Fisher Vector CNN for Image Captioning

Abhyuday Jagannatha
abhyuday@@cs.umass.edu

Aruni RoyChowdhury
arunirc@@cs.umass.edu

Huaizu Jiang
hzjiang@cs.umass.edu

Weilong Hu
weilong@cs.umass.edu

College of Information and Computer Sciences
University of Massachusetts, Amherst

Abstract

We propose a method for unsupervised domain adaptation of the deep convolutional neural network used to generate the image descriptors, which a language model uses as input in the image caption generation task. The Fisher Vector CNN model, which has been used earlier in the fine-grained visual recognition task, is used here as the domain-adapted image descriptor. A further modification appends explicit spatial information into this descriptor, which results in higher accuracy. The proposed method is simple, intuitive and results in better BLEU scores on the Flickr8k dataset as compared to using regular CNN features.

1. Introduction

The problem of describing images in natural language – generating captions for images – has received a lot of attention from the machine learning, NLP and vision research communities. In part this has been fueled by the advent of deep learning and massive labeled image datasets such as ImageNet.

Most of these approaches have the following basic structure [1]

- **Image descriptor:** Usually a pre-trained Convolutional Neural Network (CNN) is used as the image descriptor. CNNs have had great success in the ImageNet object classification challenge and mid-level features from these pre-trained networks act as excellent general- purpose image descriptors. The CNN is pre-trained for classifying 1000 object categories in the Imagenet dataset, consisting of a million images.
- **Language model:** Usually a Recurrent Neural Network (RNN) or some variant (like an LSTM), that is capable of modeling the sequence of words used to describe an image. This language model is trained on a set of images (the CNN image descriptor is used here as the input to the language model) and their corresponding captions. The RNN or LSTM eventually learns to generate captions provided with an image descriptor.

It is to be noted that *only* the language model is trained on the dataset of images and annotations for captions. The CNN, which provides the image descriptor, is usually held constant.

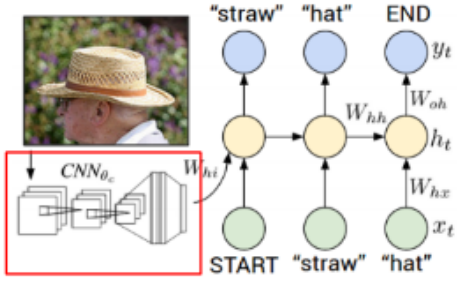

Figure 1. The model used by Karpathy et al. in NeuralTalk. The language model is trained, while the CNN providing the image descriptors remain constant. A domain adaptation of the CNN to the target domain should result in better quality image descriptors being provided to the language model, which would in turn generate better quality captions.

We hypothesize that being able to ***adapt the CNN*** to the target domain (e.g. Flickr8K) should yield better image descriptors for the task of generating captions for Flickr8k images. The distribution of images for the target task may be somewhat different from ImageNet and this domain adaptation would improve the specificity of the image descriptor being input to the LSTM, which in turn should result in better captions being generated.

Figure 2. Example image of an image from the Flickr8k dataset with caption: "black and white dog jumps over a bar."

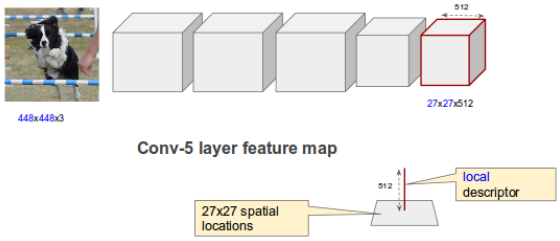

Figure 3. Schematic diagram for the extraction of local CNN features for the FV-CNN model.

2. Fisher Vector CNN model

We propose to adopt an ***unsupervised domain adaptation*** approach to get more target-database-specific image descriptors from the existing CNN. ***Fisher Vectors*** using CNN features have been shown to be very effective in the field of fine-grained visual recognition [3].

The CNN layers are truncated after the last convolutional layer. Using the standard 16-layer VGG-16 net, this gives a $13 \times 13 \times 512$ feature map as the output. We can think of this as 512-dimensional local image descriptors at each point on a 13×13 spatial grid over the input image.

A Gaussian Mixture Model (GMM) is learned on these 512-dimensional local descriptors. In an unsupervised manner, this GMM is able to model the descriptors obtained from the target domain images. The Fisher Vector (FV) representation [4] is very similar to the Bag-of-Words approach to aggregate a number of descriptors using a set of cluster centers. In case of FV, the average first and second order differences of data points to each Gaussian mode is stacked to form the descriptor. This gives a single $2KD$ dimensional vector describing an image, where K is the number of Gaussians in the GMM and D is the dimensionality of the local descriptor.

In place of the usual CNN descriptor, we use this FV-CNN feature as the input to the RNN. The rest of the setup is identical to the standard setup for caption generation models.

As another modification, we attach the (x, y) coordinates of each local descriptor to itself, resulting in a 514 dimensional local descriptor. This is done to encode some spatial information into the FV-CNN representation, such that it is able to be informative about spatial relationships when used for generating captions with words such as *"above"*, *"beside"*, *"over"*, etc.

3. Experimental Results

Our implementation was segmented into two parts, the Fisher Vector Convolutional Neural Network (FV-CNN), and the Long Short Term Memory(LSTM) network. We used MatConvNet [5] toolbox to generate the CNN (VGG-16) and FV-CNN descriptors. We used the Python library NeuralTalk [2] to implement the Long Short Term Memory Network. We adapt a VGG-16 CNN network pretrained on Image Net for this task. As described in previous sections, we use the CNN outputs of the last hidden layer to estimate the Gaussian Mixture Model for fisher encodings. For our spatially augmented Fisher CNN, we also augment the spatial co-ordinates to the CNN output vector. The LSTM Network is single layer with a dimensionality of 256. A softmax activation is used to predict words at a position one by one.

For the work reported in this paper, we chose Flickr8k dataset of eight thousand images from `http://nlp.cs.illinois.edu/HockenmaierGroup/Framing_Image_Description/KCCA.html`. The images in this data set focus on people or animals (mainly dogs) performing some action. The images were chosen from six different Flickr groups (Kids in Action, Dogs in Action, Outdoor Activities, Action Photography, Flickr-Social) and tend not to contain any well-known people or locations, but were manually selected to depict a variety of scenes and situations. We divide the data set into 6k training images, 1k validation images and 1k testing images. Each image is annotated with five different reference sentences. We evaluated original CNN (CNN), FV-CNN (FV) and spatially-augmented FV-CNN (saFV) on the Flickr8k data set. The evaluation metric used was BLEU (Bilingual Evaluation Understudy) scores. We use batch size of 100 sentences for the LSTM training. Each model is trained for 50 epochs. The training loss for each batch for LSTM training is plotted in Figure 4. We calculate perplexity on the validation dataset for each epoch. The trained model after a particular epoch is saved if the validation perplexity for that epoch is lower than any of the previous epochs. Finally, we use the model with the lowest recorded validation perplexity to calculate the BLEU scores on the test set.

The Table 3 shows the results of BLEU scores for the original CNN (CNN), FV-CNN (FV) and spatially-

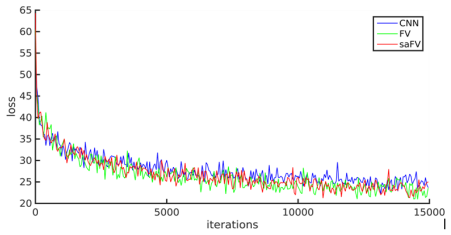

Figure 4. Training loss per batch over iterations

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
CNN	56.3	38.0	24.6	16.2
FV	55.2	37.1	24.1	15.8
saFV	**57.8**	**39.0**	**25.3**	**16.4**

Table 1. BLEU-n scores for each model evaluated over the test set.

augmented FV-CNN (saFV) models. BLEU-n evaluate n-gram similary in addition to the word based BLEU-1 score. They use exact word and n-gram matching to calculate the similarity between the predicted and referenced sentence. Since each image in Flickr8k dataset is labeled with five sentences the BLEU score for each predicted sentence is averaged over five reference sentences. As showed in Table 3, the spatially-augmented FV-CNN model achieves the highest scores in all BLEU scores, followed by the original CNN and FV-CNN. This validates our intuition behind augmenting CNN vectors with (x,y) co-ordinates. Since fisher vector encodings do not explicitly preserve spatial information, augmenting the (x,y) co-ordinates help provide additional spatial cues to the LSTM. We believe these cues are crucial to predict sentences describing spatial relations betwen objects. An example of such sentence would be 'The dog jumped over the fence'. Here, the cue for predicting 'over the' depends mainly on the spatial information provided by the vector representation that is fed in the LSTM network. Figure. 5 gives two examples of the performances of spatially-augmented FV-CNN model and original CNN (CNN) model. As showed in the top picture, both models are able to catch the "running dog" part, while spatially-augmented FV-CNN model in addition precisely describes "grass" when original CNN model only vaguely mentions "field". However for certain pictures, especially for pictures where only small partial of the object is observed, there are some issues in all models mentioned above. This can be seen in the second picture where both, the original CNN model and spatially-augmented FV-CNN model miss the swimming pool in the background.

Figure 5. Example of image captions generated by spatially augmented FV-CNN (saFV) and the baseline CNN with the ground truth (GT)

4. Conclusion

We show the usefulness of Fisher Vector Encodings for adapting pretrained Neural Network models like CNN to new tasks. We use the imagenet CNN trained for object recognition, and adapt it to the task of image captioning. For this specific task, we show that using spatially augmented Fisher Vectors for domain adaptation improves performance when compared to the baseline approch of using the unadapted CNN representation.

5. Future Work

We worked on varying the dimensionality of hidden layer from our initial value of 256. Our initial experiments at trying 400, 500 hidden layer size did not result in successful runs. The total cost seemed to explode while using reasonable learning rates. Future work would include increasing the hidden layer size to make sure we can take advantage of the larger dimensionality of the Fisher Encoding. Attention models can also be used to focus attention on certain parts of the Fisher Vector to improve the performance.

References

[1] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. 2015.

[2] A. Karpathy and J. Johnson. neuraltalk2. *https://github.com/karpathy/neuraltalk2*.

[3] C. M. S. Maji and A. Vedaldi. Deep filter banks for texture recognition and segmentation. 2015.

[4] S. J. P. F. M. T. and V. J. Image classification with the fisher vector: Theory and practice. 2013.

[5] A. Vedaldi and K. Lenc. Matconvnet – convolutional neural networks for matlab. 2015.