Supplementary Materials: Video Activity Localisation with Uncertainties in Temporal Boundary

Jiabo Huang[1,4], Hailin Jin[2], Shaogang Gong[1], and Yang Liu*[3,5]

[1] Queen Mary University of London
{jiabo.huang, s.gong}@qmul.ac.uk
[2] Adobe Research
hljin@adobe.com
[3] WICT, Peking University
yangliu@pku.edu.cn
[4] Vision Semantics Limited
[5] Beijing Institute for General Artificial Intelligence

In this supplementary, we provide additional evaluations of EMB using metrics aware of uncertainty [3] in Section. 1 and ablation studies on the importances of alignment and bounding branches (Section. 2) as well as the optimisation of elastic boundary (Section. 3) for more in-depth analysis.

1 Evaluation using uncertainty-aware metrics

Beyond the standard metrics which is agnostic to the uncertainty in annotations in inference, we adopted also the uncertainty-aware metrics introduced in [3] for evaluation. Given five boundaries of each MoI annotated by different annotators, we have an 'NN' metric to evaluate predicted moments w.r.t.the nearest-neighbour reference and a 'Rep.' metric to take the most representative boundary as the ground-truth. Specifically, for the 'Rep.' metric, a 5×5 IoU matrix is computed given 5 annotations of an activity, then the boundary with the greatest average IoU to the others is considered as the ground-truth. The determined boundary (Eq. (11)) predicted by our model was tested.

As shown in Table. 3, the 'NN' performances of all methods are better than their uncertainty-agnostic counterparts (*cf.* Table. 2) while the 'Rep.' results are inferior to. The 'NN' metric provides multiple boundaries as the ground-truth of every MoIs, therefore, the model's predictions are more likely to be considered correct regardless of the biased visual-textual correlations it derived from the uncertain annotations. On the other hand, the 'Rep.' is a stricter metric that reduces the uncertainty in temporal boundary according to the agreements among different annotators. This requires models to be robust to labelling uncertainty during training, so as to learn universally interpretable video-text alignment. The consistently lower performance drops of our 'Rep.' results compared to the standard metrics in Table 2 further demonstrate the effectiveness of explicitly modelling the uncertainties in annotations during training.

* Corresponding author.

Table 3. Performance comparisons using uncertainty-aware metrics. Average recall at IoU > 0.5 with respect to the nearest neighbour reference (NN) and representative (Rep.) references. Results of other models are from [3].

Method	Charades-STA [1]		ActivityNet-Captions [2]	
	NN	Rep.	NN	Rep.
2D-TAN [5]	51.7	24.9	60.2	36.6
SCDM [4]	59.5	34.5	50.4	31.2
EMB	**64.1**	**38.6**	**61.2**	**39.9**

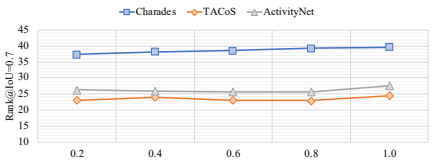

Fig. 8. An ablation study on the weight of the bounding branch.

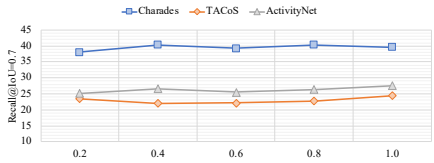

Fig. 9. An ablation study on the weight of the alignment branch.

2 Importances of branches

We further study the importance of the alignment and bounding branches to the model training by tunning the weights of the two loss terms in Eq. (14). By default, both weights are set to be 1. We then fix one of the weights and train the models with another being set to $\{0.2, 0.4, 0.6, 0.8\}$. As shown in Fig. 8 and Fig. 9, our EMB model is fairly robust to the weights of the two branches, which indicates its scalability without exhaustive parameters tunning per datasets. Besides, the best results of EMB were always obtained when the two branches are considered equally important (both weights of 1). This demonstrates the effectiveness of EMB to benefit from both the segment-wise content alignment (the spirit of proposal-based strategy) and frame-wise boundary identification (the spirit of proposal-free strategy) as well as their mutual complementary.

3 Candidate endpoints optimisation

Whilst our loss function (Eq. (10)) is formulated to optimise the sum of candidate's predicted probabilities, there are a few alternatives including optimising the most ("Max") or top-k ("Top3") confident candidates. We study the effects of them in Fig. 10 by comparing to the models learned with the fixed single manual endpoints ("Fix"). The persistent performance advantages yielded by the models learned with our elastic boundary over the ones fitting rigid temporal endpoints ("Fix") demonstrates the effectiveness of modelling explicitly the label uncertainty in the temporal boundaries, *i.e.*, the essence of our EMB model.

Moreover, the strategies optimising the predictions of multiple candidates ("Sum" and "Top3") are usually superior to the one explicitly selecting a single candidate ("Max"). This implies the potential advantages of encouraging smoother prediction distributions over single-peak ones due to the intrinsic ambiguity of video activity's temporal boundary.

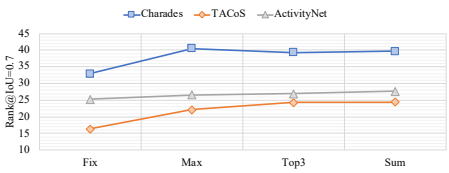

Fig. 10. Effects of different formulations for candidate endpoints optimisation.

References

1. Gao, J., Sun, C., Yang, Z., Nevatia, R.: Tall: Temporal activity localization via language query. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 5267–5275 (2017)
2. Heilbron, F.C., Escorcia, V., Ghanem, B., Niebles, J.C.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 961–970 (2015). https://doi.org/10.1109/CVPR.2015.7298698
3. Mayu Otani, Yuta Nakahima, E.R., Heikkilä, J.: Uncovering hidden challenges in query-based video moment retrieval. In: Proceedings of the British Machine Vision Conference (BMVC) (2020)
4. Yuan, Y., Ma, L., Wang, J., Liu, W., Zhu, W.: Semantic conditioned dynamic modulation for temporal sentence grounding in videos. In: Proceedings of the Conference on Neural Information Processing Systems (NeurIPS). pp. 534–544 (2019)
5. Zhang, S., Peng, H., Fu, J., Luo, J.: Learning 2d temporal adjacent networks for moment localization with natural language. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). vol. 34, pp. 12870–12877 (2020)