Sequence to Sequence Models for Generating Video Captions

Rafael A. Rivera-Soto and Juanita Ordóñez

Department of Engineering at Stanford University

Abstract

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 metric.

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, pre-defines a set of templates following specific grammar rules. [5]
- 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 .[1]

Video Captioning Example

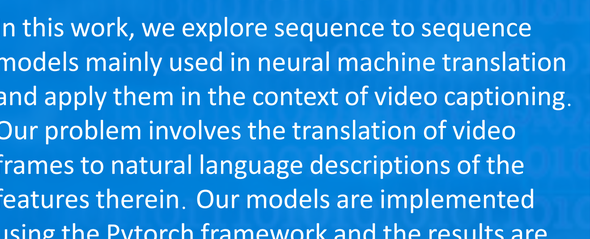

Fig 1. Example of Video Captioning objective.

Dataset

Experiments were conducted on the Microsoft Research Video Description Corpus or MSVD. [3]

- 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.

MSVD Examples

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.

Fig 2. MSVD dataset examples along with some of the English annotations.

Experiments

We experimented with two models:

- Frames to Video Caption wherein visual features are extracted with RESNET-50
- Mean pooling of features with VGG-16 and fed as hidden state to the model

We evaluated our results using the METEOR [2]. This score is computed based on the alignment between a given hypothesis sentence and a set of candidate reference sentences.

I. $$Fmean = \frac{10PR}{R+9P}$$

II. $$Penalty = 0.5 * \left(\frac{\#chunks}{\#unigrams\_matched}\right)$$

III. $$Score = Fmean * (1 - Penalty)$$

Fig 5. I) Weighted harmonic mean. II) METEOR penalty. III) METEOR score

Video Features Extraction

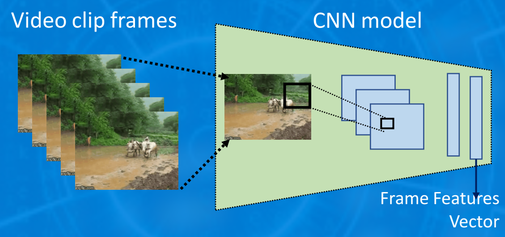

Fig 3. Video clip frames feed in CNN model pre-trained models

Sequence to Sequence

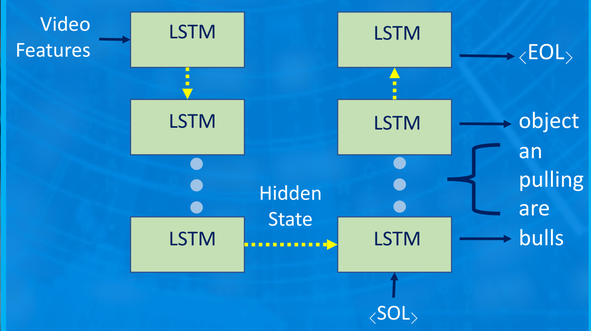

Fig 4. Sequence to sequence LSTM cells encodes the video features and decodes into a sentence.

Results

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

Table 1. METEOR scores on the test and validation set. For the model that used all the frames and mean pooling on the frames

Ground Truth: two elephants wash themselves in a river
Model: a panda climbs up a threetiered cake and another panda is sitting and looking at the cake

Ground Truth: a man slices tomatoes on a kitchen counter
Model: a man is slicing a peeled onion into four uniform slices of about 12 thickness each

Fig 5. Mean Pool example output along with ground truth.

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. This in turn increases the complexity making it more difficult for the model to learn to perform the task. Moving forward we will experiment with different LSTM cells and pre-train our models on the COCO dataset prior to running the experiments.

References

[1] Venugopalan, Subhashini, et al. "Translating videos to natural language using deep recurrent neural networks." *arXiv preprint arXiv:1412.4729* (2014).

[2] Banerjee, Satanjeev, and Alon Lavie. "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments." *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. Vol. 29. 2005.

[3] https://www.microsoft.com/en-us/download/details.aspx?id=52422

[4] N. Krishnamoorthy, G. Malkarnenkar, R. J. Mooney, K. Saenko, and S. Guadarrama. Generating natural-language video descriptions using text-mined knowledge. In *AAAI*, July 2013.