COMPUTER VISION AND IMAGE PROCESSING (CS676A)

Project Report

Image Captioning Using Object Proposals

Authors
Pramod Chunduri (13221)
Kriti Joshi (13358)

Supervisor: Prof. Vinay P. Namboodiri

Contents

1 Abstract **2**

2 Motivation **2**

3 Literature Review **3**

4 Methodology **3**
4.1 Architecture 3
4.2 LSTM model 4
4.3 Object Proposals 5
4.4 Methods 5

5 Dataset **6**

6 Results **6**

7 Generated Captions **7**

8 Conclusions **8**

9 Acknowledgment **9**

10 References **9**

1 Abstract

The aim of this project is to automatically describe content of an image. We achieved this target using a variant of the CNN + LSTM framework proposed by Vinyals et al. This model is trained to maximize the likelihood of target description given the image. Instead of sequentially giving word embedding as input to LSTM, we input the best object proposal of an image as embedding at certain places in the sequence. We observe that, depending on the location of sequence at which the proposal is inserted, the performance of captioning varies considerably. Finally, we compare the results obtained by using object proposals at different points with those obtained without using them.

2 Motivation

Image captioning task involves many non-trivialities. To describe any particular image, the usual tasks involved can be summarized as detecting the objects in the image, finding the correlation between those objects, finding attributes of these objects, retrieving the activities these objects are involved in and finally formation of sentences from above knowledge. To do these tasks in a step-by-step process would be a herculean task and also would not produce expected results, because there is large information loss in each of these steps. With the advent of deep neural networks, the step-by-step process has shifted to a generative approach, combining computer vision and natural language processing effectively, to produce captions to even some complex images.

But again, most of recent work in captioning has been done in Natural Language Processing, trying to improve upon the structure of sentences, creating more efficient vocabularies etc. That gives the primary motivation to work on Computer Vision side of the problem, by trying to give a more efficient representation of the image, rather than just their CNN features. So, we try to improvise the CNN features of an image by providing the object proposals to that image, which will help the model better understand the primary features (objects) around which the caption should be generated.

3 Literature Review

Automatic image captioning has been tried using various approaches. One of the method is to train model on human-written captions of an image. Recent retrieval techniques use neural-networks, where images and text are mapped on same feature space [3] and using some similarity metric [2] or fine-tuning[4], captions are generated. This method returns grammatically correct sentences but usually fails to work on the combination of new objects. Recently people have started working on another approach, where they extract all the important objects from the image and then generate captions[6]. Various attention based models are now used which can be trained in a deterministic manner and are good at describing an image [8] [1].

4 Methodology

4.1 Architecture

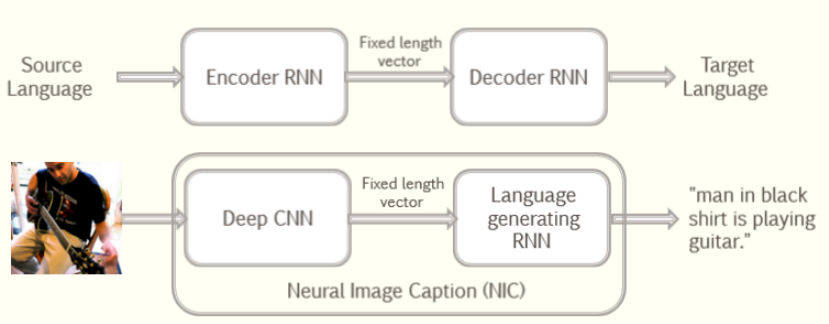

Figure 1: A high level interpretation of the model

- We use a pre-trained CNN which produces a feature vector of 512 size, when given a resized image of 256x256 as input.
- For the decoder RNN, we use a network of single-layered LSTMs, each of which takes a 512 dimension vector as new input and return the log probabilities of each word in the vocabulory while updating the memory element.

4.2 LSTM model

- We need to compute the log probabilities of occurrence of each word at every particular position of a sentence.
- The probabilities that we compute are based on the relation

$$\log p(S/I) = \sum_{t=0}^{N} \log p(S_t|I, S_0, \cdots, S_{t-1}) \tag{1}$$

  where N is the length of any particular sequence, and $S_0, \cdots, S_{t-1}$ are the first t words of the sequence.

- After seeing any new input x_t in the sequence, the memory state h_{t+1} is updated as

$$h_{t+1} = f(h_t, x_t) \tag{2}$$

- For this function f, an LSTM is used, with the definitions of different gates being same as a general LSTM.

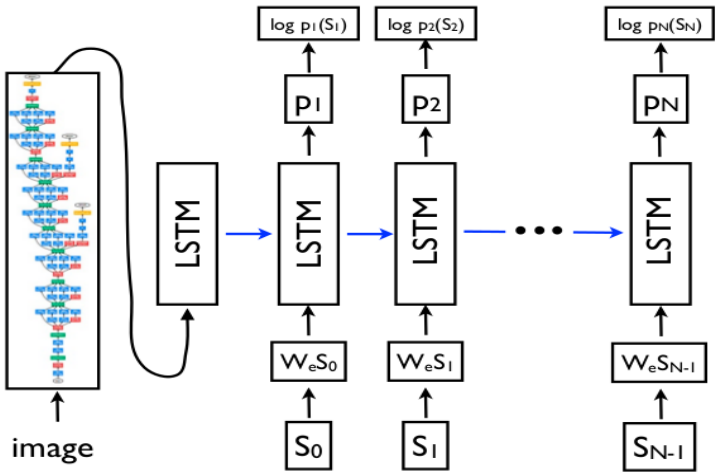

Figure 2: LSTM model

- The input x_t is a word embedding of size 512 constructed from the word vector of current word in training.

- The output of the LSTM is the log probabilities vector containing probabilities of occurrence every word in the vocabulary as the next word of the caption.

4.3 Object Proposals

- Object proposals are a set of candidate objects in a given image, known to be an efficient start point, for object recognition, segmentation, and other object-based image parsing tasks.
- These object proposals are returned with a confidence value, suggesting that we could retrieve the best candidate object in the image.
- Instead of just giving the word vectors as input embedding to LSTM, we try to use the feature vectors obtained by running a pre-trained CNN on object proposals, as the input embedding at certain places in the training sequence.
- The intuition is that, we try to bias the output caption towards the best object proposal, by using the proposal at appropriate places in the word sequence.

Figure 3: Object proposal for an image

4.4 Methods

We use three different methods of captioning.

1. Normal LSTM network is used with no object proposals.
2. The best object proposal is inserted as the first input embedding. The intuition being, we specify the best candidate object right away, so that the relevance to image is maintained.
3. The best object proposal is inserted at the second position in the sequence. The first word of caption is usually seen to be an article. The word following this word (frequently the subject) is usually the most informative. By specifying the best candidate object before this word, we hope to train the model more efficiently to predict the subject of the image.

5 Dataset

- We use the ImageNet model containing 33 layers as pre-trained CNN, used to extract features from both, the image and the best proposal.
- We currently train the model on 5000 images from the COCO dataset, validation is done on 500 images and testing is done on 500 images.
- We currently run the model till the validation error is stabilized.
- We obtained the proposals for the images by generating binary images from the COCO proposals dataset available online. [7]

6 Results

We have obtained the following results using the three methods mentioned above.

Metric	**Method 2**	**Method 3**
Val Loss	4.348(150)	3.514 (2)
Val Loss	4.618(290)	4.48(290)
Error	1.126 (150)	3.65 (2)
Error	1.041 (290)	1.10 (290)

Table 1: Validation loss

Metric	Method 1	Method 2	Method 3
Meteor	14.4	10.8	10.9
BLEU1	53.2	42.6	42.2
BLEU2	32.8	21.4	21.0
CIDEr	29.6	8.1	9.8
ROUGEl	39.5	32.5	32.5

Table 2: Percentage accuracy

7 Generated Captions

Best captions obtained from all three models:

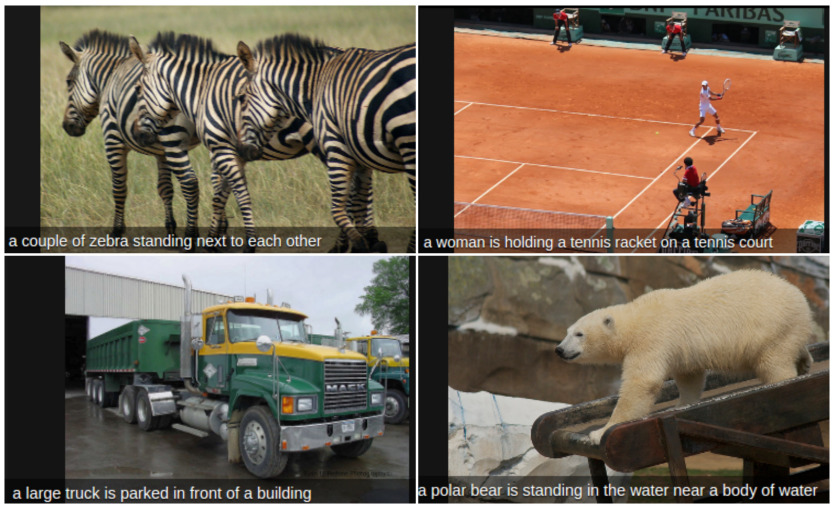

Figure 4: Model 1

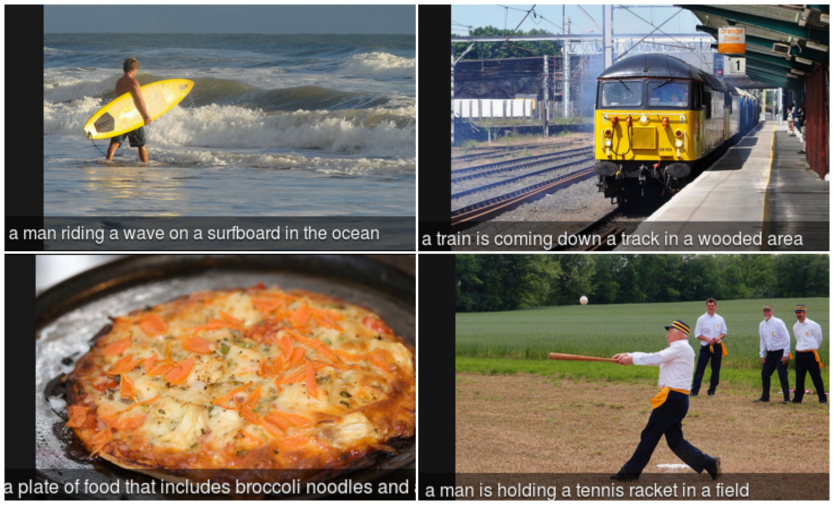

Figure 5: Model 2

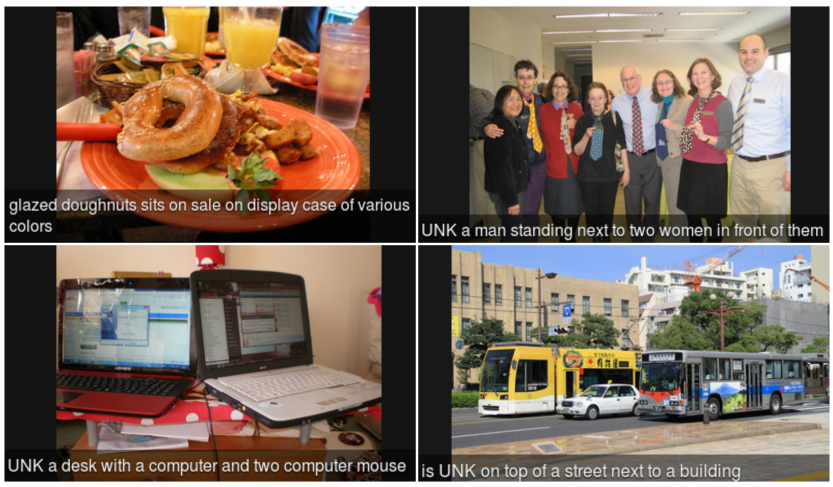

Figure 6: Model 3

8 Conclusions

The use of object proposals for captioning is seen to be improving the performance. Also, the position in which the proposal is input also effects the results. This suggests that captioning task responds greatly to changes in the

LSTM network. Apart from the methods used, we can try to input object proposals in order of confidences at every alternate step of the LSTM network, by keeping in account the structure of the caption generated (subject-verb-object).

9 Acknowledgment

We are grateful to Prof. Vinay P. Namboodri, Department of Computer Science and Engineering for providing us this opportunity to work on this project. Without his valuable suggestions and support, this project would not have been possible.

10 References

[1] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1473–1482, 2015.

[2] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *Computer Vision–ECCV 2010*, pages 15–29. Springer, 2010.

[3] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, pages 2121–2129, 2013.

[4] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015.

[5] Philipp Krähenbühl and Vladlen Koltun. Geodesic object proposals. In *Computer Vision–ECCV 2014*, pages 725–739. Springer, 2014.

[6] Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daumé III. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756. Association for Computational Linguistics, 2012.

[7] J. Pont-Tuset, P. Arbeláez, J. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping for image segmentation and object proposal generation, March 2015.

[8] Kelvin Xu, Jimmy Ba, Ryan Kiros, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2015.