Advanced Institute of Manufacturing with High-tech Innovations
National Chung Cheng University

Generating Personalized Hashtag Recommender for Instagram

Anmol Singh

Yu-Ling Hsueh

July 9th, 2018

Generating Personalized Hashtag Recommender for Instagram

Anmol Singh
Department of Computer Science and Information Engineering (CSIE)
Advanced Institute of Manufacturing with HI-tech Innovations
National Chung Cheng University, Chia-Yi 621, Taiwan, R.O.C.

Abstract

With the increased use of photo-sharing social media platforms such as Instagram, Facebook and Twitter, there has been a rise in the demand for automated caption and hashtag suggesting systems. This project deals with the same and makes use of state-of-the-art captioning model as described in the research paper, '*Show, Attend and Tell* [1]' to generate much more natural and human-like captions and hashtags by taking visual attention into account.

In the following project, we've jointly trained the CNN and RNN parts of the model [2] on MS COCO dataset using Tensorflow r1.0 in Python 2.7 unlike the code in Theano and have processed the captions generated to produce hashtags using *phrasemachine* [3] package of Python.

Having made use of the soft attention model for caption generation the performance of the captions generated matches the Theano implementation for the BLEU metric. However, since the hashtags are generated from the captions itself and do not make use of the personalized tags for training, there is much work that could be done to improve the evaluation score and personalization for them.

1. Introduction

Hashtag suggesting is a task of automatically generating tags for an image. Since this task deals with not just detecting the objects but analyzing the situation and context of the image, generating a hashtag hence is deeply connected to one of the frontier AI problems of image captioning. It requires an algorithm not only to understand the image content in depth beyond category or attribute levels, but also to connect its interpretation with a language model to create a natural sentence.

Personalized Hashtag Recommender would come as a rescue for a lot of users over the various social media platforms since crafting text is more cumbersome than taking a picture for general users. Photo-taking can be done with only a single tap on the screen of a smartphone, whereas text writing requires more time and mental energy for selecting suitable keywords and tags to describe theme, sentiment, and context of the image that too considering the things trending over the internet at time.

We, in this project, have tried to generate certain suitable hashtags for images by extracting the different noun, verb, adjective phrases, etc. from the captions that are generated based on the '*Show, Attend and Tell* [1]' model. We haven't taken into account the criteria of trending hashtags as is available over Instagram.

The contribution of this project is as follows:

(1) Tensorflow implementation of the soft attention model [1] for the purpose of image captioning
(2) Generating a list of hashtag suggestions for images by extracting specific phrases from the captions generated

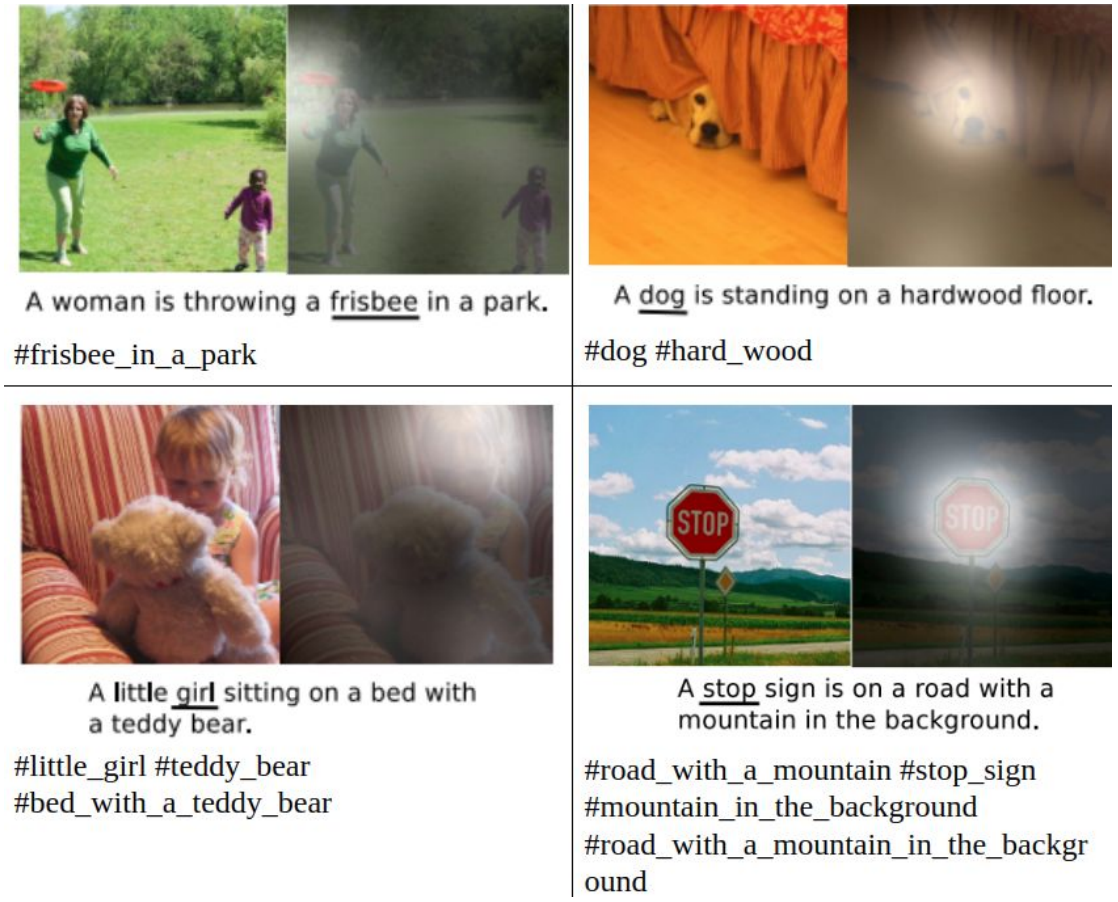


Fig. 1. Examples of image captions generated using visual attention and subsequently extracted hashtags (*white* indicates the attended regions, *underlines* indicates the corresponding word)

1.1. Related Work

Though there has been a lot of progress and research in the field of image captionioning. However, there hasn't been much of an in-depth look into the hashtag recommending system.

The only significant work in this field is the model described in the 'Attend to You [4]' research paper wherein they make use of the Instagram Personalized Image Captioning dataset for generating hashtags and post-generation text. It exploits memory as a repository for multiple types of context information giving them an added advantage of better personalization. However, unlike them we are trying to extract the useful phrases out of the captions generated by the soft visual attention model for creating a list of hashtag recommendations.

For the image captioning there has been quite extensive research going on of which the most successful one being the models defined within '*Show and Tell* [5]' and '*Show, Attend and Tell* [1]' research papers. Since the attention model provides us with better efficiency with an added aid of visual attention, we've adopted it as our base for image captioning.

2. Methodology

The hashtag recommending system is broadly divided into two subdivisions:

(1) Image Captioning
(2) Hashtag Extraction

2.1. Image Captioning with Attention Mechanism

Image Captioning is the process of generating textual description of an image. It uses both Natural Language Processing and Computer Vision to generate the captions.

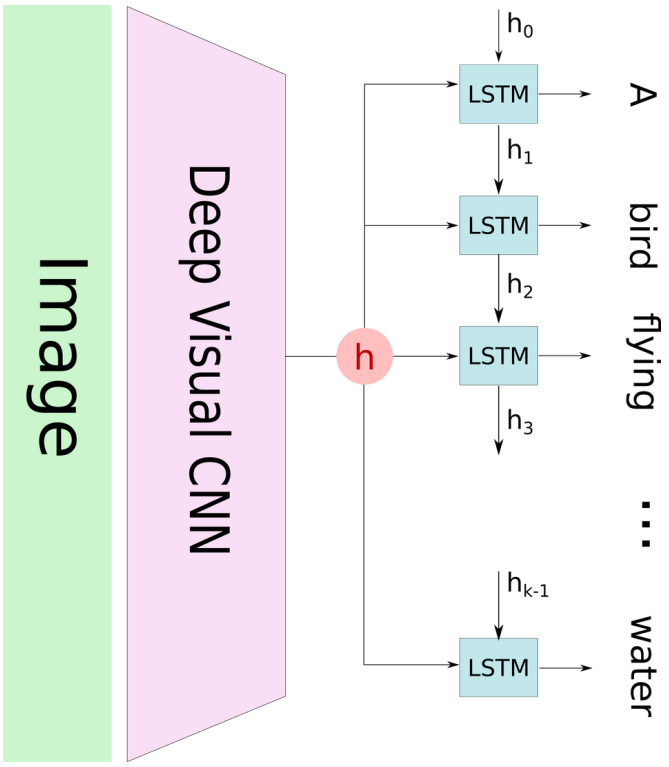


Fig. 2. A basic image-captioning model with a CNN encoder and LSTM RNN decoder (devoid of an attention model)

The input is an image, and the output is a sentence describing the content of the image. It uses a convolutional neural network to extract visual features from the image, and uses a LSTM recurrent neural network to decode these features into a sentence. The image captioning system used herein is adopted from the visual attention as in '*Show, Attend and Tell* [1].' Here's a simpler and much more intuitive explanation for the same:

2.1.1. Encoder

A convolutional neural network acts as an encoder which extracts a set of feature vectors referred to as annotation vectors. It takes a single raw image as input and generates L vectors, each being in a D-dimensional space.

$$a = \{a_1, \ldots, a_L\}, \; a_i \in \mathbb{R}^D \tag{1}$$

Unlike the models without visual attention, the extracted features are from a lower convolutional layer instead of a fully connected layer. This helps the attention model to focus on certain parts of the image.

We have made use of pretrained VGG16 net for initializing the CNN encoder.

2.1.2. Decoder

A long short-term memory (LSTM) network, a type of RNN, acts as a decoder for generating one word at every step by taking a context vector, i.e. a previous hidden state h_i, and the result of the attention model z_i as inputs.

2.1.3. Attention Model

An attention model considers all sub regions y_i and context C as its inputs and it outputs the weighted arithmetic means of these regions.

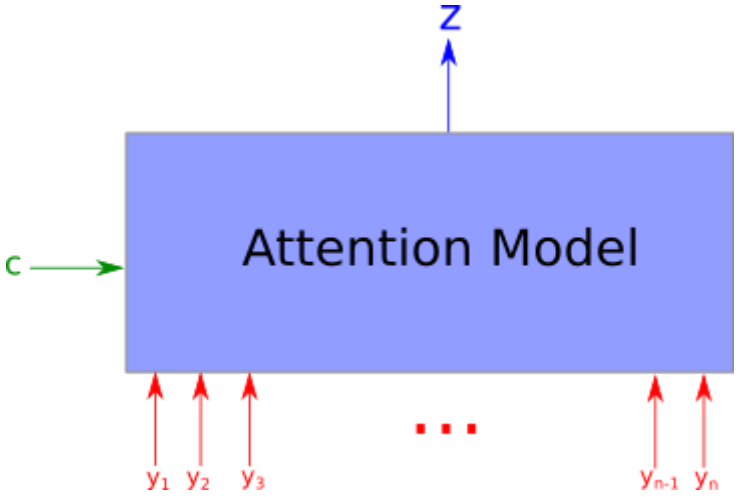


Fig. 3. A basic visualization of an attention model

$$E(X) = \sum_{n} p\,(X = X_n)\, X_n \tag{2}$$

Here, these probabilities p are determined using the context C which represents everything the RNN has outputted until then.

These inputs y_i and C are applied to weights which constitute the learnable parameters of the attention unit. i.e. the weight vectors update as we get more training data.

$$m_i = \tanh\ (y_i W_{y_i} + C W_C) \tag{3}$$

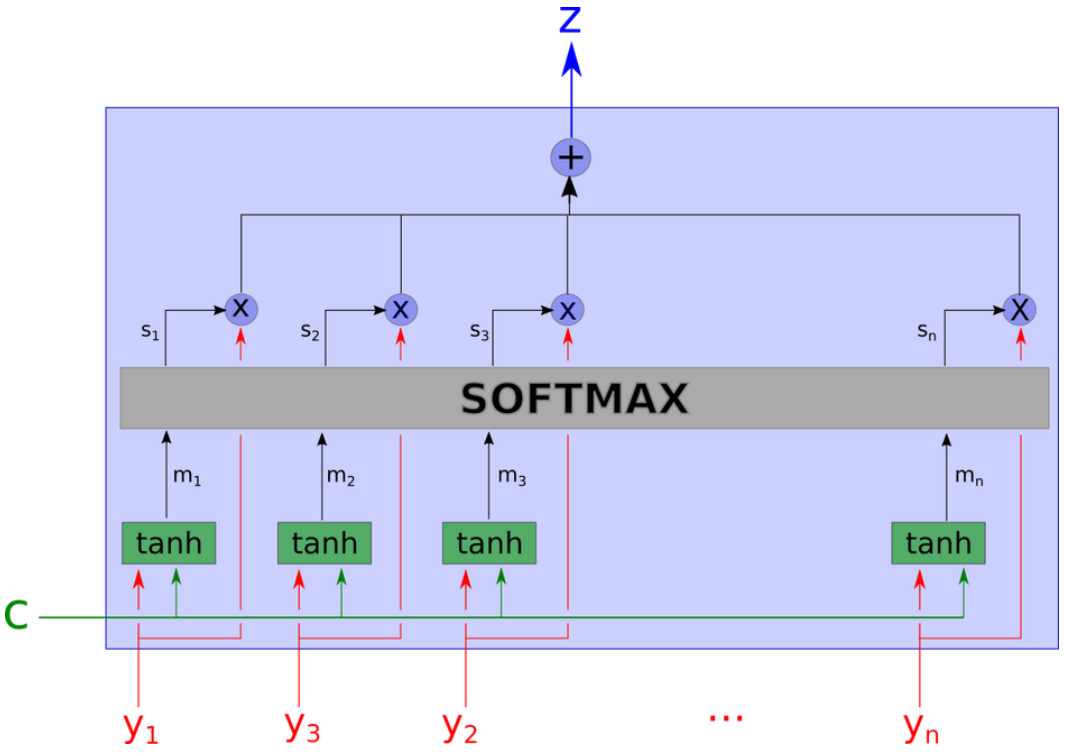


Fig. 4. Visual description of a soft attention model

We apply a tange activation so that very high values tend to have very small differences and be close to 1 and very low values also have very small differences and be close to -1. This leads to a much smoother choice of regions of interest (ROIs) within each sub region. It is more fine grained.

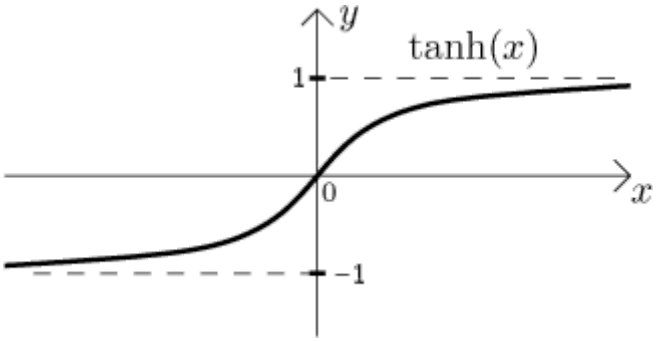


Fig. 5. $\tanh(x)$ function

These resulting m_i are then passed through a softmax function which outputs them as probabilities s.

$$s_i = \frac{e^{m_i}}{\sum_n e^{m_n}}, \; s_i \in [0,1], \; \sum_n s_i = 1 \tag{4}$$

Finally in soft attention model, we take an inner product of s_i and y_i to get the final output z of relevant regions of the entire image. The probabilities s_i correspond to the relevance of the sub regions y_i given the context C.

$$z = \sum_n s_i y_i \tag{6}$$

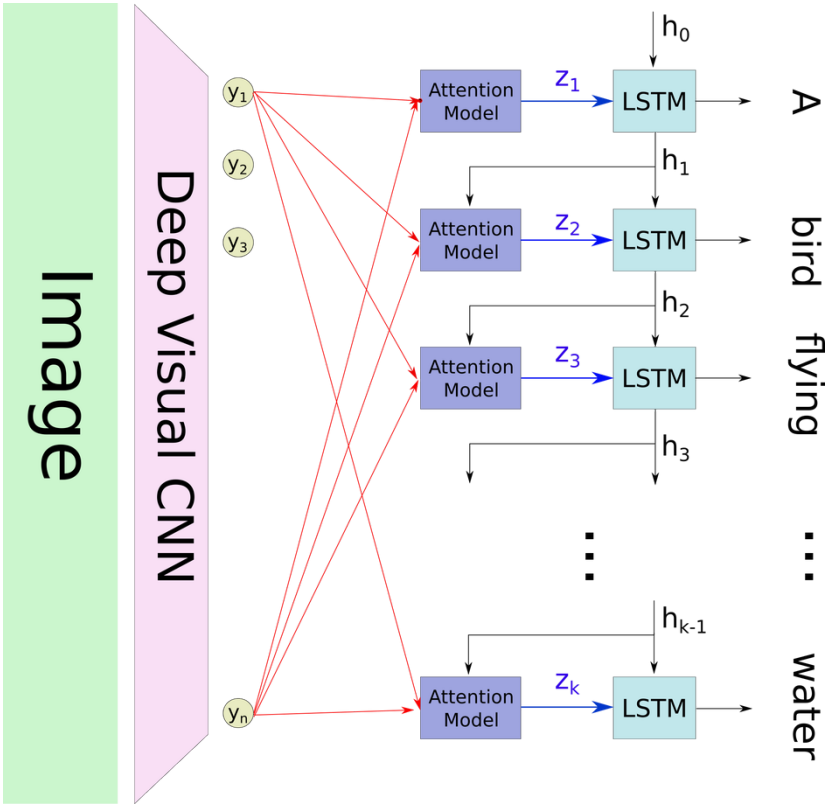
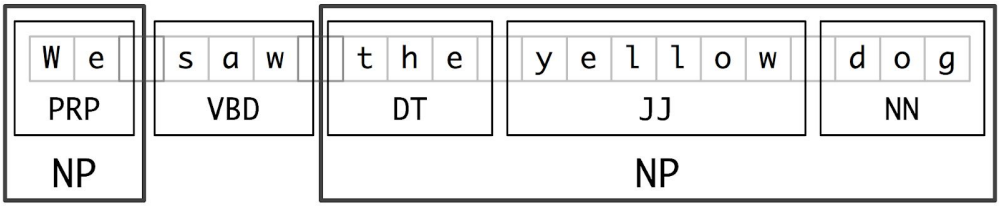


Fig. 6. Visual attention model for image captioning

2.2. Hashtag Extraction

The hashtags are extracted from the captions by means of a package *phrasemachine* [3], which generates different sets of useful phrases, be it noun, verbs, adverbs, adjectives, etc by trimming down the captions.

The *phrasemachine* package itself uses the NLTK package and a brief explanation for the grammar set defined for phrase extraction could be found below:



Fig, 7. An examples of phrase extraction from sentences for the purpose of trimming down the captions to generate hashtags.

3. Results

The model was trained on COCO train2014 dataset using Tensorflow for the purpose of image captioning and the BLEU score achieved outperforms the state-of-the-art Theano implementation of the visual attention model [1].

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
BRNN (Karpathy & Li, 2014) [6]	64.2	45.1	30.4	20.3
Google NIC	66.6	46.1	32.9	24.6
Log Bilinear	70.8	48.9	34.4	24.3
Soft-Attention (Theano)	70.7	48.9	34.4.	24.3
Hard-Attention (Theano)	**71.8**	50.4	35.7	25.0
Soft-Attention (Tensorflow)	**70.3**	**53.6**	**39.8**	**29.5**

Table 1. BLEU-1,2,3,4 metrics of Tensorflow implementation of soft visual attention for image captioning compared to other image captioning models

Though due to time shortage we were not able to evaluate the hashtag generation model but however defining an appropriate metric for hashtags and then evaluating it also poses as another task that will require a sincere effort.

Also the hashtags generated by our model are not personalized enough owing to the MS COCO dataset which does not consider the personalized tags and captioning of the images.

4. Discussion

As inferred by the image captioning model defined in '*Show, Attend and Tell* [1]', visual attention makes the image translation to text much more intuitive and natural as can be concluded by the BLEU score. Consequently, extracting hashtags from the captions would also yield better results but however training the dataset over personalized images like Instagram PIC would definitely help generate more natural results.

Improvements in the mode of extraction of hashtags from the captions could also help improve the results obtained since presently the hashtags are only a subset of the captions and not an independent text produced on interpretation of the image to captions.

5. Conclusion

We propose a Hashtag Recommendation System which could be used over Instagram and other photo-sharing social media platforms. Being based on visual attention model it's suitable for image translation to text providing state-of-the-art quality in image-captioning. However, the quality and personalization of the hashtags could be improved by training the model over Instagram PIC (Personalized Image Captioning) dataset.

We hope that further improvements could be made in the technique of hashtag extraction form captions and in turn the image. We also expect to take into consideration the trending tags for recommendation to generate better and practical hashtags for usage.

Acknowledgements

This research has been financially supported by the Ministry of Education, Taiwan, Republic of China, through the Advanced Institute of Manufacturing with High-tech Innovations (AIM-HI) program. The idea and initiative of this work belongs to Prof. Yuling Hsueh, director of the Data Management + Laboratory of National Chung Cheng University, Taiwan ROC. The author would also like to thank Chia Chun Lin (Jim Lin) for the assistance provided throughout the research period.

Appendix

A. Visualizations from 'soft' attention model [1]

Fig. A.1. A woman is throwing a frisbee in a park.

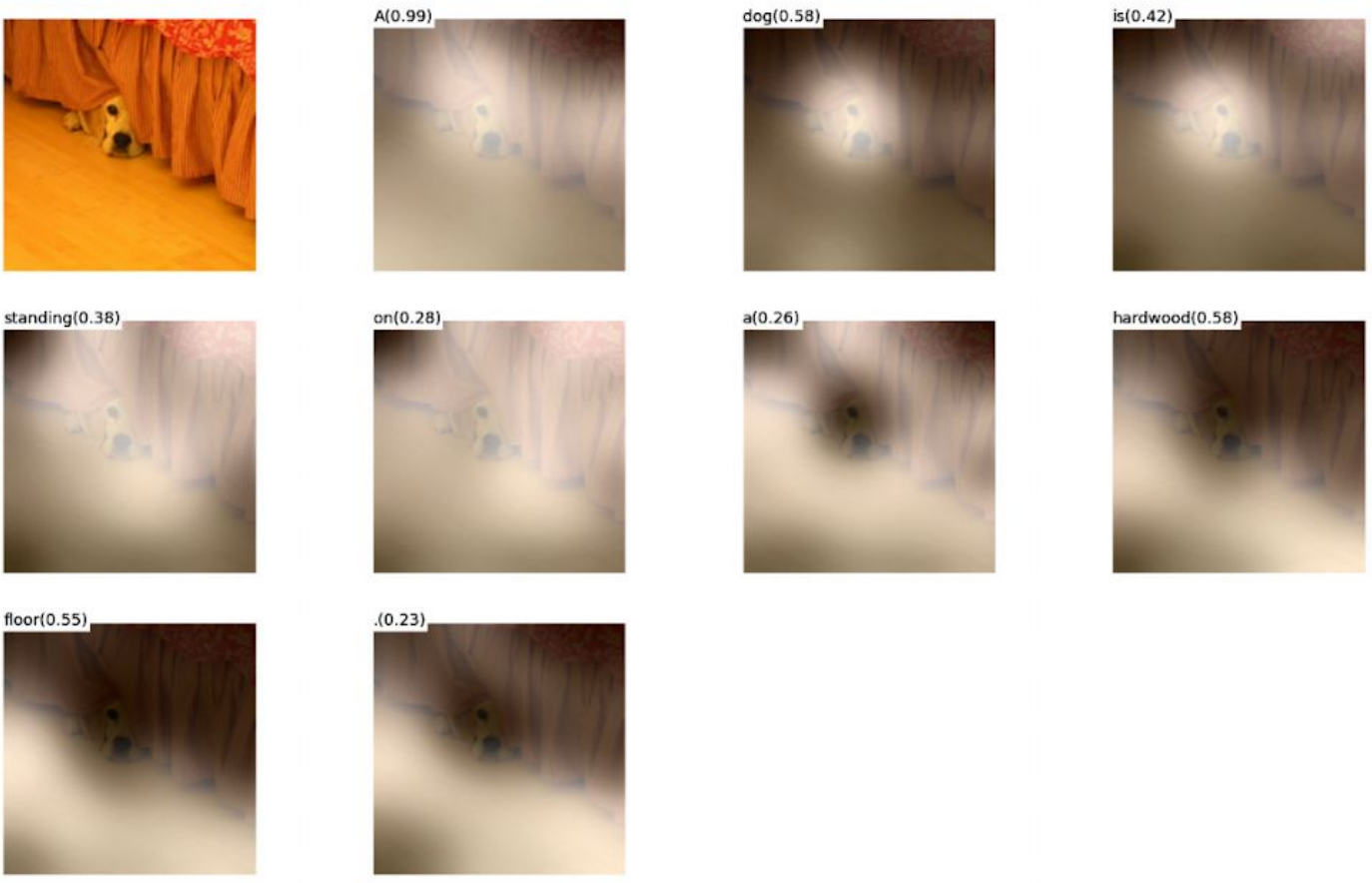


Fig. A.2. A dog is standing on a hardwood floor.

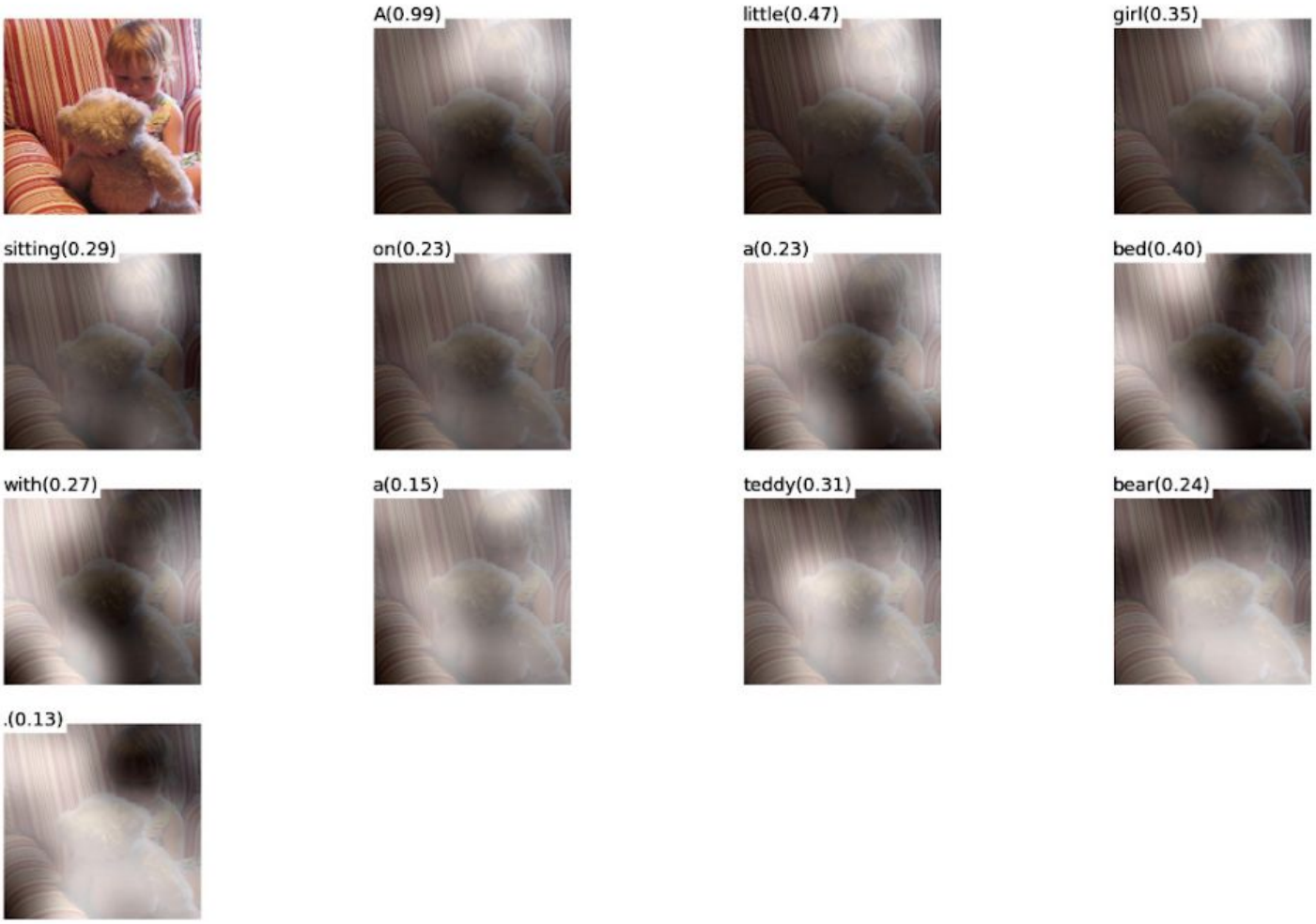


Fig. A.3. A little girl sitting on a bed with a teddy bear.

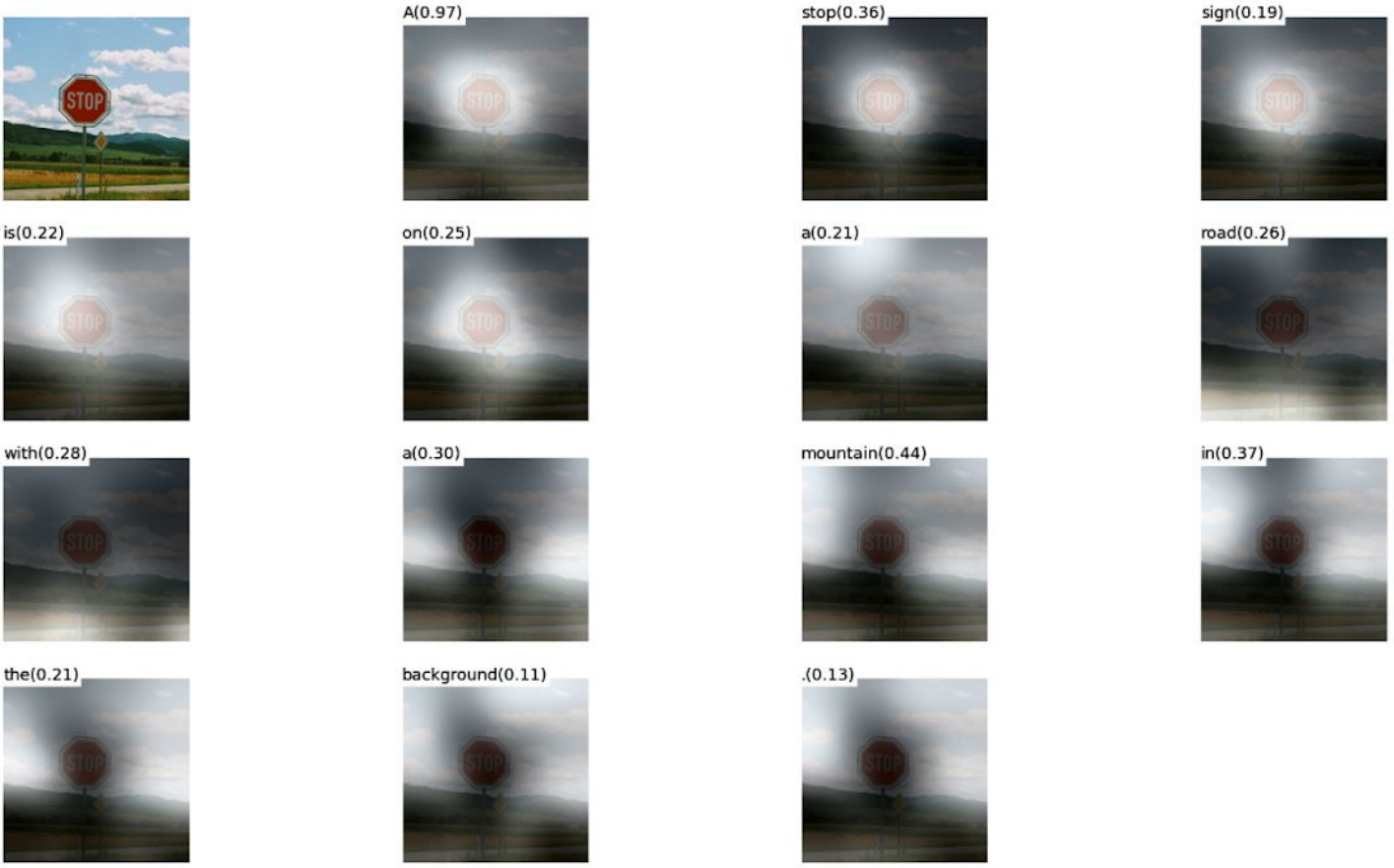


Fig. A.4. A stop sign is on a road with a mountain in the background.

B. Examples of wrong caption predictions by 'soft' attention

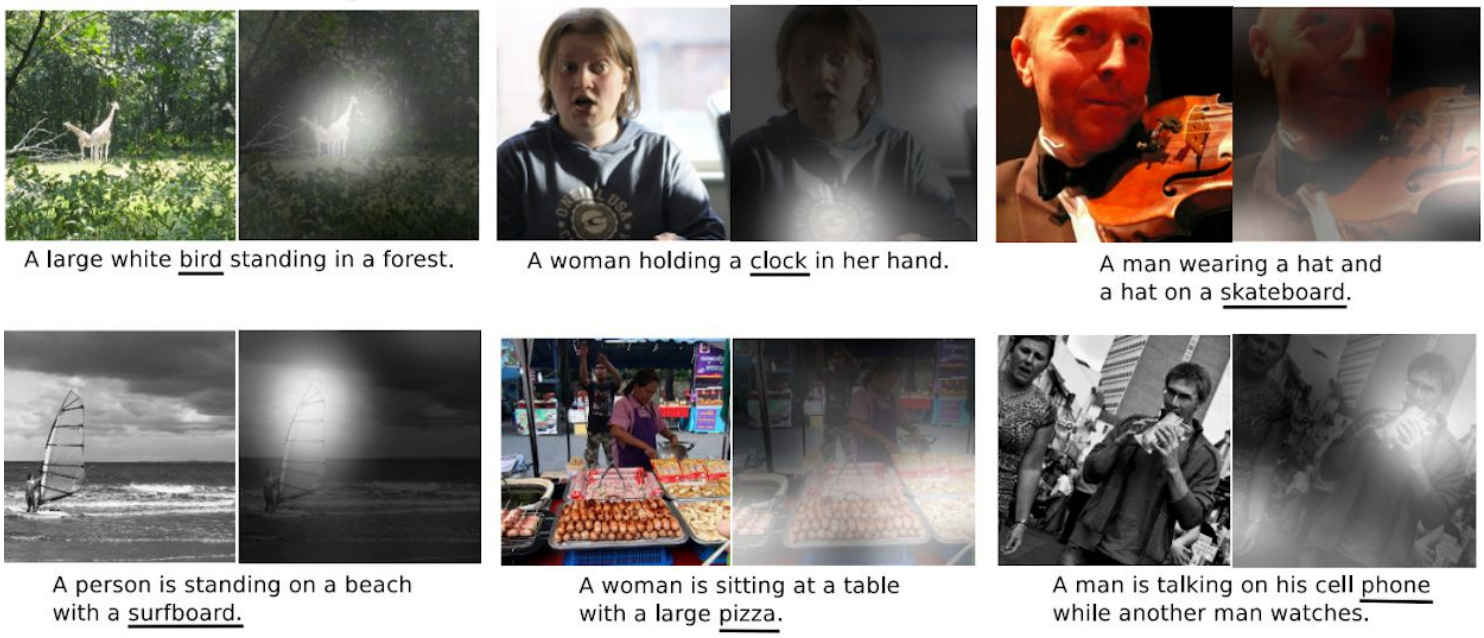


Fig. B.1. An example from '*Show, Attend and Tell* [1]'

Fig. B.2. A wrong prediction by our tensorflow implementation

C. Inaccurate and inefficient hashtags generated

Fig. C. Inaccurate hashtag recommendation for one of my pictures

D. Tree representation of *phrasemachine* implementation

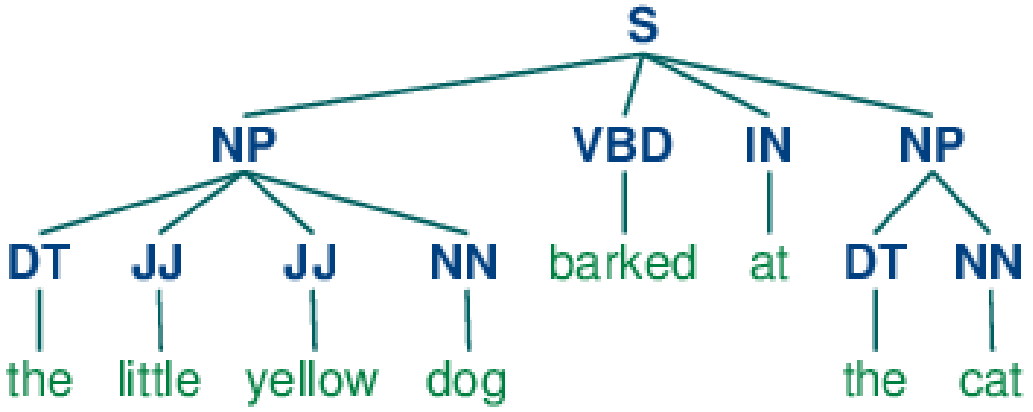


Fig. D. Tree representation for phrase extraction for the sentence - "the little yellow dog barked at the cat"

References

[1] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention*, ICML, 2015 (This is a journal reference)

[2] DeepRNN, *image_captioning*, Tensorflow implementation of "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention". (This is a code reference)

[3] A. Handler, M. Denny, H. Wallach, B. O'Connor, *Bag of What? Simple Noun Phrase Extraction for Text Analysis*, 2016 (This is a journal reference)

[4] C. Park, B. Kim, G. Kim, *Attend to You: Personalized Image Captioning with Context Sequence Memory Networks*, CVPR, 2017 (This is a journal reference)

[5] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, *Show and Tell*, 2015 (this is a journal reference)

[6] Karpathy, Andrej, Li, Fei-Fei, *Deep visual-semantic alignments for generating image descriptions*, 2014 (this is a journal reference)