Supplementary Material

Daochang Liu[1], Tingting Jiang[1], Yizhou Wang[1,2,3]
[1]NELVT, Cooperative Medianet Innovation Center, School of EECS, Peking University
[2]Peng Cheng Lab, [3]Deepwise AI Lab
{daochang, ttjiang, yizhou.wang}@pku.edu.cn

In this supplementary material, more details about the hard negative generation and model implementation are presented first. Then we report additional experimental results to further validate our model design. At last, more plots and video demos of prediction examples are attached as separate files.

1. Hard Negative Generation Details

When selecting the static frames from a training video, we first compute mean optical flow intensity in every frame and obtain an intensity sequence for the video. Then we apply thresholding on the intensity sequence so that the predefined selection ratio is met. The thresholding outputs a 1D binary mask, indicating which frames are picked. Then the mask is smoothed using a 1D dilation operation with the kernel size of two seconds. For each connected component in the mask, it is removed if the length is shorter than two seconds. Remaining selected frames are concatenated into a pseudo video. Finally, the generated video is discarded if its length is too short. Therefore, the number of generated videos does not equal the number of training videos. In practice, for efficiency, we concatenate snippet-wise features rather than raw frames. These two approaches are the same except at some boundaries.

Another strategy for action-context separation is also tried out, called *simply masking*. Instead of from the training videos, we select static frames from each testing video and get a binary mask similarly. During the test time, we directly rule out predictions on selected video frames by simply masking the output CAS. As the results shown in Table 1, this strategy makes no improvement. On the contrary, the proposed hard negative mining scheme is more flexible and handles the action-context separation effectively.

Methods	AVG (0.1:0.5)
Ours (UNT), Baseline	28.8
Ours (UNT), Baseline + Masking	28.6
Ours (UNT), Baseline + HN	32.7

Table 1. Simply masking makes no improvement. The average mAP on THUMOS'14 testing set at IoU thresholds from 0.1 to 0.5 is reported.

2. More Implementation Details

Before feature extraction, the videos are preprocessed first. Videos are sampled to 25fps and scaled to the resolution of 340x256. For the cases where multiple classes of activities occur in one video, we replicate such video several times in training set and each time it has a single category label. In practice, we stack two temporal convolutional layers sequentially in the classification and attention module to enlarge the model depth. A dropout layer is added after feature extraction. The dropout rates are 0.5 for I3D and 0.8 for UNT. The batch size is set as 32 and the weight decay is chosen as 0.001. The learning rates are set as 0.0001 for THUMOS'14 and 0.001 for ActivityNet. We set the maximum training iterations as 7500 for THUMOS'14 and 30000 for ActivityNet. At each training iteration, we loop through each video in the current training batch and accumulate the gradient to deal with the memory constraint and variable video length. When localizing action instances, the average CAS is first upsampled to the original frame rate via linear interpolation. Action instances are localized by thresholding on the upsampled CAS and then are smoothed using a 1D temporal dilation operation.

3. About ActivityNet

ActivityNet only provides Youtube URLs for videos. In our experiments, 9283, 4555 and 4655 videos are available from Youtube in the training, validation and testing set respectively for ActivityNet1.3. And 4441, 2198 and 2268 videos are accessible respectively for ActivityNet1.2.

4. More Experiments

In this section, more experimental results are presented to further evaluate our model design.

Methods	AVG (0.1:0.5)
Ours (UNT), Single + $\mathcal{L}_{mil}$ (Baseline)	28.8
Ours (UNT), Multiple + $\mathcal{L}_{mil}$	29.1
Ours (UNT), Baseline + HN	32.7
Ours (UNT), Baseline + HN (Wide)	32.9
Ours (UNT), Full (GAP)	36.3
Ours (UNT), Full (KernelSize=1)	35.1
Ours (UNT), Full	37.4

Table 2. More ablation studies on the network architecture. The average mAP on THUMOS'14 testing set at IoU thresholds from 0.1 to 0.5 is reported.

4.1. More Ablation Studies on the Architecture

Results of several following additional architecture configurations are provided in Table 2. In the second row, the multi-branch design is incorporated into the baseline without the diversity constraint. In the fourth row, we increase the filter numbers four times in the classification module of the single-branch model, so that this single-branch model has the same amount of parameters as the multi-branch one. Both two rows show no evident gain, demonstrating the gain of our multi-branch design does not result from a larger model capacity. In the fifth row, the attention module is replaced with a simple global average pooling (GAP). In the sixth row, we set the kernel size of filters in the classification module as 1, which exploits no temporal information. Both above two modifications lower the performance.

4.2. Ablation Studies on the Scoring Function

In the main paper, we devise a scoring function when localizing action instances:

$$q_i = m_{inner} - m_{outer} + \gamma \overline{p}_c \tag{1}$$

To measure the impact of each component in this function, a set of ablation studies are performed. Our full scoring function is compared to configurations with each of the following components removed: 1) both global score $\gamma \overline{p}_c$ and outer score m_{outer} 2) only global score $\gamma \overline{p}_c$ 3) only outer score m_{outer}. Results in Table 3 show that all components are important.

4.3. Experiments on Modality

Since I3D and UntrimmedNet take different forms of snippet as input, different fusion strategies should be adopted. As the results shown in Fig. 1, early-fusion works better for UntrimmedNet while late-fusion is better for I3D. UntrimmedNet takes a single frame as input in the RGB stream, which contains no motion information. Hence the performance of the single RGB stream is very low, and it is better to fuse the two streams early. In contrast, I3D takes 16-frame chunks as input in the RGB stream, which already carry temporal information. So it is preferred to fuse lately because of lower feature dimensionality and smoother training process.

Methods	AVG (0.1:0.5)
Ours (UNT), Only Inner	36.6
Ours (UNT), Full (No Global)	36.8
Ours (UNT), Full (No Outer)	37.0
Ours (UNT), Full	37.4

Table 3. Ablation studies on the scoring function. The average mAP on THUMOS'14 testing set at IoU thresholds from 0.1 to 0.5 is reported.

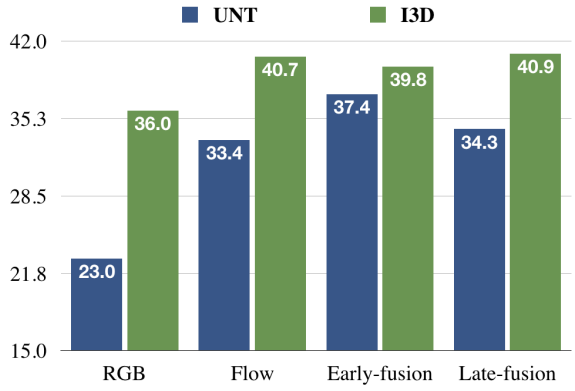

Figure 1. Experiments on modality. The average mAP on THUMOS'14 testing set at IoU thresholds from 0.1 to 0.5 is reported.

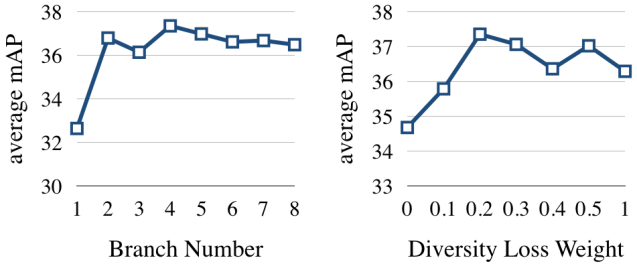

Figure 2. **Left:** Experiments on the branch number. **Right:** Experiments on the diversity loss weight. The average mAP on THUMOS'14 testing set at IoU thresholds from 0.1 to 0.5 is reported.

4.4. Class-Specific Improvement

To deeper investigate the effectiveness of handling the two challenges, class-specific average precision (AP) gains are computed. We first take the AP differences between our full model and the one without hard negative generation (Multiple + $\mathcal{L}_{sum}$ + HN *vs*. Multiple + $\mathcal{L}_{sum}$), with results shown in Fig. 3. There are evident gains in some classes such as *Billiards* and *GolfSwing*, while the performance drops greatly on several other classes such as *BaseballPitch* and *HammerThrow*. When mining the hard negatives, it is assumed that context clips are motionless. We argue that the effectiveness of hard negative mining relies on

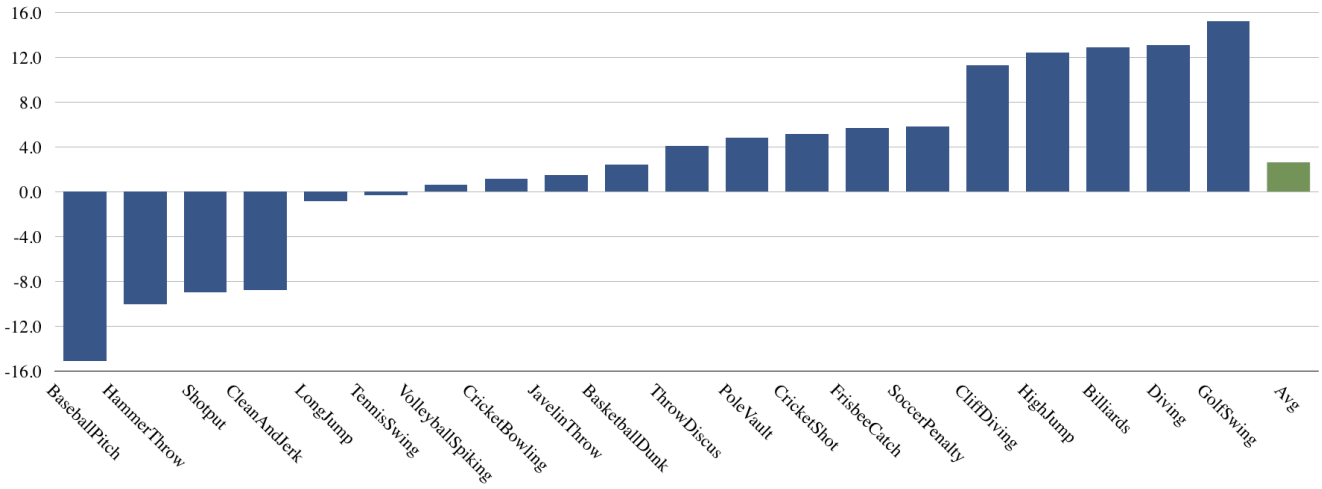

Figure 3. Class-specific gain resulting from the hard negative generation. Performance differences between our full model and the one without hard negative generation are shown. The average AP on THUMOS'14 testing set at IoU thresholds from 0.1 to 0.5 is reported.

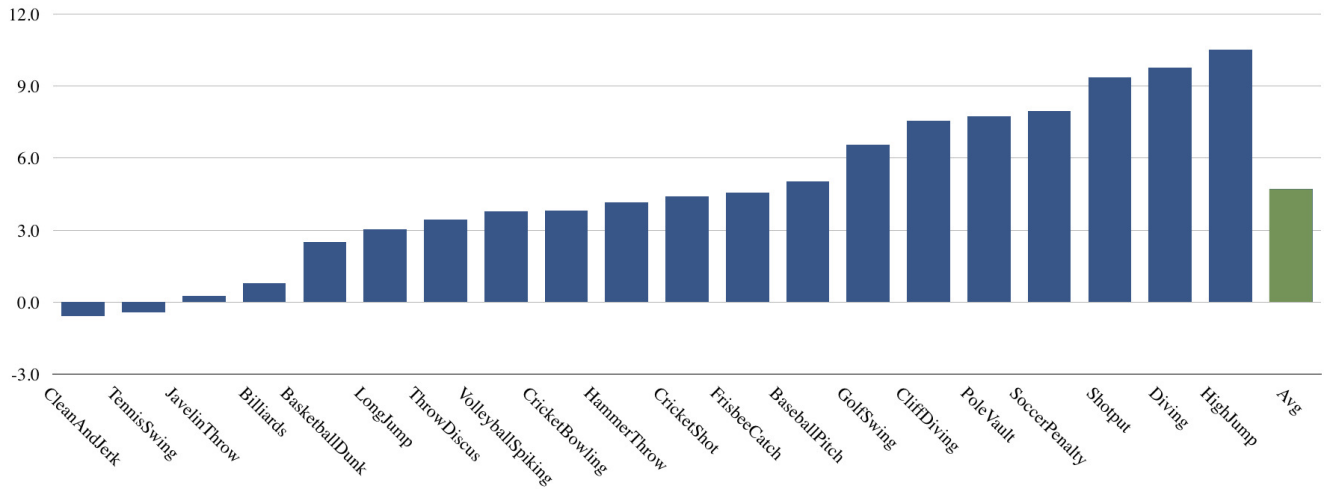

Figure 4. Class-specific gain resulting from the multi-branch design. Performance differences between our full model and the one without multi-branch design are shown. The average AP on THUMOS'14 testing set at IoU thresholds from 0.1 to 0.5 is reported.

whether this assumption holds for each action class. However, on average, the proposed method for mining hard negatives effectively promotes the mAP from 34.8 to 37.4.

Then we take the AP differences between our full model and the one without multi-branch design (Multiple + $\mathcal{L}_{sum}$ + HN *vs*. Single + $\mathcal{L}_{mil}$ + HN), with results shown in Fig. 4. The performance gains resulting from the multi-branch design are consistently evident across most categories. And on average, the multi-branch design remarkably improves the performance from 32.7 to 37.4, showing its excellence.

4.5. Others

In the main paper, we conduct comparative experiments on the branch number and the diversity loss weight on THUMOS'14 validation set. For completeness, we also post results on THUMOS'14 testing set in Fig. 2, which reflect similar trends as in the main paper.

5. More Examples

More plots and video demos of prediction examples are provided as separate files in the folders *MorePlots* and *VideoDemos* respectively. In addition, we also attach several generated hard negative videos in the folder *HN-VideoExamples*. Note that we concatenate features in practice and the attached hard negative videos are only generated for visualization.