Knowledge Based Video Retrieval (KBVR)
TRECVID 2016

Etter Solutions

Introduction

During TRECVID 2016 *[1]* our group participated in the Video-To-Text Description (VTT) showcase and the Multimedia Event Detection (MED) task. For the VTT matching task, we developed a system that automatically generates captions using four facets of classes and then rank the annotations using natural language processing techniques. In the MED task we developed a system for the pre-specified zero example sub-task. We use a rank learning approach to generate models that fuse the modalities of the metadata extracted from the video sound and image content.

Approach

Video-To-Text Description - Matching Task

In the Video-To-Text Description (VTT) task, a system attempts to automatically generate a natural language text description for a video. This task requires the system to identify objects within a video and describe their interaction over time. The evaluation dataset consisted of 2000 Vine videos with a duration of approximately 6 seconds, and two sets of single sentence truth descriptions for each video. The VTT task was split into two separate subtasks consisting of a matching/ranking task and a description generation task. The matching/ranking task attempts to match each video with its corresponding text description and is evaluated using the mean inverted rank. The description generation tasks requires the system to generate a single sentence text description for each video and is evaluated using a METEOR score.

Our approach to the VTT task was to develop a system based on the four facets of Who, What, When, and Where. We first, automatically identify classes within the video and then apply natural language processing techniques to match the automatic annotation to the previously unseen truth description. To identify classes representing the four facets, we use a series of deep learning models. These models include a subset of the ImageNet 21,841 synsets *[2]* and the 205 scene categories from the Places database *[3]*. The models were applied to a video frame at the rate of one frame per second, using a weighted average over a sliding window, to generate a four facet description. To compare our facet description with the annotator sentence, we map both descriptions into word vectors using word embeddings *[4]* and generate a document similarity score to provide a final ranking.

Multimedia Event Detection – Pre-specified 0 Examples

The Multimedia Event Detection (MED) task attempts to identify events in video consisting of objects and their interactions over time. A MED system takes an event query as input and returns a ranked list of videos most likely to contain that event. The event query consists of the text metadata describing the event and zero to one-hundred positive examples of the event. The MED task includes sub-task for pre-specified events, where the system is trained on those events, and a sub-task for ad-hoc, where the events are previously unseen.

Our metadata generation module created text metadata for each video using sound, OCR, and object detection, using deep learning models *[2,3]* . The event query component used natural language processing techniques to break the query metadata into modality specific sub-queries. A rank learning model was then trained to learn modality weightings for the pre-specified events.

Conclusions

One of the primary challenges of the VTT task is the vocabulary gap that exists between the annotator's description and the caption generation of the system. It is also possible that the annotator vocabulary is influenced by the informal structure of text often used in social media. In the future, we plan to investigate approaches to training the four facet models using short social media videos such as those found in Vine or Instagram.

References

[1] George Awad and Jonathan Fiscus and Martial Michel and David Joy and Wessel Kraaij and Alan F. Smeaton and Georges Quénot and Maria Eskevich and Robin Aly and Gareth J. F. Jones and Roeland Ordelman and Benoit Huet and Martha Larson, TRECVID 2016: Evaluating Video Search, Video Event Detection, Localization, and Hyperlinking, Proceedings of TRECVID 2016.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database. IEEE Computer Vision and Pattern Recognition (CVPR), 2009.

[3] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. “Learning Deep Features for Scene Recognition using Places Database.” Advances in Neural Information Processing Systems 27 (NIPS), 2014.

[4] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. ICLR Workshop, 2013.