Weakly-Supervised Spatio-Temporally Grounding Natural Sentence in Video
--Supplementary Material

Zhenfang Chen[1*] Lin Ma[2†] Wenhan Luo[2†] Kwan-Yee K. Wong[1]

[1]The University of Hong Kong [2]Tencent AI Lab
{zfchen, kykwong@cs}.hku.hk
{forest.linma, whluo.china}@gmail.com

We provide more descriptions of baseline methods and implementation details in this supplementary material.

Baseline Details. We consider three methods to encode visual instance features $\mathbf{F}_p \in \mathbb{R}^{t_p \times d_p}$ including averaging (Avg), NetVLAD (Arandjelovic et al., 2016), and LSTM. For Avg, we simply average all the t_p segments and forward to two fully connected layers. For NetVLAD, we treat $\mathbf{F}_p$ as t_p independent d_p-dimension features and use a fully connected layer to obtain output with the desired dimension. For LSTM, we take $\mathbf{F}_p$ as a sequence features of t_p time steps and use the last hidden state of LSTM as the embedded visual representation.

For models based on DVSA, we evaluate the similarity between the spatio-temporal instance and the query sentence with cosine similarity. For models based on GroundeR, we concatenate the representations from the visual encoder and the sentence encoder as the input for the attention network and reconstruction network. For the variant of (Zhou et al., 2018), we densely predict each frame in the video to generate a spatio-temporal instance. This baseline is carefully implemented by modifying the original method (Zhou et al., 2018) with two aspects. On one hand, we replace the noun encoder with an LSTM to encode natural sentences, since we focus on grounding with natural sentences. On the other hand, we remove the frame-wise loss weighting term as it degrades the performance on the VID-sentence dataset. Such loss term is proposed to penalize the uncertainty of the existence of objects, which is not necessary as the video in our dataset contains the target instances in all frames.

The output of Avg and Net-VLAD (Arandjelovic et al., 2016) is also set as 512 by a fully connected layer. The number of centers and the dimension of cluster-center for Net-VLAD are 32 and 128, respectively.

Implementation Details. We give more details on how to generate instances from videos and extract the corresponding visual feature for each instance. We use the region proposal network from Faster-RCNN (Ren et al., 2015) to extract 30 region proposals for each video frame. The Faster-RCNN model is based on ResNet-101 (He et al., 2016) pretrained on MSCOCO (Lin et al., 2014). For the frame-level RoI pooled feature, we use the 2048-dimensional feature from the last fully connected layer of the same Faster-RCNN model. For the I3D features (Carreira and Zisserman, 2017), we use the model pretrained on Kinetics to extract the RGB sequence features I3D-RGB and the flow sequence features I3D-Flow. For every 64 consecutive frames, we extract a set of (eight) 1024-dimensional I3D-RGB features and (eight) 1024-dimensional I3D-Flow features by the output of the last average pooling layer and dropping the last temporal pooling operation. We compute optical flow with a TV-L1 algorithm (Zach et al., 2007). We crop the region proposals from the RGB images and flow images and then resize them to 224 × 224 before feeding to I3D.

References

Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. 2016. Netvlad: Cnn architecture for weakly supervised place recognition. In *CVPR*, pages 5297–5307.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 4724–4733.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian

[*] Work done while Zhenfang Chen was a Research Intern with Tencent AI Lab.

[†] Corresponding authors.

Sun. 2016. Deep residual learning for image recognition. In *CVPR*, pages 770–778.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99.

Christopher Zach, Thomas Pock, and Horst Bischof. 2007. A duality based approach for realtime tv-l 1 optical flow. In *Joint Pattern Recognition Symposium*, pages 214–223.

Luowei Zhou, Nathan Louis, and Jason J Corso. 2018. Weakly-supervised video object grounding from text by loss weighting and object interaction. *BMVC*.