SELF SUPERVISED REPRESENTATION LEARNING FOR CAPTION GENERATION

Anindya Chakravarti, Rahul Vadaga & Sungwoo Chun
Department of Computer Science
New York University

ABSTRACT

In recent times, there has been a lot of progress in the area of unsupervised visual representation learning. Especially, self-supervised learning has shown that Convolutional Neural Networks (CNN) learn visual representation of an image by training on the pretext tasks such as context or rotation of partial images. This works explores the idea of using self supervised representation learning as a pretext task for generating captions of labeled images from MS COCO dataset. The argument that we are trying to prove is; the pretext task will require the model to learn to recognize and their different parts.

1 INTRODUCTION

The number of images we come across everyday is huge and the scale at which images are being uploaded to different social media websites runs in millions. The increasing amount of smartphones and pocket cameras is only going to increase the overall number by a few multiples. These images can be interpreted by humans easily but for a machine to be able to generate some meaning out of them requires a lot of labelled data. This labelling requires a lot of capital and man hours for annotating the images individually to make a rich dataset. This is the reason why using unsupervised learning for caption generation can be of benefit.

Image captioning is one of the fundamental problems in deep learning that connects computer vision and natural language processing. In order for a computer to communicate with humans, they have to understand the environment with a vision and communicate with the human language. There have been many approaches to generate a sentence from an image. One way is to find out the objects in the image and convert it to a word and make a sentence from it (Karpathy & Fei-Fei (2015)). The other way is to use the fact that understanding and generating a caption for an image largely depends on understanding the underlying features of the images. And we know that, CNNs are good at understanding the features of an image. The idea is to use these features and feed it into Long Short-Term Memory (LSTM) (S Hochreiter (1997)) based model and train it. The intuition is that the last fully connected layer will give out visual information to the LSTM. This idea of using CNN to generate features for LSTM was taken from Show and Tell (Vinyals et al. (2015)), which uses InceptionV3 (Ioffe & Szegedy (2015)) trained on ImageNet as the underlying CNN to generate captions for images in MS COCO. The results are very promising (shown in figure 1).

There are multiple models that use unsupervised learning to learn image representation by using multiple patches(Doersch et al. (2015)), using rotation(Spyros Gidaris (2018)), using jigsaw puzzle(Noroozi & Favaro (2016)) and colorize gray scale images (Zhang R. (2016)). We use the pretrained models from (Doersch et al. (2015)).

2 RELATED WORK

Image captioning is a very important problem, solutions to which largely depend on labeled data. The problem of creating descriptions from visual data has been studied in Computer Vision for a while now and recent advances in the understanding of image features have allowed to drive these systems and make more robust models. There are papers like Show and Tell (Vinyals et al. (2015)), which uses CNN (Inception v3 trained on ImageNet) as a base to learn image features and then uses

Figure 1: Sample results from Show and Tell.

these features to train an LSTM. There is Show, Attend and Tell (Kelvin Xu (2005)) which uses convolutional features and LSTM with attention over parts of images fed to LSTM to create a word by word description of the visual data.

The self-supervised learning framework is a framework which defines an unlabeled pretext task, using only the visual information present in the images to provide a supervisory signal for feature learning. The pretext task should be designed in such a way that will force the CNN to learn semantic image features in order to solve the task. As a result, high level semantic features will be encoded in the intermediate layers of these CNN.

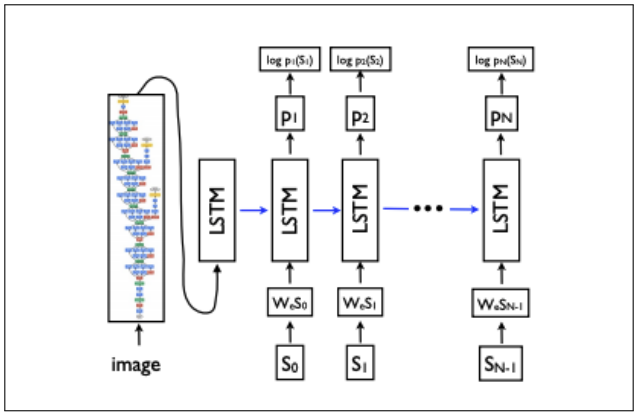

Figure 2: Show and tell model architecture, LSTM model combined with a CNN image encoder and word embeddings.

In this work, we combine the unsupervised learning from the models (Doersch et al. (2015); Spyros Gidaris (2018)) and the LSTM defined by Show and Tell. The authors of the paper(Doersch et al. (2015)) argue that to learn how to find the correct the position of the patches of an image, the model has to learn the underlying similarities across the images and be robust enough to understand the different objects in the images. We are trying to see if, given the model learns the similarities, can it be used as a good pretext task for image caption generation.

3 MODELS

In this work, we propose to use a self-supervised CNN as an image encoder for training a LSTM to generate description of visual data.

3.1 CNN ARCHITECTURE

The CNN architecture is based on ResNet50 and is trained on self-supervised pretext tasks of Context Prediction and Rotation.

Context Prediction (Relative Patch Location) Spyros Gidaris (2018): This technique divides a region in the image into 9 small patches and lets the model find out the configuration of two patches. This technique has been inspired from natural language processing technique which predicts words from the context which is words before and after each words. In the same sense, the model has to know the visual context of an image to find out the relative location of two patches. One interesting thing that the authors of the original paper noted is that the model is lazy and will always try to find shortcuts to learn image patterns that are invisible to the human eye. In this case, the model learns the chromatic aberration of the image, which happens because of the way a camera lens captures an image, to find out the general location.

Rotation Spyros Gidaris (2018): It is a technique to randomly rotate an image in 4 degrees $\{0, 90, 180, 270\}$ and let the neural network find out the correct orientation. It is a simple 4-class classification task. The model will have to know the image features to know the context of the image and the right orientation.

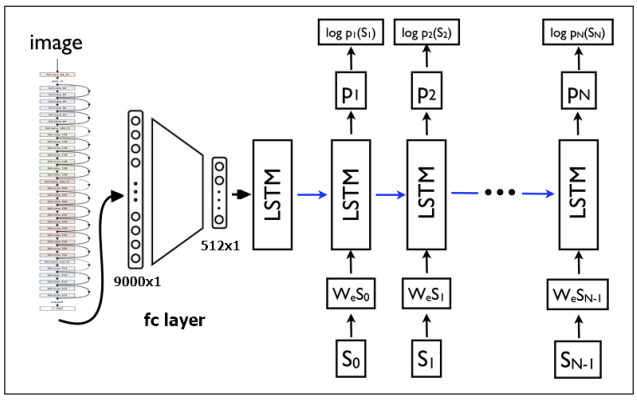

Figure 3: Sample results from Show and Tell.

3.2 LSTM MODEL

We use an LSTM model, described in the Show and Tell paper (Vinyals et al. (2015)), to predict the captions of an image. LSTMs are good at capturing long range dependencies in textual data, and are therefore a decent choice for decoding captions from the image representation. The most common LSTM variant consists of a cell and 3 gates – input, forget and output – which control the flow of information through the cell. A cell at time step t computes the current cell state from the input at time t, cell state at time $t-1$ and also the hidden state from time $t-1$. Once the cell state for time t is computed, it is used to generate the hidden state for time t, which is then further fed into the next time step $t+1$.

In our model, we generate visual representation of the input image from the network that was trained using the self supervised learning tasks discussed above. This representation is used as the first hidden state in the LSTM, along with the START symbol as its first input. The LSTM cell generates the first word in the caption, which is then fed as the input in the next time step. This process is repeated till the STOP symbol is generated.

There are multiple ways of generating the caption for an image using the LSTM. One method is **Sampling** where we will find the most probable candidate and each stage and use this as the output for the next stage. Another method, **BeamSearch**, maintains a list of best k candidates at every stage. For all of our experiments, we used $k = 20$. In practice, BeamSearch produces better results compared to Sampling method.

4 DATASET

For evaluation of image captions, we need a dataset which consists of images and sentences. Show and Tell paper was trained on MSCOCO 2014 dataset (T.-Y. Lin & Zitnick (2014)) which has largest and highest quality dataset for image captioning task. It contains 82k train images, 40k validation images and 40k test images.

The pre-training on self-supervised training was done on ImageNet and Places205. ImageNet contains 1.3 million natural images that represent 1k various semantic classes. There are 50k images in the official validation and test sets, and official test set is private. The Places 205 dataset consists of 2.5million images depicting 205 different scenes. These images are depicting different scene from the Imagenet which makes it a good candidate for evaluating whether the model has learned general visual representation from the ImageNet.

5 EXPERIMENTS

We were able to replicate the results created by the Show and Tell model for various images. We used a pre-trained tensorflow model(Kolesnikov et al. (2019)) and were able to get satisfactory descriptions of the the visual data. We also used the model from Unsupervised Representation Learning by Context Prediction to create the visual representations of different patches of the images.

This task of getting pre-trained models and the ensemble as discussed above working with Tensorflow turned out to be very time consuming. We are still going through the possible issues in our code to get the LSTM working.

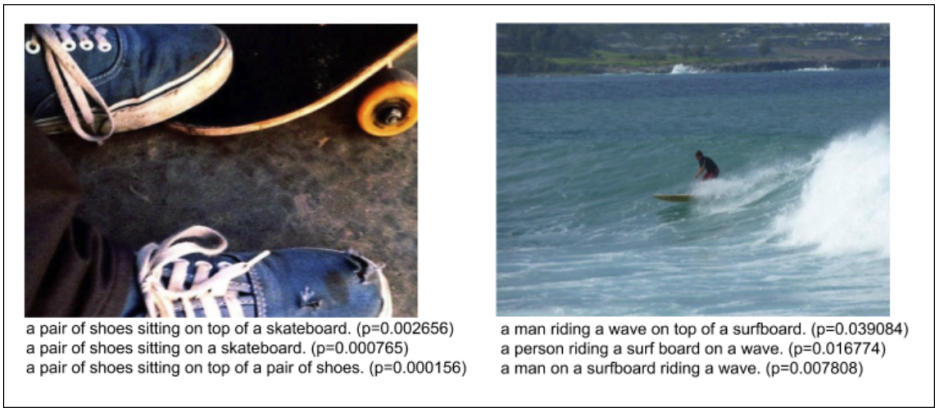

Figure 4: Results from Show and Tell replicated.

6 CONCLUSION

Although, we did not get any results from the project and were unable to get the ensemble of a CNN based on Unsupervised Learning and a LSTM to train end-to-end, we did get good insights into different quirks of deep learning. Some of them are:

1. CNNs try to learn shortcuts and generate trivial solutions. This was pointed out by the authors of Unsupervised Visual Representation Learning by Context Prediction(Doersch

et al. (2015)). The paper discusses how the CNN was able to learn the chromatic aberration of the camera lenses and perform very well for the similar locations in the image.

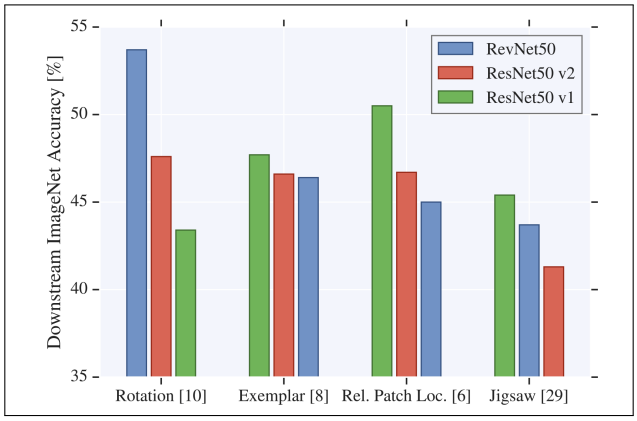

Figure 5: Performance of different self-supervised learning techniques on ImageNet recognition as downstream task.(Alexander Kolesnikov (2019))

2. Self-supervised learning performs very well as pretext task for downstream tasks. The Revisiting Self-Supervised Visual Representation Learning paper talks about some the best Self-Supervised techniques and how the representation learnt by these models perform in the downstream tasks. Figure 5 shows how the different self-supervised model perform very well in the ImageNet object recongition as a downstream task.

REFERENCES

Lucas Beyer Alexander Kolesnikov, Xiaohua Zhai. Revisiting self-supervised visual representation learning. *arXiv:1901.09005*, 2019.

Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. abs/1502.03167, 2015.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Ryan Kiros Kyunghyun Cho Aaron Courville Ruslan Salakhudinov Rich Zemel Yoshua Bengio Kelvin Xu, Jimmy Ba. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedgins of Machine Learning Research (PMLR)*, volume 35, July 2005.

Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. *arXiv preprint arXiv:1901.09005*, 2019.

M. Noroozi and P. Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. *In European Conference on Computer Vision (ECCV)*, 2016.

J Schmidhuber S Hochreiter. Long short-term memory. *Neural computation*, 1997.

Nikos Komodakis Spyros Gidaris, Praveer Singh. Unsupervised representation learning by predicting image rotations. *International Conference on Learning Representations (ICLR)*, 2018.

S. Belongie J. Hays P. Perona D. Ramanan P. Dollar T.-Y. Lin, M. Maire and C. L. Zitnick. Microsoft coco: Common objects in context, 2014.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Efros A.A. Zhang R., Isola P. Colorful image colorization. *European Conference on Computer Vision*, 2016.