Supplementary Material: Few-Shot Temporal Action Localization with Query Adaptive Transformer

Sauradip Nag[1,2]
s.nag@surrey.ac.uk
Xiatian Zhu [1]
xiatian.zhu@surrey.ac.uk
Tao Xiang[1,2]
t.xiang@surrey.ac.uk

[1] Centre for Vision Speech and Signal Processing (CVSSP)
University of Surrey, UK
[2] iFlyTek-Surrey Joint Research Centre on Artificial Intelligence

1 Meta-Training Algorithm

To facilitate understanding of our method we summarize the training of task-specific snippet classification in Algorithm 1.

2 Ablation Study

2.1 Qualitative Analysis

For visual analysis, we provide two qualitative examples in Figure 1. To visualize the intra-class variation challenge which our method in particular the proposed query adaptive Transformer aims to address, we show some common examples in Figure 2.

2.2 Choice of Video Embedding Layer

We examine which layer of GTAD [1] is a good choice for video snippet embedding. In particular, we test five GTAD layers. The result curve in Figure 3 shows that deeper layers are usually better than shallow ones, suggesting that snippet-level contextual information is useful for action localization. We select the layer-5 as our embedding layer as it has best cost-effectiveness.

Algorithm 1 Pseudo code for Snippet Classification

Input: Training dataset D_{base}, video embedding $\mathbb{VE}$.
Output: A query adaptive Transformer with the optimized parameters ψ^*.
Initialize params: iterations $\rightarrow N_{it}$, episodes $\rightarrow N_{eps}$, shot $\rightarrow K$, epochs $\rightarrow N_{ech}$, learning rate $\rightarrow \eta_\phi, \eta_\psi$.
Training:
for $i = 0$ to $(N_{ech} - 1)$ **do**
    for $j = 0$ to $(N_{eps} - 1)$ **do**
        $\{S, S_{label}, Q, Q_{label}\} \leftarrow Task(\mathcal{D}_{base}, K)$
        $\mathbb{F}_S \leftarrow \mathbb{VE}(S)$, $\mathbb{F}_Q \leftarrow \mathbb{VE}(Q)$

        Step 1: ....
        for $l = 0$ to $(N_{it} - 1)$ **do**
            Obtain logits: $p \leftarrow h_\phi(\mathbb{F}_S)$
            Compute loss: $\mathcal{L}_{ce}(p; \phi)$
            UPDATE $\phi : \phi_{l+1} = \phi_l - \eta_\phi \nabla_\phi \mathcal{L}_{ce}$
        end for
        $\phi^* \leftarrow \phi_{N_{it}}$
        Freeze $\phi^* \rightarrow \phi^*.detach()$

        Step 2 ....
        ADAPT ϕ^*: $\phi^{**} \leftarrow Trans(\phi^*, X^q_{se}, X^q_{se})$
        Obtain logits: p' $\leftarrow h_{\phi^{**}}(\mathbb{F}_Q)$
        Compute loss: $\mathcal{L}_{ce}(p'; \phi^{**})$
        META-UPDATE ψ: $\psi^* = \psi - \eta_\psi \nabla_\psi \mathcal{L}_{ce}$
    end for
    Save the best ψ^* during meta-training.
end for

References

[1] Mengmeng Xu, Chen Zhao, David S. Rojas, Ali Thabet, and Bernard Ghanem. G-tad: Sub-graph localization for temporal action detection. In *CVPR*, June 2020.

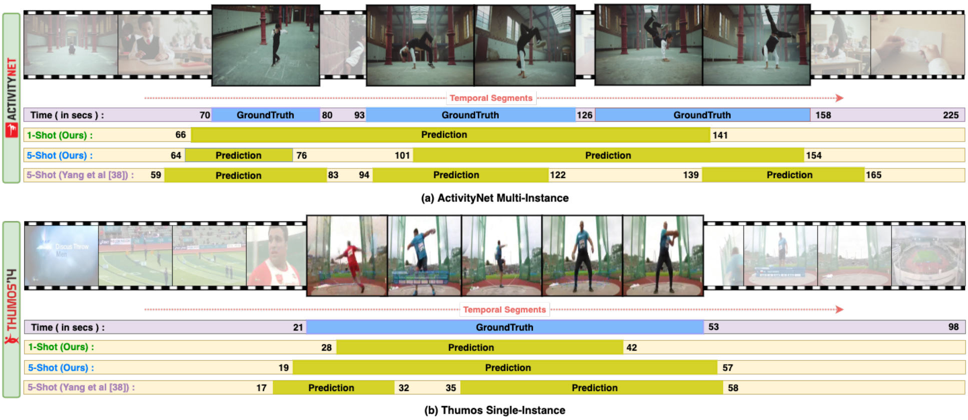

Figure 1: **Qualitative results** of (a) “BreakDancing“ class on ActivityNet and (b) “Throw Discuss“ class on THUMOS.

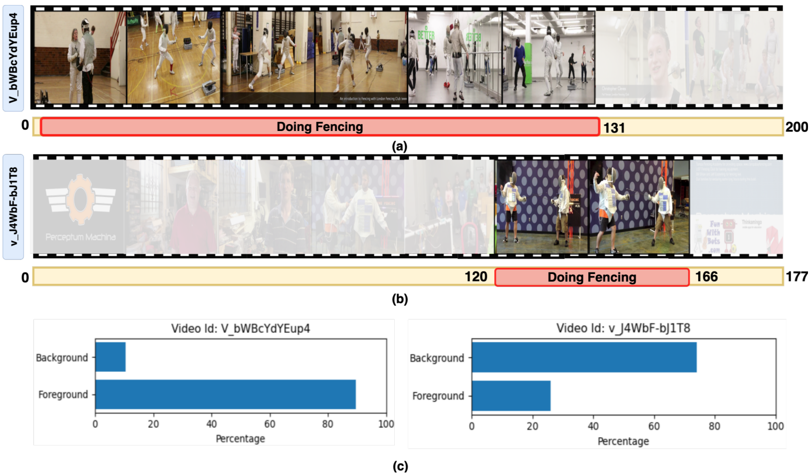

Figure 2: **Intra-class variation** example in the “Doing Fencing” class on ActivityNet-v1.3. As can be seen, the two videos present clear difference in viewpoint, scene setup, background, illumination, as well as instance length (c).

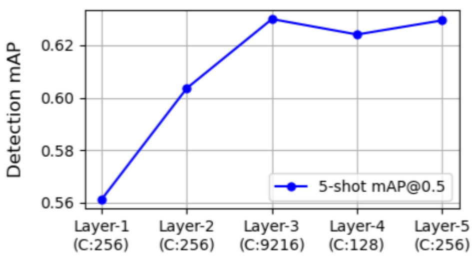

Figure 3: **Ablation of GTAD video embedding layer** in the single-instance setting on ActivityNet-v1.3. The number in round bracket is the embedding dimension.