Attention-Based Multimodal Fusion for Video Description Chiori Hori

Current methods for video description are based on encoder-decoder sentence generation using RNNs.

Temporal attention mechanism gives more weight to encoded features from specific time frames. Use two different types of features : image features and motion features.

**IN this paper they found that combining multimodal (image feature, motion feature and audio feature) attention model with standard temporal attention outperforms state-of-the-art methods on two standard datasets.**

Benefits of attentional multimodal fusion:

* Modalities that are most helpful to discriminate *emotion* can dynamically receive a stronger weight
* Network can detect interference e.g. noise and other sources of uncertainty in each modality and dynamically down-weight the modalities that are less certain.

***GOOD FOR IN THE WILD THEN AMIRITE?***