LoSA: Long-Short-range Adapter for Scaling End-to-End Temporal Action Localization

Akshita Gupta*[1,2,3], Gaurav Mittal*[1], Ahmed Magooda[1], Ye Yu[1], Graham W. Taylor[2,3], Mei Chen[1]
[1]Microsoft, [2]University of Guelph, [3] Vector Institute for AI

S1. Additional Results

In this document, we provide additional analysis and details for our work LoSA. Section S2 provides qualitative analysis of LoSA by visualizing and comparing the action snippets localized in the videos. Section S3 provides error analysis of LoSA to highlight additional aspects of the method. Finally, Section S4 expands on the limitations and future work of the LoSA.

S2. Visualizations

In Fig S1, we provide additional visualizations of the action snippets localized by LoSA compared to the baseline of head-only transfer learning in videos from THUMOS-14 using VideoMAEv2 (ViT-g). We can observe that across all the visualizations (Fig S1a-d), LoSA is able to localize action snippets with action boundaries significantly closer to the ground truth than the baseline while also predicting the action class for the snippets more accurately than the baseline. Fig S1a shows a video of "Basketball Dunk". We can observe that, compared to head-only, LoSA is able to localize the action boundaries for "Basketball Dunk" more precisely with respect to the ground truth. We believe this is due to LoSA's ability to induce untrimmed temporal video understanding at different temporal ranges in the intermediate layers via the long-range and short-range adapters. This enhances the informativeness of the adapted features of the intermediate layers, contributing towards directly improving TAL and allows to make fine distinctions between foreground and background around action boundaries. This effect is further visible around 160 s, where LoSA correctly predicts the snippet action but head-only, due to insufficient temporal context, misclassifies the action as "Volleyball Spiking", which has similar temporal motion as "Basketball Dunk".

In Fig S2, we provide visualizations of the action snippets localized by LoSA compared to the baseline of head-only transfer learning in videos from ActivityNet-v1.3 using VideoMAEv2 (ViT-g). We can observe that across all the visualizations (Fig S2a-d), LoSA is able to localize action snippets with action boundaries significantly closer to the ground truth than the baseline. In Fig S2a, where the video shows a kid playing Hopscotch, while the baseline misses the action between 16-24s (false negative) and incorrectly predicts the background as action between 32-40s (false positive), LoSA is able to mitigate both false negative and false positive and accurately predict the start and end timestamps of the action. We believe that this is due to LoSA's ability to induce untrimmed temporal video understanding at different temporal ranges in the intermediate layers via the long-range and short-range adapters. This improves the adapted feature sequence at each intermediate layer with respect to TAL, allowing the TAL head to perform better action localization.

Figure S1. Visualizations of LoSA vs. baseline (Head-only Transfer Learning) for THUMOS-14 on VideoMAEv2 (ViT-g). Across all the visualizations (a-d), LoSA is able to localize action snippets (in green) with action boundaries significantly closer to the ground truth than the baseline, leading to fewer false positives and false negatives. LoSA also predicts the action class for the snippets more accurately than the baseline (seen by incorrect class predictions in red by the baseline in (a) and (c)).

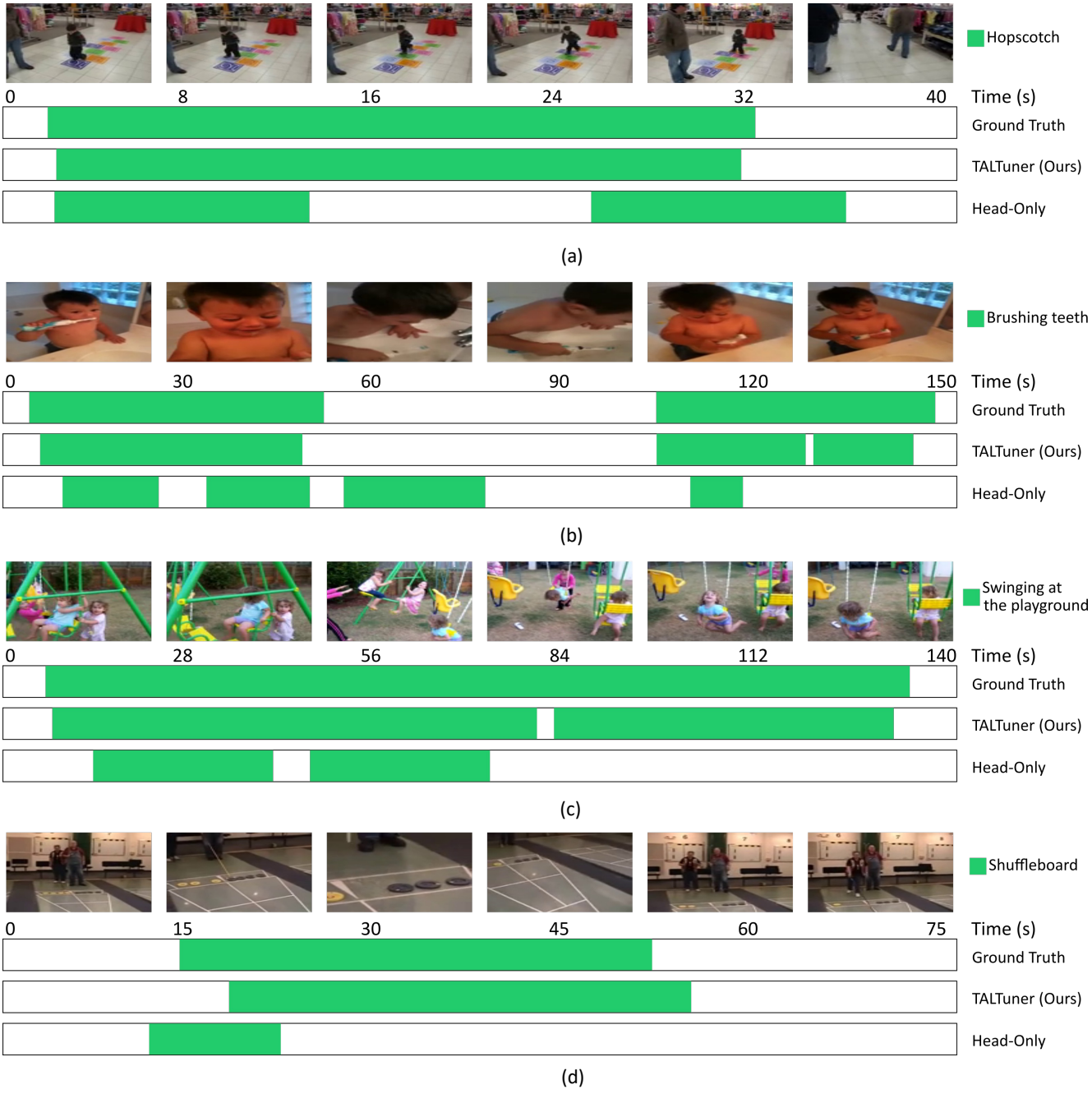

Figure S2. Visualizations of LoSA vs. baseline (Head-only Transfer Learning) for ActivityNet-v1.3 on VideoMAEv2 (ViT-g). Across all the visualizations (a-d), LoSA is able to localize action snippets (in green) with action boundaries significantly closer to the ground truth than the baseline, leading to fewer false positives and false negatives.

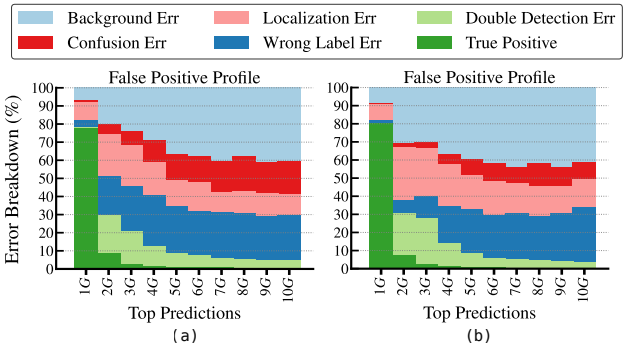

Figure S3. False positive (FP) profiling on THUMOS-14 using [**?**]. FP error breakdown for top-10 ground-truth (GT) predictions comparing (a) LoSA w/o Long-Short-range Adapter and (b) LoSA (ours). Wrong label prediction error significantly drops with LoSA compared to LoSA w/o Long-Short-range Adapter.

S3. Additional Analysis

In Fig S3, we conduct a False Positive (FP) analysis at tIoU=0.5 for THUMOS-14 using VideoMAEv2 (VIT-g). We show comparison between the baseline, LoSA w/o Long-Short-range Adapter (Fig S3a) and our method LoSA (Fig S3b). We can see a drop in the wrong label prediction error with LoSA compared to LoSA w/o Long-Short-range Adapter. This shows the significance of incorporating untrimmed temporal video understanding while adapting the intermediate layers for TAL. The chart shows FP error breakdown for top-10 ground truth (GT) predictions. For more details regarding the chart, we refer the readers to [**?**].

S4. Limitations, Negative Impact, and Future Work

To our best knowledge, we do not perceive a potential negative impact that is specific to our proposed method. While LoSA's memory-efficient design allows to leverage billion-parameter-plus models like VideoMAEv2 (ViT-g) for end-to-end TAL, the memory requirement is still linearly dependent (asymptotically) on the number of frames, frame resolution, and model size to a certain degree. In future, we can explore reducing the memory usage to sublinear while continuing to improve performance as we leverage larger foundation models. Further interesting directions include extending to end-to-end spatio-temporal localization, end-to-end video object segmentation, end-to-end video grounding, and other multi-modal video understanding tasks involving audio, text, and other modalities.

S5. Acknowledgment

Resources used in preparing this research were provided by Microsoft, and, in part, by the Province of Ontario, the Government of Canada through CIFAR, and partners of the Vector Institute.