UEC at TRECVID 2016 AVS task

Natsagdorj Choijilsuren and Keiji Yanai

Department of Informatics, The University of Electro-Communications, Tokyo, JAPAN

Abstact

In this report we describe our approach and results for the Ad-hoc Video Search (AVS) task at TRECVID2016. We used deep convolutional neural network features extracted with the VGG-16 network, and took very simple approach just by training SVM detector for each event. As a result we achieved the 0.005 mean average precision.

1. Introduction

We started participating TRECVID in 2005, and we have been continuously submitting the results to TRECVID for ten years. For those years we usually participate in semantic indexing task (SIN) and MED tasks. Because in TRECVID2016[1] the SIN task was replaced with the AVS task, this year we participated in the Ad-hoc Video Search (AVS) task. AVS is a new and very challenging task as among the TRECVID tasks, since in the AVS task no training data is provided and instead participants require to collect training data on their own. Event queries of the AVS tasks were very complicated and the number of queries is small compared to SIN task.

For the AVS this year, we collected our training dataset using the Bing image search engine, and used an off-the-shelf feature extractor, VGG-16 fc7, and linear SVM as a classifier. As a result we achieved the 0.005 mean average precision.

2. Overview of our approach

This year we collected our datasets using the Bing image search engines serviced by Microsoft by providing words in AVS event queries as query words. We prepare query words for Bing based on the sentences of the AVS event queries by hand. That is why our runs are classified as “Manually-assisted”. We used only higher-ranked still images to train SVM event detectors. We used the ImageNet1000-pretrained VGG-16 fc7 layer features[3] to learn SVM event detectors. For every query we collected 600 positive images. For negative images we used other query images.

3. Method

For this year AVS task we took very simple approach that just collects still images from BING and extracted VGG-16 fc-7 layer features and SVM to detect events.

3.1 VGG-16

In recent years it has proved that Deep Convolutional Neural Networks is very effective for large-scale object recognition. However it needs lot of training images. In fact, one of the reasons why DCNN won the Image Net Large-Scale Visual Recognition Challenge (ILSVRC) 2013 is that the ILSVRC dataset contains one thousand training images per category. This situation does not fit common visual recognition tasks including the TRECVID AVS task. Then, to make the best use of DCNN for common image recognition tasks, Donahue et al. [2] proposed the pre-trained DCNN with the ILSVRC 1000-class dataset was used as a feature extractor. Specifically, we used VGG-16 network[3], trained on Imagenet-1000 category dataset. We used fc7 layer feature L2 normalized vector. This feature vector has 4096 dimensions.

3.2 Image retrieval and re-ranking

As we metioned earlier we collected images from Bing using query words prepared by hand based on event query phrases. We used the top 10 ranked images as positive and re-ranked each images by using SVM. For each query we collected around 400-600 images. For some event our collected images were unrelated and it was the main reason of our bad result.

3.3 SVM event detector

As event detector we used linear SVM trained on the collected event still images. When we train SVM detector we used event of interest as positive and chosen three other events used as negative. Then we used those detectors for all frames to decide either frame is related to the event or not.

4. Results

Figure 1 shows this year's AVS task result. Our results are tagged with "UEC_16". We submitted two runs. We supposed that both runs were different regarding negative training data. However, it was turned out that the both runs were the same by mistake at the time of compiling the results. The Mean Average Precision of both runs were 0.005.

This year we took really simple approach. Just training a SVM detector for every event. As you can see this approach was too simple. Also we used chosen only other three event images as negative samples, but using random images as negative samples could've lead to better results.

5. Conclusions

We proposed a very simple approach just train SVM detector for every events. One of reasons that our result's bad mean average value was our training data was not sufficient enough. In addition, we did not use any motion feature. Our next work will focus more on the gaining sufficient training dataset and train CNNs as event detector. We plan to use other existing pre-trained models such as the MIT Places.

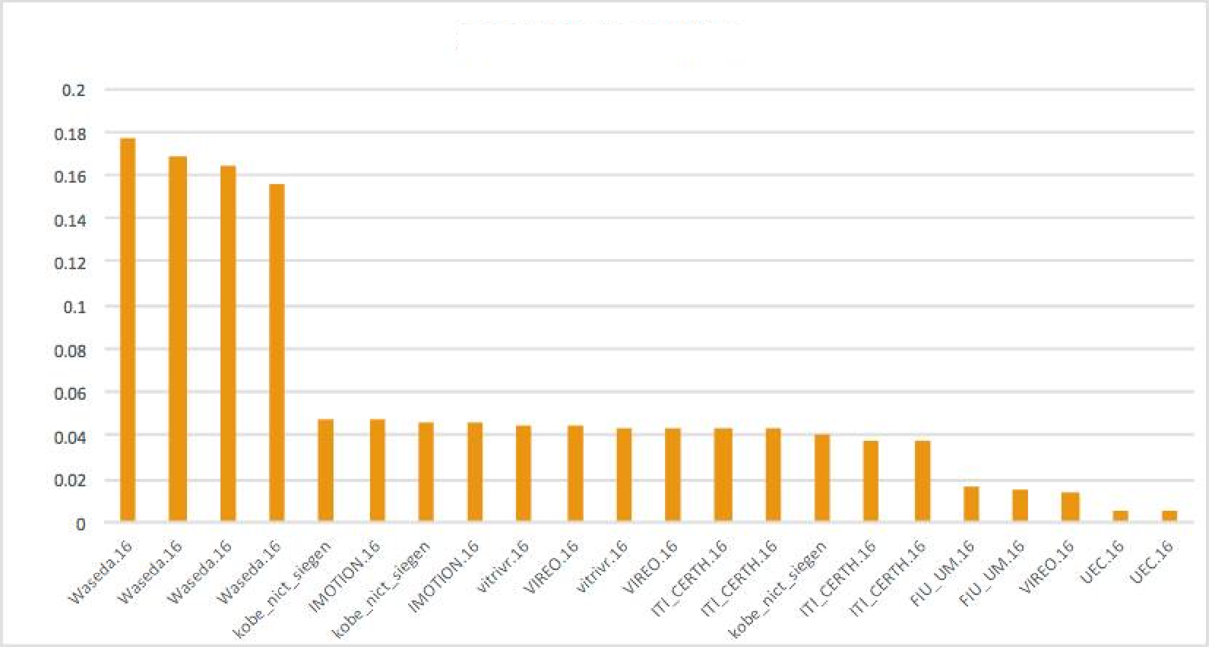

Figure 1. The comparison on the mean average precision (MAP) of all the runs submitted to the TRECVID2016 AVS task.

References

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

[2] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, (2014) DeCAF: A deep convolutional activation feature for generic visual recognition. In Proc. of International Conference on Machine Learning.

[3] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556.