1 Proposal

Video Captioning and Summarization have become very popular in the recent years due to advancements in Sequence Modelling, with the resurgence of Long-Short Term Memory networks (LSTMs) and introduction of Gated Recurrent Units (GRUs). Existing architectures \cite{1, 2} extract spatio-temporal features using CNNs and utilize either GRUs or LSTMs to model dependencies with soft attention layers. These attention layers do help in attending to the most prominent features and improve upon the recurrent units, however, these models suffer from the inherent drawbacks of the recurrent units themselves. Although these techniques have helped in capturing long term dependencies, a 30fps 5 minute video can contain about 9000 frames, making it difficult for the gradients to backpropagate through time. Each training example has to be modelled sequentially and hence, prohibits parallelization within training examples. Given that videos can be very large in size, training a recurrent model for Video captioning can be a tedious and time-consuming task. Another problem with existing techniques is that they take into account only the video features, while video’s text descriptions are completely ignored \cite{3}.

The introduction of the Transformer \cite{4} model has driven the Sequence Modelling field into a new direction. They have shown that it is possible to work on Sequences without the use of either RNNs or any of its sub-categories. They have proposed a new model, based purely on self attention, making them more parallelizable on modern GPUs as compared to their recurrent counterparts. They are also able to handle test sequences with sizes greater than any of the training sequences making them more real-world ready. We feel that Transformers can be applied to Video captioning as well, and can help in addressing the drawbacks of short term dependencies and slower training and running times of existing techniques.

2 Implementation Details

We plan to use all the available features that are present: (1) text which is used to describe it, (2) motion information available from frame to frame, and (3) the information extracted from the individual frames. For the first, we plan to use an embedding layer to get the features, while for the last two, we plan to use pre-trained image and video feature extractor networks. Common CNN architectures like DenseNet and ResNet have proven to be good frame level feature extractors, whereas, architectures like C3D \cite{5} and 3D ConvNet \cite{6} can capture spatio-temporal features. These multi-faceted features, after being projected onto a uniform space, will then be fed into a Transformer which, using self attention, would recognize the important aspects and enable video summarization.

We will test our model on standard datasets like MSR Video-to-text dataset (MVTT) \cite{7} and Microsoft Research Video Description Corpus (MSVD) \cite{8}, which have become the benchmark for video captioning techniques. MVTT consists of 10,000 clips obtained from the web with each video having a 20 words long caption and MSVD consists 1,970 video clips from YouTube with about 41 descriptions per video. Our model will be compared to the previously discussed state-of-the-art techniques on these datasets.

References


