Self-Supervised Learning on Wikipedia Articles for Multimodal Retrieval

Sree Lakshmi Addepalli
Computer Science Department
New York University
USA
sla410@nyu.edu

Nisarg Thakkar
Computer Science Department
New York University
USA
nisargthakkar@nyu.edu

Srishti Grover
Computer Science Department
New York University
USA
sg5783@nyu.edu

Abstract—**Training deep architectures fully for daily computer vision algorithms requires large-scale Image-Net datasets. These labelled datasets are useful for the neural network to learn different visual features. Training this huge dataset is very time consuming and needs lots of resources. Similarly, finding and annotating datasets needs a huge amount of manual efforts for annotations which are nearly limited to common set of classes. In this project we model a method that takes advantage of free multi-modal content namely Wikipedia articles to train vision algorithms without any manual requirement. We plan to perform self-supervised learning of visual features by using a large-scale corpus of multimodal namely images and text documents. Learning visual features by making auxiliary work which make use of available self-supervision has become a demand in the community of computer vision.**

Web and various media sources platforms have an unlimited amount of multimodal data. In this paper, we propose an technique to use the multi-modal context to provide self-supervision for the training of vision algorithms. Self-Supervised learning performed on multimodal image and text data helps neural networks to get the powerful features with no need of manual annotated data. We find that discriminative visual features are learnt by training a CNN to predict the semantic context in which a particular image is most probable to be present. we find the hidden semantic structures in the text corpus with LDA mallet topic modeling technique. We display that the implemented pipeline can learn from images with corelated text without supervision and see the semantic structure of the trained joint image and text embedding space.

Our report shows state of the art performance in multimodal retrieval compared to current self-supervised approaches. Then, we use another Wikipedia dataset, composed of images and their respective texts that can be used for good comparison of image-text embeddings.

Keywords— Self-Supervised Learning, Text embeddings, Topic-Modeling, Multi-Modal Retrieval, CNN

1. Introduction

Deep learning techniques, powerful hardware, and large annotated datasets are resulting in getting outstanding machine learning result in more challenging tasks such as image captioning or language translation. Deep learning has largely been successful because of the following two requirements being met:

- Computation Power
- Tons of Data

Hardware power is evolving fast with GPU's and TPU's providing high computation power. But with tons of data being available the main issue with deep learning for specific problems is it is difficult or economically non-viable to get proper annotations. Annotating lots of data requires training supervised deep learning models which is a very costly and manual task. Therefore, Annotation has become a Bottleneck for Training Deep Neural Networks.

Alternative solutions and strategies for annotated data requirements of supervised deep learning techniques are not using fully supervised techniques. Here, Self- Supervised learnings can be used to exploit multimodal data to learn relations between various data modalities. Web data offers large amounts of images with other information such as the image description title or date. This data is unstructured and noisy but is unlimited while being free. The basic idea of self-supervision here is can we exploit co-relations between various modalities(like here the case is images and text) and automate the annotation part.

Current work in full-supervised or self-supervised learning for computer vision has shown success in using non-visual data as a form of self-supervision for visual feature learning as referenced in papers[6,7,8,9]. Text as a modality has not been well researched till now in self-supervised methods for CNN training. Web data has been used to build classification datasets, where queries are made to search engines using class names and the retrieved images are labeled with the querying class. Here learning is limited to known classes, hence it cannot generalize to new classes. Working with image labels gives convenience for training conventional visual models, the semantics in such a discrete space are very limited in comparison with richness of human language expressiveness when outlining an image.

In this project we try to exploit distributional semantics in a given text corpus and learn from every word associated to an image. Here by leveraging the richer semantics encoded within the learnt embedding space, we can infer previously unseen concepts even though they might not be explicitly present in the training set. The noisy and unstructured text linked to web images gives information about the content in the image content and we can use to train visual features. A way to do is to encode the multimodal data (text and image) in the known vectorial space. The textual information is represented at the topic level, by using the hidden semantic structures by applying the Latent Dirichlet Allocation (LDA) topic modeling framework and use this representation to train a CNN. The reason for using topic-level text embeddings is that the number of visual data available about certain objects say an animal "cow" is less in our data collection, while it is easy to find more images of broader object topics. By using LDA, the presumed visual features that are going to learnt will be generic for a topic but would be useful for more detailed computer vision tasks. The main aim is to find how strength of language semantics in unstructured text articles act as a supervisory signal to learn visual features. For the project we use large scale corpus of multi-modal web documents namely Wikipedia articles.

2. Related Work

2.1 Self-Supervised Learning

We do not have a generic unsupervised method that works well with real-world images, in spite of the success of several unsupervised learning benchmark datasets. There has been a rise in self-supervised methods that use non-visual signals, that are correlated to the image, to supervise visual feature training.

As described in paper [6] they make use of ego motion information obtained by odometry sensors mounted on a vehicle to pre-train a CNN model. They train the CNN network using contrastive loss formulation to get the camera transformations predictions between two pairs of images.

In paper [7] the authors find the relative motion of objects in videos by using the output of their tracking algorithm. The authors in paper [8] learnt the visual features by predicting the relative position of image patches within the image. Authors in [9] used sound as the supervisory signal that is complementary to vision. They do by training a deep CNN to predict hand-crafted statistical summary of sound associated with a video frame.

In our project we work with text modality, for self-supervision of CNN feature training. Text is the most common choice for annotating images in Computer Vision task from image classification, annotation and captioning. here, we extend to a higher-level abstraction by handling text semantics with topic models. Moreover, we do not have any kind of human supervision and find the correlation between text and images in a largely huge corpus of given web articles.

2.2 Deep Learning Image-Text Embeddings

Multimodal text embeddings and images have been trending lately in various research areas. This field of study has been motivated by possibility of learning from various modalities of data. In paper

[10] the authors proposed a pipeline that, inferred the Word2Vec representations of the class labels instead of learning to predict the ImageNet classes. This results in a model that generalizes outside the labeled classes present in the training set and it makes semantical relevant predictions even while it makes errors.

In the paper [11] the authors used the captions that were linked to the images and learnt a same embedding space for text and images from which they performed semantic image retrieval. The authors used TF-IDF based Bag of Words representation over the captions of the image to find a semantic similarity measure between images as they train the CNN to minimize loss based on the distances of triplets of query-similar-dissimilar images. Authors in [2] used LDA to extract topic probabilities from Wikipedia articles and trained a CNN to embed their associated images in the same topic space.

In application specific papers authors [12] proposed joint embeddings of food images and their recipes to identify ingredients, using LSTM representations to encode ingredient names and cooking instructions, Word2Vec and a CNN to extract visual features from the associated images.

In our project we try to learn discriminative and generic features on Wikipedia articles in a self-supervised manner without using any annotated dataset.

2.3 Topic Modelling

Various image retrieval and annotation algorithms use topic modeling framework in order to embed text and images in common space. Variants like Multi-modal LDA (mmLDA) and correspondence LDA (cLDA) methods learn joint distribution of text captions and image features by finding correlations between the two sets of hidden topics.

Many of the cross-modal retrieval methods have to work on the idea of presenting data of various modalities into same space where data related to similar topic of interest mostly appears together. In paper [13] proposed a method for cross modal retrieval by representing text using LDA, and image using Bag of Words and Canonical Correlation Analysis (CCA) for finding correlation across different modalities.

The suggested method in our project is to use LDA topic-probabilities as a common representation for both image and text. We use the topic level representations of text articles to supervise and train the visual feature learning of a Convolutional Neural Network. Our CNN model is trained to predict the semantic context in which images look as illustrations, as well as takes generic visual features that can be used for other visual tasks.

3. Dataset

We require a large-scale dataset of multimodal content to find the semantic correlation of text and image pairs for self-supervised learning of visual features. Hence, we use Wikipedia web site as for the source of dataset. Wikipedia is a web-based, multilingual encyclopedia project currently having over 40 million articles across 299 different languages. Wikipedia articles are usually composed by text and various kinds of multimedia objects like image, audio, and video files, and can hence be treated as multimodal documents.

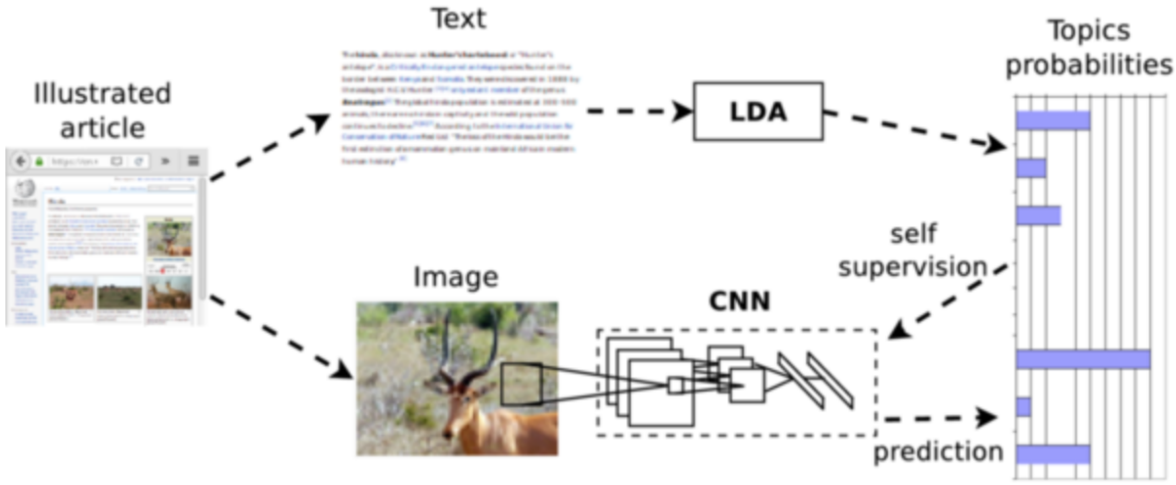
We used a modified version of Image CLEF 2010 Wikipedia collection available publicly as Image-Text Co-occurrence and consists of 4.2 million image-text pairs obtained from entire English Wikipedia. The modified version[14] considers only English articles of the Image CLEF Wikipedia collection. It also filters small images that are less than 256 pixels and images formats other than JPG. This way the modified Image CLEF training dataset consists of 100, 785 images and 35, 582 unique articles. Throughout the paper, we refer to this dataset.

4. Methodology

In order to train a self-supervised model to predict the semantic context from the given images our project proposes a two-fold method:

1. First, we apply the topic model on a text corpus of a dataset containing pairs of correlated images and texts.

Figure 1: Given a Wikipedia article we project the text into the topic-probability space provided by the topic modeling framework. We used the semantic level representation as the supervisory signal for CNN training. Our CNN learns to predict the semantic context in which images appear as illustration.



2. Second, we trained a CNN model (ALEXNET) to get the text representations (topic-probabilities) directly from the image pixels.

4.1. LDA Topic Modeling

The self-supervised learning framework has the assumption that the text data associated with the images in the dataset is made by amalgamation of hidden topics. The LDA mallet algorithm has been used for identifying these topics and presenting the textual information associated with the given image as a **probability distribution over the set of discovered topics.**

Showing the text at topic level instead of at word level (Bag of Words) provides following:
(1) Dimensionality Reduction gives a compact representation.
(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.

Let us assume we have a text corpus consisting of M documents and dictionary with N words, LDA define the generative process 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,
- 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 z1:K and word probabilities over β.
- The probability of a document d in a corpus is defined as

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

LDA on a document corpus provides two sets of parameters:

- Word probabilities given topic P(w | z1:K)
- Topic probabilities given document P(z1:K | d)

Therefore, each document is represented in terms of topic probabilities z1:K (K the number of topics) and word probabilities over the 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. [1]

4.2.Training a CNN to predict semantic topics

We trained a CNN to predict semantic context or topic probability distributions from images. We make use of Alexnet model as most of the existing self-supervised methods make use of this same architecture and hence, we can make a direct comparison with them.

To predict the target topic probability distribution, we use a categorical cross-entropy loss with soft classes to minimize the loss on the Wikipedia dataset. A Stochastic Gradient Descent (SGD) optimizer, with learning rate of 0.001, every 12,000 iterations. The batch size is set to 64 With these settings the network converges after 200,000 iterations. Please find below the layers of the network.

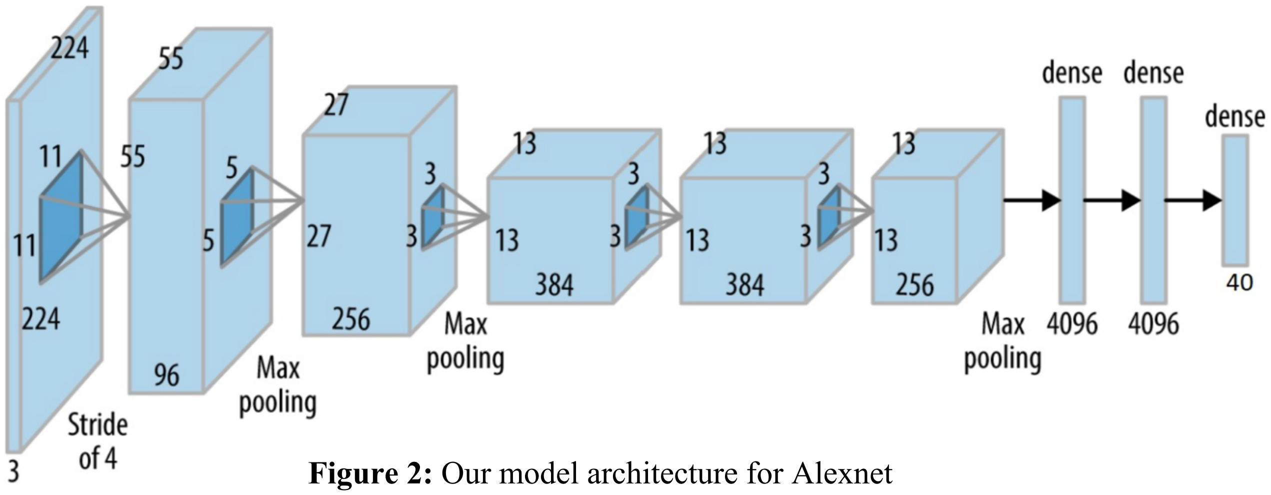


Figure 2: Our model architecture for Alexnet

A given input image will mostly never correspond to one semantic topic as the topics discovered by LDA are semantically overlapped. Hence predicting topic probabilities is a multi-label classification problem where all classes exhibit high intra-class variation. Hence, we choose a last layer as SoftMax with 40 classes for target prediction resulting in topic probabilities for a single topic.

4.3 Self-supervised learning of visual features

Once the model has been trained as above the self-supervised model can be used in an image retrieval application. It can also be used for image annotation or captioning system by accessing the common topic space in which images and text can be projected by the LDA and CNN models.

5. Experiments, Analysis and Results

We perform various analysis and benchmark comparisons of our model in the following manner:

- Comparisons of various Text Embeddings with the purpose of self-supervised learning from pairs of texts and images.
- Figure out the optimal number of topics our model will need.
- we experiment with our model in being used in image retrieval tasks from textual and visual queries.

5.1 Comparisons of various Text Embeddings

Here we reviewed the state of art of over several text-image embedding pipelines used to provide self-supervision for CNN training. In this paper [2] they trained one vs. rest SVMs on PASCAL VOC 2007 dataset using the image representation obtained by different layers on LDA, Word2Vec, Doc2Vec, Glove and Fast Text for evaluation.

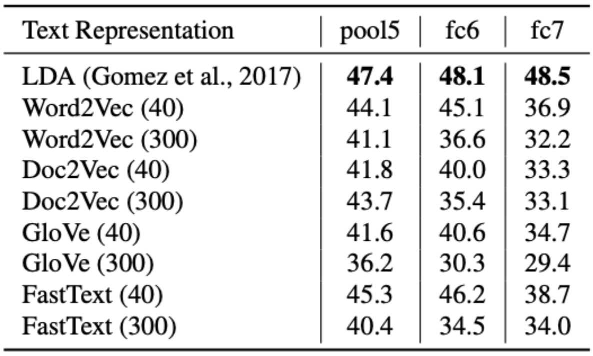
The pascal VOC 2007 dataset consisted of 9,963 images where every image has been marked with a bounding box and the object has been tagged as a class label in one of the twenty classes. This dataset has been a standard benchmark for image classification and well-suited for evaluation of self-supervised algorithms. Below are the text embeddings used to train:

- **Word2Vec:** Word2Vec understands relationships between words from a feed-forward neural network. It makes a distributed semantic representation of words using the context of words both before and after the target word.
- **Glove :** Glove learns vectors by doing dimensionality reduction on the co-occurrence counts matrix which is generated by a count-based model.
- **Doc2Vec:** Here it learns sentence, documents representations of the words.
- **Fast Text :** Each word is treated as a character ngrams, and here it learns the representations for ngrams instead of words. The vector here generated for a word is a sum of character n grams, such that it can make embeddings from the vocabulary words.

Doc2Vec and LDA directly generate text-article level embeddings while Word2Vec, Glove and FastText generate word level embeddings. Here to use these three types of embeddings for full text article supervision, we calculate the mean embeddings of all words within the articles are calculated.

They train the CNN with above text-embeddings, and one-vs-all SVMs on features were obtained from different layers in their network. Table 1 shows the results they obtained on the PASCAL VOC2007 image classification performance.

As LDA based embeddings performed the best for self-supervised learning of visual features we went ahead and incorporated a variant of it, LDA mallet in our project in the perspective that LDA Mallet gives a better quality of topics.



Text Representation	pool5	fc6	fc7
LDA (Gomez et al., 2017)	**47.4**	**48.1**	**48.5**
Word2Vec (40)	44.1	45.1	36.9
Word2Vec (300)	41.1	36.6	32.2
Doc2Vec (40)	41.8	40.0	33.3
Doc2Vec (300)	43.7	35.4	33.1
GloVe (40)	41.6	40.6	34.7
GloVe (300)	36.2	30.3	29.4
FastText (40)	45.3	46.2	38.7
FastText (300)	40.4	34.5	34.0

Table 1: PASCAL VOC2007 %mAP image classification[2]

5.2 Finding the optimal number of topics for LDA

The LDA mallet topic model was trained on the corpus of 35, 582 English Wikipedia articles from the raw articles of Image CLEF Wikipedia collection. First the raw words are removed followed by removal of stop-words, punctuation and lemmatization of words. The word dictionary generated is made to process text corpus by filtering words that appear in less than 20 articles or in more than 50% of the articles. As the number of topics increase, the documents get partitioned into finer collections from training corpus. Increasing the number of topics causes an increment on the perplexity of the model during the time of selecting the number of topics in our model. Hence, number of topics plays an important parameter for the topic model. we take a look of an approach to empirically find the optimal number of topics for our model.

In our project we find the optimal number of topics to build the LDA models with different values of number of topics (k) and picked the one that gave the highest coherence value. Topic Coherence score provides an easy way to measure how good the given topic model has been[16].

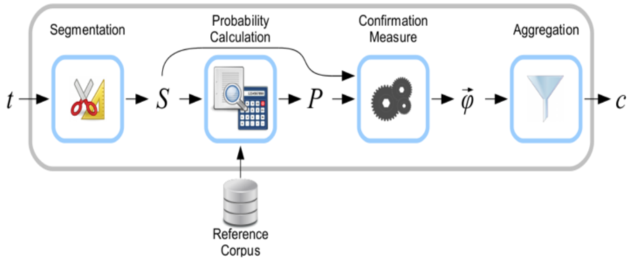


Figure 3: Topics for LDA

Where

t : Topics coming in from the topic model

S : Segmented topics

P : Calculated probabilities

Phi vector : A vector of the “confirmed measures” coming out from the confirmation module

c : The final coherence value

By selecting a ‘k’ value that tags the end of a huge growth of topic coherence values gives understandable and interpretable topics. Selecting a higher value can provide more granular subtopics, but everything would be repeating. Hence according to our model and the graph plotted below, we got 40 topics to be the number of models.

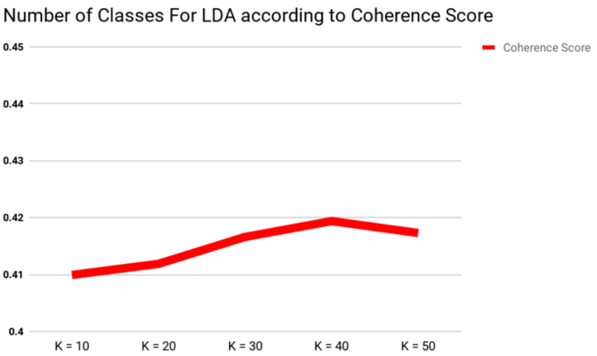


Figure 4: Number of classes selected based on coherence score.

5.3 Multi-modal image retrieval

Multimodal Image retrieval is the process where in we pass in a text query and retrieve Images similar in context and vice-versa. Here we train the text embedding LDA model on the data composed by pairs of images and related texts (I, x). Then we use we the LDA text embedding model to generate the vectorial representations of those texts. Let the text be x and let it be denoted by its embedding φ(x) ∈ R. Then we train the CNN on the text embeddings directly from the correlated images. Let an Image I be represented in the embedding space as ψ(I) ∈ RD. We train the CNN to embed the image model space into the vectorial space defined by the text embeddings model. Then for the test set images we generate the visual embeddings for the given Images. Below is the architecture diagram for multimodal retrieval.

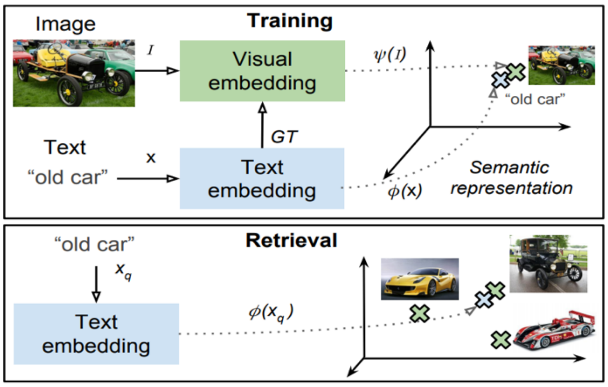


Figure 5: : Pipeline of the visual embedding model training and the image retrieval by text

We perform the self-supervised visual features for two types of multi-modal retrieval tasks namely Image Query and Text Query on a custom Wikipedia Dataset . It has about 2, 866 image-document pairs split into train set of 2173 pairs and test set of 693 pairs. To evaluate our model, we project the images and documents into learned topic space and find out the KL-divergence distance of the image query or the text query. We compared our results with the supervised and unsupervised models as mentioned in paper 1. All the below models were trained in LDA for text representation and pretrained CaffeNet for CNN training. Our Self-Supervised model performed well on text querying to retrieve images and was better than unsupervised learning methods while it performed decently on supervised techniques. We retrieved top 10 images from the given text queries. We performed an text query to retrieve nearby top 10 images obtaining an accuracy of 33.67%, and an Image query to retrieve the topics semantically related to it and achieved an accuracy of 30.01%.

Model	Type	Text Query Accuracy
JFSSL	Supervised Model	39.57%
SCM	Supervised Model	28.23%
CCA	Unsupervised Model	17.84%
PLS	Unsupervised Model	28.03%
Our model	Self - Supervised Model	33.67%

Table 2: Results while performing a text query to retrieve semantically nearer Images.

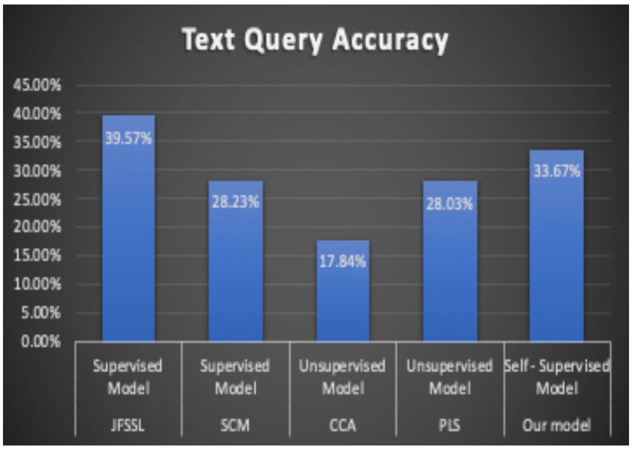


Chart 1: Results while performing a text query to retrieve semantically nearer Images.

6. Conclusion

In this project we used a method that takes advantage of free available Wikipedia multi-modal content to train vision algorithms without any human intervention.

The text used in the given articles noisy image annotations learns visual features by training a CNN to find the semantic context of the image in which it is more notably going to appear as an illustration. The model tends to become generic and has learnt the product of joint image and text embeddings for semantic concepts.

Even though we learned the visual features for a broader context in terms of we can use them in multimodal retrieval, object detections and other problems.

Our results were comparable to supervised models and were better than unsupervised models.

Acknowledgment

We gratefully acknowledge the support of the HPC with the resources of prince cluster used for this research. We also acknowledge Authors[1] of the paper Self-supervised learning of visual features through embedding images into text topic spaces for providing us the base idea and code repository[15] for reference. Lastly, I would like to thank professor Rob Fergus, for guiding us throughout the project in the Fall Semester of 2019 during the Computer Vision course.

References

[1] Gomez, Lluis, et al. "Self-supervised learning of visual features through embedding images into text topic spaces." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[2] Patel, Yash, et al. "TextTopicNet-Self-Supervised Learning of Visual Features Through Embedding Images on Semantic Text Spaces." arXiv preprint arXiv:1807.02110 (2018).

[3] Jenni, Simon, and Paolo Favaro. "Self-supervised feature learning by learning to spot artifacts." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[4] Dhruv Mahajan and Ross Girshick and Vignesh Ramanathan and Kaiming He and Manohar Paluri and Yixuan Li and Ashwin Bharambe and Laurens van der Maaten "Exploring the Limits of Weakly Supervised Pretraining" arXiv: 1805:00932.

[5] Raul Gomeza,b,∗ , Lluis Gomezb , Jaume Giberta , Dimosthenis Karatzasb "Self-Supervised Learning from Web Data for Multimodal Retrieval" arXiv:1901.02004.

[6] P. Agrawal, J. Carreira, and J. Malik. Learning to see by moving. In ICCV, 2015.

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

[8] X. Wang and A. Gupta. Unsupervised learning of visual representations using videos. In CVPR, 2015.

[9] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba. Ambient sound provides supervision for visual learning. In ECCV, 2016.

[10] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, J. Dean, Zero-Shot Learning by Convex Combination of Semantic Embeddings, in: NIPS, 2013.

[11] A. Gordo, D. Larlus, Beyond Instance-Level Image Retrieval: Leveraging Captions to Learn a Global Visual Representation for Semantic Retrieval, in: CVPR, 2017

[12] A. Salvador, N. Hynes, Y. Aytar, J. Marin, F. Ofli, I. Weber, A. Torralba, Learning Cross-Modal Embeddings for Cooking Recipes and Food Images, in: CVPR, 2017.

[13] Rasiwasia N, Costa Pereira J, Coviello E, Doyle G, Lanckriet GR, Levy R, Vasconcelos N (2010) A new approach to cross-modal multimedia retrieval. In: ACM-MM

[14]https://github.com/lluisgomez/TextTopicNet/tree/master/data/ImageCLEF_Wikipedia

[15] https://github.com/lluisgomez/TextTopicNet

[16]https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/

[17] https://rare-technologies.com/what-is-topic-coherence/