A Review: Sequence to Sequence – Video to Text

This is the first approach to video description that uses a general sequence-to-sequence model. This research offers a unique end-to-end sequence-to-sequence model for producing captions for videos that are sensitive to temporal structure and enable both input (series of frames) and output (sequence of words) of varying lengths. The model outperforms prior work on two large movie description datasets and reaches state-of-the-art performance on the MSVD dataset. The paper is well-motivated. The primary contribution of this study is the development of S2VT, a novel model that learns to directly map a sequence of frames to a sequence of words. The authors are impressed by the advancement of image-to-text models. In prior video captioning research variable length input has been handled by holistic video representations, pooling across frames, or sub-sampling on a fixed number of input frames. In contrast, in this paper, the authors are interested in a sequence-to-sequence model that can learn arbitrary temporal structure in the input sequence and is trained end-to-end.

In the novel approach, a stacked LSTM encodes the frames one by one, using the output of a Convolutional Neural Network (CNN) applied to the intensity values of each input frame as input. After reading all frames, the model constructs a phrase word for word. A parallel corpus is used to learn the encoding and decoding of the frame and word representations. The authors also compute the optical flow between pairs of consecutive frames to mimic the temporal characteristics of typical video activity. The flow images are additionally processed by a CNN before being fed into the LSTM. This enables the model to manage a varied amount of input frames, learn and apply the video's temporal structure, and build a language model to create meaningful, grammatical phrases.

What makes this paper unique and distinguishes it from previous work is that, unlike previous approaches, this one avoids the separation of content identification and sentence generation by learning to directly map videos to full human-provided sentences while simultaneously conditioning a language model on visual features. Furthermore, because the variable-length video is used as input, the authors employ LSTMs as sequence-to-sequence transducers, in accordance with language translation models. The authors also offer a simpler way to leverage temporal information by encoding the series of video frames into a distributed vector representation that is adequate for constructing a sentential description using an LSTM. As a result, according to the authors, the direct sequence-to-sequence model does not necessitate the use of an explicit attention mechanism.

However, there are specific weaknesses in the paper. To begin, we observe that the authors employ visual representations that are pre-trained, but not text representations. As a result, it must acquire a general linguistic structure from the ground up. The inclusion of flow features appears to be an afterthought. It also begs the question of whether we can concatenate them with RGB characteristics instead of training a separate model. Furthermore, the authors state that their model requires no attention mechanism; yet, it is unclear how the model captures the local structure in the video. It's understandable that the model captures the global temporal interaction in the video, but does it ignore the local temporal structure? The evaluation also has certain limitations. The authors only supplied METEOR results, no other metrics. They also do not provide human judgment of the results, which should be mandated given how frequently metrics are inaccurate and human evaluation, albeit pricey, provides the finest insights. We can also notice that the videos in different datasets are subsampled at different rates; MSVD has twice as many

timesteps as the other datasets (on average). MSVD annotations in other languages have also been removed, raising the question of whether this architecture is only useful for English. It would also be interesting to examine how the model performs in low-resource languages.

Some intriguing future studies would be to investigate improved visual and textual representation models. It would also be interesting to test if the model represents both the local structure of the video and the global temporal interaction. Experimenting with adding attention may help enhance performance. It would also be interesting to do a human evaluation to acquire a better understanding of the model's performance and to investigate developing a person-in-the-loop video captioning process.