MEID: Mixture-of-Experts with Internal Distillation for Long-Tailed Video Recognition (Supplementary Material)

Xinjie Li, Huijuan Xu
Pennsylvania State University, University Park, USA
{xql5497, hkx5063}@psu.edu

Method	Overall mAP	Head (500,+∞)	Medium (100,500]	Tail (0,100]
Summation	0.618	0.751	0.659	0.525
Mean	0.619	0.750	0.661	0.524

Table 1: Ablation studies on how to reserve the static information in the motion feature.

1 More Ablation Studies on Motion Feature

Different methods on how to reserve static information in motion feature. As we mentioned in Eq. 10, we calculate the mean value of all frames' visual features to maintain the static object information in the final motion feature. In Table 1, we analyse the performance difference between two operations to maintain the static information, i.e., mean value and summation value. The results show that the performance gap is small and the mean value performs better.

Different fusion methods of the motion feature. We also conduct an experiment on how to fuse the motion feature in the second expert. Different methods are shown in Fig. 1. From Table 2, we can see that our proposed medium fusion method performs the best.

Per-class AP analysis for the additional motion feature. In our work, we incorporate the motion feature into the second expert to tackle difficult frames with motion property and boost the performance. Here, we analyse the effectiveness of this additional information. Specifically, we select the classes with AP equaling to 1 and count the number of classes, among which we further count the number of tail classes, and the number of motion classes in these tail classes with AP equaling to 1. The results before and after the incorporation of motion feature are listed in Table 3. The results show that the additional motion feature helps the model focus more on tail classes and motion classes.

2 More Results on VideoLT Dataset

In this section, we add the Accuracy@1 and Accuracy@5 as the evaluation metrics for experiments with ResNet50 (He et al. 2016) and ResNet101 (He et al. 2016) as pre-trained visual encoders. The results are shown in Table 4.

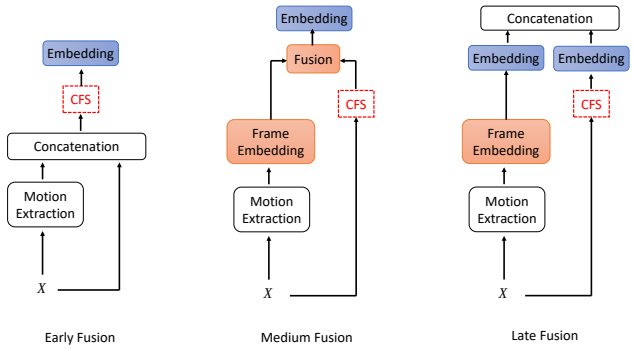

Figure 1: Illustration of different fusion methods. The medium fusion is used in our final model.

Fusion	Overall mAP	Head (500,+∞)	Medium (100,500]	Tail (0,100]
Early Fusion	0.597	0.727	0.654	0.474
Medium Fusion	0.619	0.750	0.661	0.524
Late Fusion	0.607	0.737	0.656	0.500

Table 2: Ablation studies on how to fuse the motion feature into the second expert.

We also add the experiments with TSM-ResNet-50 (Lin, Gan, and Han 2019) as the pre-trained visual encoder. The results are shown in Table 5.

3 The CharadesLT and CharadesEgoLT Benchmarks

In this section, we first introduce the construction process of the two benchmarks. For the CharadesLT dataset, we first create the training set by sampling a subset of Charades

	#All	#Tail (Ratio)	#Motion in Tail (Ratio)
Before	84	43 (51.2%)	21 (48.8%)
After	125	68 (54.4%)	35 (51.5%)

Table 3: The number of classes whose AP equaling to 1 before and after adding the motion feature.

Methods	Visual Encoder	Aggregation	Overall mAP	Head (>500)	Medium (100-500)	Tail (<100)	Acc@1	Acc@5
LDAM	ResNet50	Nonlinear	0.502	0.680	0.557	0.378	0.656	0.811
EQL			0.502	0.679	0.557	0.378	0.653	0.829
CBS			0.491	0.649	0.545	0.371	0.640	0.820
CB Loss			0.495	0.653	0.546	0.381	0.643	0.823
Mixup			0.484	0.649	0.535	0.368	0.633	0.818
FrameStack			0.516	0.683	0.569	0.397	0.658	0.834
MEDC (Hu, Gao, and Xu 2022)			0.567	0.720	0.607	0.436	0.667	0.839
Ours			**0.619**	**0.750**	**0.661**	**0.524**	**0.698**	**0.840**
LDAM	ResNet101	Nonlinear	0.518	0.687	0.572	0.397	0.664	0.820
EQL			0.518	0.690	0.571	0.398	0.664	0.838
CBS			0.507	0.660	0.559	0.390	0.652	0.828
CB Loss			0.511	0.665	0.561	0.398	0.656	0.832
Mixup			0.500	0.665	0.550	0.386	0.646	0.828
FrameStack			0.535	0.697	0.587	0.419	0.672	0.844
MEDC (Hu, Gao, and Xu 2022)			0.603	0.737	0.657	0.499	0.670	0.845
FrameStack		NetVLAD	0.670	0.780	0.707	0.590	0.710	0.858
Ours			**0.716**	**0.811**	**0.745**	**0.652**	**0.728**	**0.878**

Table 4: Experimental results on the VideoLT dataset with ResNet-50 and ResNet-101 models as pre-trained visual encoders. The results of comparison methods are copied from VideoLT (Zhang et al. 2021) and MEDC (Hu, Gao, and Xu 2022).

Methods	Model	Aggregation	Overall mAP	Head (>500)	Medium (100-500)	Tail (<100)
LDAM	TSM-R-50	Nonlinear	0.565	0.750	0.620	0.436
EQL			0.567	0.757	0.623	0.439
CBS			0.558	0.733	0.612	0.435
CB Loss			0.563	0.744	0.616	0.440
Mixup			0.548	0.736	0.602	0.425
FrameStack			0.580	0.759	0.632	0.459
LDAM		NetVLAD	0.627	0.779	0.675	0.519
EQL			0.665	0.808	0.713	0.557
CBS			0.662	0.806	0.708	0.558
CB Loss			0.666	0.801	0.712	0.566
Mixup			0.659	0.800	0.706	0.556
FrameStack			0.667	0.806	0.713	0.566
MEDC (Hu, Gao, and Xu 2022)			0.670	0.811	0.716	0.569
Ours			**0.695**	**0.814**	**0.739**	**0.600**

Table 5: Experimental results on the VideoLT dataset with TSM-ResNet-50 model as the pre-trained visual encoder. The results of comparison methods are copied from VideoLT (Zhang et al. 2021) and MEDC (Hu, Gao, and Xu 2022). We do not report the results of Accuracy under this setting following VideoLT (Zhang et al. 2021).

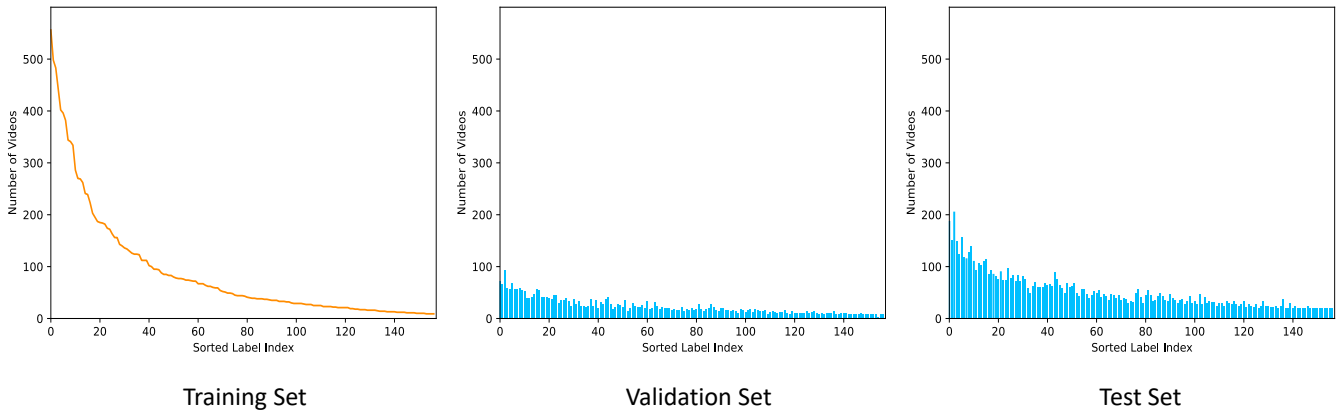

Figure 2: Distributions for the training set, validation set and test set of CharadesLT dataset. The sorted video number of the training set is used as the x-axis index for validation and test set.

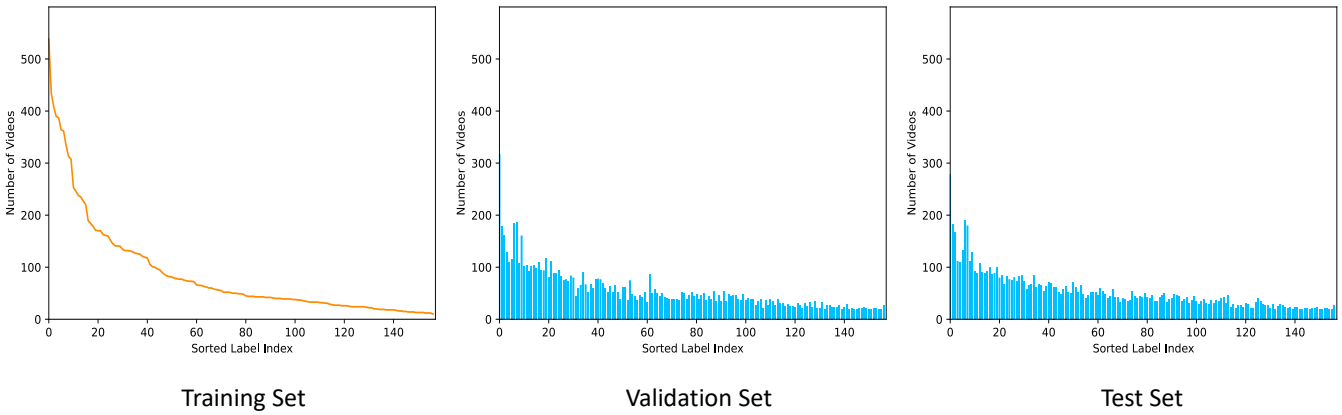

Figure 3: Distributions for the training set, validation set and test set of CharadesEgoLT dataset. The sorted video number of the training set is used as the x-axis index for validation and test set.

dataset (Sigurdsson et al. 2016) with the number of samples for each class following the Pareto distribution $pdf(x) = \alpha \frac{x_{\min}^{\alpha}}{x^{\alpha+1}}$ where the power value α is set as 10. Specifically, all categories are ranked by the number of samples, and for each class from head to tail, we increase or reduce samples in that class, and force the distribution to fit the expected Pareto distribution, following the work in long-tailed multi-label image recognition (Wu et al. 2020).

The rest of data are first used to construct the test set of CharadesLT though sampling following the approximate uniform distribution. The validation set are also sampled following an approximate uniform distribution. The CharadesEgoLT is transformed from the CharadesEgo dataset (Sigurdsson et al. 2018) and the construction procedure is similar to CharadesLT with the power value $\alpha = 17$.

The officially-provided 24-fps RGB frame images are used as raw data. Then, we resize them to 256x256 and crop the center of them. Next, features are extracted by the ResNet-101 ImageNet pre-trained model. Specifically, we adopt the output of the penultimate layer as our final feature. In the end, 200 frames are uniformly sampled from each video and each frame is represented by a 2048-dimension feature vector.

We also show dataset distributions of CharadesLT and CharadesEgoLT, in Fig. 2 and Fig. 3, respectively.

Last, we attach the training list, validation list and test list of CharadesLT and CharadesEgoLT in the zip file. The raw videos and RGB frames can be found in the official websites of the Charades (Sigurdsson et al. 2016) [1] and CharadesEgo (Sigurdsson et al. 2018) [2] datasets.

References

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 770–778. IEEE Computer Society.

Hu, Y.; Gao, J.; and Xu, C. 2022. Learning Muti-expert Distribution Calibration for Long-tailed Video Classification. *CoRR*, abs/2205.10788.

Lin, J.; Gan, C.; and Han, S. 2019. TSM: Temporal Shift Module for Efficient Video Understanding. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 7082–7092. IEEE.

Sigurdsson, G. A.; Gupta, A.; Schmid, C.; Farhadi, A.; and Alahari, K. 2018. Actor and Observer: Joint Modeling of First and Third-Person Videos. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 7396–7404. Computer Vision Foundation / IEEE Computer Society.

Sigurdsson, G. A.; Varol, G.; Wang, X.; Farhadi, A.; Laptev, I.; and Gupta, A. 2016. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. In Leibe, B.; Matas, J.; Sebe, N.; and Welling, M., eds., *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, volume 9905 of *Lecture Notes in Computer Science*, 510–526. Springer.

Wu, T.; Huang, Q.; Liu, Z.; Wang, Y.; and Lin, D. 2020. Distribution-Balanced Loss for Multi-label Classification in Long-Tailed Datasets. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J., eds., *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IV*, volume 12349 of *Lecture Notes in Computer Science*, 162–178. Springer.

Zhang, X.; Wu, Z.; Weng, Z.; Fu, H.; Chen, J.; Jiang, Y.-G.; and Davis, L. S. 2021. VideoLT: Large-scale Long-tailed Video Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7960–7969.

[1] https://prior.allenai.org/projects/charades

[2] https://prior.allenai.org/projects/charades-ego