# Supplementary Materials:Tripping through time: Efficient Localization of Activities in Videos

Meera Hahn[1]
meerahahn@gatech.edu
Asim Kadav[2]
asim@nec-labs.com
James M. Rehg[1]
rehg@gatech.edu
Hans Peter Graf[2]
hpg@nec-labs.com

[1] College of Computing
Georgia Institute of Technology
Atlanta, GA

[2] Machine Learning Department
NEC Labs America
Princeton, NJ

## 1 Datasets

We evaluate the TripNet architecture over three video datasets, Charades-STA [9], ActivityNet Captions [4] and TACoS [5]. Charades-STA was created specifically for the moment retrieval task and the other datasets were created for the video captioning task but are commonly used to evaluate the moment retrieval task. Note that we chose not to include the DiDeMo [2] dataset because in previous work, the evaluation is based off splitting the video into 21 pre-defined segments, instead of specific start and end times. This would mean changing the set of actions for our agent and we wanted the set of actions to be consistent across datasets. We do, however, compare against the method from [2] on other datasets.

All the datasets that we use contain untrimmed videos and natural language descriptions of specific moments in the videos. These language descriptions are annotated with the corresponding start and end time of the corresponding clip.

**Charades-STA [9].** This dataset takes the original Charades dataset, which contains video annotations of activities and video descriptions, and transforms these annotations to temporal sentence annotations which have a start and end time. This dataset was made for the task of temporal activity localization based on sentence descriptions. There are 13898 video to sentence pairs in the dataset. For evaluation, we use the dataset's predefined test and train splits. On average, the videos are 31 seconds long and the described temporally annotated clips are 8 seconds long.

**ActivityNet Captions [4].** In order to test the robustness of our system with longer video lengths, we use ActivityNet Captions that contains 100K temporal description annotations over 20k videos. This dataset was originally created for video captioning but is easily adaptable to our task and showcases the efficient performance of our architecture on longer videos.

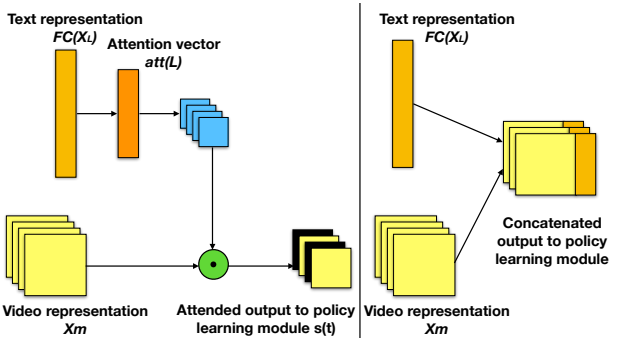


Figure 1: **This figure shows on the left TripNet-GA and on the right TripNet-Concat, where gated-attention over text features and simple concatenation are explored, respectively.**

On average, the videos are 2.5 minutes long and the described temporally annotated clips are 36 seconds long.

**TACoS [5]**. This dataset contains both activity labels and natural language descriptions, both with temporal annotations for 127 videos. Following previous work, for evaluation we randomly split the dataset into 50% for training, 25% for validation and 25% for testing. We choose this dataset because of its long videos, which are 4.5 minutes long on average and the temporally annotated clips are 5 seconds long on average.

## 2 Implementation details.

During training, we take a video and a single query sentence that has a ground truth temporal alignment in the clip. At time $t = 0$ we set the bounding window $[W_{start}, W_{end}]$ to be [0, X] where X is the average length of ground truth clips in the dataset. This means that this is the initial clip in the sequential decision process. Furthermore, it also means that the first actions selected will most likely be skipping forward in the video. The input to the system is $X$ sequential video frames and a sentence query. The sentence is first encoded through a Gated Recurrent Unit of size 256 and then through a fully-connected linear layer of size 512 with sigmoid activation. We run the video frames within the bounding window through a 3D-CNN [6] which is pre-trained on the Sports-1M dataset and extract the 5th convolution layer. The A3C reinforcement learning method is then used for the policy learning module and is trained with stochastic gradient descent (SGD) with a learning rate of .0005. The first fully-connected (FC) layer of the policy learning module is 256 dimensions and is followed by an long short term memory (LSTM) layer of size 256. During training, we set A3C to run 8 parallel threads.

In addition to comparing against prior works, we run TripNet without the gated attention mechanism, shown as TripNet-Concat in Figure 1. TripNet-Concat does self attention over the mean pooled C3D features of the video frames and concatenates the output with the Skip-Thought [3] encoding of the sentence query to produce the state representation. Testing this method allows us to explore the performance of the state processing module separately from the policy learning module.

| **Method** | A | B | C |
|---|---|---|---|
| CTRL | 44.19ms | 218.95ms | 342.12ms |
| TripNet-GA | 5.13ms | 6.23ms | 11.27ms |

Table 1: **The average time in milliseconds that it takes to localize a moment on different datasets from CTRL [1] verus our method. A is Charades-STA, B is ActivityNet Caps, and C is TACoS.**

# References

[1] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. *arXiv preprint arXiv:1705.02101*, 2017.

[2] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, pages 5803–5812, 2017.

[3] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *NIPS*, pages 3294–3302, 2015.

[4] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, 2017.

[5] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *ACL*, 1:25–36, 2013.

[6] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. C3d: generic features for video analysis. *CoRR, abs/1412.0767*, 2:7, 2014.