CS 747: PROJECT PROPOSAL

Vamsi Krishna Reddy	Yaswanth Kumar	Vighnesh Reddy
160050064	160050066	160050090

1 Proposal

We propose to utilize Policy Gradient optimization in the task of Video Captioning.

Prior work of utilizing policy gradients on captioning have shown promising results specifically in the face of Image Captioning. We plan to extend this idea by making suitable changes to existing Image Captioning works using to support Video Captioning. To the best of our knowledge, this attempt has been limitedly explored and we consider the papers [1] and [2] as our baseline inspiration.

We describe the details of our baseline work below in 2.1 and then describe our extension towards Video Captioning in 2.2. We hereby refer to Policy Gradient as PG.

2 Details

2.1 Prior Work

The pipeline suggested in the paper [1] for images is summarized here. We encode the image using a pretrained backbone network as a feature extractor to get an image encoding. We then pass it to an RNN along with ground truth captions to make a sequence of outputs. Traditional captioning systems interpret the outputs as word predictions and optimize through a loss over them.

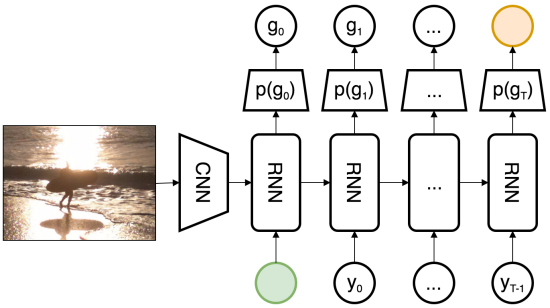

Figure 1: CNN-RNN Architecture from [3]

The paper [1] suggests to interpret the history of sequence until this point *i.e.* g_1:g_{t-1} as the `state` and the `actions` being the next word in the sequence g_t. Then the RNN decoder thus acts like a stochastic policy. The PG method then uses Monte Carlo rollouts to directly optimize SPICE + CIDEr metrics commonly used analysing performance of captioning systems.

Thus in this case, the value function is defined as the expected future reward:

$$V(g_{1:t}|x^n, y^n) = E_{g_{t+1:T} \sim \pi_\theta(\cdot|g_{1:t}, x^n)}[R(g_{1:t}; g_{t+1:T}|x^n, y^n)]$$

where x_n is ground truth image and y_n is ground truth caption.

The advantage of using Policy Gradient based methods is the fact that these metrics including the BLEU score are non-differentiable and hence traditional systems cannot directly optimize over such metrics during training. In our case, we consider these metrics as our `rewards` to our `agent`. The eventual goal would be to maximize average reward starting from empty state s_0, *i.e.*

$$J(\theta) = \frac{1}{N} \sum_{n=1}^{N} V_\theta(s_0|x^n, y^n)$$

2.2 Our Plan

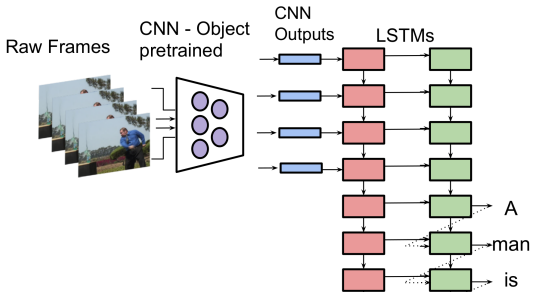

Figure 2: Traditional architecture for Video Captioning

The work from [2] suggests using CNN-LSTM model to encode the the video features by a similar frame by frame processing through a backbone network and then through the LSTM to account for temporal information in the video. We take a similar inspiration to extend the current PG optimization towards Video Captioning systems.

We plan to work on the MSR-VTT dataset [4] for training and evaluation of various experiments we do over the same metrics used in baseline paper *i.e.* SPICE and CIDEr.

3 References

References

[1] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, "Improved Image Captioning via Policy Gradient optimization of SPIDEr," 2016.

[2] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence – video to text," 2015.

[3] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," 2014.

[4] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language." IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), June 2016. [Online]. Available: https://www.microsoft.com/en-us/research/publication/msr-vtt-a-large-video-description-dataset-for-bridging-video-and-language/