UPC at ActivityNet Challenge 2016

Alberto Montes, Santiago Pascual de la Puente, Amaia Salvador, Ignasi Esquerra and Xavier Giró-i-Nieto,
Universitat Politècnica de Catalunya
{al.montes.gomez, santi.pdp}@gmail.com, {amaia.salvador, ignasi.esquerra, xavier.giro}@upc.edu

Abstract

This notebook describes our proposed solution for both the classification and detection tasks of the ActivityNet Challenge 2016. We propose a system consisting of two different stages. First, the videos are organized in 16-frame clips, for which we individually extract both audio and visual features. Visual features were extracted from a pre-trained 3D convolutional network (C3D), while MFCC coefficients were extracted for audio. On top of these features, we train a recurrent neural network to predict the activity sequence of each video at the granularity of the 16-frames clip.

1. Introduction

Recognizing activities in videos has become a hot topic over the last years due to the continuous increase of video cameras devices and online repositories. This large amount of data requires an automatic indexing to be accessed after capture. The recent advances in video coding, storage and computational resources have boosted research in the field towards new and more efficient solutions for organizing and retrieving video content.

The techniques described in this document have been tested on the video dataset defined by the ActivityNet Challenge 2016. This dataset contains 640 hours of video and 64 million frames. The ActivityNet dataset offers untrimmed videos, which means that has temporal annotations for the given ground truth class labels. Nearly half of the video hours (311 hours of video) contain a label among the 200 activity classes defined by the dataset. This dataset also give the temporal regions where activities occurs. For the details of the ActivityNet dataset please refer to the dataset description[1].

The architecture proposed is composed of two stages. First, we extract spatio-temporal features with a 3D convolutional neural network, which exploits temporal correlations in short video clips. The second stage of our proposed architecture is a Recurrent Neural Network (RNN), which exploits long term dependencies in the feature sequence. The recurrent neural network generates a sequence of predictions that naturally allows the temporal localization of the activities within a video shot.

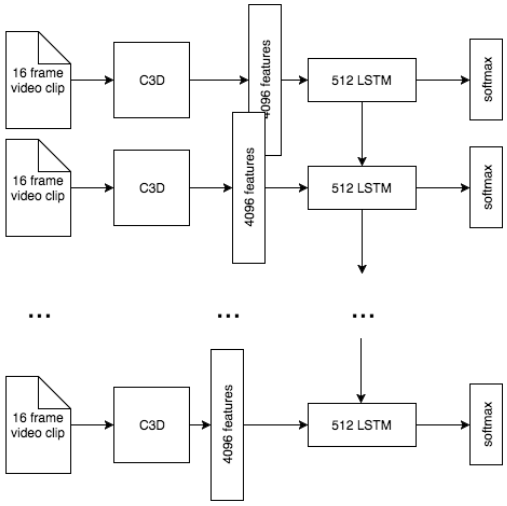

Figure 1. The proposed architecture. The network receives as input the features from the 3DS network, and trains an LSTM to output the class probability for each video clip.

2. Architecture

This section explains in detail the two stages of our proposed architecture, which is depicted in Figure 1. This architecture allows solving both the classification and detection tasks formulated in the ActivityNet challenge.

2.1. Audiovisual Feature Extraction

In order to extract spatio-temporal correlations on short clips of 16 frames, we adopted the C3D features proposed in [5], which have been proven to be well suited for video classification tasks [4] [6]. We use the network proposed in the original C3D network which was trained with the Sports1M dataset[3] and extract the features from the first fully connected layer (fc6), which was chosen based on the previous results reported in [5]. For visual feature extraction, the videos were split in clips of 16 frames each without overlap,

ending up with a total of 4 million clips. Videos clips were resized to 112x112 pixels for feature extraction, in order to match the original input size for which the C3D network was originally trained. This way, for each 16-frame clip, we obtain a visual feature of dimension 4096.

In addition to the video features, audio features were also explored as additional information for activity recognition. The audio features chosen were 40 MFCC coefficients (20 MFCC + 20 Delta-MFCC coefficients) for 20ms window length and 10ms window shift. In addition 8 Spectral coefficients for global audio track where added. The MFCC coefficients were grouped together to match video features in length and duration. The grouping of the MFCC coefficients was made computing the mean and the standard deviation. In total 88 audio features were computed. They where used in addition to the visual features in order to test if this could improve results. When used it, the audio features were concatenated to the video features out of the C3D before training the recurrent neural network.

2.2. Recurrent Neural Network

As a second stage, a Recurrent Neural Network aims at exploiting the long term dependencies in time of the extracted audiovisual features. Our RNN is based on LSTM cells, which control the flow of information that goes through them with gating mechanisms, retaining the necessary information for long periods of time, making them exploit the long-term dependencies better than classic RNNs[2]. We also proposed a sequence to sequence approach, where the model is fit with the video features as a sequence and returns a sequence of the activity class for each clip.

In addition, during our tests we explored an architecture with *feedback*, where the output predicted at the previous time step is added as an input to the LSTM. This approach aimed at smoothing the output sequence of predictions.

3. Experiments

The presented model was trained with the training partition provided by the ActivityNet challenge, and the results reported were obtained based on the predictions over the validation set.

3.1. Classification Task

For the classification task and knowing that each video has a single activity on it, we obtain the activity probabilities for the whole video as the mean of each activity output through the whole video sequence. Then, we get the maximum among all classes (excluding the background) and sort them by probability. Testing different architectures, we obtained the results given on Table 1 and Table 2. The best configuration was obtained with a single layer of LSTM with 512 neurons using only video features as input, without audio features nor feedback from the previous timestep.

Architecture	mAP	Hit@3
3 x 1024-LSTM	0.5635	0.7437
2 x 512-LSTM	0.5492	0.7364
1 x 512-LSTM	**0.5938**	**0.7576**

Table 1. Results for classification task comparing different deep architectures. All values with only video features on the validation dataset.

Features used	mAP	Hit@3
Only video	**0.5938**	**0.7576**
Video w/ audio	0.5755	0.7352
Only video & feedback	0.5210	0.6982
Video w/ audio & feedback	0.5652	0.7319

Table 2. Results for classification task with the model made by one 512-LSTM. Compare between features and feedback on the validation dataset.

α	$k = 0$	$k = 5$	$k = 10$
0.2	0.207324	**0.225138**	0.221362
0.3	0.198542	0.220776	0.221001
0.5	0.190353	0.219376	0.213029

Table 3. mAP with an IOU threshold of 0.5 over validation dataset. Here there is a comparison between values on post processing.

3.2. Detection Task

In order to solve the detection task, we post-process the output of the network with the assumption that videos only contain a single activity. This way, in this task we only focus on detecting the class with the highest probability throughout the video. To achieve this, we compute the activity probability as the sum of probabilities from all the activities except the background. A threshod α was learned and then applied along the sequence of predictions over the 16-frames clips, so that only the predictions with a probability over the threshold were considered. The best results were obtained with $\alpha = 0.2$.

Finally, a post-processing was required to improve the temporal localization of the activities. A mean filter with a window of $k = 10$ at the output of our recurrent network provided the best results, as seen in Table 3. Figure 2 shows an example of the output of our model.

References

[1] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015.

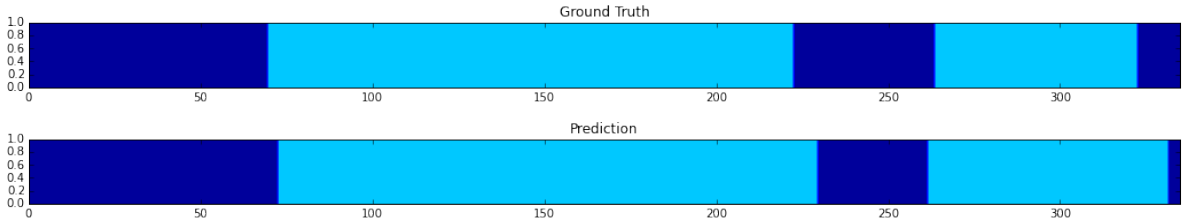

Figure 2. Example of prediction. The dark blue represents background and the light blue represents the *Rafting* activity on video `K3sJnHGHQHM`.

[2] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[3] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

[4] Z. Shou, D. Wang, and S.-F. Chang. Temporal action localization in untrimmed videos via multi-stage cnns.

[5] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. *arXiv preprint arXiv:1412.0767*, 2014.

[6] H. Zhang, M. Xua, C. Xu, and R. Jain. Modelling temporal information using discrete fourier transform for video classification. *arXiv preprint arXiv:1603.06182*, 2016.