ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

Generating Captions for Images Using Multimodal Neural Networks

Dhoomil Sheta[1], Parth Parekh[1], Nishant Shah[1], Rutvik Parekh[1], Manan Shah[1], Prof S. C. Shrawne[2]

B.Tech Student, Department of Computer Engineering, VJTI, Mumbai, India[1]

Assistant Professor, Department of Computer Engineering, VJTI, Mumbai, India[2]

ABSTRACT: Automatically generating captions for images has always been an area of research in Artificial Intelligence. We propose a model to train the image representation to generate captions for the images using Recurrent Neural Networks. Image representations are extracted using highly efficient Deep Residual Network (ResNet-50)[0]. We also present an extension to traditional LSTM that improves performance of caption generation. We use the dataset Flickr8k and validate the performance using widely accepted metrics such as BLEU, CIDEr, METEOR.

I. INTRODUCTION

Automatically generating captions for an image is a challenging task, and one of the primary goals of Computer Vision. The problem is particularly difficult because it requires correctly recognizing different objects in images and how they interact with each other. Another challenge is that an image description generator needs to express these interactions in a natural language. Therefore, a different language model is required in addition to extracting features from the image. Hence, we need to use a multimodal structure.

Recently, this problem has been studied by many different authors. Despite the challenging nature of this task, there has been considerable work in this field due to improvement in machine learning techniques. Neural networks create a model the same way in which a human brain perceives an image and generates a caption for the same. Many well-known companies, such as Microsoft and Google, have published models recently that attempt to solve this problem.

The main inspiration for our work comes from advancement in research in neural network models. For many years, language modelling was achieved by translating words individually, aligning them and reordering etc., but now this can be achieved easily using Recurrent Neural Networks(RNN).

We use a multimodal neural network structure, using Convolutional Neural Network (CNN) and Recurrent Neural Networks(RNN) in this project. Over the last few years, it has been proved that CNN is the state of the art model for providing a rich vectorial representation of an input image. Therefore, we use CNN for extracting features from an input image and use RNN to create a description of the image in a particular natural language.

We propose the following solutions:

- Modifying LSTM to create f-LSTM, where each step in LSTM is provided with information about the image, also known as focus.
- Creating a hybrid model with 2 different variations of LSTM, where one is standard LSTM and the second is f-LSTM, and the output from each instance is compared and the best one is selected.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

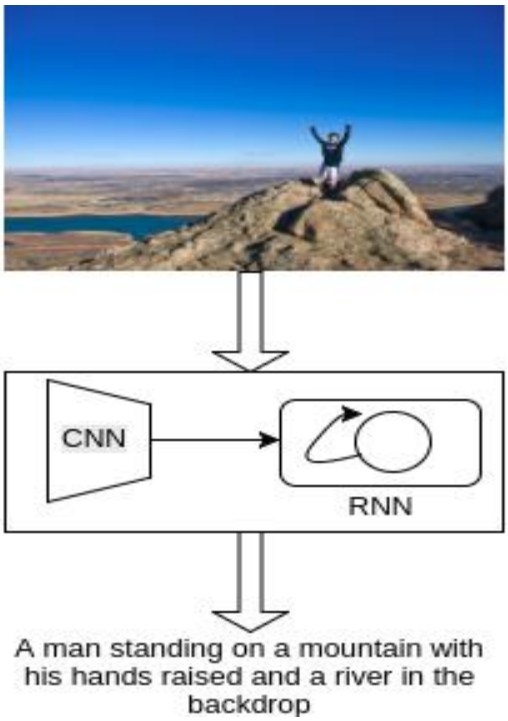

Fig1. Generation of sentence from image using the proposed model

II. RELATED WORK

In this section we highlight the efforts taken by various researchers in the domain of image captioning. A large number of other works have shared the similar higher level goal. Earlier papers started with smaller tasks of classifying images into a fixed number of classes. A classical approach was to treat it as a retrieval problem where the test image would be simply described using annotations that was ranked highest [1],[2],[3]. Certain ranking models use the idea of co-embedding the images and texts in the same vector space, thus retrieving text that is close to the image in the vector space [1].

Later works started with holistic scene approach that classifies the overall scene, recognizes and segments the object components [4].

The first approach to use neural networks for caption generation was Kiros et al.[5], who proposed a multimodal Log Bilinear model. Further works in this field replaced a feed-forward neural language model with a recurrent one. These works have some common key structures, and accordingly, we call those structures simply as dominated model or general model. In details, two of major parts, i.e., the CNN and RNN, play core roles in general model respectively.

Especially, for Flickr 8k dataset, Mao et al.[6] present a multimodal Recurrent Neural Network (m-RNN) model which contains a VGG-net CNN and a vanilla RNN. Besides, Vinyals et al.[7] use LSTM instead of other RNNs in their model and unlike [4], [5] wisely show the image to RNNs at the beginning, leading to performance improvement. Vinyals et al.

All of these works represent images as a single feature vector from the top layer of a pre-trained convolutional network. Karpathy[8] instead proposed to learn a joint embedding space for ranking and generation whose model learns to score sentence and image similarity as a function of R-CNN object detections with outputs of a bidirectional RNN. Show and Tell[7] presents a generative model that uses a similar architecture but uses Inception-v3 model for CNN and vanilla LSTM for the caption generation part. They proposed a system with encoder-decoder to maximize probability of the target description sentence given a training image. Show, Attend and Tell[8] introduced an attention based model that automatically learns to describe the content of images. They introduce two attention-based image caption

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

generators under a common framework: 1) a "soft" deterministic attention mechanism trainable by standard back-propagation methods and 2) a "hard" stochastic attention mechanism trainable by maximizing an approximate variational lower bound or equivalently by REINFORCE

Our model is based on similar multimodal architecture. Other works like Fang et al.[9] propose a model that does not make use of the RNN. Instead it uses a visual detector for words that commonly occur and is trained using multiple instance learning. The model which has been used for achieving our goal is similar in spirit to the log-bilinear model (LBL) (Kiros et al, 2014). Devlin[10] compares two highly accurate methodologies for generating image descriptions. The comparison finds the empirical process of using the penultimate activation layer of CNN as an input to the RNN providing better results over the model which uses a pipelined process where a CNN generates a set of candidate words which are arranged into coherent sentences by maximum entropy (ME) language models.

III. PROPOSED MODEL

This paper presents a model which tries to focus on generating high level descriptions of images using a convolutional neural network to extract features which will then be provided as an input to a recurrent neural network using an improvement to the existing Long Short Term Memory(LSTM) cells.

A. *Feature Extraction:*

Convolutional Neural Networks is a variant of artificial neural networks that uses many identical copies of the same neuron. This allows the network to have lots of neurons and express computationally large models while keeping the number of actual parameters – the values describing how neurons behave – that need to be learned fairly small. In the early days, people have to design features manually. This step is called feature engineering. The greatest advantage of convolutional neural networks is they can learn appropriate features by themselves automatically.

For feature extraction, we use pretrained ResNet-50[11] as provided by Kaiming He. It is common to pre train a CNN model on very large dataset(eg. ImageNet, with 1.2 million images) to extract features for the task of interest.

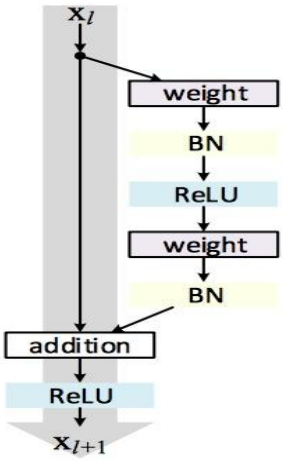

Fig. 2 High level representation of a single residual block[2]

The intuition behind using ResNet-50 for feature extraction came due to their superior performance in ImageNet. The feed forward shortcut connection of identity mapping do not add any extra parameter or computational complexity. VGGNet is very expensive computationally as it uses 140M parameters whereas in ResNet-50, the parameters used are

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

23M. Originally ResNet included 152 layers but we use 50 as it provides a trade-off between rich features and computational complexity.

B. *Caption Generation:*

Traditional Neural Networks lack the ability of persistence. Thus they throw away previous information and start thinking from scratch again. These shortcomings can be overcome by Recurrent Neural Networks that have loops in them, which allows the information to persist. RNNs can be thought of as multiple copies of the same network, each network passing messages to the next one.

Long Short Term Memory[12] or LSTM in short is a variation of RNN which solves the problem of vanishing gradients. They are capable of learning long term dependencies through the use of gates.Vanilla LSTMs use 3 gates, namely the forget gate, input gate and the output gate. These gates are usually composed of a sigmoid neural net layer and a pointwise multiply operation.

$$f_t = \sigma (W_f . [h_{t-1} , x_t] + b_f) \quad (1)$$

$$i_t = \sigma (W_i . [h_{t-1} , x_t] + b_i) \quad (2)$$

$$\check{C}_t = tanh (W_c . [h_{t-1} , x_t] + b_c) \quad (3)$$

$$C_t = f_t * C_{t-1} + i_t * \check{C}_t \quad (4)$$

$$o_t = \sigma (W_o . [h_{t-1} , x_t] + b_o) \quad (5)$$

$$h_t = o_t * tanh (C_t) \quad (6)$$

One problem of LSTM is that each step does not have a reference to the image given in the input. Only the first step receives the image information. As a result of the gates, the information about the images passed on after each step fades away eventually.We try to improve the original LSTM by adding extra focus to the input image at each step. Now, the equations at each changes at each gates including focus is given as:

$$i_t = \sigma(W_{ix} x_t + W_{ih} h_{t-1} + W_{iq} f) \quad (7)$$

$$f_t = \sigma(W_{fx} x_t + W_{fh} h_{t-1} + W_{fq} f) \quad (8)$$

$$o_t = \sigma(W_{ox} x_t + W_{oh} h_{t-1} + W_{oq} f) \quad (9)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot h(W_{cx} x_t + W_{ch} h_{t-1} + W_{cq} f) \quad (10)$$

$$h_t = o_t \odot c_t \quad (11)$$

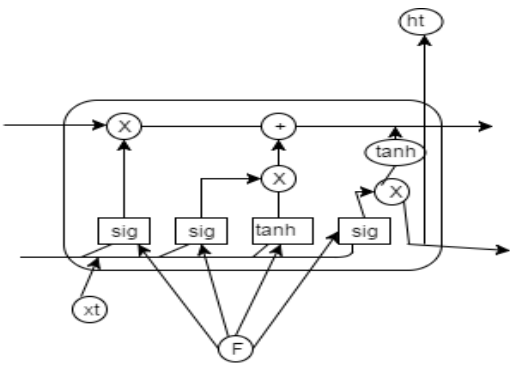

Fig 3.Focussed LSTM

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

Our strategy of altering model divides the hidden layer into two parts and these two parts stay uncorrelated until the output unit. In forward-propagation process, the normal LSTM receives the previous cell state, whereas the focussed LSTM receives the feature matrix as well, and then transmit outputs of RNN unit to y_t with corresponding ratios. We adopt fully recurrent network as example to illustrate our method. The corresponding formulae are shown as follows:

$$h_{1t} = relu(W_{x1}x_t + W_{h1}h_{1t-1} + b_{h1}) \quad (12)$$

$$h_{2t} = relu(W_{x2}x_t + W_{h2}h_{2t-1} + b_{h2})(from\ f\text{-}LSTM) \quad (13)$$

$$y_t = softmax(r_1 W_{d1}h_{1t} + r_2 W_{d2}h_{2t} + b_d) \quad (14)$$

$$dy1 = r_1 \times dy \quad (15)$$

$$dy2 = r_2 \times dy \quad (16)$$

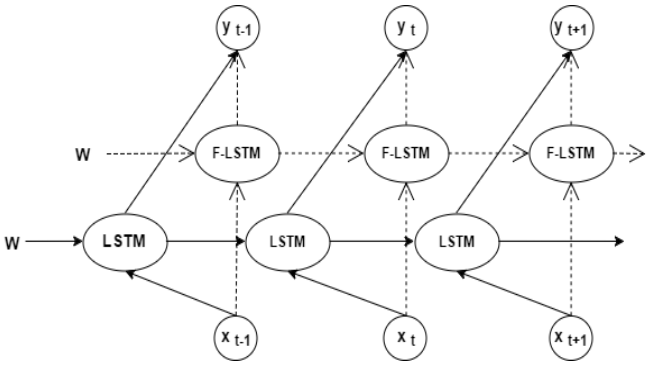

Fig 4. Hybrid LSTM

IV. EXPERIMENTS

Dataset: The Image Captioning dataset we use for our experiments is Flickr8k [0]. The Flickr8k dataset is a famous dataset consisting of a total of 8000 images selected from Flickr.com. These 8000 images are divided into three sets, namely training (6000), validation (1000) and testing(1000). Each image in Flickr8k dataset is described using 5 captions which have been collected from humans.

Evaluation Metrics: We use three of the most widely used measures in machine translation and image captioning literature namely, BLEU[0], CIDEr[0], METEOR[0].

BLEU is a precision-based metric which measures a score that is a precision of word n-grams between the caption generated by our model and the ground truth caption. BLEU is often criticized to favour short sentences The metric has some obvious drawbacks, then too it is widely used and so we consider this metric for evaluation.

Since, no single metric can be perfect report on METEOR and CIDEr. METEOR score is calculated by taking harmonic mean of unigram precision and recall. It gives more importance to word matches than precision. It has three levels of matching namely, exact, stemmed and synonym. It is seen that this metric shows higher correlation with human judgement.

CIDEr stands for Consensus-based Image Description Evaluation. It basically measures 'how likely it is that the generated caption sounds just like a human'. It captures consensus based on similarity of novel captions with the ones already available i.e human generated captions.

All scores are computed using the coco-caption code1.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

Data Preprocessing: We convert all the sentences to lower case and ignore all non-alphanumeric characters in every sentence. We also filter words such that we keep only those which appear at least 5 times in the actual training dataset captions resulting in a vocabulary consisting of 2537 unique words encompassing all the captions of images in the training set.

V. SIMULATION RESULTS

Several related models have been developed in parallel to this work. We include these in the table below for comparison. We improve on some metrics. Compared to these approaches, our model prioritizes simplicity and speed at a slight cost in performance.

Figure 5. Correct Predictions

Figure 6. Partially Correct Predictions

Figure 7. Incorrect Predictions

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

	B@1	B@2	B@3	B@4	CIDEr	METEOR
LogBilinear	65.6	42.4	27.7	17.4	-	17.31
Show and Tell[1]	63	-	-	27.7	0.855	23.7
Aligning Where to See and What to Tell: Image Captioning[2]	63.9	45.9	31.9	21.7	0.538	20.4
Karpathy et al[4]	57.9	38.3	24.5	16.0	0.69	-
Guided LSTM[5]	66.8	46.9	31.6	21.3	0.7	-
Show,Attend and Tell[6]	67	45.7	31.4	21.3	0.893	20.30
Hybrid LSTM	**67.9**	**47.2**	28.0	23.5	0.72	**24.2**

Figure 8. Results

VI. CONCLUSION

In this work, we have proposed an extension of the LSTM model for image caption generation. By adding the additional focussed LSTM and creating a hybrid model for caption generation, we show how the model can stay 'on track' rather than drifting away to unrelated yet common phrases. We successfully infer that our model performs better than similar models based on BLEU-1, BLEU-2 and METEOR scores. We achieved our aim to find a model which provides a reasonable trade-off between speed, complexity and accuracy.

VII. ACKNOWLEDGEMENT

We would like to acknowledge the support provided by our institute, Veermata Jijabai Technological Institute (VJTI), for supporting us and providing GPUs.

REFERENCES

1. M. Hodosh, P. Young and J. Hockenmaier (2013) "Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics", Journal of Artificial Intelligence Research, Volume 47, pages 853-899
2. R Socher et al (2013), Recursive deep learning for natural language processing and computer vision.
3. Srivastava & Salakhutdinov, (2014), Dropout: A Simple Way to Prevent Neural Networks from Overfitting
4. Li Jia Li et al:Towards Total Scene Understanding: Classification, Annotation and Segmentation in an Automatic Framework
5. Ryan Kiros et al (2010), Multimodal Neural Language Models.
6. Mao et al (2014), Multimodal Recurrent Neural Network
7. Vinyals et. al, "Show and Tell: A Neural Image Caption Generator"
8. Andrej Karpathy and Li Fei Fei (2014), Deep Visual-Semantic Alignments for Generating Image Descriptions.
9. Fang et al (2015), Language Models for Image Captioning: The Quirks and What Works.
10. Devlin, J. et al (2015),Exploring Nearest Neighbor Approaches for Image Captioning
11. Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun (2015), Deep Residual Learning for Image Recognition
12. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory
13. Hinton, Coursera, "RMSProp and equilibrated adaptive learning rates for non-convex Optimization"
14. Felix A. Gers, Jürgen Schmidhuber, and Fred Cummins, "Learning to Forget: Continual Prediction with LSTM"
15. Sergey Ioffe,Christian Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift"
16. Christopher Olah, "Understanding LSTMs" , http://colah.github.io/posts/2015-08-Understanding-LSTMs
17. Stanford University : "Convolutional Neural Networks for Visual Recognition"
18. F.Chollet, Keras Library, https://github.com/fchollet/keras
19. Xinlei Chen,C. Lawrence Zitnick, "Mind's Eye: A Recurrent Visual Representation for Image Caption Generation"
20. Google Research Blog, https://research.googleblog.com/2016/09/show-and-tell-image-captioning-open.html
21. PDollar/ImageCaptioning,https://pdollar.wordpress.com/2015/01/21/image-captioning/
22. Junhua Mao,Wei Xu,Yi Yang,Jiang Wang,Alan L. Yuille, "Explain Images with Multimodal Recurrent Neural Networks"
23. Ryan Kiros, Ruslan Salakhutdinov, Richard S. Zemel, "Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models"
24. Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, Trevor Darrell, "Long-term Recurrent Convolutional Networks for Visual Recognition and Description"
25. Xinlei Chen,C. Lawrence Zitnick, "Learning a Recurrent Visual Representation for Image Caption Generation"

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

26. Junqi Jin, Kun Fu, Runpeng Cui, Fei Sha2, Changshui Zhang, "Aligning where to see and what to tell: image caption with region-based attention and scene factorization"
27. Kenneth Tran, Xiaodong He, Lei Zhang, Jian Sun Cornelia Carapcea, Chris Thrasher, Chris Buehler, Chris Sienkiewicz, " Rich Image Captioning in the Wild"
28. Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, Jiebo Luo, "Image Captioning with Semantic Attention"Lin Bai, Kan Li, "Predicting Image Captioning by a unified hierarchical model",2015 IEEE International Conference on Multimedia and Expo (ICME)
29. Ramakrishna Vedantam,C. Lawrence Zitnick,Devi Parikh, "CIDEr: Consensus-based Image Description Evaluation"
30. A. Karpathy, A. Joulin, and L. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping",[2014].
31. Kishore Papineni, Salim Roukos, Todd Ward,Wei-Jing Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation"
32. Satanjeev Banerjee,Alon Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments"