TVNet: Temporal Voting Network for Action Localization

1、Motivation

在localization中引入voting的概念。

2、Approach

overview：

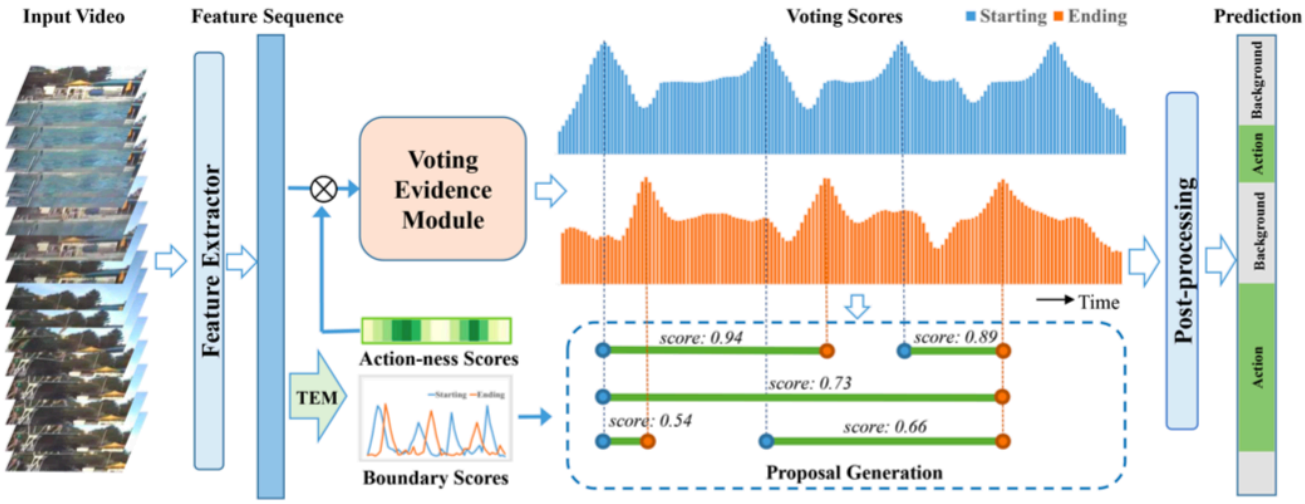


Figure 1: Overview of our proposed TVNet. Given an untrimmed video, frame-level features are extracted. Our main contribution, the Voting Evidence Module, takes in this feature sequence and outputs sequences of starting and ending voting scores. Local maxima in these voting scores are combined to form action proposals, which are then scored and classified.

2.1 Problem definition

希望模型能够对动作的boundary进行预测，同时对划分的动作进行classification。

2.2 Voting Evidence Module（VEM）

首先作者提出了sign distance的概念。具体描述为：

Assume a frame is 5 frames past the start of the action. We denote this relative dis- tance ‘-5’, indicating the start of the action is 5 frames ago. In contrast, when a frame precedes the start of the action by say 2 frames, we denote ‘+2’.

如果一个模型能够很好地进行动作定位，那么它也应该能够比较好地预测出对应帧与前后动作边界的sign distance。为了学习这种距离计算机制，作者设计了VEM。

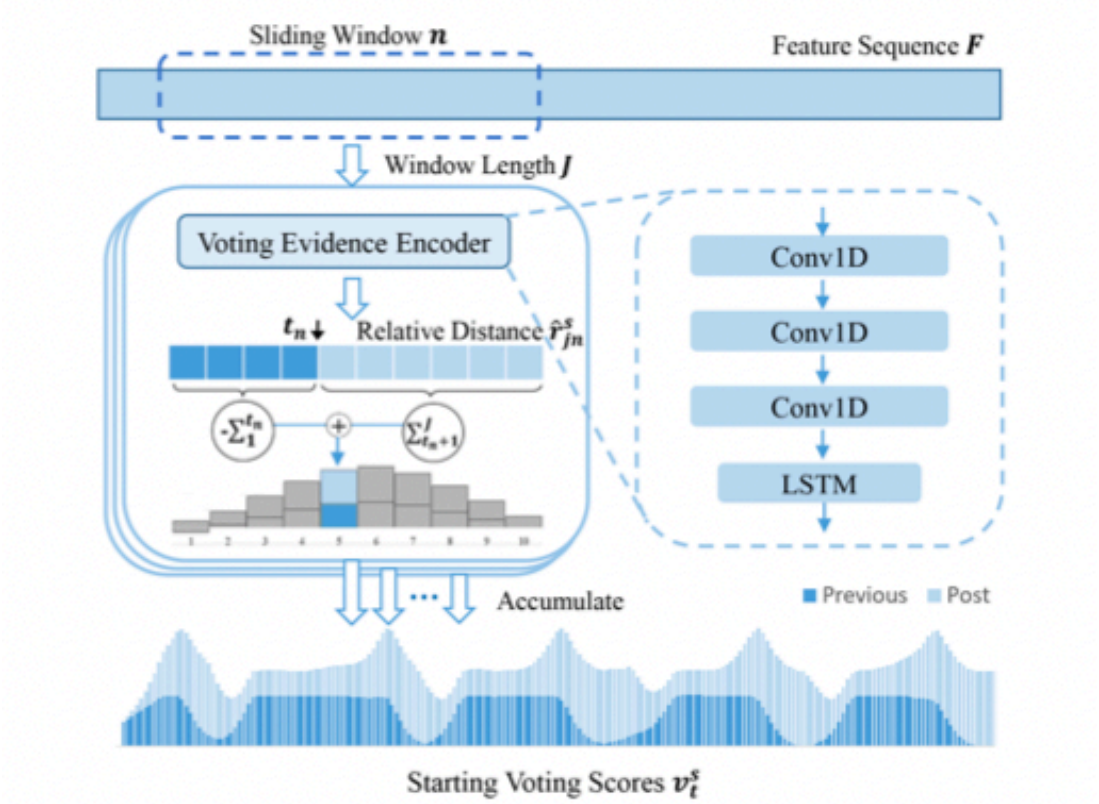


Figure 2: Illustration of our proposed Voting Evidence Module. We accumulate evidence from all frames to calculate boundary scores for starting.

①Voting evidence encoder

对于由视频得到的特征序列F，使用一个长度为J的滑动窗口对每一帧对应的特征进行编码。通过三层Conv1D和LSTM，每个滑动窗口得到输出$\hat{R} = \{(\hat{r}^s_j, \hat{r}^e_j)\}^J_{j=1}$，表示窗口中的每个帧j与**最近**的start/end frame之间的相对距离。关于relative distance的理解可以参考下图：

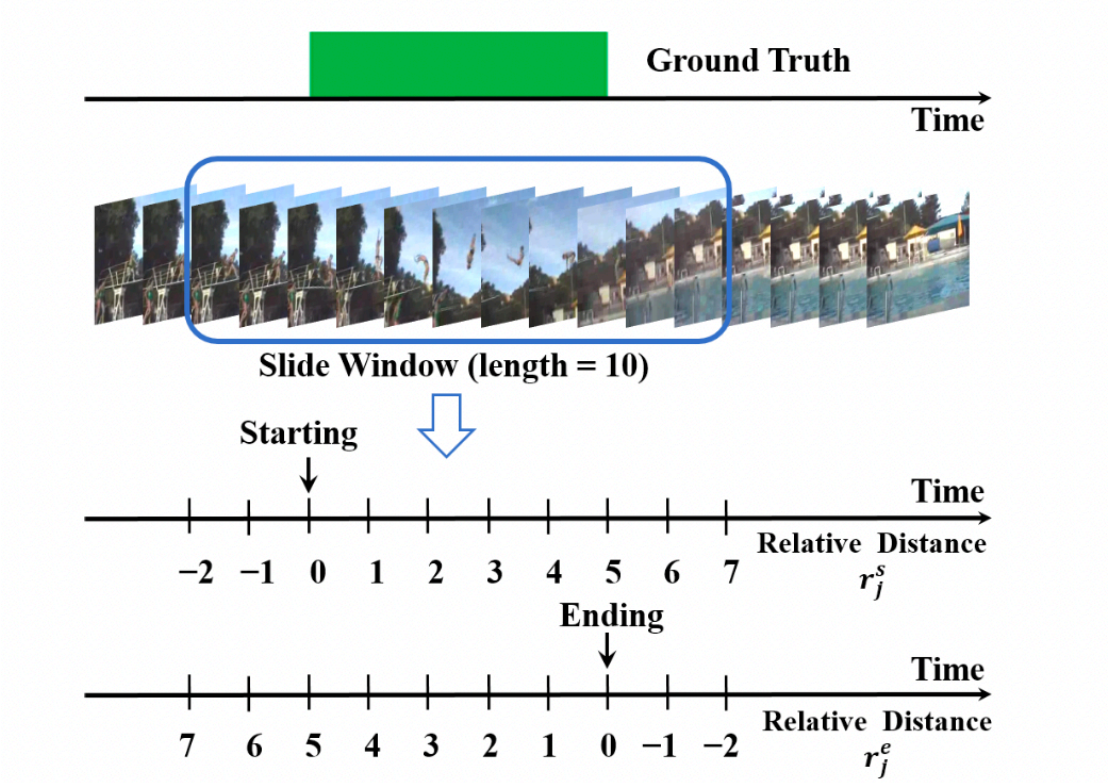


Figure 3: Supervision for relative distances of starting and ending action boundaries within a sliding window, from ground truth.

在训练的时候，用ground truth生成的distance $R = \{(r_j^s, r_j^e)\}_{j=1}^J$来指导encoder的学习。其中，$r_j^s = j - s^*, r_j^e = e^* - j$。将relative distance归一化到-1~1之间后用MES Loss来更新参数：

$$L^s = \frac{1}{J}\sum_{j=1}^{J}(\hat{r}_j^s - r_j^s)^2$$

$$L^e = \frac{1}{J}\sum_{j=1}^{J}(\hat{r}_j^e - r_j^e)^2$$

②Voting accumulation

计算到每个window对应每一帧的distance后，需要将信息进行整合、

$$v_t^s = \sum_{n=1}^{N}(\sum_{j=1}^{t_n}(-\hat{r}_{jn}^s) + \sum_{j=t_n+1}^{J}\hat{r}_{jn}^s)$$

$$v_t^e = \sum_{n=1}^{N}(\sum_{j=1}^{t_n}\hat{r}_{jn}^e + \sum_{j=t_n+1}^{J}(-\hat{r}_{jn}^e))$$

v表示voting score，其计算的是在涵盖某一位置的所有滑动窗口的relative distance关系。越高表示越可能是start或end point。可以将每个部分的score整合起来得到$V^s = \{v_t^s\}_{t=1}^T$和$V^e = \{v_t^e\}_{t=1}^T$。进一步可以得到porposals。

2.3 Proposal getneration and post-processing

①proposal generation

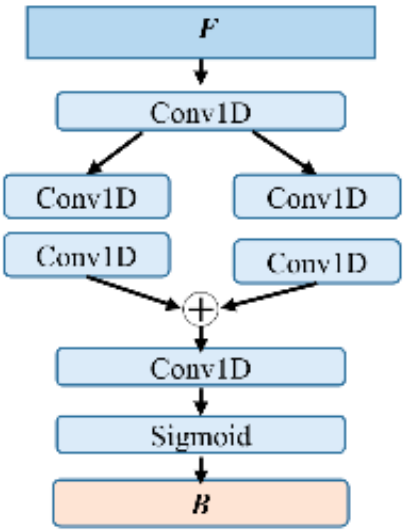
首先预定一个分数阈值ξ，将分数在这个阈值之上的局部最高点作为可能的start point和end point。将时间长度小于预定的时间阈值τ的start point和end point之间的区间作为proposal。得到proposal后，用每个proposal中包含的特征，在前人提出的Proposal Evaluation Module中计算直接置信度p($\hat{s}$,$\hat{e}$)

参考文献：<Lin, T., Liu, X., Li, X., Ding, E., and Wen, S. (2019). BMN: Boundary-matching network for temporal ac- tion proposal generation. In *International Conference on Computer Vision*.>

②proposal confidence scores

作者在前人的基础上改进了Temporal Evaluation Module。该模块以从视频中提取到的特征作为输入，输出的是基本starting score（B^s）、基础ending score（B^e）和actionness score（B^a）。

比较好奇为什么可以从某一帧来判断动作?



根据模块的输出，结合前面得到的porposal得分V和p，计算每个proposal的置信度：

$$\hat{c} = (v^s_{\hat{s}} + \alpha b^s_{\hat{s}})(v^e_{\hat{e}} + \alpha b^e_{\hat{e}})p(\hat{s}, \hat{e})$$

③redundant proposal suppression

使用Soft-NMS移除冗余的proposals。

④classification

对每个候选的proposal获取类标，这部分应该是比较简单，作者也没有做很多阐述。

3、Experiment

dataset：ActivityNet1.3、THUMOS14

3.1 Main results

Table 1: Action localization results on ActivityNet-1.3 and THUMOS14. **Bold** for best model and underline for second best.

Method	Publication	ActivityNet-1.3 (mAP@IoU)				THUMOS14 (mAP@IoU)				
		0.5	0.75	0.95	Average	0.3	0.4	0.5	0.6	0.7
SSN (Zhao et al., 2017)	ICCV2017	43.26	28.70	5.63	28.28	51.9	41.0	29.8	-	-
TAL-Net (Chao et al., 2018)	CVPR 2018	38.23	18.30	1.30	20.22	53.2	48.5	42.8	33.8	20.8
BSN (Lin et al., 2018)	ECCV 2018	46.45	29.96	8.02	30.03	53.5	45.0	36.9	28.4	20.0
BMN (Lin et al., 2019)	ICCV 2019	50.07	34.78	8.29	33.85	56.0	47.4	38.8	29.7	20.5
MGG (Liu et al., 2019)	CVPR 2019	-	-	-	-	53.9	46.8	37.4	29.5	21.3
GTAN (Long et al., 2019)	CVPR 2019	**52.61**	34.14	8.91	34.31	57.8	47.2	38.8	-	-
G-TAD (Xu et al., 2020)	CVPR 2020	50.36	34.60	9.02	34.09	54.5	47.6	40.2	30.8	23.4
BC-GNN (Bai et al., 2020)	ECCV 2020	50.56	34.75	9.37	34.26	57.1	49.1	40.4	31.2	23.1
BSN++ (Su et al., 2021)	AAAI 2021	51.27	**35.70**	8.33	**34.88**	59.9	49.5	41.3	31.9	22.8
TVNet	-	51.35	34.96	**10.12**	34.60	**64.7**	**58.0**	**49.3**	**38.2**	**26.4**

Table 2: Action localization results on THUMOS14 for methods combined with proposal-to-proposal relations from PGCN (Zeng et al., 2019) and MUSES (Liu et al., 2021). **Bold** for best and underline for second best.

Method	Publication	THUMOS14 (mAP@IoU)				
		0.3	0.4	0.5	0.6	0.7
BSN + PGCN (Zeng et al., 2019)	ICCV 2019	63.6	57.8	49.1	-	-
Uty + PGCN (Chen et al., 2020)	BMVC 2020	66.3	59.8	50.4	37.5	23.5
G-TAD + PGCN(Xu et al., 2020)	CVPR 2020	66.4	60.4	51.6	37.6	22.9
TVNet + PGCN	-	**68.3**	**63.7**	**56.0**	**39.9**	**24.2**
BSN + MUSES (Liu et al., 2021)	CVPR 2021	68.9	64.0	56.9	46.3	31.0
TVNet + MUSES (Liu et al., 2021)	-	**71.1**	**66.4**	**59.1**	**47.8**	**32.1**

上图说明了TVNet的优越。下面一幅图关于proposal-to-proposal的实验没太理解，可能需要阅读一下P-GCN那篇文章。

3.2 quanlitative results

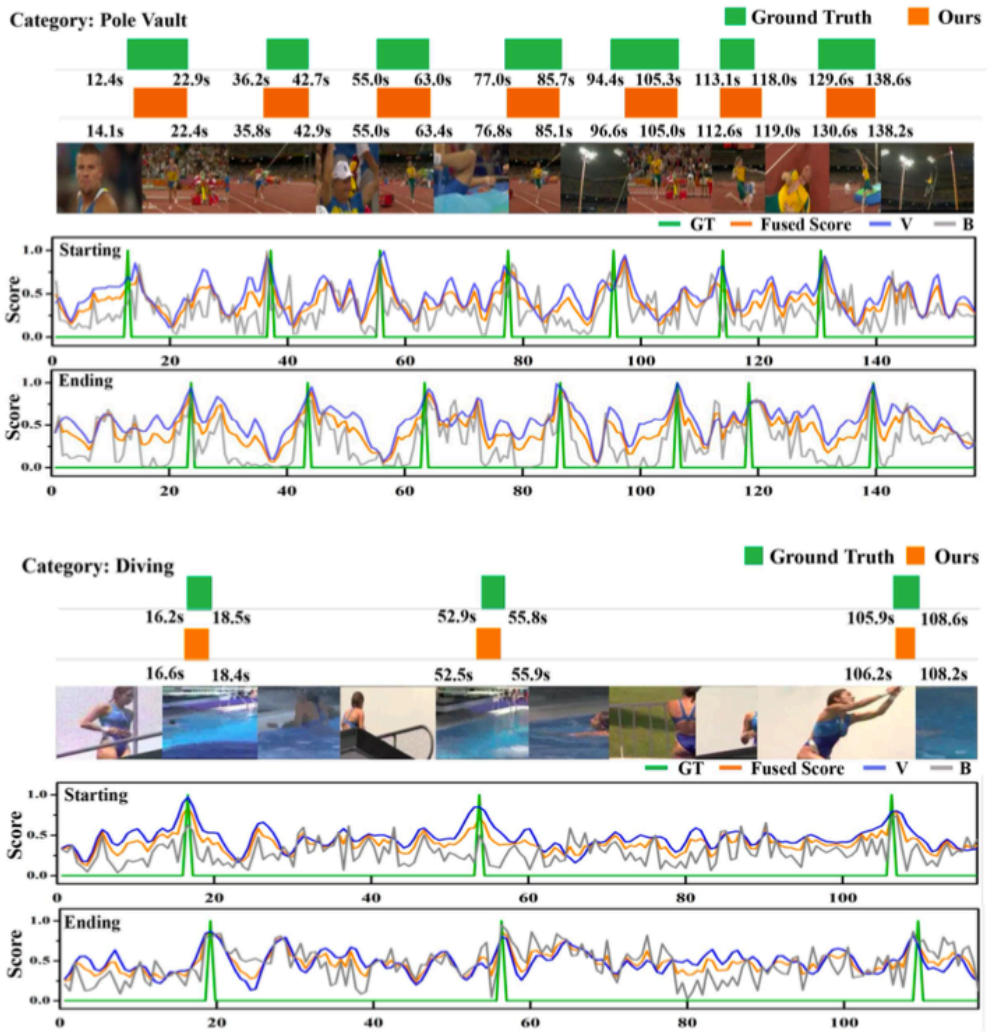


Figure 5: Qualitative results on THUMOS14, where TVNet detects multiple dense (top) and sparse (bottom) actions with accurate boundaries. The green bars indicate ground truth instances and the orange bars indicate TVNet detections. The green, orange, blue and grey lines are ground truth boundaries, weighted boundaries scores, voting scores and boundary scores respectively.

在密集和稀疏的动作视频上验证了效果。

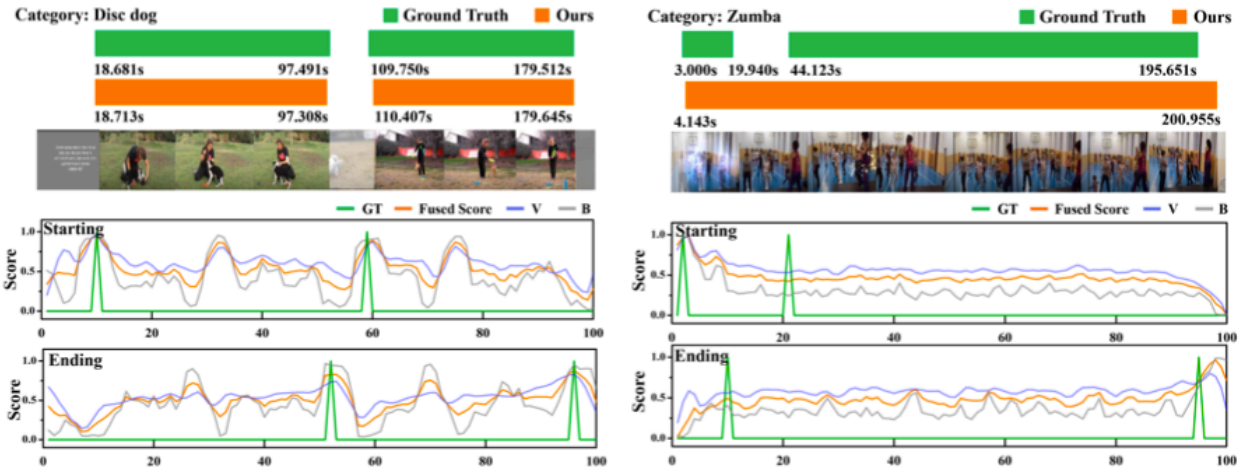


Figure 6: Qualitative TVNet results on ActivityNet. Left: a success case, where the actions are detected with boundaries closely matching the ground truth. Right: a failure case, where the wrong start/end times are matched forming one long action.

success case和failure case的对比。文章好像缺少了对这种失败现象的解释。只是说“起码起始时间和结束时间是对的”。

3.3 Ablation study

①Effectiveness of TEM

Table 3: The effect of actionness scores (B^a) and boundary scores (B^s, B^e) on ActivityNet-1.3. The top row indicates neither, which equates to the model with the TEM removed.

B^a	B^s/B^e	mAP@IoU			
		0.5	0.75	0.95	Average
✗	✗	50.45	34.26	7.97	33.55
✓	✗	51.30	34.89	8.90	34.40
✗	✓	51.14	34.74	9.68	34.35
✓	✓	**51.35**	**34.96**	**10.12**	**34.60**

②voting and boundary scores

Table 4: The effect of different combinations of boundary score (B), voting scores (V) and proposal generation (G) based on V on ActivityNet-1.3. All results are from our implementation, apart from *, which denotes the original B from (Lin et al., 2019), and can be considered as TVNet without the VEM.

B	G	V	mAP@IoU			
			0.5	0.75	0.95	Average
✓*	✗	✗	50.13	33.18	9.50	33.15
✓	✗	✗	50.64	34.30	8.93	33.84
✓	✓	✗	50.68	34.17	9.73	33.87
✗	✓	✓	51.30	34.89	8.90	34.40
✓	✓	✓	**51.35**	**34.96**	**10.12**	**34.60**

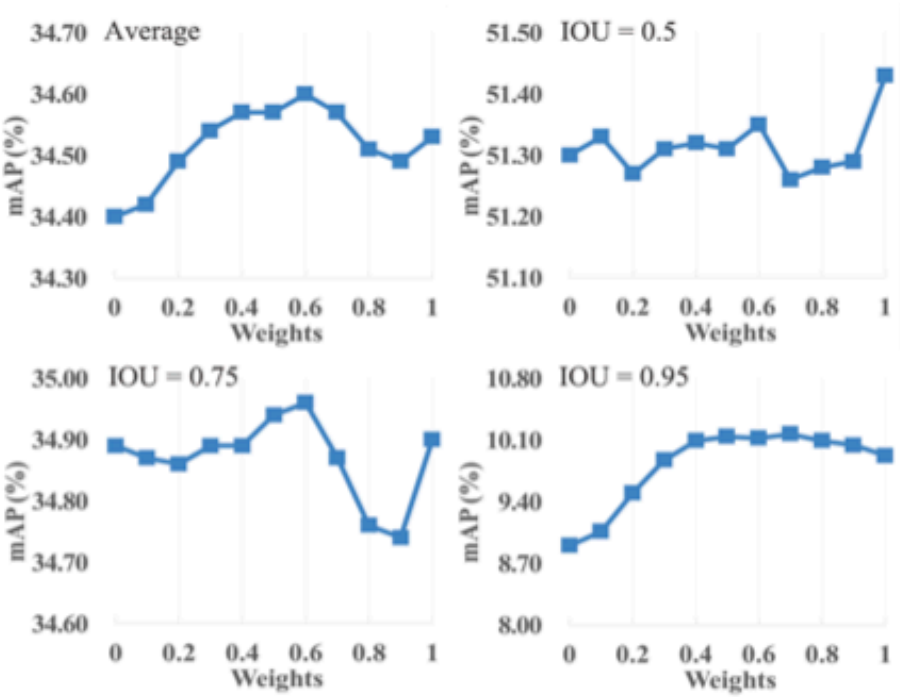


Figure 7: Performance of different weights used to fusion on ActivityNet1.3.

后面还有一些超参的消融实验，在这里不做过多阐述。