IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-5, Issue-10 E-ISSN: 2347-2693

# Discriminatory Image Caption Generation Based on Recurrent Neural Networks and Ranking Objective

**Geetika[1*], Tulsi Jain[2]**

[1*] Dept. of CSE, National Institute of Technology, Kurukshetra, India
[2]Dept. of CSE, Indian Institute of Technology, Delhi, India

**Corresponding Author: geetika.jain220694@gmail.com*

**Available online at: www.ijcseonline.org**

Received: 15/Sep/2017, Revised: 28/Sep/2017, Accepted: 19/Oct/2017, Published: 30/Oct/2017

***Abstract*—** This paper proposes a novel approach for image caption generation. Being able to describe the content of an image in natural language sentences is a challenging task, but it could have great impact because great amount of resources is required to meet the demands of vast availability of image dataset. The growing importance of image captioning is commensurate with requirement of image based searching, image understanding for visual impaired person etc. In this paper, we develop a model based on deep recurrent neural network that generates brief statement to describe the given image. Our models use a convolutional neural network (CNN) to extract features from an image. We used ranking objective to pay attention to subtle difference between the similar images to generate discriminatory captions. MS COCO dataset is used, nearly half of the dataset for training and one fourth of dataset for each validation and testing. For every image five captions are provided to train the model Our model consistently outperforms other models with on ranking objective. We evaluated our model based on BLEU, METEOR and CIDEr scores.

***Keywords*—** Visual Geometry Group, Long Short Term Memory, Ranking Objective, Image Captioning

## I. INTRODUCTION

Image captioning is one of the cardinal goals of computer vision. Despite many challenges, this is very active research area. Not only focusing on just image classification or object recognition tasks but a caption should express how these objects are relate to each other. A language model is also needed as captain has to be express in natural language like English. This is a difficult problem because of various reason primarily, model needs to mimic what human being does when they see the images. As human beings learn by viewing enormous amount of graphics content in day to day life, we carry out the similar approach by training machine with millions of images with annotated image caption generation. Recent performance by neural networks and availability of big GPUs were conducive for many researchers to implement many novel techniques. It leads to increase in performance of image captioning generator.

Many approaches have been developed to address the problem statement. In widely used approach, single vector feature is extracted using Convolutional Neural Networks (CNN) like VGG-16 Network and Google Inception Model. Furthermore, Recurrent Neural Networks (RNN) is used to generate automatic image caption. Both networks are trained end-to-end as a single joint model. In these models, log-likelihood of the target image caption is maximized of the training dataset. But it does not pay attention to subtle difference between the images hence generate general captions. Figure 1 shows same captions is generated by models even though subtle difference is present in images. The model is able to describe the scene in a very general way (i.e. "A moving bus on the road"), it is not able to capture subtle difference between the images i.e. in the last image caption can be "People are waiting for a bus at the stand". In this paper, we focus to develop a model that can take subtle difference present in the images into consideration and for similar images is able to generate discriminatory captions. We are motivated by the [1], we propose a novel which incorporated ranking objective. In which, misaligned image-sentence pairs are supported to have a lower score than aligned pairs by a margin.

Section I starts with the important of image captioning in computer vision and provides brief introduction of approach for image captioning generator. Section II contains the related work on image captioning Moving forward, Section III provides details to the baseline approaches and continues developing methodology of revised versions of Recurrent Neural Network by incorporating Ranking Objective part. Section IV describes the Experiments performed and discusses results achieved by the proposed model. Finally,


*Figure 1. Example of repetitive captions for four different images – a moving bus on the road.*

Section V concludes research work with future directions.

## II. RELATED WORK

We discuss the related work done in the field of image captioning. Recently, many approaches have been adopted for automatic image caption generation. For machine translation use of sequence-to-sequence training with neural networks is very successful. This inspired many approaches for image captioning. Because translating an image to a sentence is analogous with encoder-decoder framework of machine translation [2]. Two approaches have been primarily used: 1) bottom-up and 2) top down. In the bottom-up approach, items are observed independently in an image followed by combining of the item to identified into a caption. Due recent advances in statistical machine translation state-of-the-art models achieved by top-down approach. In this approach, a semantic representation of an image is created then decoded into a caption using deep learning model, such as recurrent neural networks.

Kiros et al. [3] proposed the first approach to use neural networks for caption generation and used a multimodal log-bilinear model that was biased by features from the image. In the expanded approach natural way of doing both generation and ranking was allowed explicitly [5]. A similar approach was used by Mao et al. [4] but recurrent neural model was used instead of feedforward neural language model. Vinyals et al. [6] used long short-term memory (LSTM), which is based on the recurrent neural network. As the name suggests, LSTM is good in retaining memory. All these models contain some variations from each other. For example, in [3] image is shown to the model at each time step of the output word but with Vinyals et al. [6] image is shown at the beginning.

In most of these approaches single feature vector is obtained from the pre-trained convolutional network. Karpathy & Li [1] instead proposed a different approach. This focuses on learning a joint embedding space for generation and ranking. The model learns to score similarity between sentence and image as a function of R-CNN object detections with outputs of a bidirectional RNN. By incorporating object detections, a three-step process is proposed for image captioning by Fang et al. Based on a multi-instance learning framework models first learn object detection then to the detected areas a model trained on captions is applied, afterwards rescore is assigned from an image-text embedding space.

Two main methodologies were widely used for image captioning before neural networks. The first approach is based on object detection and attributes discovery. With the help of these results caption templates is generated, which were filled in. In the second approach, from a big database images with similar captioned were retrieved. These captions were modified to fit the query. Intermediate "generalization" step was involved in both of these approaches. This step is incorporated to eliminate the specifics parts that are only relevant to the retrieved image, such as the name of a city. Due to the success of dominant neural network methods and availability of big computational machines both of these approaches have fallen out.

## III. METHODOLOGY

The model we used to automatically determine the short caption of an image is deep recurrent architecture. Model is consists of two units: Convolutional Neural Network(CNN), used to extract the image feature vector, which was already pre-trained on ImageNet [7]. Recurrent Neural Network (RNN), used as language model to determine the caption of an image in English in which output words at time-step *(t-1)* will be input at time-step *t* along with CNN extracted image features.

**Image Feature Extraction Using CNN**

We are using CNN for image feature extraction. CNN has been widely used to analyze visual imagery, image classifier, object detection. For all images, we extracted the features using VGG-16 [8] network, pre-trained by Oxford's renowned *Visual Geometry Group* (VGG), which achieved great performance on the ImageNet dataset. As a result, we get 4096-D feature vector, with the help of Principal Component Analysis we reduced this feature vector to K-dimensional vector, where K is the word embedding size, which will feed as the input to language or LSTM [9] model.

**Caption Generator Using LSTM**

We are using Long Short-Term Memory (LSTM) model for sentence generation as our language model. As special type of Recurrent Neural Network (RNN), the activation function is the *identity* function so, the back propagated gradient neither vanishes nor explodes when passing through, but remain constant. This is the most common shortcoming of Vanilla RNNs. LSTM has memory unit that allows network to learn when to update hidden states and when to forget the previous hidden states over time while supplying the new inputs.

Like vanilla RNN at each time-step, we have an input $x_t \in \mathbb{R}^D$ and the previous step hidden state $h_{t-1} \in \mathbb{R}^{\mathrm{H}}$, as LSTM architecture has H-dimensional memory cell, so we have previous step cell state $c_{t-1} \in \mathbb{R}^{\mathrm{H}}$. Along with this, LSTM learnable parameters are a hidden-to-hidden matrix $W_h \in \mathbb{R}^{4HxH}$, an input-to-hidden matrix $W_x \in \mathbb{R}^{4HxD}$ and a bias vector $b \in \mathbb{R}^{4H}$. We compute an activation vector $a \in \mathbb{R}^{4H}$ at each time step using following equation

$$a = W_x x_t + W_h h_{t-1} + b \ldots (1)$$

After this, we divide activation vector into four vectors $a_i$, $a_f$, $a_o$, $a_g$ where $a_i$ has first H elements of $a$, $a_f$ has next H elements of $a$ and so on. Now we evaluate the forget gate $f \in \mathbb{R}^H$ which controls whether to forgot the current cell , input gate $g \in \mathbb{R}^H$, if it should read its input $i \in \mathbb{R}^H$, and output gate $o \in \mathbb{R}^H$ as

$$i = \sigma(a_i) \ldots (2)$$
$$f = \sigma(a_f) \ldots (3)$$
$$o = \sigma(a_o) \ldots (4)$$
$$g = \tanh(a_g) \ldots (5)$$

Where tanh is hyperbolic tangent, and $\sigma$ is sigmoid function; both operations are applied element-wise. Lastly, we have to compute the next hidden state $h_t$ and next cell state $c_t$ as

$$c_t = f \circ c_{t-1} + i \circ g \ldots (6)$$
$$h_t = o \circ \tanh(c_t) \ldots (7)$$

Where $\circ$ is Hadamard product of vectors and $h_t$ represents hidden state at any time *t.* Image feature vector $\mathrm{I}_t$ will be the first hidden state to LSTM along with a series of input vectors $(x_1, \ldots, x_D)$. At each time-step it outputs a series of log probabilities:

$$\mathcal{Y} = \{\mathcal{Y}_1, \mathcal{Y}_2, \ldots, \mathcal{Y}_D\}, \mathcal{Y}_i \in \mathbb{R}^M \ldots (8)$$

Where D is the length of sentence and M is the size of vocabulary.

**Ranking Objective**

N image-sentence pairs have been passed to our model at each forward pass. To find the similarity between the $i$-th image and $j$-th sentence we use the dot product $I_i^T S_j$. As to ensure that the generated caption is uniquely in accordance with an image, we take $I_i^T S_j$ to be larger than $I_i^T S_j$, where $i \neq j$, by some margins, thus we able to add discriminatory power to our model. A batch of 32 images features $\mathrm{I} \in \mathbb{R}^{N \times K}$ has been passed to ranking model along with log probabilities after transforming log probabilities to probabilities:

$$P = \exp(\mathcal{Y}) \in \mathbb{R}^{D \times N \times M} \ldots (9)$$

In procedure to compute the word embedding, we have used these computes probabilities as soft indices to corresponds into same word embedding table as used the same way in the language model. And another LSTM is used for learning sentence embedding

$$S = \{s_1, s_2, \ldots, s_N\}, s_i \in \mathbb{R}^{N \times K} \ldots (10)$$

where word embedding has been passed to LSTM at each time-step, and the output corresponds to the last time-step is represented as sentence-embedding. Similarity matrix computation has been computed as follows:

$$Sim(\mathrm{I}, S) = S \cdot I^T \in \mathbb{R}^{N \times N} \ldots (11)$$

Ranking objective for one batch can be defined as the summation of max-margin loss of both rows and columns:

$$J\big(Sim(\mathrm{I}, S)\big) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} \max(0, Sim[i,j] - Sim[i,i] + 1) + \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{N} \max(0, Sim[i,j] - Sim[i,i] + 1) \ldots (12)$$

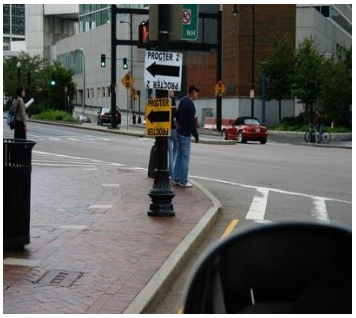
For next word prediction, the language model is trained to combine the previous hidden state $(h_{t-1})$ and word embedding $(x_t)$. Initial hidden state vector $h_0$ has been set to the image feature vector and initial word embedding $x_1$ has been set to special token. The cost function can be defined as to minimize the value of negative log probability such as :

$$L(\mathrm{I}, Y) = -\frac{1}{N}\sum_{i=1}^{N} y_i \ldots (13)$$

And total loss during training can be defined as sum of the softmax loss and ranking objective:



| | | | |
|---|---|---|---|
| **Baseline** a man is sitting on a chair | **Baseline** a dog is jumping in a park | **Baseline** two people standing on road | **Baseline** a person is crossing road |
| **Our Model** a man is sitting on a chair with laptop | **Our Model** a dog is playing with frisbee | **Our Model** a man and a women is standing | **Baseline** a car and a person standing on the road |

*Figure 2. Comparison of baseline model and our model with Qualitative results where text in red shows error and text in green shows discriminatory captions.*

$$Loss = w_j J\big(Sim(\mathrm{I}, S)\big) + w_L L(\mathrm{I}, Y) \dots (14)$$

## IV. EXPERIMENT AND RESULTS

We used Microsoft COCO [10] dataset, an image dataset mainly designed for image captioning, object detection and segmentation, for training and testing our model. The dataset is consists of nearly 82,000 training images, 40504 validation images where each image has 5 written caption descriptions and nearly 40,000 testing images. Descriptions words that occur less than 5 times are mapped to special token <UNK>.

**Evaluation Metrics**

Human evaluation is most reliable and efficient metric for image captioning, which may take few months to evaluate and efficient is a major concern because it involves human labor that can not be reused. In this experiment, we are using the several metrics to evaluate the effectiveness of the model. BLEU Score (Bilingual Evaluation Understudy)[6] is the most common metric to evaluate the performance. It can be computed by counting the matches between the n-gram candidate translation and n-grams of reference translations. METEOR is advanced and designed to fix few problems found in BLEU, and also produces a good correlation between human judgment at segment and sentence level. Besides BLEU and METEOR, we also use CIDEr (Consensus-based Image Description Evaluation), is a popular metric for evaluating the quality of descriptions. All three metrics (BLEU, METEOR, CIDEr) follows the same relations as higher the score better the candidate caption is.

**Baseline Model**

We use NeuralTalk2 model as baseline model. It uses the Torch library which has the same language and image model as of ours. Single vector features are extracted using pre-trained VGG-16 network. A 512-dimensional vector is used for both word embedding and LSTM hidden state of language model. The Initial learning rate is set to $4 \times 10^{-4}$, which decreases after every 40,000 iterations. When gradient exceeds 0.1, we use a 0.5 dropout. The batch size is 32 for both image and language model. Adam optimizer is used with alpha and beta are set to 0.8 and 0.999 respectively.

**Experiment Model and Results**

Keeping in mind computational cost we train both image and language model for 20 epochs. While training model same set of hyper-parameters are used and for ranking model, RMSProp optimizer with alpha is set to 0.8 and a learning rate of $1 \times 10^{-5}$. We initialize the weight $w_j$ to $1 \times 10^{-6}$ for ranking loss, and doubles $w_j$ after every 5000 iterations. Intuitively, mostly random captions have been generated at initial stages. Ranking loss has been enforced more strongly by making $w_j$ larger when model started outputting sensible image captions.

To prove ranking model effectiveness, we train baseline model, which is without ranking loss, and our model using the same set of hyper-parameters for 20 epochs. Using the different sets of hyper-parameters, we cross-validate these models and observed that our model outperforms the baseline model consistently. Due to the limitations of computational resources, we observed that the validation and loss scores have not fully converged hence accuracy can be further improved if training is done for more number of epochs.

**Quantitative Results**

Lack of discriminatory power in evaluation metrics led to the failure of most of the existing models to capture the subtle differences between similar images. So, we do not expect a significant boost in validation scores on these metrics. Visualization of results is mentioned in the following graph:


*Figure 3. Quantitative Results of validation data for BLEU, METEOR, CIDEr metrics*

Figure 3 and Table 1 shows BLEU/CIDEr/METEOR scores on validation data. In particular, there is a 9% increase in CIDEr score, which shows that including ranking model does not help in generating more discriminatory captions, but also helps to increase the overall performance of the experiment.

| Model | BLEU | METEOR | CIDEr |
|---|---|---|---|
| **Karpathy et al. [12]** | 62.5 % | 19.5% | 60.0 % |
| **CHEN and Zitnick [13]** | - | 20.4% | - |
| **LRCN [14]** | 62.8 % | - | - |
| **Our LSTM Model** | 62.5 % | 19.4% | 65.8 % |

*Table 1. Comparison of various model for test images of the Microsoft COCO dataset.*

**Qualitative Results**

As seen in the figures 2, our model generates more sensible, descriptive and differentiable captions as compared to caption generated by the baseline model. As can be visualized, our model able to interpret visual-semantic correspondences, even for small objects such as phones and is able to determine the less salient objects such as "laptop", "frisbee" and "car" as shown in figure 2.

## V. CONCLUSION AND FUTURE SCOPE

A thorough qualitative and quantitative analysis of the metrics shows that model is able to generate more sensible and descriptive captions. Discriminatory power is added to the model by incorporating ranking objective loss. However, our model does not show significant improvement quantitatively. Better results could be achieved by adding more LSTM hidden layers and running model for more epochs. For better results, instead of grouping random images in a batch, we can group similar images. This configuration makes the ranking objective to perform more effective because there is no further need to push down the misaligned image-sentence pair if all the images are different. Image captioning has many applications in market value, public opinion etc. This work can be extended towards captioning video directly to sentence.

## VI. REFERENCES

[1] 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.

[2] Kyunghyun Cho, Bart Van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "*Learning phrase representations using rnn encoderdecoder for statistical machine translation*", arXiv preprint arXiv:1406.1078, 2014.

[3] Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel, "*Multimodal neural language models*". In Proceedings of the 31st International Conference on Machine Learning (ICML-14), pages 595–603, 2014

[4] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille, "*Deep captioning with multimodal recurrent neural networks (m-rnn)*", arXiv preprint arXiv:1412.6632, 2014.

[5] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel, "*Unifying visual-semantic embeddings with multimodal neural language models*", arXiv preprint arXiv:1411.2539, 2014.

[6] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, "*Show and tell: A neural image caption generator*". In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3156–3164, 2015.

[7] Jia Deng,Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "*Imagenet: A large-scale hierarchical image database*". In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 248–255. IEEE, 2009.

[8] Karen Simonyan and Andrew Zisserman, "*Very deep convolutional networks for large-scale image recognition*", arXiv preprint arXiv:1409.1556, 2014.

[9] Hochreiter, Sepp, and Jrgen Schmidhuber, "*Long Short-Term Memory*", Neural Computation 9.8 (1997): 1735-780. Web. 23 Apr. 2016

[10] Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollr, and C. Lawrence Zitnick, "*Microsoft COCO: Common Objects in Context*" Computer Vision ECCV 2014 Lecture Notes in Computer Science (2014): 740-55. Web. 27 May 2016

[11] Papineni, Kishore, Salim Roukos, ToddWard, Wei-Jing Zhu, *Bleu: a method for automatic evaluation of machine translation*" Proceedings of the 40th Annual Meeting on Association for

Computation Linguistics (ACL): 311-318 (2002). Web. 24 May 2016

[12] Karpathy, Andrej, and Li Fei-Fei, "*Deep Visual-semantic Alignments for Generating Image Descriptions*" 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015). Web. 29 May 2016

[13] Chen, Xinlei and C. Lawrence Zitnick, "*Learning a Recurrent Visual Representation for Image Caption Generation*", CoRR abs/1411.5654 (2014). Web. 19 May 2016

[14] Donahue, Jeff, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Trevor Darrell, and Kate Saenko, "*Long-term Recurrent Convolutional Networks for Visual Recognition and Description*", 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015). Web. 20 Apr. 2016

## Authors Profile

*Ms. Geetika* pursued Computer Science Engineering from National Institiue of Technology Kurukshetra from 2011-2015. She is currently working in the field of Applied Machine Learning, specifically Natural Language Processing and Computer Vision.

*Mr. Tulsi* Jain received his Bachelor of Technology degree from Indian Institute of Technology Delhi in the year 2015. After graduation, he joined Oracle Corporation as an application developer. At present, he is pursuing research in the field of Artificial Intelligence, Natural Language Processing and Computer Vision.