the method, the shown captions provide local image specific information, whereas the entire text article provides global subject information.

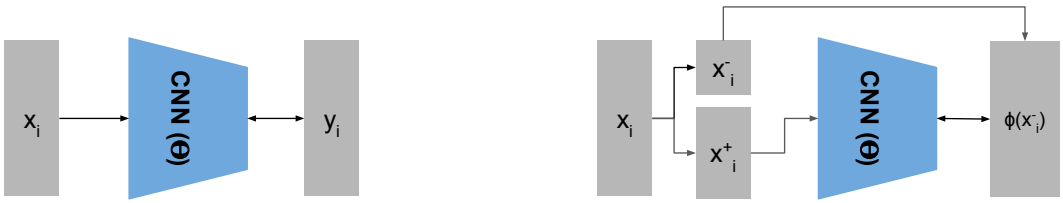

Figure 3: Supervised (left) vs. Self-Supervised Training (right). In supervised training the ground-truth labels y_i are collected by human annotation. Whereas in self-supervised training, a transformation on a part of input data is used as the target label for training.

(RBMs), auto-encoders [46], and K-means clustering [5, 8, 16]. However, despite the success of such methods in several unsupervised learning benchmark datasets, a generic unsupervised method that works well with real-world images does not exist.

Bojanowski & Joulin *et al.* [4] present an approach for unsupervised learning of visual features using Noise As Target (NAT) label for training. Their approach is domain agnostic and makes use of fixed set of target labels for training. The primary difference between our work and [4] is that, in our work the final network predictions are directly useful for a specific task - cross-modal matching and retrieval.

2.1.2 Uni-modal Self-Supervised Methods. In contrast to the purely supervised approaches, uni-modal self-supervised algorithms make use of the structure in the visual data itself for the purpose of representation learning. Agrawal *et al.* [1] make use of egomotion information obtained by odometry sensors mounted on a vehicle. They train a network using contrastive loss formulation [20] to predict the camera transformations between two image pairs.

Wang and Gupta *et al.* [40, 41] make use of videos as training data and use relative motion of objects as supervisory signal for training. The relative motion information is obtained by using a standard unsupervised tracking algorithm. A Siamese-triplet network is then trained using a ranking loss function.

Pathak and Efros *et al.* [25] take inspiration from auto-encoders and proposed a context-encoder. They train a network using a combination of L2 loss and adversarial loss to generate arbitrary image regions conditioned on their surrounding. Doersch *et al.* [7] use spatial context such as relative position of patches within an image to make the network learn object and object parts.

Our proposed method is different from all of these methods since it makes use of multi-modal axillary task for training. Further, by training the network to predict local and global contexts in which an image appears as illustration could be directly used for cross-modal retrieval. Our work is more correlated with the multi-modal self-supervised approaches as elaborated in next section.

2.1.3 Multimodal Self-Supervised Methods. Multi-modal self-supervised learning alrogithms attempt to utilize the structure in one modality to provide the training supervision for co-occuring modality.

Owens *et al.* [22] make use of sound as a modality to provide supervisory signal. They do so by training a deep CNN to predict a hand-crafted statistical summary of sound associated with a video frame.

Gomez and Patel *et al.*[12] make use of Wikipedia documents which consist of text articles and co-occuring images. First, a Latent Drichilet Allocation (LDA) [3] topic model is learned on the entire Wikipedia dataset. Second, text articles are represented in the form of topic-probabilities using learned LDA model. Finally, a convolutional neural network is trained on images in Wikipedia, where the target label is the representation of corresponding text article.

Our work is more closely related to [12, 23, 24], however, as previously mentioned, their approach makes use of same target representation for all images within a text article. This not only leads to sub-optimal performance but also completely ignores the local context of an image.

2.2 Cross-Modal Representation Learning

Two general categories of the representation learning methods for cross-modal retrieval can be: (a) *real-valued*, (b) *binary valued*. The

binary methods are more focused on efficiency and aim to map the items from different modalities on a common binary hamming space [31, 33, 43, 48].

Our approach falls in the category of *real-valued* methods. Within this category of methods the training for cross-modal retrieval could be: unsupervised [2, 11, 14, 34, 44] or supervised [13, 37, 38, 45].

Zhang *et al.*[46] propose a multimodal hashing method, called semantic correlation maximization (SCM), which integrates semantic labels into the hashing learning procedure. This method uses label vectors to get semantic similarity matrix and tries to reconstruct it through the learned hash codes.

Gong *et al.*[13] propose a novel three-view CCA (CCA-3V) framework, which explicitly incorporates the dependence of visual features and text on the underlying semantics.

Wang *et al.*[38] propose a novel regularization framework for the cross-modal matching problem, called LCFS (Learning Coupled Feature Spaces). It unifies coupled linear regressions, l_{21}-norm and trace norm into a generic minimization formulation so that subspace learning and coupled feature selection can be performed simultaneously. Furthermore, they extend this framework to more than two-modality case in [37], where the extension version is called JFSSL (Joint Feature Selection and Subspace Learning).

Wang *et al.*[35] propose an adversarial learning approach for cross-modal retrieval. The method is built around the idea of min-max game involving two different processes *players*: a modality classifier distinguishing the items in terms of their modalities, and a feature projector generating modality-invariant and discriminate representations and aiming to confuse the modality classifier.

While most of these supervised or unsupervised approaches attempts to learn a common embedding space for the prupose of cross-modal retrieval, they assume that the visual representations are provided by a pre-trained CNN (either AlexNet [17] or VGG-16 [32]) on ImageNet dataset [29]. The cost (human annotation effort) of this pre-training is not accounted by cross-modal retrieval methods. Further, the underlying assumption of is that ImageNet pre-trained features transfer well for cross-modal retrieval.

The proposed method investigates these two aspects, firstly, we do not make use of ImageNet pre-training and instead use the self-supervised visual representations. Secondly, in the experiments, we train the network just once on our dataset and no form of adaptation is done on test datasets. This demonstrates that our proposed method is capable of learning a general purpose category-agnostic cross-modal retrieval system.

3 WIKIPEDIA DATASET WITH CAPTIONS

In order to obtain a training dataset for our method, we scrapped the entire English Wikipedia while considering only articles with at least 50 words and illustrated with at-least one image. Similarly to the preprocessing of ImageCLEF dataset we filtered small images (< 256 pixels) and images with formats other than JPG. Furthermore, we only considered the images for which captions are available. With these constraints our dataset consists of 1.8 million images with captions and the associated text article that they appear with. Figure 2 shows samples from the dataset.

4 METHOD

In this section, we first elaborate over the core distinction between supervised and self-supervised trainings. Then we discuss about the Latent Dirichlet Allocation (LDA) [3], which is used for representing text articles and image captions, and thus for generating target representations for training the CNN. Finally, we go over the training of the CNN.

4.1 Self-Supervised Learning

The supervised methods learn rich visual representations from large collections of training data. This data always has human annotations, $D = \{(x_1, y_1), (x_2, y_2)...(x_N, y_N)\}$, and the deep network is trained to minimize the overall risk term:

$$R = \sum_{i=1}^{N} [loss(f(x_i, \Theta), y_i)] \tag{1}$$

Where Θ are the parameters of the deep network.

Unlike supervised approaches, self-supervised methods train without making use of any human annotations. The training data, $D = \{(x_1), (x_2)...(x_N)\}$ can be sub-divided into components and one or more components can be used to provide self-supervision for others, thus, data is represented as $D = \{(x_1^+, x_1^-), (x_2^+, x_2^-)...(x_N^+, x_N^-)\}$ and the training for one component is governed by the other changing the overall risk term to:

$$R = \sum_{i=1}^{N} [loss(f(x_i^+, \Theta), x_i^-)] \tag{2}$$

Fig. 3 shows the explicit difference between supervised and self-supervised approaches.

4.2 Latent Dirichlet Allocation

LDA [3] is a generative statistical model of a text corpus where each document can be viewed as a mixture of various topics, and each topic is characterized by a probability distribution over words. LDA can be represented as a three level hierarchical Bayesian model. Given a text corpus consisting of M documents and a dictionary with N words, Blei *et al.*define the generative process [3] for a document d as follows:

- Choose $\theta \sim Dirichlet(\alpha)$.
- For each of the N words w_n in d:
  - Choose a topic $z_n \sim Multinomial(\theta)$.
  - Choose a word w_n from $P(w_n \mid z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

where θ is the mixing proportion and is drawn from a Dirichlet prior with parameter α, and both α and β are corpus level parameters, sampled once in the process of generating a corpus. Each document is generated according to the topic proportions $z_{1:K}$ and word probabilities over β. The probability of a document d in a corpus is defined as:

$$P(d \mid \alpha, \beta) = \int_{\theta} P(\theta \mid \alpha) \left(\prod_{n=1}^{N} \sum_{z_K} P(z_K \mid \theta) P(w_n \mid z_K, \beta) \right) d\theta$$

Learning LDA [3] on a document corpus provides two set of parameters: word probabilities given topic $P(w \mid z_{1:K})$ and topic probabilities given document $P(z_{1:K} \mid d)$. Therefore each document

is represented in terms of topic probabilities $z_{1:K}$ (being K the number of topics) and word probabilities over topics. Any new (unseen) document can be represented in terms of a probability distribution over topics of the learned LDA model by projecting it into the topic space.

4.3 Network Architecture

Throughout our experiments, we make use of AlexNet architecture [17]. The choice of AlexNet is justified because most of the existing self-supervised methods make use of this same architecture [1, 12, 22, 23, 25, 36]. Further, we compare to cross-modal retrieval methods with reported performance using AlexNet [13, 14, 27, 30, 37, 38]. Thus the use of AlexNet architecture is essential for fair comparisons.

As shown in Figure 1, till the *fc7* layer the architecture is same as standard AlexNet [17], which is followed by two fully-connected branches one prediction caption level topic probabilities and other predicting article level topic probabilities.

4.4 Learning Self-Supervised Representations

Following up with the formal definition of self-supervised learning as described in Section 4.1. The multimodal document from Wikipedia can be thought as a training sample, x_i. This multimodal document consists of text article x_i^A, image captions x_i^C and images x_i^I.

Let $\Phi(x_i^A)$ and $\Phi(x_i^C)$ be the text topic probability distributions given by LDA 4.2 for the document text and the image captions accordingly. The deep CNN is trained to predict the above topic distributions given the corresponding article image, and producing as outputs: $f_A(x_i^I, \Theta)$ (for article) and $f_C(x_i^I, \Theta)$ (for caption).

The loss is computed as the cross entropy between the LDA topic distribution and the predicted distribution. The overall risk term on the training data will be:

$$R = \sum_{i=1}^{i=N} [\sum_{topic=1}^{topic=K} \Phi(x_i^A)_{topic} log(f_A(x_i^I, \Theta)_{topic}) + \sum_{topic=1}^{topic=K} \Phi(x_i^C)_{topic} log(f_C(x_{i,topic}^I, \Theta)_{topic})] \quad (3)$$

where N is total number of samples in the training data, K is the number of topics in the LDA model [3] and Θ maps to the learnt CNN parameters. Note that K is a hyper-parameter and we fix $K = 40$ throughout the experiments.

4.5 Training Details

Learning to predict the target topic probability distributions we minimize a sigmoid cross-entropy loss as shown in the overall risk term Eq. 3. We use a Stochastic Gradient Descent (SGD) optimizer, with base learning rate of 0.001, with a step decay after every 200,000 iterations by a factor of 0.1, and momentum of 0.9. The batch size is set to 128. With these settings the network converges after 500,000 iterations of training.

5 EXPERIMENTS

We will first compare the learnt visual representations with other self-supervised methods on the task on image classification on two standard benchmark datasets (Section 5.1). Next, we will compare our method with various cross-modal retrieval methods (Section 5.2).

5.1 Self-Supervised Features for Image Classification

5.1.1 PASCAL VOC. Self-supervised learned features are tested for image classification on PASCAL VOC 2007 [9] dataset. In total there are 9,963 images, and 20 semantic classes. The data has been split into 50% for training/validation and 50% for testing. The classification here is multi-label, that is, each image can be classified into multiple classes.

We extract features from the top layers of the CNN (fc7, fc6, pool5) for each image of the dataset. Then, for each class we perform a grid search over the parameter space of an one-vs-all Linear SVM classifier [1] to optimize its validation accuracy. Then, we use the best performing parameters to train again the one-vs-all SVM using both training and validation images.

In Tables 1 and 2, we compare our results on the PASCAL VOC2007 test set with different state-of-the-art self-supervised learning algorithms using features from different top layers and SVM classifiers.

Our method which leverages global and local contexts for self-supervised training achieves state-of-the-art performance as seen in Table 2. This demonstrates that a network that identifies global and local semantic contexts in which an image is more probable to appear gives better visual representations.

In Table 1, we provide a per-class comparison with various self-supervised and supervised visual representation learning algorithms. It can be clearly seen that our method performs better than other self-supervised methods for most of the classes. In the case of "bottle" class our method outperforms fully supervised network.

5.1.2 SUN 397. Table 3 compares our results on the SUN397 [42] test set with state-of-the-art self-supervised learning and supervised algorithms. SUN397 [42] consists of 50 training and 50 test images for each of the 397 scene classes. We follow the same evaluation protocol as [1, 22] and make use 20 images per class for training and remaining 30 for validation. We evaluate our method on three different partitions of training and testing and report the average performance. This scene classification dataset is suitable for the evaluation of self-supervised approaches as it contains less frequently occurring classes and thus is more challenging compared to PASCAL VOC 2007 dataset.

We appreciate that our method outperforms all other modalities of supervision in this experiment. We observe that using features from *fc6* layer gives better performance compared to using features from *fc7* layer. This indicates that *fc6* and *pool5* layers of our network are more robust towards uncommon classes.

5.2 Cross-Modal Retrieval

As seen in Fig. 1, the final layer of the network projects the images on same representation as text as obtained by the LDA model (Section 4.2). Therefore, cross-modal retrieval can be directly done by making use of LDA topic probabilities for text and network final predictions for image. We use KL-divergence as a distance metric

[1] Liblinear implementation from http://scikit-learn.org/

Method	aer	bk	brd	bt	btl	bus	car	cat	chr	cow	din	dog	hrs	mbk	prs	pot	shp	sfa	trn	tv
Ours	**73**	**56**	**49**	**65**	**26**	**50**	**73**	**46**	**48**	**38**	**45**	**42**	73	**64**	**86**	**34**	**44**	**44**	**74**	**48**
TextTopicNet (Wikipedia) [23]	71	52	47	61	26	49	71	46	47	36	44	41	72	62	85	31	40	42	72	44
TextTopicNet (ImageCLEF) [12]	67	44	39	53	20	49	68	42	43	33	41	35	70	57	82	30	31	39	65	41
Sound [22]	69	45	38	56	16	47	65	45	41	25	37	28	**74**	61	85	26	39	32	69	38
Texton-CNN	65	35	28	46	11	31	63	30	41	17	28	23	64	51	74	9	19	33	54	30
K-means	61	31	27	49	9	27	58	34	36	12	25	21	64	38	70	18	14	25	51	25
Motion [40]	67	35	41	54	11	35	62	35	39	21	30	26	70	53	78	22	32	37	61	34
Patches [7]	70	44	43	60	12	44	66	52	44	24	45	31	73	48	78	14	28	39	62	43
Egomotion [1]	60	24	21	35	10	19	57	24	27	11	22	18	61	40	69	13	12	24	48	28
ImageNet [17]	79	**71**	**73**	75	**25**	60	80	**75**	51	**45**	60	**70**	**80**	**72**	**91**	42	**62**	56	82	62
Places [47]	**83**	60	56	**80**	23	**66**	**84**	54	**57**	40	**74**	41	**80**	68	90	**50**	45	**61**	**88**	**63**

Table 1: PASCAL VOC2007 per-class average precision (AP) scores for the classification task with pool5 features.

Method	max5	pool5	fc6	fc7
Ours	-	**53.8**	**54.9**	**56.8**
TextTopicNet (Wikipedia) [23]	-	51.9	54.2	55.8
TextTopicNet (ImageCLEF) [12]	-	47.4	48.1	48.5
Sound [22]	39.4	46.7	47.1	47.4
Texton-CNN	28.9	37.5	35.3	32.5
K-means [16]	27.5	34.8	33.9	32.1
Tracking [40]	33.5	42.2	42.4	40.2
Patch pos. [7]	26.8	46.1	-	-
Egomotion [1]	22.7	31.1	-	-
ImageNet [17]	**63.6**	**65.6**	**69.6**	**73.6**
Places [47]	59.0	63.2	65.3	66.2

Table 2: PASCAL VOC2007 mAP comparison for image classification with supervised (bottom), and self-supervised (middle) methods.

Method	max5	pool5	fc6	fc7
Ours	-	**30.3**	**33.5**	**28.2**
TextTopicNet (Wikipedia) [23]	-	28.8	32.2	27.7
Sound [22]	17.1	22.5	21.3	21.4
Texton-CNN	10.7	15.2	11.4	7.6
K-means [16]	11.6	14.9	12.8	12.4
Tracking [40]	14.1	18.7	16.2	15.1
Patch pos. [7]	10.0	22.4	-	-
Egomotion [1]	9.1	11.3	-	-
ImageNet [17]	29.8	34.0	37.8	37.8
Places [47]	**39.4**	**42.1**	**46.1**	**48.8**

Table 3: SUN397 accuracy for image classification with supervised (bottom), and self-supervised (middle) methods.

to short the samples of the target modality, since both the LDA encoding and the CNN output represent probability distributions.

Note that our comparisons are made with existing methods with reported performance using ImageNet pre-trained AlexNet [17] architecture for image representations and LDA [3] or BoW representations for text.

5.2.1 Wikipedia. We use the Wikipedia retrieval dataset [27], which consists of 2,866 image-article pairs split into train and test set of 2,173 and 693 pairs respectively Further, each image-document pair is labeled with one of ten semantic classes [27].

In Table 4 we compare our results with supervised and unsupervised multi-modal retrieval methods discussed in [39] and [15]. Supervised methods make use of class or categorical information associated with each image-document pair, whereas unsupervised methods do not. All of these methods use LDA for text representation and CNN features from pre-trained CaffeNet, which is trained on ImageNet dataset in a supervised setting. We observe that the self-supervised baseline method outperforms unsupervised approaches, and has competitive performance to supervised methods without using any labeled data.

In Table 4, we also observe that our method which leverages global and local contexts for self-supervised training leads to state-of-the-art performance, even when compared to fully supervised approaches. This demonstrates that training a network to predict both the global and local semantic contexts in which it is more probable to appear leads to better learning for retrieval task. Further note that, except ours and TextTopicNet [12, 23] all the other methods use ImageNet pre-trained network.

5.2.2 Pascal Sentences. We also evaluate our method on pascal sentences dataset [10] which is a subset of pascal VOC dataset. It contains 1000 pairs of an image along with several sentences from 20 categories. While, the other methods randomly split the dataset into 600 training and 400 testing samples, we test on all 1000 samples. This is due to the fact that we do not make use of this dataset for training at any point.

Table 5 provides an extensive comparison with existing methods. Compared to other retrieval methods that use self-supervised visual representations [12, 23], our method achieves 1.6% higher MAP with $\frac{1}{4^{th}}$ the size of training data. This demonstrates the efficacy of jointly using global and local self-supervision signals.

Method	Image Query	Text Query	Average
Ours	39.10	**43.40**	**41.25**
TextTopicNet (Wikipedia) [23]	37.63	40.25	38.94
TextTopicNet (ImageCLEF) [12]	39.58	38.16	38.87
CCA [14, 27]	19.70	17.84	18.77
PLS [28]	30.55	28.03	29.29
SCM* [27]	37.13	28.23	32.68
GMMFA* [30]	38.74	31.09	34.91
CCA-3V* [13]	40.49	36.51	38.50
GMLDA* [30]	40.84	36.93	38.88
LCFS* [38]	41.32	38.45	39.88
JFSSL* [37]	**42.79**	39.57	41.18

Table 4: Mean average precision (MAP) comparison on Wikipedia dataset [27] with supervised (bottom), unsupervised (middle) and self-supervised (top) methods. Methods marked with asterisk make use of document (image-text) class category information.

Method	Image Query	Text Query	Average
Ours	32.6	**36.0**	**34.3**
TextTopicNet (Wikipedia)[23]	30.1	35.2	32.7
TextTopicNet (ImageCLEF) [12]	26.4	31.6	29.0
CCA [14, 27]	9.90	9.7	9.8
CFA [18]	18.7	21.6	20.2
KCCA (Poly) [14]	20.7	19.1	19.9
KCCA (RBF) [14]	23.3	24.9	24.1
Bimodal AE [21]	24.5	25.6	25.1
Multimodal DBN [34]	19.7	18.3	19.0
Corr-AE [11]	26.8	27.3	27.1
JRL [45]	30.0	28.6	29.3
CMDN [26]	**33.4**	33.3	33.4

Table 5: Mean average precision (MAP) comparison on pascal sentences dataset [10] with supervised image representations (bottom) and self-supervised image representations (top) methods.

5.3 Qualitative Retrieval Results

Finally, in this section we provide additional qualitative experiments for an image retrieval task.

Figure 4 shows the top-8 nearest neighbors for a given text query (from left to right and top to bottom: "car"+"fast", "car"+"slow", "aeroplane"+"passenger", "aeroplane"+"fighter", "people"+"eating", and "people"+"playing") in the learned topic space of of our model (without fine tuning). We appreciate that, by leveraging textual semantic information, our method learns rich visual representations that can disambiguate correctly between those combined queries.

Figure 5 shows the 4 nearest neighbors for a given query image (left-most), where each row makes use of features obtained from different layers of our model (again without fine tuning). Query images are randomly selected from PASCAL VOC 2007 dataset and never shown at training time. It can be appreciated that when retrieval is performed in the semantic space layers (prob-article and prob-caption), the results are semantically close, although not necessarily visually similar. As features from earlier layers are used, the results tend to be more visually similar to the query image.

6 CONCLUSION

In this article we put forward a self-supervised method that takes advantage of the natural correlation between an article's text and the images used to illustrate it, in order to learn useful visual representations.

The proposed method is capable of exploiting the rich semantics and broad coverage of illustrated articles, making use of both article-wide semantics and specific image semantics captured by the image caption.

We demonstrated that the learned visual features can transfer well to any general computer vision task such as image classification or object detection, while they can be directly used in a cross-modal retrieval framework yielding state of the art results both on the Wikipedia retrieval dataset and the Pascal Sentences dataset. Notably, the obtained model improves the state of the art not only in comparison to other self-supervised methods, but also when compared to supervised models.

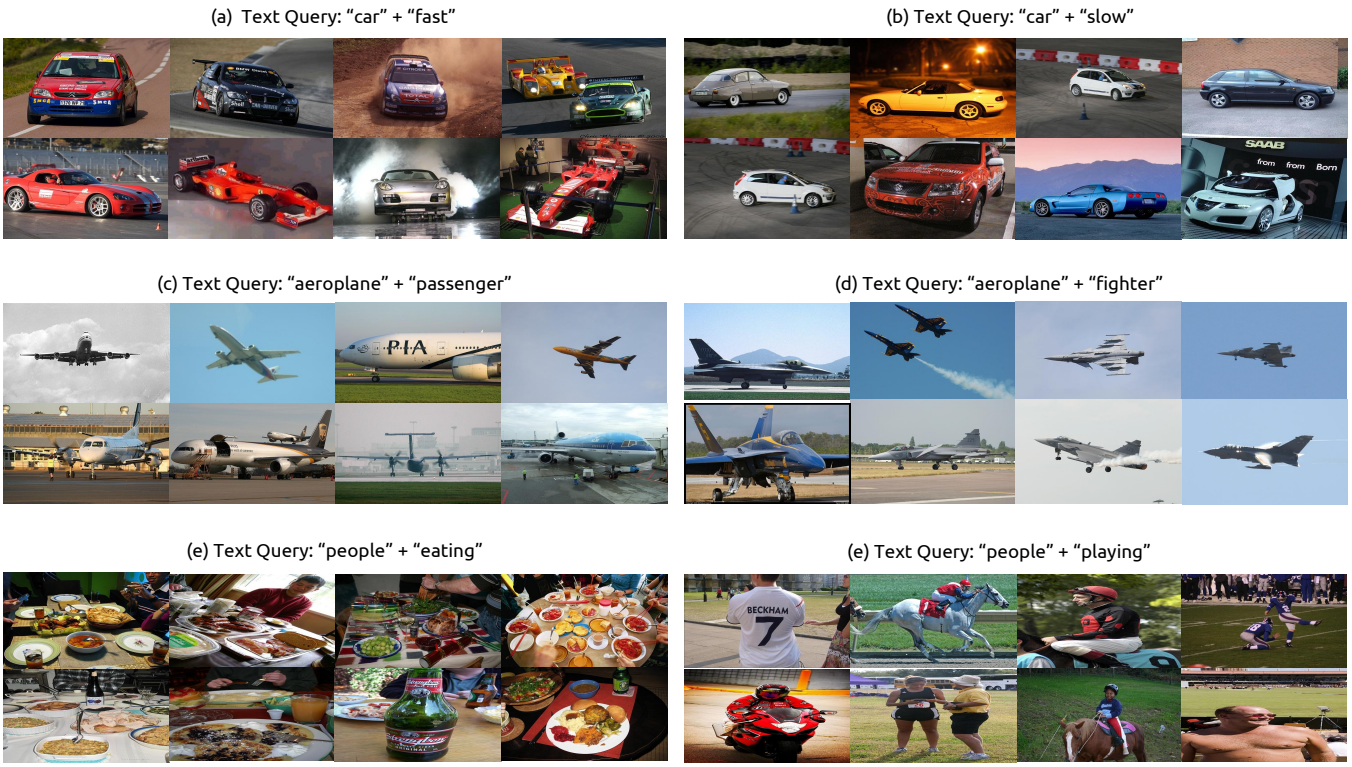

Figure 4: Qualitative examples of text query to image retrieval using nearest neighbour search by comparing network output from caption branch ($f_C(x, \Theta)$) with LDA topic probabilities ($\Phi(x^C)$).

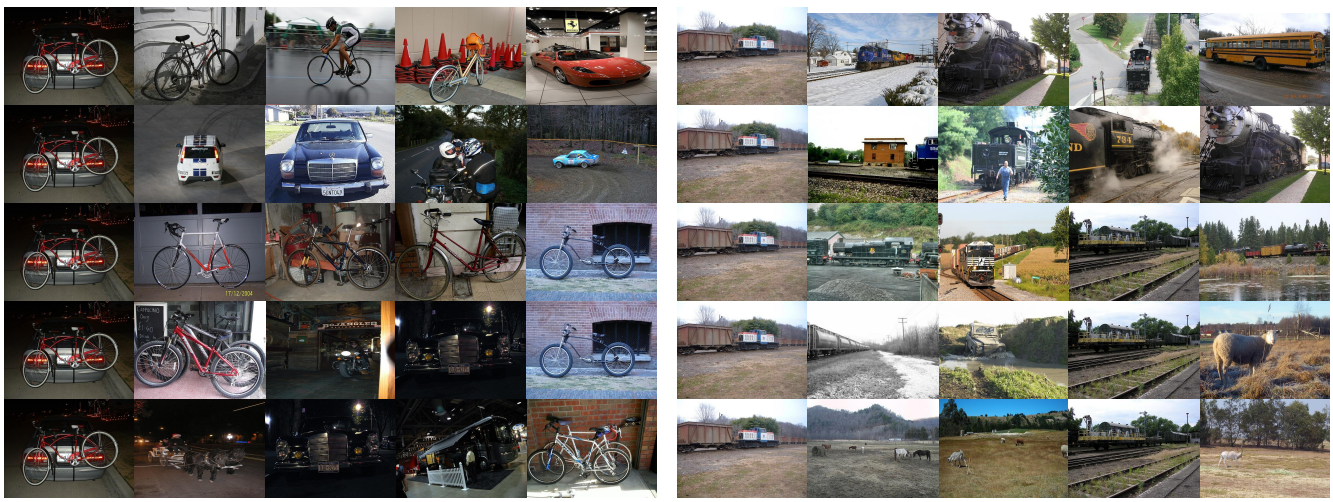

Figure 5: Top 4 nearest neighbors for a given query image image (left-most). Each row makes use of features obtained from different layers of our network (without fine tuning). From top to bottom: *prob-article* ($f_A(x, \Theta)$), *prob-caption* ($f_C(x, \Theta)$), *fc7*, *fc6*, *pool5*.

REFERENCES

[1] Pulkit Agrawal, Joao Carreira, and Jitendra Malik. 2015. Learning to see by moving. In *ICCV*.

[2] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *ICML*.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* (2003).

[4] Piotr Bojanowski and Armand Joulin. 2017. Unsupervised learning by predicting noise. *ICML* (2017).

[5] Adam Coates, Honglak Lee, and Andrew Y Ng. 2011. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS*.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

[7] Carl Doersch, Abhinav Gupta, and Alexei A Efros. 2015. Unsupervised visual representation learning by context prediction. In *ICCV*.

[8] Aysegul Dundar, Jonghoon Jin, and Eugenio Culurciello. 2016. Convolutional Clustering for Unsupervised Learning. In *ICLR*.

[9] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (VOC) challenge. *IJCV* (2010).

[10] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *ECCV*.

[11] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. 2014. Cross-modal retrieval with correspondence autoencoder. In *ACM-MM*.

[12] Lluis Gomez, Yash Patel, Marçal Rusiñol, Dimosthenis Karatzas, and CV Jawahar. 2017. Self-supervised learning of visual features through embedding images into text topic spaces. In *CVPR*.

[13] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. 2014. A multi-view embedding space for modeling internet images, tags, and their semantics. *International Journal of Computer Vision* (2014).

[14] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation* (2004).

[15] Cuicui Kang, Shengcai Liao, Yonghao He, Jian Wang, Wenjia Niu, Shiming Xiang, and Chunhong Pan. 2015. Cross-modal similarity learning: A low rank bilinear formulation. In *CIKM*.

[16] Philipp Krähenbühl, Carl Doersch, Jeff Donahue, and Trevor Darrell. 2015. Data-dependent initializations of convolutional neural networks. In *ICLR*.

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*.

[18] Dongge Li, Nevenka Dimitrova, Mingkun Li, and Ishwar K Sethi. 2003. Multimedia content processing through cross-modal association. In *ACM-MM*.

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *ECCV*.

[20] Hossein Mobahi, Ronan Collobert, and Jason Weston. 2009. Deep learning from temporal coherence in video. In *ICML*.

[21] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *ICML*.

[22] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. 2016. Ambient sound provides supervision for visual learning. In *ECCV*.

[23] Yash Patel, Lluis Gomez, Raul Gomez, Marçal Rusiñol, Dimosthenis Karatzas, and CV Jawahar. 2018. TextTopicNet-Self-Supervised Learning of Visual Features Through Embedding Images on Semantic Text Spaces. *arXiv preprint arXiv:1807.02110* (2018).

[24] Yash Patel, Lluis Gomez, Marçal Rusinol, and Dimosthenis Karatzas. 2016. Dynamic Lexicon Generation for Natural Scene Images. In *ECCV*.

[25] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. 2016. Context encoders: Feature learning by inpainting. In *CVPR*.

[26] Yuxin Peng, Xin Huang, and Jinwei Qi. 2016. Cross-Media Shared Representation by Hierarchical Learning with Multiple Deep Networks.. In *IJCAI*.

[27] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. 2010. A new approach to cross-modal multimedia retrieval. In *ACM-MM*.

[28] Roman Rosipal and Nicole Krämer. 2006. Overview and recent advances in partial least squares. In *Subspace, latent structure and feature selection*.

[29] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* (2015).

[30] Abhishek Sharma, Abhishek Kumar, Hal Daume, and David W Jacobs. 2012. Generalized multiview analysis: A discriminative latent space. In *CVPR*.

[31] Fumin Shen, Xiang Zhou, Yang Yang, Jingkuan Song, Heng Tao Shen, and Dacheng Tao. 2016. A Fast Optimization Method for General Binary Code Learning. *IEEE Transactions on Image Processing* (2016).

[32] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[33] Jingkuan Song, Yang Yang, Yi Yang, Zi Huang, and Heng Tao Shen. 2013. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In *ACM-SIGMOD*.

[34] Nitish Srivastava and Ruslan R Salakhutdinov. 2012. Multimodal learning with deep boltzmann machines. In *NIPS*.

[35] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. 2017. Adversarial cross-modal retrieval. In *ACM-MM*.

[36] Dong Wang and Xiaoyang Tan. 2016. Unsupervised feature learning with c-svddnet. *Pattern Recognition* (2016).

[37] Kaiye Wang, Ran He, Liang Wang, Wei Wang, and Tieniu Tan. 2016. Joint feature selection and subspace learning for cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2016).

[38] Kaiye Wang, Ran He, Wei Wang, Liang Wang, and Tieniu Tan. 2013. Learning coupled feature spaces for cross-modal matching. In *ICCV*.

[39] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. 2016. A Comprehensive Survey on Cross-modal Retrieval. *CoRR* (2016).

[40] Xiaolong Wang and Abhinav Gupta. 2015. Unsupervised learning of visual representations using videos. In *CVPR*.

[41] Xiaolong Wang, Kaiming He, and Abhinav Gupta. 2017. Transitive Invariance for Self-supervised Visual Representation Learning. In *ICCV*.

[42] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*.

[43] Xing Xu, Fumin Shen, Yang Yang, Heng Tao Shen, and Xuelong Li. 2017. Learning discriminative binary codes for large-scale cross-modal retrieval. *IEEE Transactions on Image Processing* (2017).

[44] Fei Yan and Krystian Mikolajczyk. 2015. Deep correlation for matching images and text. In *CVPR*.

[45] Xiaohua Zhai, Yuxin Peng, and Jianguo Xiao. 2014. Learning cross-media joint representation with sparse and semisupervised regularization. *IEEE Transactions on Circuits and Systems for Video Technology* (2014).

[46] Junbo Zhao, Michael Mathieu, Ross Goroshin, and Yann Lecun. 2016. Stacked what-where auto-encoders. In *ICLR*.

[47] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning deep features for scene recognition using places database. In *NIPS*.

[48] Xiaofeng Zhu, Zi Huang, Heng Tao Shen, and Xin Zhao. 2013. Linear cross-modal hashing for efficient multimedia search. In *ACM-MM*.