CSE628 Natural Language Processing
Image Captioning Project Report

Kiranmayi Kasarapu
111447596
kkasarapu@cs.stonybrook.edu

Aishwarya Danoji
111493647
adanoji@cs.stonybrook.edu

Rahul Bhansali
111401451
rbhansali@cs.stonybrook.edu

Abstract

In this project report, we summarize three different papers on Image Caption Generation and the techniques that were used in each paper. We will discuss briefly about the methods, improvements over earlier techniques, evaluation metrics that were used. Later, we'll talk about the how the work is related to our project, and what improvements we will make. We will also discuss what are the key things we learned from this project and present ideas for improving as future work.

1 Introduction

Image Captioning is a technique that is used to generate descriptive sentences that describe an image. It is a fundamental problem in artificial intelligence that connects computer vision and natural language processing. It is a challenging task for machine learning algorithms because it means mimicking the exceptional human ability to compress huge amount of salient visual information into descriptive language. Also a single image can have multiple correct captions. It is not just language generation, but the task of finding relationships between the objects in the image, which makes it such a difficult problem. The input to the task is a set of images and the output is a word sequence. The common approach that has been explored in most of the research is CNN + RNN based models. Convolutional Neural Networks are used to encode the image into a fixed length encoding and this encoding is fed as an input to the Recurrent Neural Network which outputs the caption for that image.The most common datasets that are used for this task are Flickr8k (8000 images comprising of 1GB), Flickr30K (31K images comprising of 6GB) and MSCOCO (123K images comprising of 18GB) datasets.The 3 papers we read i.e Show and Tell,Deep Visual-Semantic Alignments for Generating Image Descriptions and Show, Attend and Tell discuss different techniques to generate captions for an image.We improved these models by using a rich dense feature sets of images to generate captions. We also implemented different models and different loss functions for Image Captioning and analysed which gives the best result and why. We started our project by building the simplest image captioner, by using NIC's loss function and combining it with simple RNN for the base model. We then replaced the RNN by BRNN so as to improve the quality of captions generated.We further improved the model by using attension based caption generation. The problem in all papers is the evaluation metrics used. Although BLEU score is a standard metric in machine translation, it might not give accurate results for all datasets.Thus we used all the metrics mentioned in the 3 papers and saw which one works best for us. Apart from the Beam Search that is used in most papers, we also tried Argmax search for caption generation and compared the results for both.

2 Motivation

Image Captioning can be useful for physically disabled people like semi-blind or blind people if voice output is added to the generated captions. It can also be used in virtual assistants such as Siri or Cortana to help searching images of a particular type. For e.g. :- "Show me pictures of myself wearing a blue shirt." Thus, we can see that there is plenty of motivation and usefulness involved in the image captioning task.

3 Related Work

Majority of the work in Image Captioning involve hard coding of the visual concepts and sentence templates.The paper "Show and Tell:A Neural

Image Caption Generator" describes a simple model that uses CNN for vision to generate image embeddings and RNN for language generation to describe the content of the image.The motivation for this paper lies in the recent advancements in machine translation, where given a sentence in one language, we transform to another sequence of words in another language. This task can be related to machine translation, but instead of having a sequence of words as input, we have a fixed embedding of the image and input and generate a sequence of words.The key idea in this paper is to maximize the probability of the correct sentence given an image, i.e $P(S|I)$.The second paper "Deep Visual-Semantic Alignments for Generating Image Descriptions" generates image description using a model that is rich enough to simultaneously reason about contents of images and their representation in the domain of natural language. Also, their model is free of assumptions about specific hard-coded templates, rules or categories and instead rely on learning from the training data.The main motivation behind this approach is that while the previous models based on closed vocabularies of visual concepts constitute a convenient modeling assumption, they are vastly restrictive when compared to the enormous amount of rich descriptions that a human can compose.The paper present a model that aligns sentence snippets to the visual regions that they describe through a multimodal embedding. It then treats these correspondences as training data for a second, multimodal Recurrent Neural Network model that learns to generate the snippets.Basically the model is based on a combination of CNN for image regions and bidirectional RNN for descriptions, and an objective function. The main task in the third paper "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention" is understanding of the scene in the image. Rather than compressing an entire image into a static representation, attention allows for salient features to dynamically come to the forefront as needed. This is especially important when there is a lot of clutter in an image.Here they used CNN's and LSTM based models for feature and sentence generation.This paper proposes two variants which are, "hard"(trainable by maximizing an approximate variational lower bound or equivalently by REINFORCE) and "soft"(trainable by standard back-propagation methods) attention based captioning models.

4 Model Descriptions

We implemented three models:

1. CNN-RNN (Google's Implementation as our baseline)

2. CNN-BRNN

3. Attention-based model

We started with the simple CNN-RNN model as our baseline. The model takes an image I as input, and is trained to maximize the likelihood $P(S|I)$ of producing a target sequence of words S = S1, S2, . . . where each word St comes from a given dictionary, that describes the image adequately.In the caption generation model CNN is used as an image encoder , by first pre-training it for an image classification task and using the last hidden layer as an input to the RNN decoder that generates sentences.This presents an end-to-end system of neural net which is fully trainable using stochastic gradient descent.

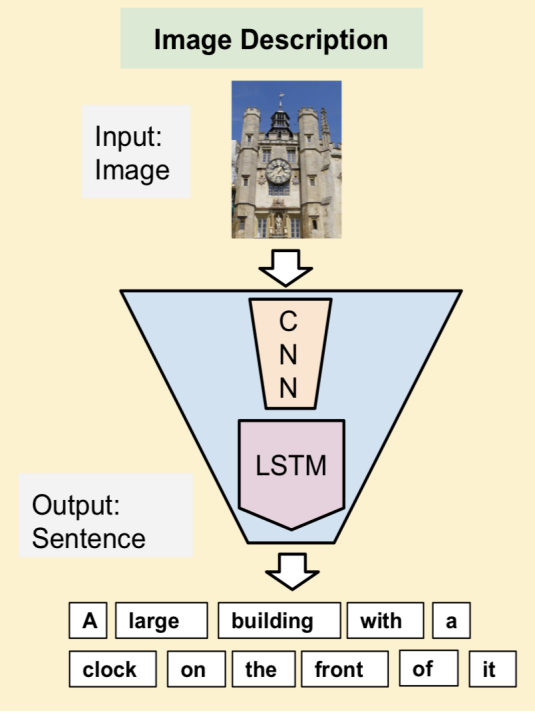

Figure 1: NIC model based on neural nets consisting of vision CNN followed by a language generating RNN.

Here we maximized the probability of the correct description given the image by using the fol-

lowing formulation:

$$\theta^* = argmax_\theta \sum_{(I,S)} \log p(S|I;\theta) \quad (1)$$

where θ are the parameters of our model, I is an image and S its correct transcription. It is an unbounded sequence. So, we use the chain rule to compute the joint probability.

$$\log P(S|I) = \sum_{t=0}^{N} \log P(S_t|I, S_0, ...S_{t-1}) \quad (2)$$

Here, we are conditioning upon up to previous t-1 words in the sentence, which can be represented as a hidden state h_{t-1} to compute h_t. For this we use Long Short-Term Memory networks, which are the current state of the art for tasks like machine translation etc. Convolutional Neural Networks are used to encode the image, as these are the current state of the art for object detection. We use the forget, input and output gates for the LSTM as studied in class. Probability distribution is produced for each word using softmax as the output of LSTM.

$$x_{-1} = CNN(I) \quad (3)$$

$$x_t = W_e S_t \quad t \in 0, ..N-1 \quad (4)$$

$$P_{t+1} = LSTM(x_t) \quad t \in 0, ..N-1 \quad (5)$$

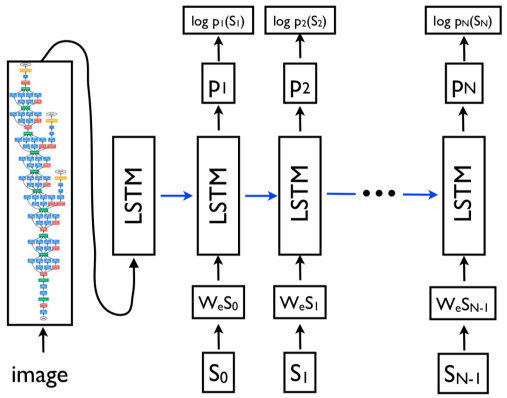

Figure 2: LSTM model combined with a CNN image embedder and word embeddings.

Both image and words are mapped to the same dimensional space using CNN and word embeddings W_e. We formulate the loss function as the negative of log likelihood of the correct word.

$$L(I, S) = -\sum_{t=1}^{N} \log p_t(S_t) \quad (6)$$

We then improved the captions generated by using the Bidirectional RNN instead of the RNN as proposed above.Here a Region Convolutional Neural Network is used to encode the image.

$$v = W_m[CNN_{\theta_c}(I_b)] + b_m \quad (7)$$

$CNN(I_b)$ acts a black box which takes pixels of the image I_b and produces a 4096-dimensional vector.Bidirectional RNN is used for word representation.The BRNN takes a sequence of N words (encoded in a l-of-k representation) and trans forms each one into an h-dimensional vector.BRNN formulation is as follow:

$$x_t = W_w I_t \quad (8)$$

$$e_t = f(W_e x_t + b_e) \quad (9)$$

$$h_t^f = f(e_t + W_f h_{t-1}^f + b_f) \quad (10)$$

$$h_t^b = f(e_t + W_b h_{t+1}^b + b_b) \quad (11)$$

$$s_t = f(W_d(h_t^f + h_t^b) + b_d) \quad (12)$$

Here, t is an indicator column vector that has a single one at the index of the t-th word in a word vocabulary. The weights W_w is a word embedding matrix of 300 dimension with random initialization. The final h-dimensional representation s_t for the t-th word is a function of both the word at that location and also its surrounding context in the sentence. The activation function f used is ReLU. The best model for caption generation was obtained for attention based model. Here again we use CNN's and LSTM based models for feature and sentence generation.

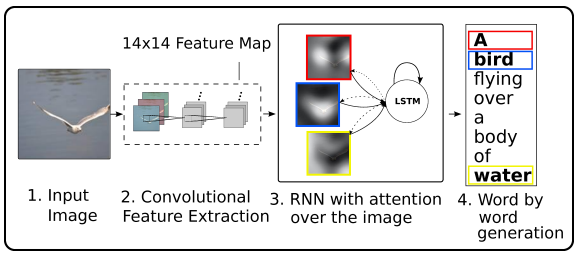

Figure 3: Steps in Neural image caption generator with visual attension.

The model takes encodings a of the image and produces a sequence y, which is 1-of-K encoded words.

$$a = \{a_1, a_2, ...a_L\} \quad a_i \in \mathbb{R}^D \quad (13)$$

$$y = \{y_1, y_2, ...y_C\} \quad y_i \in \mathbb{R}^K \quad (14)$$

Features for the images are extracted from the lower convolutional layers rather than from the fully connected layer. LSTM is used to generate a caption by generating one word y_t at every time step, which is conditioned on context vector $\hat{z}_t$, the previous hidden state h_{t-1} and previously generated words y_{t-1}.

y_{t-1} is the input to the memory cell layer at time t and W_i, W_f, W_c, W_o, U_i, U_f, U_c, U_o are the parameters of the model and E is the embedding matrix

$$i_t = \sigma(W_i E y_{t-1} + U_i h_{t-1} + Z_i \hat{z}_t) \quad (15)$$

$$f_t = \sigma(W_f E y_{t-1} + U_f h_{t-1} + Z_f \hat{z}_t) \quad (16)$$

$$o_t = \sigma(W_o E y_{t-1} + U_o h_{t-1} + Z_o \hat{z}_t) \quad (17)$$

$$C_t = f_t * C_{t-1} + i_t * \tanh(W_c E y_{t-1} + U_c h_{t-1} + Z_c \hat{z}_t) \quad (18)$$

$$h_t = o_t * \tanh(C_t) \quad (19)$$

i_t, f_t, o_t, C_t, h_t are input, forget, output, memory, hidden states of LSTM. Context Vector $\hat{z}_t$ is a dynamic represenatation of the relevant part of the image input at time t

$$\hat{z}_t = \phi(a_i, \alpha_i) \quad (20)$$

α_i are weights for each annotation vector a_i and they are computed by the attention model f_{att}, which uses the multilayer perceptron conditioned on h_{t-1}. The paper presents a deep output layer to predict the probabilities of the output words.

$$p(y_t | a, y_1^{t-1}) \propto \exp(L_o(E y_{t-1} + L_h h_{t-1} + L_z \hat{z}_t)) \quad (21)$$

where L_o, L_h, L_z, E are learned parameters.
Deterministic Soft Attention takes the expectation of the context vector $\hat{z}_t$ to formulate a deterministic attention model. It is trained end-to-end by minimizing the following negative log-likelihood function:

$$L_d = -\log(p(y|a)) + \lambda \sum_i^L \left(1 - \sum_t^C \alpha_{ti}\right)^2 \quad (22)$$

5 Evaluation Steps

We conducted extensive experiments to evalute various models implemented using different metrics and datasets in order to compare the results and finally obtain the best model for caption generation.The various evaluation metrics used are BLEU (Bilingual Evaluation Understudy)score, METEOR and Cider. Recall@k is also used to rank a set of description with respect to given image.Here the BLEU is the most commonly used metric in the image description literature, which is a form of precision of word n-grams between generated and refernence sentences. However the most reliable evaluation metric is to ask raters to give a subjective score on the usefullness of each description given the image, this is called human evaluation.After the model generation the scores for each model is obtained using the nlg-eval library. The datasets used are as given in the table named "Dataset Used".We used the flikcr8k and flikcr30k dataset to obtain results for CNN-BRNN model. MSCOCO dataset was used for the evaluation of the baseline model and the attension based model.

Datasets Used.			
Name	Train	Val	Test
MSCOCO	82783	40504	41000
Flickr8k	6000	1000	1000
Flickr30k	18000	6000	6000

6 Key Results

We report our main results on all relevant datasets in Tables below:

Evaluation Metrics for various Models.			
Methods	bleu-1	bleu-2	bleu-3
Baseline$_{mscoco}$	34.97	20.15	11.95
CNN-BRNN$_{8k}$	61.19	39.10	25.19
CNN-BRNN$_{mscoco}$	36.19	20.67	12.34
Attent$_{mscoco}$	33.28	18.50	10.51

Methods	bleu-4	meteor	rouge-l
Baseline$_{mscoco}$	7.42	13.72	33.02
CNN-BRNN$_{8k}$	16.28	18.99	47.26
CNN-BRNN$_{mscoco}$	7.84	13.43	34.62
Attent$_{mscoco}$	6.17	12.87	31.47

Methods	CIDEr
Baseline$_{mscoco}$	76.37
CNN-BRNN$_{8k}$	38.32
CNN-BRNN$_{mscoco}$	75.64
Attent$_{mscoco}$	65.93

7 Analysis

CNN-RNN model (trained on MSCOCO 80k images):
This model can be considered to be the vanilla

Figure 4: Baseline model i.e. CNN-RNN model result-1 on MSCOCO Dataset

Figure 5: CNN-RNN model result-2 on MSCOCO Dataset

Figure 6: Result-1 for CNN-BRNN based model on Flickr-8k Dataset.Normal Max search: A dog is playing with a stick in the water . Beam Search, k=3: Two dogs are playing in a pool . Beam Search, k=5: A brown dog in a swimming pool . Beam Search, k=7: A brown dog in a swimming pool .

model of our project. Some of the cases where this model worked best on generating captions containing animals such as dogs, cats and people

Figure 7: Result-2 for CNN-BRNN based model on Flickr-8k Dataset.Normal Max search: A woman in a red jacket is walking down a road . Beam Search, k=3: A woman in a red jacket is standing on a street . Beam Search, k=5: A woman in a red jacket is standing on a sidewalk . Beam Search, k=7: A group of people playing in the grass .

Figure 8: Result-3 for CNN-BRNN based model on Flickr-8k Dataset.Normal Max search: A man is climbing a rock . Beam Search, k=3: A man in a red shirt is climbing a rock . Beam Search, k=5: A man is climbing a rock . Beam Search, k=7: A man climbing a rock .

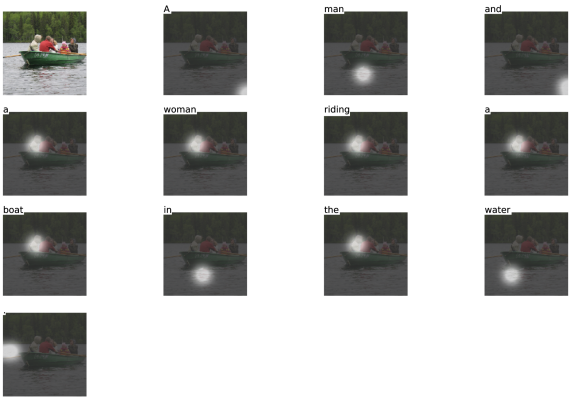

Figure 9: Attention based model, Sentence generated - A man and women riding a boat in the water.

skiing. The probable reason can be that many such instances were seen in the training set and hence the model learnt to generate good captions for this model.The model, however, does not perform well in distinguishing the gender of a person, for example in Fig. 4 it can be seen that the model was not able to succesffully tell that the person was a woman. It also unnecessarily added "on a stove" to the caption, although, there is no such thing in the image. Also, in Fig. 10 it was unable to identify most of the objects in the image, let alone a perfect description of the image. Our speculated reason is this is because of insufficient training of the model and cure can be better tuning of certain hyper-parameters and further investigation along those lines.
Key Stats: On the MSCOCO dataset the CNN-RNN i.e. base-line model gives a BLEU-1 score of 35.

CNN-BRNN model (trained on MSCOCO 80k images): This model tried to capture "both-side" contexts of the ground truth captions and outperformed the CNN-RNN model. This specific model performed well in recognizing images of sceneries and humans. Although this model performs well overall, it fails to capture some of the nuances in certain images. For example, in Figure 6 we see that it recognizes a dog as a stick (using argmax search) and generates caption "brown colored dog" (on using k¿3 beam search). The reason for this, we speculate, is because the model learnt certain "templates" during training and hence tries to fit those patterns whenever it identifies the objects it saw during training (in this case: a dog). This is further confirmed, from Fig 8, where beam search generates a caption of "a man in a red shirt" although it is not true.
Key Stats: On the Flickr8k dataset, CNN-BRNN yielded a BLEU-1score of 61, to be compared to the current state-of-the-art of 63.

Attention-Based Model (trained on MSCOCO 10k images): Ideally, this model should outperform both the other models. However, due to lack of sufficient resources (GPU power, size of the extracted features being too huge), we weren't able to scale this model to perform its best, and could only train this model on sampled 10k images from the MSCOCO dataset.This model, still, worked quite well and generated better descriptive captions compared to the other two models (on manual checking for some images). Figure 9 shows an example of an image caption generated through this model. The best thing about this model is that we can visualize the part of the image the model focuses on while generating words of the caption. This lead to a better understanding of how the hidden layers of a neural networks work (as this is generally a black box in most of the cases).
Key Stats: On MSCOCO dataset the attention-based model gives a BLEU-1 score of 33 (ideal score should have been 70 with sufficient training).

a red fire hydrant sitting in the middle of a street <END>

Figure 10: CNN+RNN model, Bad Prediction - Could only identify the street but could not identify other objects in the image correctly.

8 Future Work

Improvements for the baseline model i.e. CNN-RNN can be changed to use a different loss function like Noise Contrastive Estimation method, and sample some negative words that should not appear in the sentence for a given image. Also, we could use better evaluation metric to compare against the human results. Possible extensions for CNN-BRNN based this model could be to use multiplicative interactions with image and word embedding rather than additive. We think multiplicative interactions would capture much more information and will be able to produce better sentences in attention model. It may also predict phrases that are not present in the training set. With enough computation power, we should be able to finetune the attention model to achieve bet-

ter results.
For the sentences that did not generate a proper description in the attention model Fig 10, we can try to visually analyse the region of the image that the model is focusing on while generating the words of that caption. Also, we can try to feed in the ground truth caption for such an image to the model, to see which part of the image the model "looks on" during learning. Another experiment that can be performed is to try different activation function, like the cuboid function (in one of the assignments in class) or other activation functions in the attention-based model, to see how the BLEU scores are affected.

9 Conclusion

First and foremost, this project acted as a pathway for us to explore the field of Deep Learning through NLP and learn about the various state-of-the-art techniques being used in the specific task of Image Captioning. We explored various technologies such as TensorFlow and also how effective CNNs and RNNs can be in being "unreasonably effective" in such a task. We also learnt that properly training Deep Learning models take sufficient amount of time and needs sufficient amount of GPU resources in order to generate effective results. Some of the things that we found surprising was that the simple and straightforward CNN+RNN recepie (Google paper) with a simple loss function was so effective in generating captions and reaching state-of-the-art BLEU metric scores. Some strange observations were also found, for example, in the CNN-BRNN model, a beam search of k=3 gives a reasonable and good enough description of images in most of the cases. However, on increasing the beam size further, strange captions are generated, although we would expect the caption quality to improve further. The reason for this can be investigated further. Finally, the attention-based model does really well in generating richer and better descriptive captions, as it focuses on specific parts of the image to generate words in the captions. However, it also needs sufficient resources to train properly.

10 References

1) Translating Videos to Natural Language Using Deep Recurrent Neural Networks - Venugopalan, Subhashini and Xu, Huijuan and Donahue, Jeff and Rohrbach, Marcus and Mooney, Raymond and Saenko, Kate (for Figure-1)

Research papers:
2) Show and Tell: A Neural Image caption Generator- Oriol Vinyals ,Alexander Toshev,Samy Bengio,Dumitru Erhan ,2015 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), June 2015.

3) Deep Visual-Semantic alignments for Generating Image Description- Andrej Karpathy,Li Fei-Feiu, Department of Computer Science, Stanford University, USA,2015 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), June 2015.

4) Show, Attend and Tell: Neural Image caption Generation with Visual Attention- Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, Yoshua Bengio, Proceedings of the 32 nd International Conference on Machine Learning, Lille, France, 2015. JMLR: W& CP volume 37.

Links referred:
5) https://github.com/tensorflow/models/tree/master/research/im2txt
6) https://research.googleblog.com/2016/09/show-and-tell-image-captioning-open.html
7) https://yashk2810.github.io
8) https://github.com/yunjey/show-attend-and-tell