NMS Threshold matters for Ego4D Moment Queries

Lin Sui[1], Fangzhou Mu[2], Yin Li[2]

[1]State Key Laboratory for Novel Software Technology, Nanjing University
[2]University of Wisconsin-Madison

suilin0432@gmail.com, fmu2@wisc.edu, yin.li@wisc.edu

Abstract

This report describes our submission to the Ego4D Moment Queries Challenge 2023. Our submission extends ActionFormer [12], a latest method for temporal action localization. Our extension combines an improved ground-truth assignment strategy during training and a refined version of SoftNMS at inference time. Our solution is ranked 2nd on the public leaderboard with 26.62% average mAP and 45.69% Recall@1x at tIoU=0.5 on the test set, significantly outperforming the strong baseline from 2023 challenge. Our code is available at https://github.com/happyharrycn/actionformer_release.

1. Introduction

The Ego4D Moment Queries (MQ) task aims to localize all moments of actions in time and recognize their categories within an untrimmed egocentric video. We adopt a two-stage approach for this task, where clip-level features are first extracted from raw video frames using a pre-trained feature network, followed by a temporal localization model that predicts the onset and offset of action instances as well as their categories. Our submission last year explored the combination of a latest localization model (ActionFormer [12]) and a strong set of video features [9]. This work seeks to improve the localization model .

A limitation of the ActionFormer [12] lies in its label assignment at training time; annotated action instances are assigned to candidate moments based on *center sampling*, a heuristic that designates positive labels to moments proximal to the center of an action instance. Recent literature in object detection, however, shows that such static assignment strategy is insufficient for complex spatial configuration of objects. Inspired by this insight, we propose to adapt SimOTA [5], a dynamic label assignment strategy, for temporal action localization. SimOTA assigns ground-truth action instances to the candidate moments on the fly by solving an optimal transport problem. Further, we refine SoftNMS to account for densely overlapping actions on Ego4D.

Equipped with these modifications, our solution extends our prior work and is ranked 2nd on the public leaderboard. Specifically, our solution attains 26.62% average mAP and 45.69% Recall@1x at tIoU=0.5 on the test set, significantly outperforming the strong baseline from 2023 challenge. We hope our work will shed light on future development in temporal action localization and egocentric vision.

2. Method

Our method builds on ActionFormer [10], the state of the art for temporal action localization, yet introduces two key modifications. First, we adopt SimOTA [5], a dynamic ground truth assignment strategy in training. Second, we flatten the Gaussian penalty function in SoftNMS [2] to account for ground truth moments with significant overlap. We now present the key components of our method.

Moment Localization with ActionFormer. ActionFormer [12] takes as input a 1D sequence of clip-level video features and builds a feature pyramid using local self-attentions. This pyramid serves as a multi-scale representation of moment candidates. Each location in the pyramid defines the center of a moment, whose temporal scale is determined by the pyramid level (*i.e.*, longer moments reside on higher levels of the pyramid). A classification head subsequently assigns a confidence score to each moment, whereas a regression head predicts the distances from the center of a moment to its onset and offset. These predictions are decoded into action segments and further combined using SoftNMS [2]. We refer readers to [12] for more details on ActionFormer.

Dynamic Label Assignment. ActionFormer follows a *fixed* set of rules collectively known as center sampling to convert ground-truth action segments into point-wise classification labels. Center sampling was first seen in the literature of single-stage object detection [11], in which it has lately been superseded by more powerful, *dynamic* label assignment strategies [4, 5] that evolve in tandem with training
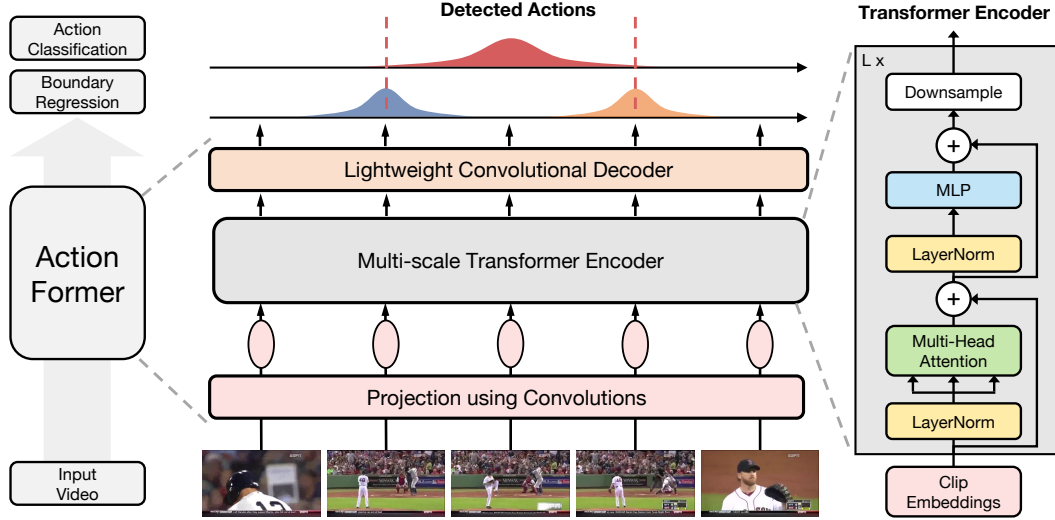
Figure 1: **Overview of ActionFormer.** Taken from [12].

losses. In this work, we report a similar finding that training ActionFormer with SimOTA [5], an efficient dynamic label assignment technique, yields a small yet significant performance gain compared to using center sampling. We provide ablation results in Section 3 and refer readers to [5] for more details on SimOTA.

Dealing with Near-Replicates. Our initial exploratory analysis revealed that 15% of ground-truth moments in the Ego4D-MQ dataset are *near-replicates* (*i.e.*, $\geq 90\%$ overlap with another moment). This presents a unique challenge to ActionFormer, which relies on aggressive non-maximum suppression (NMS) to reduce highly overlapping predictions, thereby harming precision at high recall in the presence of near-replicates. In this work, we propose to tune the standard deviation σ of the Gaussian penalty function f in SoftNMS [2] as a simple fix. Intuitively, a small σ as recommended by [2] yields a peaky f that incurs strong penalty on near-replicates, whereas f flattens out as σ increases, leaving near-replicates less affected. We empirically found that setting σ to the unusually large value of 2 brought a significant improvement in mAP by 1.8 absolute percentage point compared to using the default value of 0.9 as in [10].

3. Experiments and Results

We now present our experiments and results.

Evaluation Protocol and Metrics. We follow the official *train*/*val*/*test* splits for evaluation. Our model is trained on the *train* split when results are reported on the *val* split, and is trained on the *train* and *val* splits combined when results are submitted for final evaluation on the *test* split. In line with the official guideline, we adopt average mAP as our main evaluation metric.

Implementation Details. We use pre-extracted Omnivore [6], EgoVLP [7] and InternVideo [3] features from raw videos as input to ActionFormer, and set the embedding dimension throughout the model to 1152. Following the official code release, we train ActionFormer using the AdamW optimizer [8] for 15 epochs with a mini-batch size of 2, a learning rate of 0.0001 and a weight decay of 0.05.

Results. Our results on the *val* and *test* splits are summarized in Table 1. Replacing the official SlowFast features with InternVideo features brings a notable 2.71 absolute-percentage-point improvement on average mAP on the *val* split. This highlights the strength of InternVideo as a video foundation model for representation learning. The introduction of dynamic label assignment further boosts the average mAP by a small yet significant 0.3 absolute percentage point. Finally, the largest performance gain (>1.3 absolute percentage points on both *val* and *test* splits) is attained by tuning the spread of Gaussian penalty function in SoftNMS to account for near-replicates. With everything combined, ActionFormer reaches an average mAP of 26.07% on the *val* split and 26.62% on the *test* split. Figure 2 provide visualizations of model predictions.

Limitations. We present false negative analysis (Figure 3), false positive analysis (Figure 4) and sensitivity analysis (Figure 5) of our method. Our method demonstrates stronger performance on videos with high moment coverage and is more capable of identifying longer actions. It exhibits significantly higher error rate on videos with low moment coverage

Split	Features	SimOTA	SoftNMS σ	average mAP
Val	SlowFast + Omnivore + EgoVLP		0.9	21.40
	InternVideo + Omnivore + EgoVLP		0.9	24.11
	InternVideo + Omnivore + EgoVLP	✓	0.9	24.41
	InternVideo + Omnivore + EgoVLP	✓	1.5	25.71
	InternVideo + Omnivore + EgoVLP	✓	2.0	**26.07**
	InternVideo + Omnivore + EgoVLP	✓	4.0	25.76
Test	InternVideo + Omnivore + EgoVLP	✓	0.9	25.33
	InternVideo + Omnivore + EgoVLP	✓	2.0	**26.62**

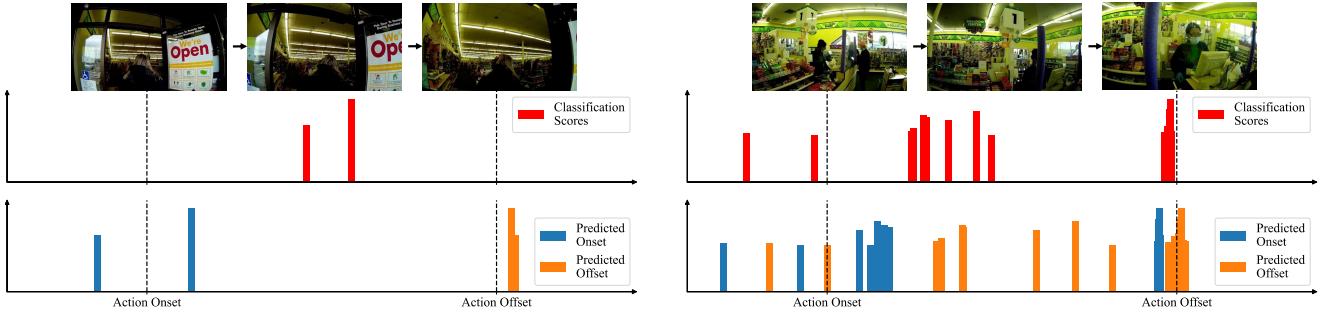Table 1: **Results on Ego4D Moment Queries dataset.**

Figure 2: **Result Visualizations.** From *top* to *bottom*: (1) input video frames; (2) action scores at each time step; (3) histogram of action onsets and offsets computed by weighting the regression outputs using action scores. *Left*: a success case for ActionFormer. *Right*: a failure case with multiple center regions and wrong onset/offset regression.
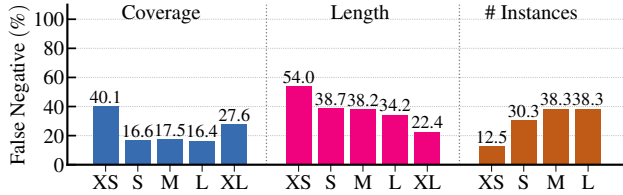

Figure 3: **False Negative (FN) Analysis with respect to different moment characteristics.** Our method tends to miss short actions as well as actions in videos with short moment coverage. It also exhibits higher FN rate on videos with multiple actions, possibly due to near-replicates.

and was not able to accurately localize short actions. This is further manifested by the surprisingly high background error rate.

4. Conclusion

In this report, we described our solution to the Ego4D Moment Queries Challenge 2023. Our solution is based on ActionFormer, yet introduces two key modifications to training and post-processing that brought substantial performance gain without change to the model architecture. We provided extensive analysis of our results which highlights
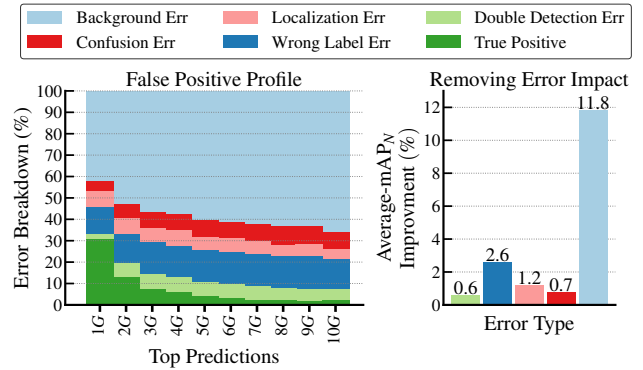

Figure 4: **False Positive (FP) Analysis using DETAD [1].** *Left*: FP error breakdown when considering the predictions for the top-10 ground-truth (G) instances. *Right*: The impact of error types. Background error and wrong label error are the top two error types.

the strength and weakness of our approach. We hope our solution and results can offer new insights to the Ego4D MQ task.
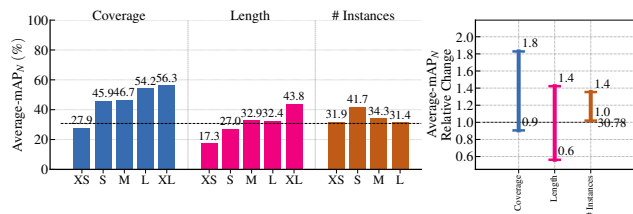
Figure 5: **Sensitivity Analysis with respect to different moment characteristics.** *Left:* Normalized mAP at tIoU=0.5. Our method performs better on videos with high moment coverage and is more capable of detecting long actions. *Right:* The relative normalized mAP change at tIoU=0.5. Performance of our method is most sensitive to moment coverage and duration.

References

[1] Humam Alwassel, Fabian Caba Heilbron, Victor Escorcia, and Bernard Ghanem. Diagnosing error in temporal action detectors. In *Eur. Conf. Comput. Vis.*, pages 256–272, 2018. 3

[2] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-NMS–improving object detection with one line of code. In *ICCV*, 2017. 1, 2

[3] Guo Chen, Sen Xing, Zhe Chen, Yi Wang, Kunchang Li, Yizhuo Li, Yi Liu, Jiahao Wang, Yin-Dong Zheng, Bingkun Huang, et al. InternVideo-Ego4D: A pack of champion solutions to Ego4D challenges. *arXiv preprint arXiv:2211.09529*, 2022. 2

[4] Zheng Ge, Songtao Liu, Zeming Li, Osamu Yoshie, and Jian Sun. Ota: Optimal transport assignment for object detection. In *CVPR*, 2021. 1

[5] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 1, 2

[6] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *CVPR*, 2022. 2

[7] Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. In *NeurIPS*, 2022. 2

[8] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 2

[9] Fangzhou Mu, Sicheng Mo, Gillian Wang, and Yin Li. Where a strong backbone meets strong features–actionformer for ego4d moment queries challenge. *arXiv preprint arXiv:2211.09074*, 2022. 1

[10] Fangzhou Mu, Sicheng Mo, Gillian Wang, and Yin Li. Where a strong backbone meets strong features–actionformer for ego4d moment queries challenge. *arXiv preprint arXiv:2211.09074*, 2022. 1, 2

[11] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *ICCV*, 2019. 1

[12] Chen-Lin Zhang, Jianxin Wu, and Yin Li. ActionFormer: Localizing moments of actions with transformers. In *ECCV*, 2022. 1, 2