

# Self-supervised learning of visual features through embedding images into text topic spaces

Lluís Gomez\*

Computer Vision Center, UAB, Spain  
lgomez@cvc.uab.es

Marçal Rusiñol

Computer Vision Center, UAB, Spain  
marcal@cvc.uab.es

Yash Patel\*

CVIT, KCIS, IIIT Hyderabad, India  
yash.patel@students.iiit.ac.in

Dimosthenis Karatzas

Computer Vision Center, UAB, Spain  
dimos@cvc.uab.es

C.V. Jawahar

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

## Abstract

*End-to-end training from scratch of current deep architectures for new computer vision problems would require Imagenet-scale datasets, and this is not always possible.*

*In this paper we present a method that is able to take advantage of freely available multi-modal content to train computer vision algorithms without human supervision. We put forward the idea of performing self-supervised learning of visual features by mining a large scale corpus of multi-modal (text and image) documents. We show that discriminative visual features can be learnt efficiently by training a CNN to predict the semantic context in which a particular image is more probable to appear as an illustration. For this we leverage the hidden semantic structures discovered in the text corpus with a well-known topic modeling technique.*

*Our experiments demonstrate state of the art performance in image classification, object detection, and multi-modal retrieval compared to recent self-supervised or natural-supervised approaches.*

## 1. Introduction

A picture is worth a thousand words. When we read an article about an unknown object, event, or place we greatly appreciate that it is accompanied by some image that supports the textual information. These images complement the textual description and at the same time provide context to our imagination. Illustrated texts are thus ubiquitous in

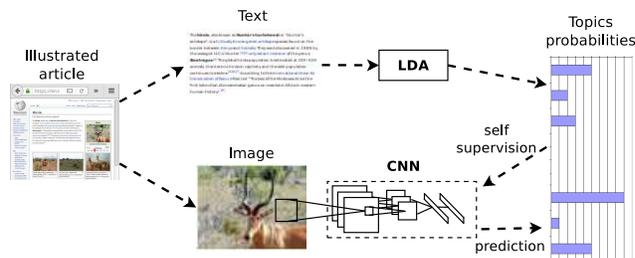


Figure 1: Our CNN learns to predict the semantic context in which images appear as illustration. Given an illustrated article we project its textual information into the topic-probability space provided by a topic modeling framework. Then we use this semantic level representation as the supervisory signal for CNN training.

our culture: newspaper articles, encyclopedia entries, web pages, etc. Can we take advantage of all this available multi-modal content to train computer vision algorithms without human supervision?

Training deep networks requires a significant amount of annotated data. The emergence of large-scale annotated datasets [5] has undoubtedly been one of the key ingredients for the tremendous impact deep learning is having on almost every computer vision task. However, the amount of human resources needed to manually annotate such datasets represents a problem. The goal of this paper is to propose an alternative solution to fully supervised training of CNNs by leveraging the correlation between images and text found in illustrated articles.

In most cases human generated data annotations consist of textual information with different granularity depending on the visual task they address: a single word to identify an

\*These authors contributed equally to this work

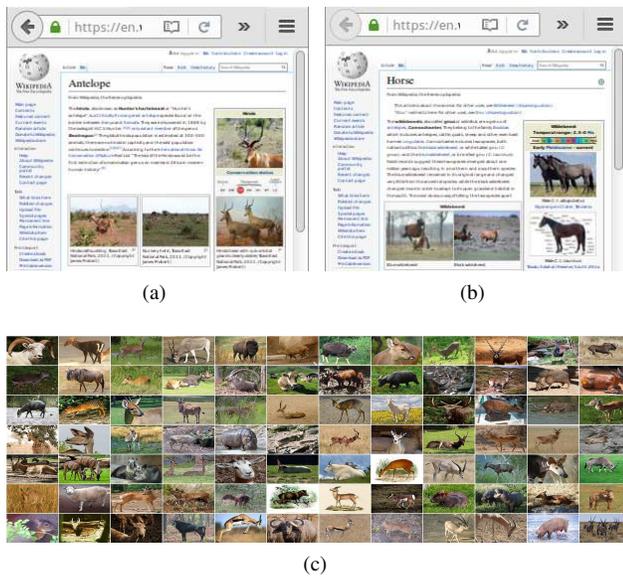


Figure 2: Illustrated Wikipedia articles about specific entities, like “Antelope” (a) or “Horse” (b), typically contain around five images. The total number of images for broader topics, e.g. “herbivorous mammals” (c), can easily reach hundreds or thousands.

object/place (classification), a list of words that describe the image (labeling), or a descriptive phrase of the scene shown (captioning). In this paper we consider that text found in illustrated articles can be leveraged as a type of image annotation, albeit being a very noisy one. The key benefit of this approach is that these annotations can be obtained for “free”.

Recent work in self-supervised or natural-supervised learning for computer vision has demonstrated success in using non-visual information as a form of self-supervision for visual feature learning [1, 41, 6, 25]. Surprisingly, the textual modality has been ignored until now in self-supervised methods for CNN training.

In this paper we present a method that performs self-supervised learning of visual features by mining a large scale corpus of multi-modal web documents (Wikipedia articles). We claim that it is feasible to learn discriminative features by training a CNN to predict the semantic context in which a particular image is more probable to appear as an illustration. For this we represent textual information at the topic level, by leveraging the hidden semantic structures discovered by the Latent Dirichlet Allocation (LDA) topic modeling framework [3], and use this representation as supervision for visual learning as shown in Figure 1.

As illustrated in Figure 2, the intuition behind using topic-level text descriptors is that the amount of visual data available about specific objects (e.g. a particular animal) is limited in our data collection, while it would be easy to find enough images representative of broader object cate-

gories (e.g. “mammals”). As a result of this approach the expected visual features that we are going to learn will be generic for a given topic, but still useful for other, more specific, computer vision tasks.

Our main motivation is to explore how strong are language semantics as a supervisory signal to learn visual features. In this paper we demonstrate that CNNs can learn rich features from noisy and unstructured textual annotations. By training a CNN to directly project images into a textual semantic space, our method is not only able to learn visual features from scratch without a large annotated dataset, but it can also perform multi-modal retrieval in a natural way without any extra annotation or learning efforts.

The contributions of this paper are the following: First, we present a method that performs self-supervised feature learning of visual features by leveraging the correlation between images and the semantic context in which they appear. Second, we experimentally demonstrate that the learned visual features provide comparable or better performance to recent self-supervised and unsupervised algorithms in image classification, object detection, and multi-modal retrieval tasks on standard benchmarks.

## 2. Related Work

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 (RBMs), auto-encoders [44], and K-means clustering [4, 8, 20]. 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.

As an alternative to fully-unsupervised algorithms, there has recently been a growing interest in self-supervised or natural-supervised approaches that make use of non-visual signals, intrinsically correlated to the image, as a form to supervise visual feature learning. Agrawal *et al.* [1] make use of egomotion information obtained by odometry sensors mounted on a vehicle to pre-train a CNN model. Wang & Gupta [41] use relative motion of objects in videos by leveraging the output of a tracking algorithm. Doersch *et al.* [6] learn visual features by predicting the relative position of image patches within the image. In Owens *et al.* [25] the supervisory signal comes from a modality (sound) that is complementary to vision.

In this paper we explore a different modality, text, for self-supervision of CNN feature learning. As mentioned earlier, text is the default choice for image annotation in many computer vision tasks. This includes classical image classification [5, 10], annotation [9, 17], and captioning [24, 23]. In this paper, we extend this to a larger level of

abstraction by capturing text semantics with topic models. Moreover, we avoid using any human supervision by leveraging the correlation between images and text in a largely abundant corpus of illustrated web articles.

Our method is closely related with various image retrieval and annotation algorithms that also use a topic modeling framework in order to embed text and images in a common space. Multi-modal LDA (mmLDA) and correspondence LDA (cLDA) [2] methods learn the joint distribution of image features and text captions by finding correlations between the two sets of hidden topics. Supervised variations of LDA are presented in [30, 42, 28] where the discovered topics are driven by the semantic regularities of interest for the classification task. Sivic *et al.* [33] adopt BoW representation of images for discovering objects in images using pLSA [16] for topic modelling. Feng *et al.* [11] uses the joint BoW representation of text and image for learning LDA. Most cross-modal retrieval methods work with the idea of representing data of different modalities into a common space where data related to same topic of interest tend to appear together. The unsupervised methods in this domain utilize co-occurrence information to learn a common representation across different modalities. Verma *et al.* [36] do image-to-text and text-to-image retrieval using LDA [3] for data representation. Methods such as those presented in [29, 13, 27, 22] use Canonical Correlation Analysis (CCA) for establishing relationships between data of different modalities. Rasiwasia *et al.* [29] proposed a method for cross-modal retrieval by representing text using LDA [3], image using BoW and CCA for finding correlation across different modalities.

Our method is related to these image annotation and image retrieval methods in the sense that we use LDA [3] topic-probabilities as common representation for both image and text. However, we differ from all these methods in that we use the topic level representations of text to supervise the visual feature learning of a convolutional neural network. Our CNN model, by learning to predict the semantic context in which images appear as illustrations, learns generic visual features that can be leveraged for other visual tasks. A similar idea is explored in the work of Gordo and Larlus [14] in these same proceedings, where image captions are leveraged to learn a global visual representation for semantic retrieval.

### 3. TextTopicNet

In order to train a CNN to predict semantic context from images (TextTopicNet) we propose a two-fold method: First, we learn a topic model on a text corpus of a dataset composed by pairs of correlated texts and images (i.e. illustrated articles). Second, we train a deep CNN model to predict text representations (topic-probabilities) directly from the image pixels. Figure 1 shows a diagram of the method.

### 3.1. LDA topic modeling

Our self-supervised learning framework assumes that the textual information associated with the images in our dataset is generated by a mixture of hidden topics. Similar to various image annotation and image retrieval methods discussed in 2, we make use of the Latent Dirichlet Allocation (LDA) algorithm [3] for discovering those latent topics and representing the textual information associated with a given image as a probability distribution over the set of discovered topics.

Representing text at topic level instead of at word level (BoW) provides us with: (1) a more compact representation (dimensionality reduction), and (2) a more semantically meaningful interpretation of descriptors.

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

where  $\theta$  is the mixing proportion and is drawn from a Dirichlet prior with parameter  $\alpha$ , and both  $\alpha$  and  $\beta$  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  $\beta$ . The probability of a document  $d$  in a corpus is defined as :

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

Learning LDA [3] on a document corpus provides two sets of parameters: word probabilities given topic  $P(w | z_{1:K})$  and topic probabilities given document  $P(z_{1:K} | d)$ . 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 the topics of the learned LDA model by projecting it into the topic space.

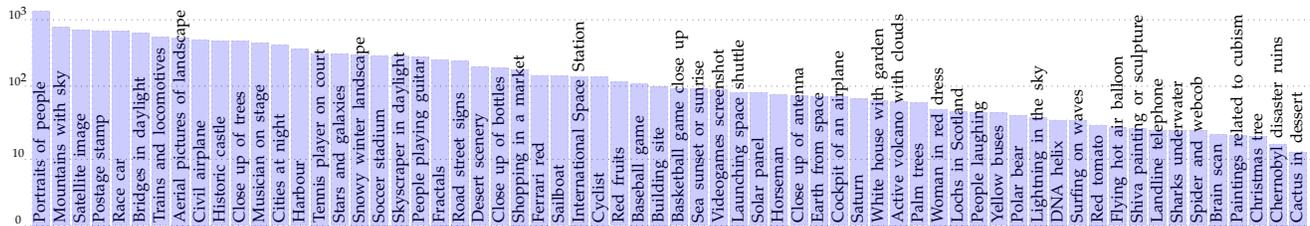


Figure 3: Number of relevant images (log scale) for a variety of semantic queries on the ImageCLEF Wikipedia collection [35].

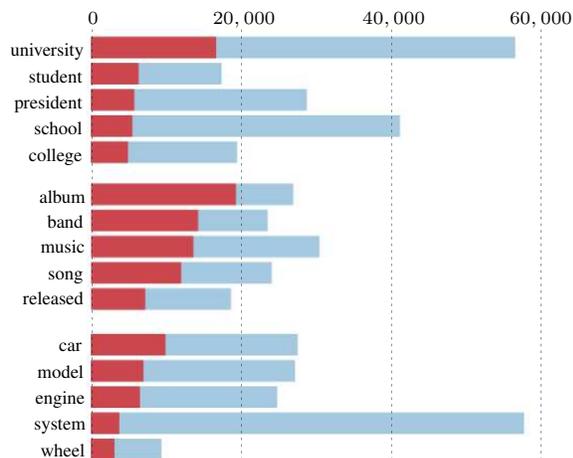


Figure 4: Top-5 most relevant words for 3 of the discovered topics by LDA analysis (left), and top-5 most relevant images for the same topics (right). Overall word frequency is shown in blue, and estimated word frequency within the topic in red.

### 3.2. Training a CNN to predict semantic topics

We train a CNN to predict text representations (topic probability distributions) from images. Our intuition is that we can learn useful visual features by training the CNN to predict the semantic context in which a particular image is more probable to appear as an illustration.

For our experiments we make use of two different architectures. One is the 8 layers CNN CaffeNet [18], a replication of the AlexNet [21] model with some differences (it does not train with the relighting data-augmentation, and the order of pooling and normalization layers is switched). The other architecture is a 6 layers CNN resulting from removing the 2 first convolutional layers from CaffeNet. This smaller network is used to do experiments with tiny images.

For learning to predict the target topic probability distributions we minimize a sigmoid cross-entropy loss on our image dataset. We use a Stochastic Gradient Descent (SGD) optimizer, with base learning rate of 0.001, multiplied by 0.1 every 50,000 iterations, and momentum of 0.9. The batch size is set to 64. With these settings the network converges after 120,000 iterations.

We train our models on a subset of Wikipedia articles provided in the Wikipedia ImageCLEF dataset [35]. The ImageCLEF 2010 Wikipedia collection consists of 237,434

Wikipedia images and the Wikipedia articles that contain these images. An important observation is that the data collection and filtering is not semantically driven. The original ImageCLEF dataset contains all Wikipedia articles which have versions in three languages (English, German and French) and are illustrated with at least one image in each version. Thus, we have a broad distribution of semantic subjects, similar as to the entire Wikipedia or other general-knowledge data collections. A semantic analysis of the data, extracted from the ground-truth of relevance assessments for the ImageCLEF retrieval queries, is shown in Figure 3. Although the dataset provides also human-generated annotations in this paper we train CNNs from scratch using only the raw Wikipedia articles and their images.

We consider only the English articles of the ImageCLEF Wikipedia collection. We also filter small images ( $< 256$  pixels) and images with formats other than JPG (Wikipedia stores photographic images as JPG, and uses other formats for digital-born imagery). This way our training data is composed of 100,785 images and 35,582 unique articles. We use data augmentation by random crops and mirroring.

Figure 4 shows the top-5 most relevant words for three of the discovered topics by LDA analysis, and the top-5 most relevant images for such topics. We appreciate that the discovered topics correspond to broad semantic categories for

which, a priori, it is difficult to find the most appropriate illustration. Still we observe that the most representative images for each topic present some regularities and thus allow the CNN to learn discriminative features, despite the noise introduced by other images that appear in articles from the same topic.

On the other hand, a given image will rarely correspond to a single semantic topic. Because by definition the discovered topics by LDA have a certain semantic overlap. In this sense we can think of the problem of predicting topic probabilities as a multi-label classification problem in which all classes exhibit a large intra-class variability. These intuitions motivate our choice of a sigmoid cross-entropy loss for predicting targets interpreted as topic probabilities instead of a one hot vector for a single topic.

### 3.3. Self-supervised learning of visual features

Once the TextTopicNet model has been trained following the steps in Section 3.1 and Section 3.2 it can be straightforwardly used in an image retrieval setting. Furthermore, it can be easily extended to an image annotation or captioning system by leveraging the common topic space in which text and images can be projected by the LDA and CNN models.

However, in this paper we are more interested in analyzing the qualities of the visual features that we have learned by training the network to predict semantic topic distributions. We claim that the learned features, out of the common topic space, are not only of sufficient discriminative power but also carry more semantic information than features learned with other state of the art self-supervised and unsupervised approaches.

The proposed self-supervised learning framework will have thus a broad application in different computer vision tasks. With this spirit we propose the use of TextTopicNet as a convolutional feature extractor and as a CNN pre-training method. We evaluate these scenarios in the next section and compare the obtained results in different benchmarks with the state of the art.

## 4. Experiments

In order to demonstrate the quality of the visual features learned by our text topic predictor (TextTopicNet) we have performed several experiments. First we analyze the quality of TextTopicNet top layers features for image classification on the PASCAL VOC2007 dataset [10]. Second we compare our method with state of the art unsupervised learning algorithms for image classification on PASCAL and STL-10 [4] datasets, and for object detection in PASCAL. Finally, we perform qualitative experiments on image retrieval from visual and textual queries.

For all our experiments we make use of the same LDA topic model learned on a corpus of 35,582 English

Wikipedia articles from the ImageCLEF Wikipedia collection [35]. From the raw articles we remove stop-words and punctuation, and perform lemmatization of words. The word dictionary (50,913 words) is made from the processed text corpus by filtering those words that appear in less than 20 articles or in more than 50% of the articles. At the time of choosing the number of topics in our model we must consider that as the number of topics increase, the documents of the training corpus are partitioned into finer collections, and increasing the number of topics may also cause an increment on the model perplexity [3]. Thus, the number of topics is an important parameter in our model. In the next section we take a practical approach and empirically determine the optimal number of topics in our model by leveraging validation data.

### 4.1. Unsupervised feature learning for image classification

In this experiment we evaluate how good are the learned visual features of the 6 layer CNN (CaffeNet) for image classification when trained with the self-supervised method explained in Section 3. Following [25] we extract features from top layers of the CNN and train one vs. rest linear SVMs for image classification in PASCAL VOC2007 dataset.

First of all, we perform model selection and parameter optimization using the standard train/validation split of the dataset. Figure 5 shows validation accuracy of SVM classification using *fc7* features for different number of topics in our model. Best validation performance is obtained for 40 topics. This configuration is kept for the rest of the experiments in this section.

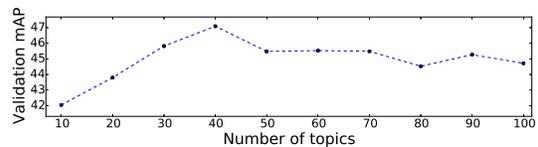


Figure 5: One vs. Rest linear SVM validation %mAP on PASCAL VOC2007 by varying number of topics of LDA [3] in our method.

Tables 1 and 2 compare our results on the PASCAL VOC2007 test set with different state of the art self-supervised learning algorithms. Scores for all other methods are taken from [25]. We appreciate in Table 2 that using text semantics as supervision for visual feature learning outperforms all other modalities in this experiment. In Table 1, attention is drawn to the fact that our pool5 features are substantially more discriminative than the rest for the most difficult classes, see e.g. “bottle”, “pottedplant” or “cow”.

TextTopicNet (COCO) in Table 2 corresponds to a model trained with MS-COCO [23] images and their ground-truth caption annotations as textual content. Since MS-COCO

Method	aer	bk	brd	bt	btl	bus	car	cat	chr	cow	din	dog	hrs	mbk	prs	pot	shp	sfa	trn	tv
TextTopicNet (Wiki)	67	44	39	53	<b>20</b>	<b>49</b>	<b>68</b>	42	43	<b>33</b>	41	<b>35</b>	70	57	82	<b>30</b>	31	<b>39</b>	65	41
Sound [25]	69	<b>45</b>	38	56	16	47	65	45	41	25	37	28	<b>74</b>	<b>61</b>	<b>85</b>	26	<b>39</b>	32	<b>69</b>	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 [41]	67	35	41	54	11	35	62	35	39	21	30	26	70	53	78	22	32	37	61	34
Patches [6]	<b>70</b>	44	<b>43</b>	<b>60</b>	12	44	66	<b>52</b>	<b>44</b>	24	<b>45</b>	31	73	48	78	14	28	<b>39</b>	62	<b>43</b>
Egomotion [1]	60	24	21	35	10	19	57	24	27	11	22	18	61	40	69	13	12	24	48	28
ImageNet [21]	79	<b>71</b>	<b>73</b>	<b>75</b>	<b>25</b>	60	80	<b>75</b>	51	<b>45</b>	60	<b>70</b>	<b>80</b>	<b>72</b>	<b>91</b>	42	<b>62</b>	56	82	62
Places [46]	<b>83</b>	60	56	<b>80</b>	23	<b>66</b>	<b>84</b>	54	<b>57</b>	40	<b>74</b>	41	<b>80</b>	68	90	<b>50</b>	45	<b>61</b>	<b>88</b>	<b>63</b>

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

Method	max5	pool5	fc6	fc7
TextTopicNet (Wiki)	-	<b>47.4</b>	<b>48.1</b>	<b>48.5</b>
Sound [25]	<b>39.4</b>	46.7	47.1	47.4
Texton-CNN	28.9	37.5	35.3	32.5
K-means [20]	27.5	34.8	33.9	32.1
Tracking [41]	33.5	42.2	42.4	40.2
Patch pos. [6]	26.8	46.1	-	-
Egomotion [1]	22.7	31.1	-	-
TextTopicNet (COCO)	-	<b>50.7</b>	<b>53.1</b>	<b>55.4</b>
ImageNet [21]	<b>63.6</b>	<b>65.6</b>	<b>69.6</b>	<b>73.6</b>
Places [46]	59.0	63.2	65.3	66.2

Table 2: PASCAL VOC2007 %mAP image classification.

annotations are human generated, this entry can not be considered a self-supervised method, but rather as a kind of weakly supervised approach. Our interest in training this model is to show that having more specific textual content, like image captions, helps TextTopicNet to learn better features. In other words, there is an obvious correlation between the noise introduced in the self supervisory signal of our method and the quality of the learned features. Actually, the ImageNet entry in Table 2 can be somehow seen as a model with a complete absence of noise, i.e. each image corresponds exactly to one topic and each topic corresponds exactly to one class (a single word). Still, the TextTopicNet (Wiki) features, learned from a very noisy signal, perform surprisingly well compared with the ones of the TextTopicNet (COCO) model.

As an additional experiment we have calculated the classification performance of the combination of TextTopicNet (Wiki) and Sound entries in Table 2. Here we seek insight about how complementary are the features learned with two different supervisory signals. By using the concatenation of *fc7* features of those models the mAP increases to 54.81%, indicating a certain degree of complementarity.

We further analyze the qualities of the learned features by visualizing the receptive field segmentation of TextTopicNet convolutional units using the methodology

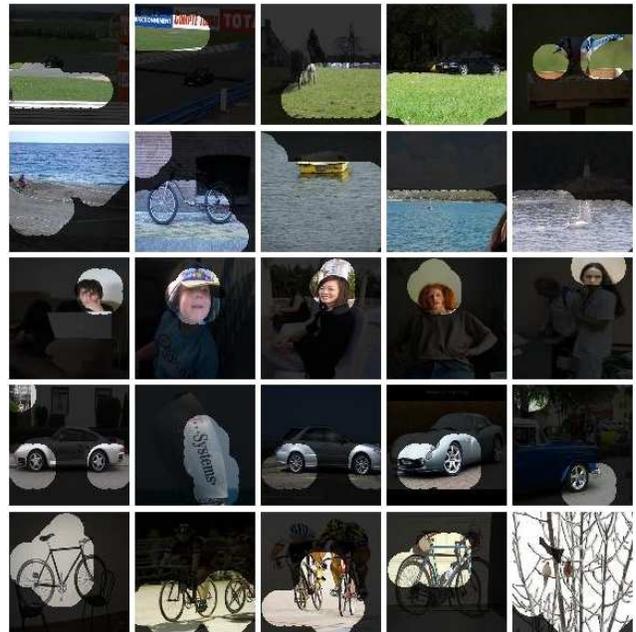


Figure 6: Top-5 activations for five units in *fc7* layer of TextTopicNet (Wiki) model. While most TextTopicNet units are selective to generic textures, like grass or water, some of them are also selective for specific shapes, objects, and object-parts.

of [45, 25]. The purpose of this experiment is to gain insight in what our CNN has learned to detect.

Figure 6 shows a selection of neurons in the *fc7* layer of our model. We appreciate that our network units are quite generic, mainly selective to textures, shapes and object-parts, although some object-selective units are also present (e.g. faces).

## 4.2. Comparison to unsupervised pre-training and semi-supervised methods

In this experiment we analyze the performance of TextTopicNet for image classification and object detection by fine-tuning the CNN weights to specific datasets (PASCAL and STL-10) and tasks.

For fine-tuning our network we use the following op-

timization strategy: we use Stochastic Gradient Descent (SGD) for 120,000 iterations with an initial learning rate of 0.0001 (reduced by 0.1 every 30,000 iterations), batch size of 64, and momentum of 0.9. We use data augmentation by random crops and mirroring. At test time we follow the standard procedure of averaging the net responses at 10 random crops. For object detection we fine-tune our classification network using Fast R-CNN [12] with default parameters for 40,000 iterations.

Table 3 compares our results for image classification and object detection on PASCAL with different self-supervised learning algorithms.

Method	classif.	detection
TextTopicNet	55.7	43.0
Sound [25]	-	44.1
K-means [20]	56.6	45.6
Tracking [41]	<b>62.8</b>	<b>47.4</b>
Patch pos. [6]	55.3	46.6
Egomotion [1]	52.9	41.8
ImageNet [21]	<b>69.6</b>	<b>73.6</b>
Egomotion [1] + K-means [20]	54.2	43.9
Tracking [41] + K-means [20]	63.1	47.2
Patch pos. [6] + K-means [20]	65.3	51.1

Table 3: PASCAL VOC2007 finetuning %mAP for image classification and object detection.

Table 4 compares our classification accuracy on STL-10 with different state of the art unsupervised learning algorithms. In this experiment we make use of the shortened 6 layers network in order to adapt better to image sizes for this dataset ( $96 \times 96$  pixels). We do fine-tuning with the same hyper-parameters as for the 6 layer network.

The standard procedure on STL-10 is to perform unsupervised training on a provided set of 100,000 unlabeled images, and then supervised training on the labeled data. While our method does not directly compare with unsupervised and semi-supervised methods in Table 4, because of the distinct approach (self-supervision), the experiment provides insight about the added value of self-supervision compared with fully-unsupervised data-driven algorithms. It is important to notice that we do not make use of the STL-10 unlabeled data in our training.

### 4.3. Multi-modal image retrieval

We evaluate our learned self-supervised visual features for two types of multi-modal retrieval tasks: (1) Image query vs. Text database, (2) Text query vs. Image database. For this purpose, we use the Wikipedia dataset [29], which consists of 2,866 image-document pairs split into train and test set of 2173 and 693 pairs respectively. For retrieval we project images and documents into the learned topic space

Method	Acc.
TextTopicNet (Wiki) - CNN-finetuning *	<b>76.51%</b>
TextTopicNet (Wiki) - fc7+SVM *	66.00%
Semi-supervised auto-encoder [44]	<b>74.33%</b>
Convolutional k-means [8]	74.10%
CNN with Target Coding [43]	73.15%
Exemplar convnets [7]	72.80%
Unsupervised pre-training [26]	70.20%
Swersky <i>et al.</i> [34] *	70.10%
C-SVDDNet [37]	68.23%
K-means (Single layer net) [4]	51.50%
Raw pixels	31.80%

Table 4: STL-10 classification accuracy. Methods with an asterisk mark make use of external (unlabeled) data.

and compute the KL-divergence distance of the query (image or text) with all the entities in the database. In Table 5 we compare our results with supervised and unsupervised multi-modal retrieval methods discussed in [40] and [19]. 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 [18], which is trained on ImageNet dataset [5] in a supervised setting. We appreciate that our self-supervised method outperforms unsupervised approaches, and has competitive performance to supervised methods without using any labeled data.

Method	Image query	Text query	Average
TextTopicNet	39.58	38.16	38.87
CCA [15, 29]	19.70	17.84	18.77
PLS [31]	30.55	28.03	29.29
SCM [29]	37.13	28.23	32.68
GMMFA [32]	38.74	31.09	34.91
CCA-3V [13]	40.49	36.51	38.50
GMLDA [32]	40.84	36.93	38.88
LCFS [39]	41.32	38.45	39.88
JFSSL [38]	42.79	39.57	41.18

Table 5: MAP comparison on Wikipedia dataset [29] with supervised (bottom) and unsupervised (middle) methods.

Finally, in order to analyze better what is the nature of learned features by our self-supervised TextTopicNet we perform additional qualitative experiments for an image retrieval task.

Figure 7 shows the 4 nearest neighbors for a given query image (left-most), where each row makes use of features obtained from different layers of TextTopicNet (without fine tuning). From top to bottom: prob, fc7, fc6, pool5. Query

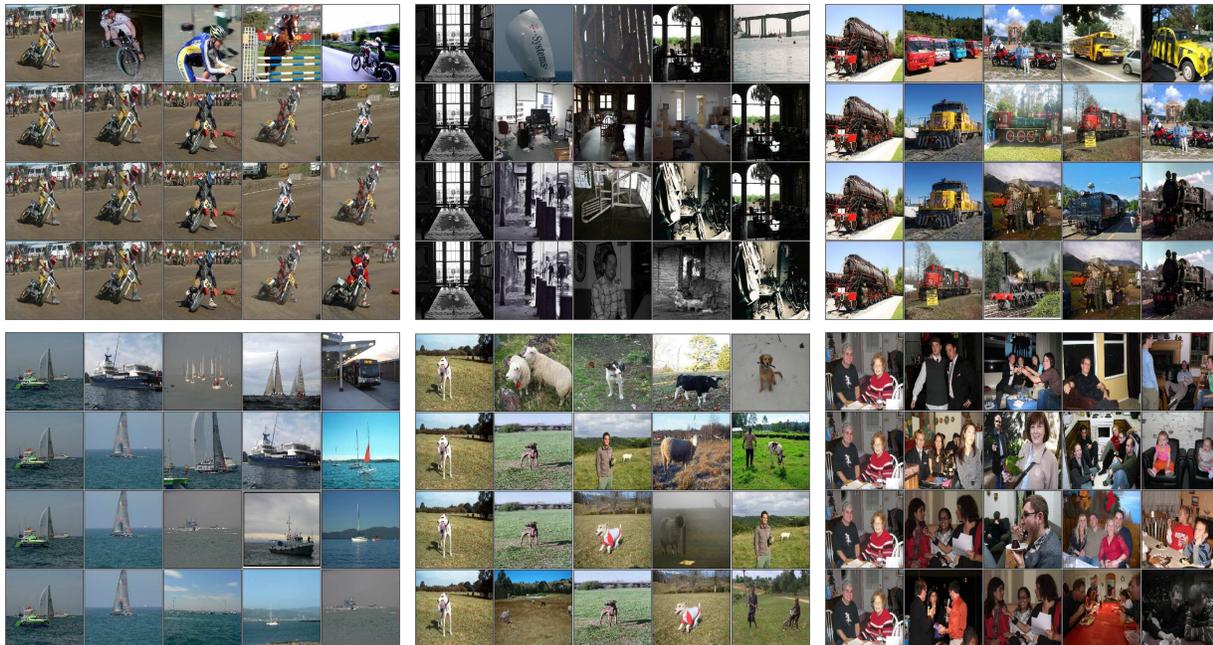


Figure 7: Top 4 nearest neighbors for a given query image image (left-most). Each row makes use of features obtained from different layers of TextTopicNet (without fine tuning). From top to bottom: prob, fc7, fc6, pool5.



Figure 8: Top 10 nearest neighbors for a given text query (from left to right: “airplane”, “bird”, and “horse”) in the topic space of TextTopicNet.

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 topic space layer (prob, 40 dimensions, top row), 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.

Figure 8 shows the 10 nearest neighbors for a given text query (from left to right: “airplane”, “bird”, and “horse”) in the topic space of TextTopicNet (again, without fine tuning). Interestingly, the list of retrieved images for the first query (“airplane”) is almost the same for related words and synonyms such as “flight”, “airway”, or “aircraft”. By leveraging textual semantic information our method learns a polysemic representation of images.

## 5. Conclusion

In this paper we have presented a method that is able to take advantage of freely available multi-modal content to train computer vision algorithms without human supervision. By considering text found in illustrated articles as

noisy image annotations the proposed method learns visual features by training a CNN to predict the semantic context in which a particular image is more probable to appear as an illustration.

The contributed experiments show that although the learned visual features are generic for broad topics, they can be used for more specific computer vision tasks such as image classification, object detection, and multi-modal retrieval. Our results are comparable with state of the art self-supervised algorithms for visual feature learning.

TextTopicNet source code and pre-trained models are publicly available at <https://git.io/vSotz>.

## Acknowledgment

We gratefully acknowledge the support of the NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research. This work has been partially supported by the Spanish research project TIN2014-52072-P and the CERCA Programme/Generalitat de Catalunya.

## References

- [1] P. Agrawal, J. Carreira, and J. Malik. Learning to see by moving. In *ICCV*, 2015. 2, 6, 7
- [2] D. M. Blei and M. I. Jordan. Modeling annotated data. In *SIGIR*, 2003. 3
- [3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 2003. 2, 3, 5
- [4] A. Coates, H. Lee, and A. Y. Ng. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS*, 2011. 2, 5, 7
- [5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1, 2, 7
- [6] C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, 2015. 2, 6, 7
- [7] A. Dosovitskiy, J. T. Springenberg, M. Riedmiller, and T. Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *NIPS*, 2014. 7
- [8] A. Dundar, J. Jin, and E. Culurciello. Convolutional clustering for unsupervised learning. In *ICLR*, 2016. 2, 7
- [9] P. Duygulu, K. Barnard, J. F. de Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV*, 2002. 2
- [10] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *IJCV*, 2010. 2, 5
- [11] Y. Feng and M. Lapata. Topic models for image annotation and text illustration. In *HLT*, 2010. 3
- [12] R. Girshick. Fast R-CNN. In *ICCV*, 2015. 7
- [13] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *International journal of computer vision*, 2014. 3, 7
- [14] A. Gordo and D. Larlus. Beyond instance-level image retrieval: Leveraging captions to learn a global visual representation for semantic retrieval. In *CVPR*, 2017. 3
- [15] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 2004. 7
- [16] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 2001. 3
- [17] M. J. Huiskes and M. S. Lew. The MIR flickr retrieval evaluation. In *MIR*, 2008. 2
- [18] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ICM*, 2014. 4, 7
- [19] C. Kang, S. Liao, Y. He, J. Wang, W. Niu, S. Xiang, and C. Pan. Cross-modal similarity learning: A low rank bilinear formulation. In *CIKM*, 2015. 7
- [20] P. Krähenbühl, C. Doersch, J. Donahue, and T. Darrell. Data-dependent initializations of convolutional neural networks. In *ICLR*, 2015. 2, 6, 7
- [21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 4, 6, 7
- [22] A. Li, S. Shan, X. Chen, and W. Gao. Face recognition based on non-corresponding region matching. In *ICCV*, 2011. 3
- [23] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 2, 5
- [24] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011. 2
- [25] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba. Ambient sound provides supervision for visual learning. In *ECCV*, 2016. 2, 5, 6, 7
- [26] T. L. Paine, P. Khorrami, W. Han, and T. S. Huang. An analysis of unsupervised pre-training in light of recent advances. In *ICLR*, 2015. 7
- [27] J. C. Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. R. Lanckriet, R. Levy, and N. Vasconcelos. On the role of correlation and abstraction in cross-modal multimedia retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014. 3
- [28] D. Putthividhy, H. T. Attias, and S. S. Nagarajan. Topic regression multi-modal latent Dirichlet allocation for image annotation. In *CVPR*, 2010. 3
- [29] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACM-MM*, 2010. 3, 7
- [30] N. Rasiwasia and N. Vasconcelos. Latent Dirichlet allocation models for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013. 3
- [31] R. Rosipal and N. Krämer. Overview and recent advances in partial least squares. In *Subspace, latent structure and feature selection*. 2006. 7
- [32] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs. Generalized multiview analysis: A discriminative latent space. In *CVPR*, 2012. 7
- [33] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering objects and their location in images. In *ICCV*, 2005. 3
- [34] K. Swersky, J. Snoek, and R. P. Adams. Multi-task bayesian optimization. In *NIPS*, 2013. 7
- [35] T. Tsirikika, A. Popescu, and J. Kludas. Overview of the Wikipedia image retrieval task at ImageCLEF 2011. In *CLEF (Notebook Papers/Labs/Workshop)*, 2011. 4, 5
- [36] Y. Verma and C. Jawahar. Im2Text and Text2Im: Associating images and texts for cross-modal retrieval. In *BMVC*, 2014. 3
- [37] D. Wang and X. Tan. Unsupervised feature learning with c-svddnet. *Pattern Recognition*, 2016. 7
- [38] K. Wang, R. He, L. Wang, W. Wang, and T. Tan. Joint feature selection and subspace learning for cross-modal retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 2016. 7
- [39] K. Wang, R. He, W. Wang, L. Wang, and T. Tan. Learning coupled feature spaces for cross-modal matching. In *ICCV*, 2013. 7
- [40] K. Wang, Q. Yin, W. Wang, S. Wu, and L. Wang. A comprehensive survey on cross-modal retrieval. *CoRR*, 2016. 7

- [41] X. Wang and A. Gupta. Unsupervised learning of visual representations using videos. In *CVPR*, 2015. 2, 6, 7
- [42] Y. Wang and G. Mori. Max-margin latent Dirichlet allocation for image classification and annotation. In *BMVC*, 2011. 3
- [43] S. Yang, P. Luo, C. C. Loy, K. W. Shum, and X. Tang. Deep representation learning with target coding. In *AAAI*, 2015. 7
- [44] J. Zhao, M. Mathieu, R. Goroshin, and Y. Lecun. Stacked what-where auto-encoders. In *ICLR*, 2016. 2, 7
- [45] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Object detectors emerge in deep scene CNNs. In *NIPS*, 2015. 6
- [46] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *NIPS*, 2014. 6