Sequence to Sequence Models for Generating Video Captions

Sachin Verma (216cs1147), M.TECH 2nd Year
Department of Computer Science and Engineering
NIT Rourkela

Introduction

In this work, we explore sequence to sequence models mainly used in neural machine translation and apply them in the context of video captioning. Our problem involves the translation of video frames to natural language descriptions of the features therein. Our models are implemented using the Pytorch framework and the results are quantified using the METEOR [1] metric.
METEOR computes a penalty for a given alignment, to take into consideration longer matches. Where the penalty increases as the number of chunks from model generated sentence increases.

Applications and Previous Approaches

Applications of video captioning include but are not limited to: human-robot interaction, description of videos to the blind, video indexing and information retrieval. Past attempts to this problem have been the following:

- Template-based language models, predefines a set of templates following specific grammar rules.
- Average frames features in a video resulting a single vector. Shortcoming of this approach is that this representation ignores the ordering of the video frames [3].

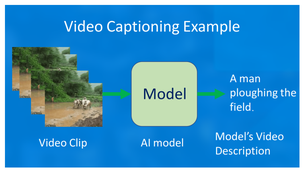

Figure 1: Example of Video Captioning Objective.

Dataset

Experiments were conducted on the Microsoft Research Video Description Corpus or MSVD[4].

MSVD Examples

- 1970 YouTube Video snippets. Typically single activity and no dialogue (10 to 30 seconds)
- Each video has descriptions in multiple languages. About 40 English descriptions per video. Descriptions and videos collected on AMT. Size of vocabulary 12,000 words.

1. A man is plowing a mud field.
2. Bulls are pulling an object.
3. Tow oxen are plowing a field

1. A man is walking on a rope.
2. A man is balancing on a rope.
3. A man is balancing on a wire.

Figure 2: Some samples from the MSVD Dataset.

Experiments

We experimented with two models:

- Frames to Video Caption wherein visual features are extracted with RESNET-50 [2]

- Mean pooling of features with VGG-16 and fed as hidden state to the model.

We evaluated our results using the METEOR. This score is computed based on the alignment between a given hypothesis sentence and a set of candidate reference sentences.

$$\mathbf{F_{mean}} = \frac{10 \times \mathbf{P} \times \mathbf{R}}{\mathbf{R} + 9 \times \mathbf{P}}$$

$$\mathbf{Penalty} = \frac{0.5 \times \#\mathbf{chunkes}}{\#\mathbf{unigrams} - \#\mathbf{matched}}$$

$$\mathbf{Score} = \mathbf{F_{mean}} \times (1 - \mathbf{Penalty})$$

Video Features Extraction

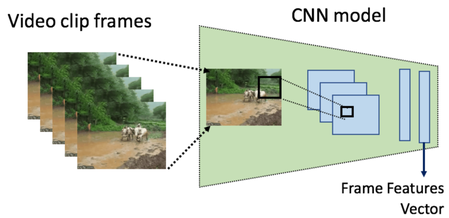

Figure 3: Extraction of Visual Features using a Pre-trained CNN

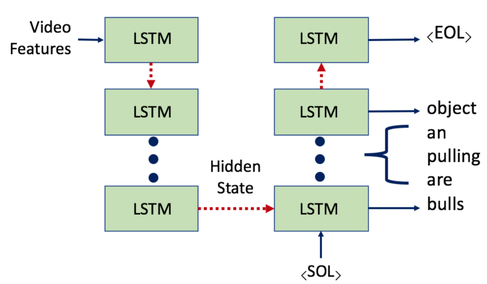

Figure 4: Encoder-Decoder Network

Result

Model	Test Score	Validation Score
Frames to Text	0.171	0.180
Mean Pool to Text	0.174	0.179

Figure 5: METEOR scores on the test and validation set. For the model that used all the frames and mean pooling on the frames

Ground Truth: one woman is cooking some kind of coated meat in a pan on a hotplate
Model: a woman is standing on a bicycle on a bicycle on a block of a car off a car off

Figure 6: METEOR scores on the test and validation set. For the model that used all the frames and mean pooling on the frames

Conclusion

Comparing the METEOR results, we can see that the mean pool model performed better than the frames to text model. This might be due to the fact that the mean pool model only has to associate one vector with a descriptive caption whereas the video to text model must associate many frames to one caption.

References

[1] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[3] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[4] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pages 4534–4542, 2015.