Ensemble Modeling for Multimodal Visual Action Recognition

Jyoti Kini, Sarah Fleischer, Ishan Dave, and Mubarak Shah

Center for Research in Computer Vision, University of Central Florida, USA
{jyoti.kini, sa420832, ishandave}@ucf.edu, shah@crcv.ucf.edu

Abstract. In this work, we propose an ensemble modeling approach for multimodal action recognition. We independently train individual modality models using a variant of focal loss tailored to handle the long-tailed distribution of the MECCANO [21] dataset. Based on the underlying principle of focal loss, which captures the relationship between tail (scarce) classes and their prediction difficulties, we propose an exponentially decaying variant of focal loss for our current task. It initially emphasizes learning from the hard misclassified examples and gradually adapts to the entire range of examples in the dataset. This annealing process encourages the model to strike a balance between focusing on the sparse set of hard samples, while still leveraging the information provided by the easier ones. Additionally, we opt for the *late fusion* strategy to combine the resultant probability distributions from RGB and Depth modalities for final action prediction. Experimental evaluations on the MECCANO dataset demonstrate the effectiveness of our approach.

1 Introduction

Amidst the surge of data in recent times, multimodal learning has emerged as a transformative approach, leveraging heterogeneous cues from multiple sensors to enhance the learning process. Both early and late multimodal fusion mechanisms have demonstrated the ability to effectively harness complementary information from diverse sources. However, real-world multimodal data associated with action occurrences suffer from an inherent skewness, giving rise to the long-tailed action recognition scenario. In such cases, some action classes are prevalent and well-represented in the training data, while others are scarce, leading to significant data imbalance. The inherent complexity of multimodal data, combined with such data imbalance, presents a formidable challenge for learning approaches.

In order to fuse information from different data streams, researchers in the vision community have proposed a variety of approaches, spanning from consolidating feature representations at an early stage (early fusion) [19,10] to aggregating prediction scores at the final stage (late fusion) [3,9]. Furthermore, to combat the long-tailed visual recognition issue, use of data augmentation [26,25,31,29], re-sampling [18,7,1], cost-sensitive loss [13,4,2,27], and transfer learning [24,16,33,32,12,5] strategies is highly recommended. Methods that rely on data augmentation techniques like M2M [8], ImbalanceCycleGAN[23], and

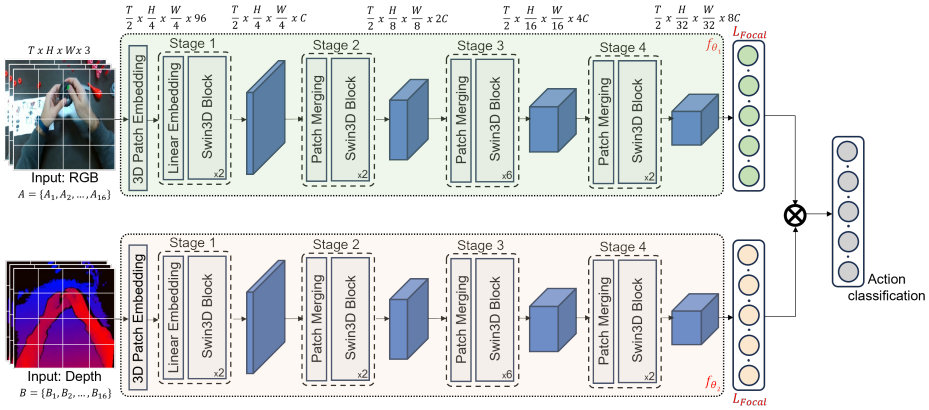

Fig. 1: **Architecture**: The RGB frames $\{A_i,\ A_{i-1},..,A_T\}$ and Depth frames $\{B_i,\ B_{i-1},..,B_T\}$ are passed through two independently trained Swin3D-B [15] encoders f_{θ_1} and f_{θ_2} respectively to generate feature tokens. The resultant class probabilities, obtained from each pathway, are averaged to subsequently yield action classes. Exponentially decaying focal loss L_{Focal} is leveraged to deal with the long-tailed distribution exhibited by the data.

MetaS-Aug[11] attempt to augment the minority classes with diverse samples. Other data augmentation-based works [14,30,28] generate pseudo labels to reduce scarcity in tail classes. However, these methods are often limited by their ability to generate realistic and diverse minority class samples. Some of the re-sampling approaches [20,22,6,1] focus on assigning larger sampling probabilities to the tail classes. Although beneficial, re-sampling models are at risk of overfitting to the tail classes.

In this paper, we introduce an ensemble training strategy that leverages multimodal RGB and Depth signals for visual action recognition, using the late fusion mechanism. We, also, allow the model to focus on *h*ard-to-classify examples using an exponentially decaying variant of the focal loss objective function. This function not only reduces the KL divergence between the predicted distribution and the ground-truth distribution but also simultaneously increases the entropy of the predicted distribution, thereby preventing model overconfidence towards majority classes.

2 Approach

2.1 Cross-Modal Fusion

Figure 1 provides comprehensive details of our proposed approach. Given a set of spatiotemporally aligned RBG and Depth sequences that extend between $[t_s, t_e]$, where t_s and t_e are the start and the end duration of the sequence, our goal is to predict the action class $\mathcal{O} = \{o_1,\ o_2,..,\ o_K\}$ associated with the sequence. In order to achieve this, we adopt an ensemble architecture comprising two dedicated Video Swin Transformer [15] backbones to process the RGB clip

$\mathcal{A} = \{A_i, A_{i-1},..,A_T\}$ and Depth clip $\mathcal{B} = \{B_i, B_{i-1},..,B_T\}$ independently. Here, i corresponds to a random index spanning between t_s and t_e. The input video for each modality defined by size $T \times H \times W \times 3$ results in token embeddings of dimension $\frac{T}{2} \times H_d \times W_d \times C$. We pass this representation retrieved from *stage-4* of the base feature network to our newly added fully connected layer and fine-tune the overall network. The final prediction is derived by averaging the two probability distributions obtained as output from the RGB and Depth pathways.

2.2 Exponentially Decaying Focal Loss

Focal loss [13] is a variant of cross-entropy loss with a modulating factor that down-weighs the impact of easy examples and focuses on the hard ones. It, therefore, tends to prevent bias towards data-rich classes and improves the performance on scarce categories.
Multi-classification cross-entropy (CE) loss is given by:

$$L_{CE} = -\sum_{j=1}^{K} y_j \log(p_j) \tag{1}$$

where, say we have K action classes, and y_j and p_j correspond to the ground-truth label and predicted probability respectively for the j^{th} class.
On the other hand, the key objective of focal loss [13] is defined as:

$$L_{Focal} = -\sum_{j=1}^{K} (1 - p_j)^{\gamma} \log p_j \tag{2}$$

In our work, we use focal loss L_{Focal} and exponentially decay γ from 2 to 0.1. When γ=0, the objective function is equivalent to cross-entropy loss. Our proposed annealing process for γ allows for the model to focus on the sparse set of hard examples in the early stage of training, and gradually shift its focus towards easy examples. This configuration is essential to ensure that the model learns meaningful representations and generalized decision boundaries.

3 Experimental Setup

3.1 Data-preprocessing

For our experiments, we resize the frames to a width of 256, without disturbing the aspect ratio of the original image, followed by a random crop of 224×224. In addition, we use 16 consecutive frames to generate a single clip for the forward pass. In the case of shorter sequences, we pad the sequence with the last frame.

3.2 Training

We use the Swin3D-B [15] backbone, which is pre-trained on the Something-Something v2 [17] dataset. We adopt focal loss [13] with exponentially decaying γ for training the classification model. For optimization, AdamW optimizer with a learning rate of 3×10^{-4} and a weight decay of 0.05 has been employed. Our model converges in about 20 epochs on the MECCANO dataset. We report the *Top-1* and *Top-5* classification accuracy as our evaluation metrics. Additionally, to demonstrate the effectiveness of employing the focal loss for this task, we present the average class *Precision*, *Recall* and *F1-score*.

Modality	Loss	Accuracy		AVG Class		AVG *F1-score*
		Top-1	*Top-5*	*Precision*	*Recall*	
RGB	*CE*	48.35	80.91	45.52	48.35	46.22
Depth	*CE*	43.32	75.38	41.79	43.32	41.88
RGB+Depth	*CE*	50.94	81.79	47.28	50.94	48.08
RGB	Focal	50.80	82.36	47.17	50.80	47.95
Depth	Focal	45.52	78.07	43.74	45.52	43.41
RGB+Depth	Focal	52.82	83.85	49.97	52.82	49.41
RGB*	Focal	53.03	85.37	50.46	53.03	50.39
Depth*	Focal	48.39	80.55	46.43	48.39	46.35
RGB+Depth*	Focal	**55.37**	85.58	52.41	55.37	52.28

Table 1: Results demonstrating the effectiveness of our ensemble modeling approach for the action recognition task on the MECCANO test dataset. *CE* implies Cross-Entropy loss. ***** refers to model trained using both train+validation set.

4 Discussion

Table 1 presents our results on the MECCANO test set. Applying cross-entropy loss to fine-tune our model, pre-trained on Something-Something v2, gives us an initial baseline accuracy of 50.94% on our multimodal setup. Introducing focal loss with exponential decay in γ boosts the overall accuracy by $\approx$ 2%. Figure 2 demonstrates the effectiveness of our approach in dealing with the long-tailed distribution of the MECCANO dataset. Furthermore, combining the train and validation data gives the best *Top-1* accuracy of 55.37%.

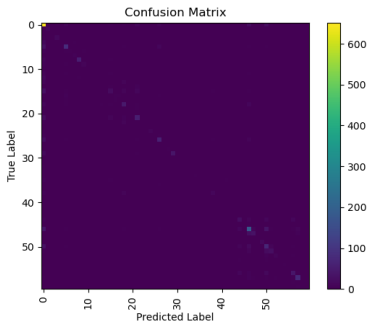

Fig. 2: The resultant confusion matrix obtained from the MECCANO test set highlights our model's proficiency in handling the long-tailed distribution.

References

1. Buda, M., Maki, A., Mazurowski, M.A.: A systematic study of the class imbalance problem in convolutional neural networks. Neural networks **106**, 249–259 (2018)
2. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9268–9277 (2019)
3. Ding, C., Tao, D.: Robust face recognition via multimodal deep face representation. IEEE transactions on Multimedia **17**(11), 2049–2058 (2015)
4. Huang, C., Li, Y., Loy, C.C., Tang, X.: Learning deep representation for imbalanced classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5375–5384 (2016)
5. Jamal, M.A., Brown, M., Yang, M.H., Wang, L., Gong, B.: Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7610–7619 (2020)
6. Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., Kalantidis, Y.: Decoupling representation and classifier for long-tailed recognition. arXiv preprint arXiv:1910.09217 (2019)
7. Kim, B., Kim, J.: Adjusting decision boundary for class imbalanced learning. IEEE Access **8**, 81674–81685 (2020)
8. Kim, J., Jeong, J., Shin, J.: M2m: Imbalanced classification via major-to-minor translation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13896–13905 (2020)
9. Koo, J.H., Cho, S.W., Baek, N.R., Kim, M.C., Park, K.R.: Cnn-based multimodal human recognition in surveillance environments. Sensors **18**(9), 3040 (2018)
10. Landi, F., Baraldi, L., Cornia, M., Corsini, M., Cucchiara, R.: Perceive, transform, and act: Multi-modal attention networks for vision-and-language navigation. arXiv preprint arXiv:1911.12377 **3**(9) (2019)
11. Li, S., Gong, K., Liu, C.H., Wang, Y., Qiao, F., Cheng, X.: Metasaug: Meta semantic augmentation for long-tailed visual recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5212–5221 (2021)
12. Li, T., Wang, L., Wu, G.: Self supervision to distillation for long-tailed visual recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 630–639 (2021)
13. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
14. Liu, B., Li, H., Kang, H., Vasconcelos, N., Hua, G.: Semi-supervised long-tailed recognition using alternate sampling. arXiv preprint arXiv:2105.00133 (2021)
15. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3202–3211 (2022)
16. Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., Yu, S.X.: Large-scale long-tailed recognition in an open world. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2537–2546 (2019)
17. Mahdisoltani, F., Berger, G., Gharbieh, W., Fleet, D., Memisevic, R.: On the effectiveness of task granularity for transfer learning. arXiv preprint arXiv:1804.09235 (2018)

18. More, A.: Survey of resampling techniques for improving classification performance in unbalanced datasets. arXiv preprint arXiv:1608.06048 (2016)
19. Nishida, N., Nakayama, H.: Multimodal gesture recognition using multi-stream recurrent neural network. In: Image and Video Technology: 7th Pacific-Rim Symposium, PSIVT 2015, Auckland, New Zealand, November 25-27, 2015, Revised Selected Papers 7. pp. 682–694. Springer (2016)
20. Peng, M., Zhang, Q., Xing, X., Gui, T., Huang, X., Jiang, Y.G., Ding, K., Chen, Z.: Trainable undersampling for class-imbalance learning. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 4707–4714 (2019)
21. Ragusa, F., Furnari, A., Livatino, S., Farinella, G.M.: The meccano dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1569–1578 (2021)
22. Ren, J., Yu, C., Ma, X., Zhao, H., Yi, S., et al.: Balanced meta-softmax for long-tailed visual recognition. Advances in neural information processing systems **33**, 4175–4186 (2020)
23. Sahoo, A., Singh, A., Panda, R., Feris, R., Das, A.: Mitigating dataset imbalance via joint generation and classification. In: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16. pp. 177–193. Springer (2020)
24. Samuel, D., Atzmon, Y., Chechik, G.: From generalized zero-shot learning to long-tail with class descriptors. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 286–295 (2021)
25. Verma, V., Lamb, A., Beckham, C., Najafi, A., Mitliagkas, I., Lopez-Paz, D., Bengio, Y.: Manifold mixup: Better representations by interpolating hidden states. In: International conference on machine learning. pp. 6438–6447. PMLR (2019)
26. Wang, J., Lukasiewicz, T., Hu, X., Cai, J., Xu, Z.: Rsg: A simple but effective module for learning imbalanced datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3784–3793 (2021)
27. Wang, P., Han, K., Wei, X.S., Zhang, L., Wang, L.: Contrastive learning based hybrid networks for long-tailed image classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 943–952 (2021)
28. Wei, C., Sohn, K., Mellina, C., Yuille, A., Yang, F.: Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10857–10866 (2021)
29. Xiang, L., Ding, G., Han, J.: Increasing oversampling diversity for long-tailed visual recognition. In: Artificial Intelligence: First CAAI International Conference, CICAI 2021, Hangzhou, China, June 5–6, 2021, Proceedings, Part I 1. pp. 39–50. Springer (2021)
30. Yang, Y., Xu, Z.: Rethinking the value of labels for improving class-imbalanced learning. Advances in neural information processing systems **33**, 19290–19301 (2020)
31. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
32. Zhong, Y., Deng, W., Wang, M., Hu, J., Peng, J., Tao, X., Huang, Y.: Unequal-training for deep face recognition with long-tailed noisy data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7812–7821 (2019)

33. Zhou, B., Cui, Q., Wei, X.S., Chen, Z.M.: Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9719–9728 (2020)