Final Project Report
CS698N: Recent Advances in Computer Vision

Visual Storytelling

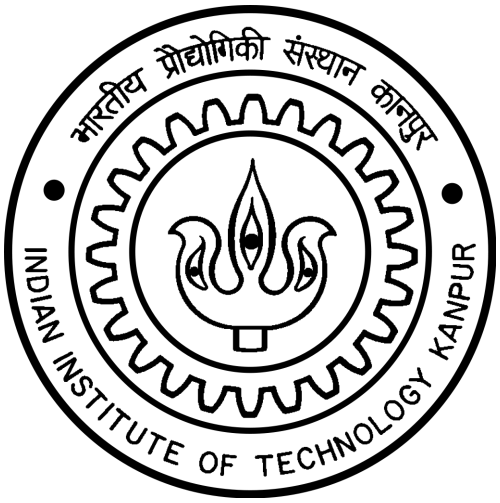
Vasu Sharma(12785)

Amlan Kar(13105)

Nishant Rai(13449)

Course Instructor: Gaurav Sharma

December 18, 2016



1 The Problem: Definition and it's Importance

The problem of Visual Storytelling is a problem of mapping sequential images to sequential, human like, narrative descriptive sentences or 'stories'. Simply put, it involves understanding a sequence of images and trying to explain it's contents in a story like manner. The problem has 3 major types:

- Description of images in Isolation (DII): This is similar to a single image caption generation task
- Descriptions of images-in sequence (DIS): This involves creating captions for a sequence of images, without incorporating any particular narrative style
- Stories for images-in sequence (SIS): This is the task of creating story like narrative descriptions for a sequence of images.

The problem is important as it involves 2 new major concepts:

1. Learning to move beyond reasoning about standalone images and be able to interpret and understand a sequence of images
2. Moving beyond simple descriptive captions to more human like narratives

The problem is a completely new one with hardly any prior work which makes it a challenging one and the novelty of the problem is what makes it an important one to us.
We believe that such a system is a step into creating smarter machines which have human like capabilities of understanding and explaining the things around us. This work could have tremendous impact in areas like virtual assistants, fully automated customer response systems, interactive learning for students etc. It's the wide practical applicability of such a system and it's importance in achieving the Artificial Intelligence dream which makes this problem an important one.

2 Related Works: A Literature review

The problem of Visual Storytelling was first introduced by Huang et al [2] from Microsoft Research at NAACL-2016 a few months back. This paper introduces a dataset for the Visual Storytelling task, proposes certain evaluation Metrics and presents some baseline approaches and results obtained using them.
We note that this problem is a completely new one and there is very little prior work on this topic so far.
However, the intersection of vision and language has been a very popular research area of late with several problems like caption generation and visual question answering receiving a lot of attention. We take inspiration and ideas from several such problems and feel that some of these problems directly relate to the problem of visual storytelling and a clever combination of such approaches may help us solve this problem.
One of the closest related works to our problem is the Neural Storyteller [4], [5] which gives narrative descriptions for a single image. They learn skip thought vector representations for each passage of a Romantic Novel dataset and train a normal image captioning network in parallel. Then they use a novel 'Style Transfer' algorithm to transfer the thought in the romantic novels to the normal captions which allows them to generate narrative, romantic captions for each image.
Zhu et al [9] propose a setup where the system learns to create story-like visual descriptions by trying to align books and movies. Other prominent works we draw inspiration from include the works of Karpathy and Fei-Fei [3] and that of Vinyals et al [8] on image captioning, Antol et al [1] and Ren et al [7] in visual question answering and Ramanathan et al [6] in video understanding.

3 Approach

We work with a dataset which consists of 81,743 unique images in 20,211 sequences with associated captions and narrative sequences. We use METEOR and BLEU scores as evaluation metrics for this task. METEOR aims to assign

scores to the candidate based on human-made ground truths. It is based on the harmonic mean of unigram precision and recall, where recall is weighted higher than precision. It also includes several features such as stemming and synonymy matching, which makes it correlate well with human judgements. The approach we use for the task is as follows:

- **Data Pipeline creation**: We first pre-process the dataset and create a TensorFlow data pipeline for the new dataset which could be fed into our encoder-decoder model
- **Feature Extraction**: We then use an Inception network pre-trained on ImageNet and fine-tuned on MS COCO to obtain useful image representations for the images in the dataset
- **Encoding Image sequences**: The image sequence is then encoded by running a Gated Recurrent Unit Network over the image representations backwards. This is used as the initial hidden state of the story decoder network.
- **Generating stories using decoder network**: A Story Decoder network, modelled as a Gated Recurrent Unit, coupled with Beam Search is used to produce the story for the image sequence, one word at a time using words from the training vocabulary.
- Initially the pre-trained Inception network parameters are frozen and only the encoder and decoder are trained. After that once the training error becomes stable, we unfreeze the entire network and all the network parameters including the inception network weights are fine tuned
- We use all the encoder GRU states while decoding and producing the stories
- We have also implemented the GRU with bidirectional connections for decoding and plan to experiment with that once the present set of experiments complete
- We also use our own implementation of Beam Search which use TensorFlow functions to make use of distributed computing when compared to the classic beam search which is implemented in a sequential manner
- We also try out heuristics like preventing duplication sentence generation and variable beam width in beam search which improve the generated stories tremendously

Note that the training is presently running with a GRU based decoder with only unidirectional connections. We expect the training to take roughly 2 weeks on the limited computational resources available to us. In the meanwhile, we have implemented a bidirectional version of the decoder GRU and will experiment with the same once the training for the unidirectional GRU decoder completes.
We first performed both training and testing on the test set itself to make the model overfit as a sanity check for the approach. After obtaining satisfactory results from the same, we train our model on the test set. As the model is still training, we present some intermediate results in this report.

4 Libraries and Tools used

Our project was implemented using the TensorFlow library with a Python interface. We used Google's Show and Tell model as the base code for our project and built upon it. The data pipeline, encoder and decoder were re-implemented based on our needs and for the new dataset.
For the METEOR score calculation we modified the METEOR score calculator provided by CMU.

5 Results

Here is an example of how our model performs:

Beam Beam Width	No repeat Heuristic	METEOR	Bleu	CIDEr	ROUGE_L
1	Yes	0.071	0.173	0.060	0.153
2	Yes	0.082	0.209	0.097	0.160
3	Yes	0.084	0.217	0.094	0.163
4	Yes	0.083	0.217	0.089	0.163
5	Yes	0.082	0.214	0.092	0.161
1	No	0.062	0.152	0.036	0.148
3	No	0.063	0.159	0.038	0.146
5	No	0.060	0.152	0.030	0.144

Table 1: Results on the Visual Storytelling task using our approach

Figure 1: Example Image sequence

Generated Captions: *The bride and groom were very happy to be getting married . the family is having a great time . the couple was excited to see their new friends . The bride and her bridesmaids looked absolutely gorgeous . the bride was happy to be there.*

6 Discussions

In this project we worked on the problem of Visual Storytelling and experimented with several novel approaches for the task. We also demonstrated the effectiveness of our model on the task for the Visual Storytelling dataset.
Due to the problem being an extremely new one and due to the lack of many existing code bases, the project involved a significant coding component and a lot of techniques and models had to be implemented from scratch. We were able to implement a decoder which looks at all encoder states while decoding and also implemented a bi-directional GRU for decoder. Access to more compute resources would have allowed us to experiment with a larger number of design choices and may have helped us improve the model even further.
We will also like to point out the inadequacies in the dataset. We note that in several places, the ground truth story is very different from what a human would usually label. Also the dataset captures only some of the possible stories which could be assigned to a given image sequence while the number of possibly correct stories could be very large. A more exhaustive dataset is hence needed to improve the performance on this task further.
Another note we make is of the ineffectiveness of the performance metrics in certain scenarios. Both METEOR and BLEU often fail to assign accurate scores to the produced stories. Hence a better evaluation metric is needed to evaluate the performance on this task in a more just manner.
Some of the obvious shortcomings we notice in our model is it's tendency to produce stories which are present with a large prior in the train set. It produces common sentences like "the man was very happy to be there" , "the family was having a great time" , "the girls were having a good night" , "we had a lot of fun" etc very often. We plan to incorporate a penalty term to penalize the occurance of such common phrases.

7 Comparison: Proposed vs Implemented

Feature	Proposed	Mid sem progress	Final Progress
Using seq2seq model for generating captions (Implemented from scratch)	✓	✓	✗
Producing descriptions for images in sequence	✓	✓	✓
Replicating the State of the art paper results	✓	✗	✗ (partial)
Decoding using all encoder states	✓	✗	✓
Bi-directional connections in LSTM	✓	✗	✓
Custom Implementation of Beam Search	✓	✗	✓

Table 2: Comparison table between proposal and mid term and final progress reports

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. *In International Conference on Computer Vision (ICCV)*, 2015.

[2] T. Huang, F. Ferraro, N. Mostafazadeh, I. Misra, A. Agrawal, J. Devlin, R. Girshick, X. He, P. Kohli, D. Batra, L. Zitnick, D. Parikh, L. Vanderwende, M. Galley, and M. Mitchell. Visual storytelling, 2016. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

[3] A. Karpathy and L. Fei-Fei. Deep visual semantic alignments for generating image descriptions. *In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2015*, 2015.

[4] R. Kiros, Y. Zhu, R. Salakhutdinov, R. S. Zemel, A. Torralba, R. Urtasun, and S. Fidler. Neural storyteller. `https://github.com/ryankiros/neural-storyteller`, 2015.

[5] R. Kiros, Y. Zhu, R. Salakhutdinov, R. S. Zemel, A. Torralba, R. Urtasun, and S. Fidler. Skip-thought vectors. *arXiv preprint arXiv:1506.06726*, 2015.

[6] V. Ramanathan, P. Liang, and L. Fei-Fei. Video event understanding using natural language descriptions. *IEEE International Conference on Computer Vision (ICCV), 2013*, 2013.

[7] M. Ren, R. Kiros, and R. Zemel. Exploring models and data for image question answering. *Advances in Neural Information Processing Systems*, 2015.

[8] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. *In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2015*, 2015.

[9] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *arXiv preprint arXiv:1506.06724*, 2015.