# Experiment 6: Optimizer & Learning Rate Tuning

In order to address the limitations that we have discern from the initial base run, this experiment intends to investigate the impact of training duration and learning rate adaptation on model generalisation. Specifically, in comparing the training for 30 epochs versus 50 epochs so as to identify if continued training leads to diminishing returns or exacerbates overfitting. In addition, we will be attempting to implement a learning rate scheduler to allow the model to begin with a higher learning rate for faster convergence and then gradually reduce it to refine learning in later epochs.

We also intend to incorporate early stopping criteria based on validation loss to terminate training once performance stagnates or deteriorates, ensuring computational efficiency and minimising overfitting. As such, with these modifications in place, the aim of this experiment is to enhance the model's ability to retain strong generalisation performance while avoiding unnecessary or harmful over-training.

### Experiment Parameters

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Optimizer** | **Learning Rate** | **Batch Size** | **Scheduler Used** | **Epochs** |
| Adam | 0.001 | 32 | None | 15 |
| Adam | 0.001 | 32 | None | 30 |
| Adam | 0.001 | 32 | ReduceLROnPlateau(factor=0.1, patience=3, min\_lr=1e-6) | 30 |
| Adam | 0.001 | 32 | ReduceLROnPlateau(factor=0.1, patience=3, min\_lr=1e-6) | 50 |
| Adam | 0.001 | 32 | ReduceLROnPlateau + Early Stopping(patience=5, delta=1e-4) | Up to 50 (may stop early) |

### Performance Metrics Comparison

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Metric** | **Initial Baseline Run (Epoch=15)** | **Tuned Run #1 (Epoch=30)** | **Tuned Run #2**  **(Epoch=30 with Scheduler)** | **Tuned Run #3**  **(Epoch=50 with Scheduler)** | **Tuned Run #4**  **(Epoch=50 with Scheduler and Early Stopping)** |
| Train Accuracy | 0.94 | 0.98 | 0.98 | 0.96 |  |
| Validation Accuracy | 0.89 | 0.91 | 0.92 | 0.91 |  |
| Test Accuracy | 0.90 | 0.90 | 0.90 | 0.87 |  |
|  |  |  |  |  |  |
| Apple F1 Score | 0.92 | 0.92 | 0.92 | 0.92 |  |
| Banana F1 Score | 0.92 | 0.92 | 0.91 | 0.89 |  |
| Orange F1 Score | 0.94 | 0.94 | 0.94 | 0.89 |  |
|  |  |  |  |  |  |
| Mixed Precision | 0.57 | 0.60 | 0.57 | 0.50 |  |
| Mixed Recall | 0.80 | 0.60 | 0.80 | 0.60 |  |
| Mixed F1 Score | 0.67 | 0.60 | 0.67 | 0.55 |  |
|  |  |  |  |  |  |
| Macro Average F1 Score | 0.86 | 0.85 | 0.86 | 0.81 |  |
| Weighted Average F1 Score | 0.90 | 0.90 | 0.90 | 0.87 |  |

### Observations & Insights

##### Tuned Run #1:

* Validation accuracy peaked at Epoch 15 (93.75%) and showed fluctuations or decline beyond that point. This is suggestive that further training will lead to diminishing returns and the possible potential of overfitting.
* Training loss the continued to decrease steadily while validation loss began increasing after Epoch 15, which is also indicative that the model has started memorising the training data.
* Despite high overall accuracy, the macro F1 score (0.95) highlights uneven performance across classes, dragged down primarily by the ‘Mixed’ class.
* With the erratic rise of validation loss from Epoch 17 onwards, this suggests the need for a learning rate schedule to slow updates during later epochs, so as to preserve generalisation.
* In all, this run serves as a strong baseline to justify the need of a learning rate scheduler and early stopping, which may help to stabilise performance and improve generalisation, particularly for the underperforming ‘Mixed’ class.

### Tuned Run #2:

* While maintaining a strong training accuracy (98%) and test accuracy (90%) that is consistent with previous runs, this particular run had showed **improved training stability** with the use of a learning rate scheduler.
* Validation accuracy peaked at **92.92% at Epoch 18** and remained stable above 90% in subsequent epochs. This is suggestive of an improved generalisation and a reduced overfitting compared to the baseline.
* The scheduler successfully reduced the learning rate four times: once at Epoch 10 (to 0.001), second at Epoch 15(to 0.0001), then again at Epoch 24 (to 0.00001) and lastly at Epoch 28 (to 0.000001), allowing the model to continue refining its weights with smaller updates during later epochs.
* Compared to the baseline 30-epoch run, **validation loss remained lower and was noted to be more consistent,** especially after Epoch 10, indicating better optimization and less overfitting.
* The ‘Mixed’ class F1 score had increased from 0.60 to **0.67**, with recall improving from 0.60 to **0.80**. This shows that the model became more sensitive to detecting true positives for this class.
* The macro F1 score was also observed to have increased from 0.85 to **0.86**, driven by better performance in the Mixed class. Whereas performance in Apple, Banana, and Orange remained stable and high (~0.91 to 0.94).
* In all, these findings highly suggest that the learning rate scheduler had enhanced the model's ability to generalise across all classes, particularly benefiting underperforming class without sacrificing overall accuracy.

### Tuned Run #3:

* While maintaining an overall high training accuracy of 96.2% and test accuracy of 87%, a slightly reduced generalisation compared to the 30-epoch scheduled run (Tuned Run #2) was observed for this run.
* Validation accuracy peaked at 92.08% in Epoch 32, but had stabilised or slightly declined despite continued training in the subsequent epochs, indicating a saturation point at around 30–35 epochs.
* Unlike the 30-epoch run in tuned run #2, performance was noted to plateau for nearly 20 epochs even with LR scheduling with no significant increase in accuracy or F1 scores beyond that point.
* The learning rate dropped progressively from 0.001 to 0.0001, then to 0.00001, and finally to 0.000001 by Epoch 19. However, it was noted that smaller LR updates were insufficient to push the model toward better generalisation.
* Class-level performance saw a drop in F1 score for the ‘mixed’ class from 0.67 to 0**.**55, driven by a drop in precision 0.57 to 0.50 and a recall drop (0.80 to0.60).
* Other classes like Apple, Banana, and Orange maintained consistent F1 scores (~0.89 to 0.92), though slightly lower than the 30-epoch counterpart.
* The macro F1 score was observed to have dropped from 0.86 to 0.81, suggesting that there might be a greater class imbalance in performance, likely exacerbated by overfitting in later epochs.
* In all, this run confirms that extending training to 50 epochs under scheduled LR control does not yield any benefit. In fact, the model begins to over-specialise, leading to worse results on classes like ‘Mixed’.

### Final Recommendations