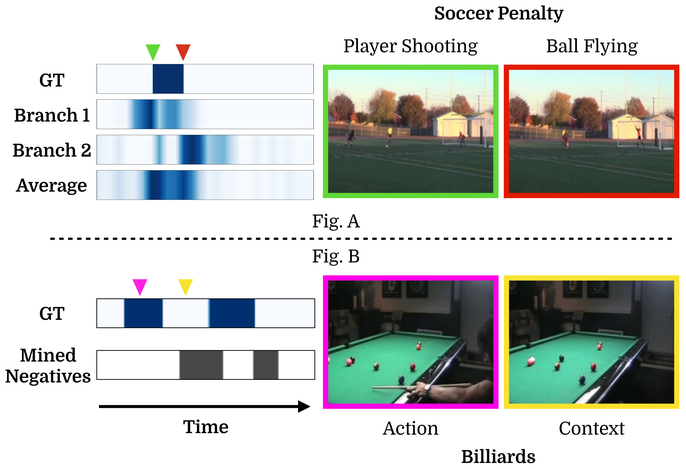
Completeness Modeling and Context Separation for Weakly Supervised Temporal Action Localization

Daochang Liu[1], Tingting Jiang[1], Yizhou Wang[1,2,3]
[1]Peking University, [2]Peng Cheng Lab, [3]Deepwise AI Lab
{daochang, ttjiang, yizhou.wang}@pku.edu.cn

Introduction

- **Temporal action localization**
  - **Input:** Untrimmed video
  - **Output:** Action instances (start, end, class, score)
- **Weakly supervised:**
  - Only video-level action label for training
  - No start and end time

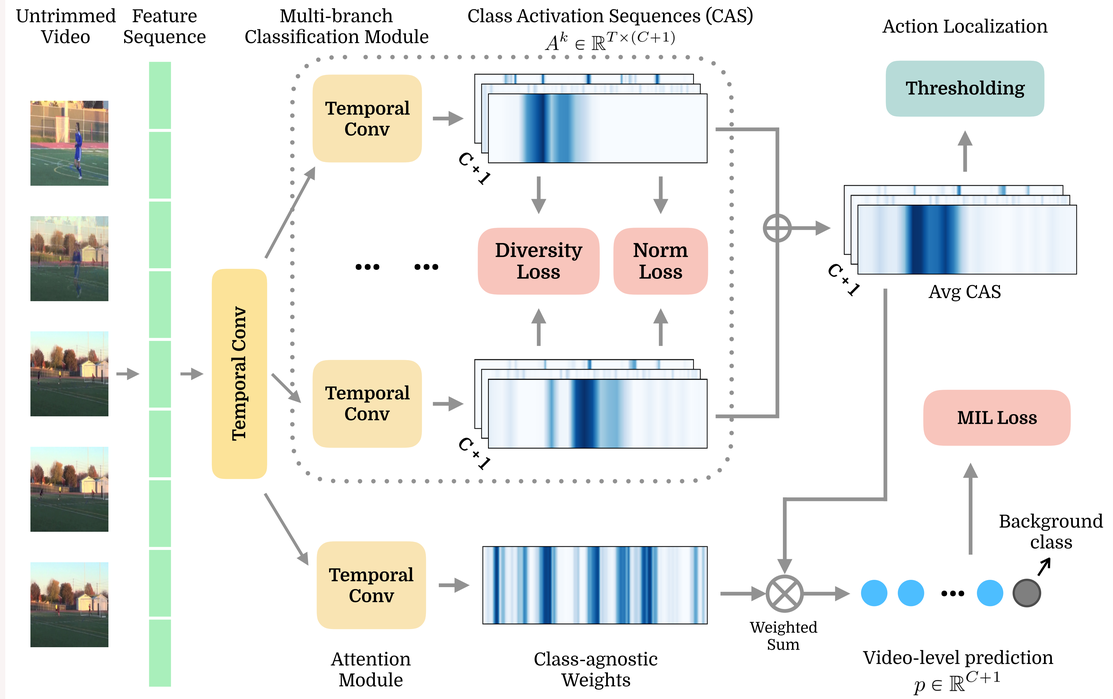
- **Two challenges posed by the weak supervision:**
  - **Completeness modeling:** How to locate not only the discriminative part but also the complete action (Fig. A)
  - **Context Separation:** How to distinguish the true action instance from its co-occurring context clips (Fig. B)



- **Solutions:**
  - A multi-branch neural network with diversity loss to find multiple action parts and therefore the full action instance.
  - Generate hard negatives using the motion prior (static clips are unlikely to be actions).

Scan for paper

Model



Diversity Loss
Enforcing branches to discover different parts

$$\mathcal{L}_{div} = \frac{1}{Z}\sum_{c=1}^{C+1}\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}\frac{A^i{}_{*,c}\cdot A^j{}_{*,c}}{\|A^i{}_{*,c}\|\|A^j{}_{*,c}\|}$$

Norm Loss
Balancing the strength of branches

$$\mathcal{L}_{norm} = \frac{1}{K(C+1)}\sum_{c=1}^{C+1}\sum_{i=1}^{K}\left|\|A^i{}_{*,c}\| - \mu_c\right|$$

MIL Loss
Cross-entropy for classification

$$\mathcal{L}_{mil} = -\sum_{c=1}^{C+1} y_c \log p_c$$

CAS for class c from branch i: $A^i{}_{*,c} \in \mathbb{R}^T$

Mean CAS norm for class c: $\mu_c = \frac{1}{K}\sum_{i=1}^{K}\|A^i{}_{*,c}\|$

Branch number: K

Class number: C

Video length: T

Ground truth: y

Experiments

Results on THUMOS14

	mAP (IoU 0.1)	mAP (IoU 0.3)	mAP (IoU 0.5)	mAP (IoU 0.1:0.5)
Hide-and-Seek	36.4	19.5	6.8	20.6
UntrimmedNet	44.4	28.2	13.7	29.0
Zhong et al.	45.8	31.1	15.9	30.9
STPN (UNT)	45.3	31.1	16.2	31.0
W-TALC (UNT)	49.0	32.0	18.8	33.7
AutoLoc (UNT)	-	35.8	**21.1**	-
Ours (UNT)	**53.5**	**37.5**	19.9	**37.4**
STPN (I3D)	52.0	35.5	16.9	35.0
W-TALC (I3D)	55.2	40.1	22.8	39.8
Ours (I3D)	**57.4**	**41.2**	**23.1**	**40.9**

Results on ActivityNet 1.2 Validation Set

	mAP (IoU 0.5)	mAP (IoU 0.75)	mAP (IoU 0.95)	mAP (AVG)
Zhong et al.	27.3	14.7	2.9	15.6
AutoLoc (UNT)	27.3	15.1	3.3	16.0
Ours (UNT)	**33.9**	**19.9**	**5.1**	**20.5**
W-TALC (I3D)	**37.0**	-	-	18.0
Ours (I3D)	36.8	**22.0**	**5.6**	**22.4**

Results on ActivityNet 1.3 (I3D)

	mAP (IoU 0.5)	mAP (IoU 0.75)	mAP (IoU 0.95)	mAP (AVG)
STPN (Val Set)	29.3	16.9	2.6	-
Ours (Val Set)	**34.0**	**20.9**	**5.7**	**21.2**
STPN (Test Set)	-	-	-	20.1
Ours (Test Set)	-	-	-	**23.1**

Conclusion

- A multi-branch network with diversity loss is proposed to model action completeness.
- A hard negative video generation scheme is devised to separate co-occurring context.
- Better performances on benchmarks.

Scan for code