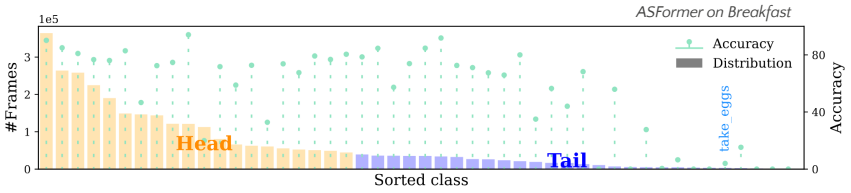
Cost-Sensitive Learning for Long-Tailed Temporal Action Segmentation

Zhanzhong Pang[1], Fadime Sener[2], Shrinivas Ramasubramanian[3], Angela Yao[1]
[1] National University of Singapore, [2] Meta Reality Labs, [3] Carnegie Mellon University

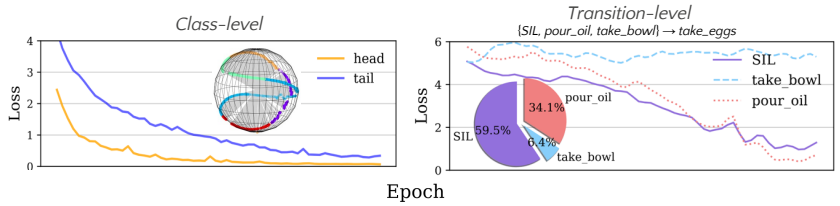
BMVC 2024

I. Motivation

Long-Tailed Distribution & Effect

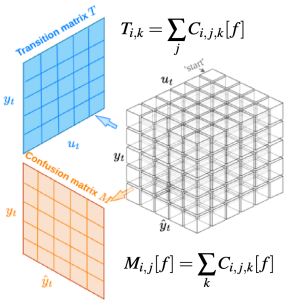


Bi-level Learning Bias



- ❑ The tail converges more slowly than the Head.
- ❑ Common transitions are better learned.

II. Confusion Tensor



$$T_{i,k} = \sum_j C_{i,j,k}[f]$$

$$M_{i,j}[f] = \sum_k C_{i,j,k}[f]$$

$$C_{i,j,k}[f] = \mathbb{E}_{(X,Y)}[\mathbb{1}(y_t = i, \hat{y}_t = j, u_t = k)]$$

prediction; classifier; ground truth; previous action

- **Class Accuracy**

$$Acc_i[f] = \frac{M_{i,i}[f]}{\pi_i} = \frac{\sum_k C_{i,i,k}[f]}{\pi_i}$$

- **Transition Accuracy**

$$Tacc_{k \to i}[f] = \frac{C_{i,i,k}[f]}{T_{i,k}}$$

III. Cost-Sensitive Learning with Constraint Optimization

➢ **Objective** – Maximize per-class accuracy, ensuring equal attention to each class

➢ **Constraint** – Reduce transition learning bias, penalizing under-learned transitions

$$\max_f \sum_{i,k} \frac{C_{i,i,k}[f]}{\pi_i} \qquad \text{s.t. } \forall (k \to i) \in V_T, \ \frac{C_{i,i,k}[f]}{T_{i,k}} \geq \varepsilon \overline{Tacc}.$$

Lagrangian min-max → **ITERATIVELY** optimize the classifier and the multipliers

$$\max_f \min_{\lambda \in \mathbb{R}_+} \sum_{i,k} \frac{C_{i,i,k}[f]}{\pi_i} + \sum_{i,k} \mathbb{1}(T_{i,k} > 0) \lambda_{i,k} \left(\frac{C_{i,i,k}[f]}{T_{i,k}} - \varepsilon \overline{Tacc} \right) \frac{T_{i,k}}{\pi_i}$$

Algorithm 1 Optimizing Per class Accuracy with Transitional Constraints

Input: Training set $\mathcal{D}$, Class prior $\pi \in \mathbb{R}_+^L$ and transition prior $T \in \mathbb{R}_+^{L \times (L+1)}$ derived from $\mathcal{D}$, Learning rate for multiplier $\gamma \in \mathbb{R}_+$, Cost-sensitive loss function l, Lagrangian objective $\mathcal{L}$

Initialize: Classifier f, Multiplier $\lambda \in \mathbb{R}_+^{L \times (L+1)}$

for epoch $l \leftarrow 0, \ldots, N$ **do**

① *// Update G*
Calculate the gain tensor G based on π, T, and λ
// Update f
$f^{l+1} \in \arg\min_f \frac{1}{|\mathcal{D}|} \sum_{(X,Y) \in \mathcal{D}} l(y_t, \hat{y}_t, G)$ *// $Y = \{y_t\}$*

② *// Update λ*
$C_{i,j,k}[f^{l+1}] = \frac{1}{|\mathcal{D}|} \sum_{(X,Y) \in \mathcal{D}} \mathbb{1}(y_t = i, \hat{y}_t = j, u_t = k)$ *// calculate confusion matrix*
Calculate $\overline{Tacc}$ based on T and $C[f^{l+1}]$
$\lambda_{i,k}^{l+1} = \max\{\lambda_{i,k}^l - \gamma \nabla_{\lambda_{i,k}} \mathcal{L}, 0\}$ *// gradients are calculated based on $\overline{Tacc}$*

Step 1. Maximizing the Lagrangian $\mathcal{L}(f)$ with fixed λ

$$\max_f \sum_{i,k} G_{i,i,k} C_{i,i,k}[f] + \text{constant}$$

$$G_{i,i,k} = \underbrace{\frac{1}{\pi_i}}_{\text{action prior}} + \underbrace{\frac{\lambda_{i,k} \mathbb{1}(T_{i,k} > 0)}{\pi_i}}_{\text{transition learning state}}$$

equivalent to minimizing a re-weighted loss

$$l_{CE}(y_t, u_t, X) = -G_{y_t, y_t, u_t} \log(p(y_t \mid X))$$

Step 2. Minimizing the Lagrangian $\mathcal{L}(\lambda)$

IV. Results

➢ **Datasets** – Breakfast, 50salads, Assembly101
➢ **Metrics** – [*per class*] Balanced accuracy, F1 score.
[*global*] accuracy, F1 score, Edit distance.
➢ **Backbones** – MSTCN, AsFormer, DiffAct

1. Per-class & global results

Model	Breakfast					50salads					Assembly101				
	Per class				G_F1	Per class				G_F1	Per class				G_F1
	F1@{10,25,50}			Acc.		F1@{10,25,50}			Acc.		F1@{10,25,50}			Acc.	
MSTCN [13]	48.1	44.8	36.9	49.1	57.9	78.8	76.4	67.6	75.6	75.9	7.5	6.6	4.8	8.3	27.2
+ CB [9]	+0.9	+0.7	+0.3	+0.6	0.0	-0.6	-0.2	-0.8	-0.3	-0.4	+1.8	+1.7	+1.2	+1.5	-0.5
+ LA [32]	+1.0	+1.1	+0.1	+1.4	0.0	-0.2	-0.7	0.0	-0.3	-0.7	+2.1	+1.4	+1.2	+1.2	-1.1
+ Focal [30]	+0.2	-0.3	-1.2	-0.5	-0.4	+0.6	+0.5	+1.0	+0.4	+0.2	+1.9	+1.6	+0.5	+1.4	-0.2
+ τ-norm [21]	-1.1	-1.0	-1.0	-0.8	-0.9	-0.6	-0.5	-0.2	+0.2	-0.6	+0.1	+0.2	+0.1	-0.2	+0.2
+ ours(S-NCM)	**+8.1**	**+8.1**	**+5.7**	**+3.7**	+6.1	**+2.8**	**+3.1**	**+3.2**	**+1.7**	+3.1	**+4.1**	**+3.3**	**+2.0**	**+2.6**	+2.3

2. Group-wise results

Model	Breakfast				50salads				Assembly101			
	Accuracy		F1@25		Accuracy		F1@25		Accuracy		F1@25	
	Head	Tail	Head	Tail	Head	Tail	Head	Tail	Head	Tail	Head	Tail
MSTCN	65.1	37.7	53.3	38.7	87.7	70.0	85.7	72.1	33.9	4.7	26.3	3.9
+ CB [9]	64.1	39.3	54.1	39.4	**88.4**	69.3	85.3	72.0	34.8	6.8	28.1	6.0
+ LA [32]	64.4	40.6	56.0	38.7	87.5	69.6	86.0	71.0	34.3	6.4	27.1	5.8
+ Focal [30]	**66.1**	36.1	53.6	38.0	88.3	70.3	84.8	73.3	**35.3**	6.6	26.3	6.4
+ τ-norm [21]	65.3	36.2	52.7	37.4	87.6	70.3	85.1	71.6	34.0	4.3	25.9	4.2
+ ours(S-NCM)	65.3	**44.0**	**64.5**	**44.6**	87.8	**72.5**	**87.7**	**75.7**	34.1	**8.7**	**31.7**	**7.6**

3. Evolution of Transition Accuracy & Multipliers