TRECVID 2017: Video to Text Notebook Pages

Areté Associates

Brian Lang, Daniel Lavery, Kelly Roman, Erford Porter

Author Note

Areté Associates 1550 Crystal Drive, Arlington VA 22202

TRECVID 2017: Video to Text Notebook Pages

The TREC Video Retrieval Evaluation series (trecvid.nist.gov) was designed to promote progress in content-based analysis of and retrieval from digital video via open, metrics-based evaluation. TRECVID was a laboratory-style evaluation that attempted to model real world situations or significant component tasks involved in such situations. In it's 17th annual evaluation cycle TRECVID evaluated participating systems on six different video analysis and retrieval tasks using various types of real world datasets. We participated in the automated Video to Text annotation task.

Automatic annotation of videos using natural language text descriptions has been a long-standing goal of computer vision. The task involved demonstrating an understanding of many temporal order of entity object events and interactions. Utilizing recent research and advances in machine learning and computer vision techniques enabled Areté to start work in this problem space. A lot of use case application scenarios can greatly benefit from such technology such as video summarization, facilitating the search and browsing of video archives using such descriptions.

Areté participated in both the Matching and Ranking and Description Generation tasks. In Description Generation, Areté sought to automatically generate a one sentence description (including Who, What, Where, When) for each Twitter Vine video segment selected by the NIST task leaders. For the Matching and Ranking task we sought to provide a similarity ranking of our generated descriptive outputs to the annotations.

NOTE BOOK PAGES

Briefly, list all the different sources of training data used in the creation of your system and its components

For the description task, Areté built upon Venugopalan et al.'s Sequence to Sequence - Video to Text (S2VT) model[i], which was trained on the Microsoft Video Description Corpus (MSVD), the MPII Movie Description Dataset (MPII-MD), and the Montreal Video Annotation Dataset (M-VAD). The S2VT model was in turn built using visual features from Simonyan and Zisserman's VGG network, which was trained over the ILSVRC-2014 dataset.

Areté then retrained our modified networks using various subsets of the TRECVID 2016 VTT annotated data collection, which resulted in varying levels of over training

Briefly, what approach or combination of approaches did you test in each of your submitted runs

Description Generation Task:

All of our entries used a modified S2VT model trained on varying vocabularies from VTT 2016

ARETE.20170901_ARETE_VTT_DESCRIPTION_ABfromorig.primary.txt was trained on the TRECVID 2016 data with both description set A and description set B used as truth.

ARETE.20170901_ARETE_VTT_DESCRIPTION_ABfromlatest_SECONDARY.txt was trained on the same data, but the model was first trained on only description set A (as a diagnostic), and then further trained on sets A and B together.

ARETE.20170901_ARETE_VTT_DESCRIPTION_A_TERTIARY.txt was trained on the same data, but was only trained on description set A.

ARETE.20170901_ARETE_VTT_DESCRIPTION_ABfromorigpost_QUATERNARY.txt is the same S2VT model as run 1, but includes optional post-processing to handle numeric characters and out-of-vocabulary tokens.

Matching and Ranking Task:

Our ranking submissions were generated by running each of the videos through our captioning model, then ranking the candidate captions by their METEOR score with the caption that we had generated.

What if any significant differences (in terms of what measures) did you find among the runs?

All of our runs used the same model architecture and algorithmic pipeline, so the only comparison we can make is between training strategies. Here, surprisingly, the tertiary run, which saw not only the least training overall, but also saw the least diversity in truth labels, generalized best. [It's possible that this is because either a.) there is better correspondence between the vocabulary (or other speech patterns?) in 2016's description set A and the truth for 2017 than there is for 2016's description set B, or b.) that punishing the network with categorical cross-entropy at each predicted token doesn't properly reflect the similarity or relationship between synonyms or alternative sentence structures, and so training the network on the same data with different truth caused it to collapse towards a sort of semantically meaningless uncertainty. I haven't examined that either of these ideas is actually the case.]

We performed markedly better when compared using the results generated by BLEU versus any of the others. Again

Based on the results, can you estimate the relative contribution of each component of your system/approach to its effectiveness?

We conducted limited analysis and chose our primary submission based on a subjective assessment of the quality of the results. We used METEOR to quantitatively assess the improvements made by adding or subtracting vocabularies from the base network. As our initial METEOR values and those of the out of the box network model were unexpectedly low, we looked for incremental global improvements. Qualitatively we considered the readability of and similarity of the prediction output when compared to truth. Additionally we considered the number of Unknown_Encoding entries in each of the descriptions on a local and global basis.

Overall, what did you learn about runs/approaches and the research question(s) that motivated them?

We learned that the task is highly sensitive to selection of training data and performance metrics and that there isn't always even strong correspondence (via METEOR or BLEU) between varying "true" human annotations of the same video.

There is significant risk to overtraining our model (the final embedding->prediction layer has 23 million parameters), which motivates further work into video data augmentation and regularization methods (or avoiding the softmax layer altogether…)

Alternatively, a model that is conservative (short sentences, mostly generic nouns, articles and prepositions) may somewhat optimize the loss function we used to train the network, but do a poor job of replicating natural language and score poorly on recall-heavy metrics

Venugopalan, S.; Rohrbach, M.; Donahue J.; Mooney, R.; Darrell, T. and Saenko, K.
2015 ,Sequence to Sequence -- Video to Text; Proceedings of the IEEE International Conference on Computer Vision (ICCV)}

Awad, G; Butt, A; Fiscus, J Joy, D.; Delgado, A; Michel, M; Smeaton, A.F.; Graham, Y; Kraaij, W; Quénot, G; Eskevich, M; Ordelman, R; Jones, G.J.F; and Benoit Huet},
2017, TRECVID 2017: Evaluating Ad-hoc and Instance Video Search, Events Detection, Video Captioning and Hyperlinking}, Proceedings of TRECVID 2017}