Rugby Scene Classification Enhanced by Vision Language Model

Supplementary Material

9. Appendix

9.1. Optimal frame interval

In the scene classification experiments for rugby, images labeled manually were extracted from match footage at intervals of 0.2 seconds. Due to the nature of rugby, where changes within 0.2 seconds are typically minimal, utilizing all labeled images could lead to high similarity among the data, potentially resulting in overfitting to the training set.

To address this concern, we investigated how the performance of image classification changes when training with sparsely picked image data at regular intervals. Labeled image data were arranged chronologically, and if consecutive frames shared the same label, data used for training were picked at intervals of X frames. Specifically, for a sequence of labeled data $(a_1, a_2, a_3, a_4, a_5, a_6, a_7, b_8, b_9, b_{10}, c_{11}, a_{12}, a_{13}, a_{14})$, with $X = 3$, we used $(a_2, a_6, b_9, c_{11}, a_{13})$ for training. Preliminary experiments were conducted using $X = 1, 5, 10$, and 15, corresponding to picking data every $0.2, 1, 2$, and 3 seconds, respectively.

For training, we employed ResNet-50 [14] pretrained on the ImageNet-1k dataset. Training was conducted with a batch size of 256 and a learning rate of 0.0001. Three out of the four split train/val sets were used for training, while the remaining set was utilized for parameter tuning and performance evaluation. Four evaluation tasks were employed: multi-class classification using all labels, binary classification for lineout, tackle and ruck and binary classification for contact (lineout, maul, ruck and tackle). Weigthed F1 score was used as the evaluation metric for multi-class classification, and F1 score for the positive class was used for binary classification. Training data were picked every X frames in prior to the actual training and was fixed during the training, while all frames were used for evaluation data.

9.2. Optimal model architecture

Subsequently, we investigated the optimal model architecture and the use of pretrained weight for rugby scene classification. We examined ten conditions: ResNet18, ResNet34, ResNet50, ResNet101, and ResNet152 architectures [14], with each architecture evaluated with and without using the pretrained weights on the ImageNet-1k dataset. We examined all ten models for five target tasks respectively.

Results are shown in Tab. 6. For multi-class classification, ResNet-18 with pre-trained weights exhibited the highest validation set score, with decreasing trend towards larger model size when using pre-trained weights. In contact scene classification, ResNet-101 with pre-trained weights demonstrated the highest validation set score. In the remaining three classification tasks, ResNet-152 with pre-trained weights achieved the highest validation set score. Given these results, we employed the best-performing architecture for subsequent experiments.

9.3. Optimal learning rate and batchsize

After selecting optimal model architecture and use of pre-trained weights, we investigated suitable learning rate and batch size for each task. For the learning rate, we examined five settings: specifically, 0.00005, 0.0001, 0.00025, 0.0005, and 0.001. Regarding batch size, we conducted tests using different values. For both multi-class classification and contact scene classification, we explored batch sizes of 32, 64, 128, 256, and 512. For the other three tasks utilizing ResNet-152, we evaluated batch sizes of 32, 64, 128, and 256. We opted not to assess a batch size of 512 with ResNet-152 due to GPU memory constraints of NVIDIA A100 GPU.

The results are shown in Tab. 7, Tab. 8, Tab. 9, Tab. 10, Tab. 11 for multi-class classification, lineout, ruck, tackle and contact respectively. For the multi-class classification, learning rate of 0.0001 with batch size of 128 showed the best validation set score. For the lineout classification, learning rate of 0.00005 with batch size of 64 showed the best validation set score. For the ruck classification, the best validation score was obtained with batch size of 32 and learning rate of 0.00025. For the tackle classification, the highest validation score was achieved with batch size of 64 and learning rate of 0.00025. For the contact classification, we observed the best validation set score with batch size of 512 and learning rate of 0.00025. We used the batch size and learning rate for prompt comparison and final evaluation based on these results.

Table 6. Scene classification with ResNet variants. 'Pretrained' indicates the model pretrained with ImageNet-1k dataset.

	ResNet18		ResNet34		ResNet50		ResNet101		ResNet152	
Pretrained	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓
Multi-class	0.443	**0.637**	0.463	0.599	0.247	0.608	0.239	0.502	0.379	0.463
Lineout	0.137	0.398	0.264	0.381	0.215	0.436	0.072	0.498	0.093	**0.705**
Ruck	0.514	0.499	0.535	0.377	0.469	0.480	0.469	0.503	0.398	**0.602**
Tackle	0.276	0.421	0.216	0.341	0.284	0.452	0.296	0.412	0.307	**0.457**
Contact	0.426	0.662	0.586	0.689	0.596	0.695	0.591	**0.736**	0.490	0.714

Table 7. Grid search result of the multi-class classification. Values are F1 score of validation set with **bold** indicating the best.

			Batch size		
	32	64	128	256	512
0.001	0.551	0.593	0.508	0.443	0.587
0.0005	0.572	0.544	0.484	0.502	0.567
0.00025	0.586	0.613	0.627	0.641	0.633
0.0001	0.592	0.611	**0.646**	0.592	0.615
0.00005	0.634	0.630	0.638	0.636	0.633

Table 8. Grid search result of the lineout classification. Values are F1 score of validation set with **bold** indicating the best.

		Batch size		
	32	64	128	256
0.001	0.176	0.471	0.190	0.360
0.0005	0.382	0.305	0.371	0.564
0.00025	0.373	0.442	0.357	0.430
0.0001	0.533	0.582	0.508	0.327
0.00005	0.471	**0.685**	0.678	0.521

Table 9. Grid search result of the ruck classification. Values are F1 score of validation set with **bold** indicating the best.

		Batch size		
	32	64	128	256
0.001	0.329	0.386	0.458	0.595
0.0005	0.444	0.564	0.616	0.448
0.00025	**0.674**	0.595	0.586	0.473
0.0001	0.653	0.612	0.589	0.531
0.00005	0.642	0.571	0.521	0.554

Table 10. Grid search result of the tackle classification. Values are F1 score of validation set with **bold** indicating the best.

		Batch size		
	32	64	128	256
0.001	0.339	0.358	0.347	0.244
0.0005	0.389	0.355	0.346	0.416
0.00025	0.419	**0.488**	0.419	0.369
0.0001	0.376	0.417	0.430	0.473
0.00005	0.435	0.436	0.434	0.278

Table 11. Grid search result of the contact classification. Values are F1 score of validation set with **bold** indicating the best.

			Batch size		
	32	64	128	256	512
0.001	0.545	0.577	0.612	0.673	0.672
0.0005	0.682	0.687	0.690	0.724	0.701
0.00025	0.677	0.734	0.715	0.717	**0.748**
0.0001	0.639	0.728	0.678	0.706	0.648
0.00005	0.716	0.716	0.692	0.680	0.693