**Experiment Run**

**Experiment Run Report**

**Experiment Title:** Numerosity-Based Categorization - Experiment Run 4

**Date:** 27/02/2025

**Researcher:** Karoki Evans Njogu

**1. Experiment Details**

| **Parameter** | **Value** |
| --- | --- |
| Seed | 42 |
| Dataset Size | 5000 samples |
| Image Size | 128x128 pixels |
| Categories | Few (1-5), Medium (6-15), Many (>16) |
| Batch Size | 256 |
| Learning Rate | 0.0002 |
| Epochs | 20 |
| Optimizer | AdamW |
| Dropout Rate | 0.3 |
| Weight Decay | 5e-4 |
| Loss Function | CrossEntropyLoss |
| Early Stopping | Yes (Patience = 5) |
| Device Used | GPU – NVIDIA L4 |
| eps | 1e-6 |
| betas | 0.9, 0.98 |

**2. Experiment Setup**

* **Dataset:** Synthetic Dot Patterns
* **Model Architecture:** Residual CNN with three convolutional layers and fully connected layers.
* **Training Strategy:**
  + Train on 70% of data.
  + Validate on 15%.
  + Test on 15%.
* **Evaluation Metrics:**
  + Accuracy
  + Loss Curves
  + Confusion Matrix
  + Precision, Recall, and F1-Score

**3. Training & Validation Performance**

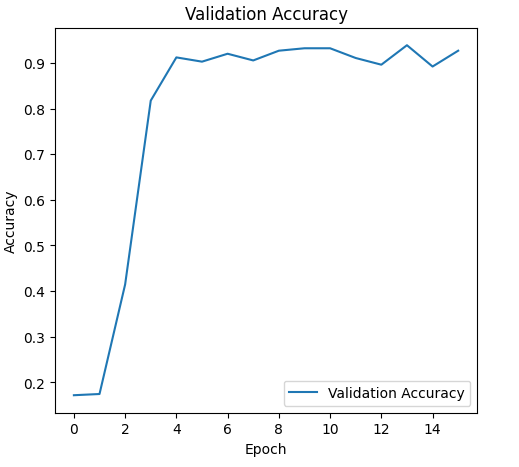
**3.1 Loss and Accuracy Trends**

| **Epoch** | **Train Loss** | **Validation Loss** | **Validation Accuracy (%)** |
| --- | --- | --- | --- |
| 1 | 1.5640 | 1.5103 | 17.20% |
| 2 | 0.3902 | 0.6055 | 17.47% |
| 3 | 0.2664 | 0.1795 | 41.47% |
| 4 | 0.2319 | 0.4067 | 81.73% |
| 5 | 0.2099 | 0.2170 | 91.20% |
| 6 | 0.2056 | 0.2481 | 90.27% |
| 7 | 0.1862 | 0.2137 | 92.00% |
| 8 | 0.1744 | 0.2082 | 90.53% |
| 9 | 0.1410 | 0.1712 | 92.67% |
| 10 | 0.1263 | 0.1765 | 93.20% |
| 11 | 0.1119 | 0.1662 | 93.20% |
| 12 | 0.0960 | 0.2009 | 91.07% |
| 13 | 0.0923 | 0.2743 | 89.60% |
| 14 | 0.0861 | 0.1719 | 93.87% |
| 15 | 0.0615 | 0.2913 | 89.20% |
| 16 | 0.0641 | 0.2138 | 92.67% |

**3.2 Loss Curve & Accuracy Plot**

A graph with blue line and orange line

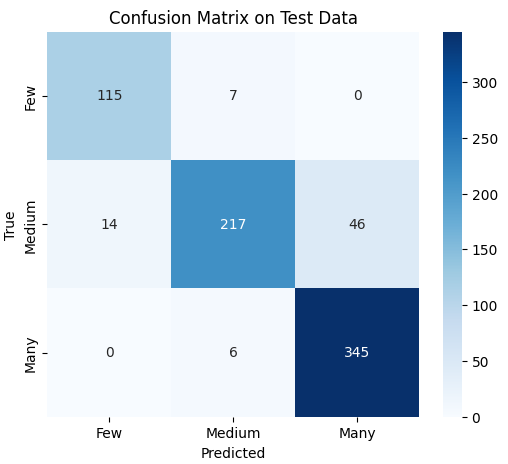
AI-generated content may be incorrect.



**4. Test Set Evaluation**

**Final Test Accuracy:** 90.27%

**4.1 Confusion Matrix**

****

**4.2 Classification Report**

| **Class** | **Precision** | **Recall** | **F1-Score** | **Support** |
| --- | --- | --- | --- | --- |
| Few | 0.89 | 0.94 | 0.92 | 122 |
| Medium | 0.94 | 0.78 | 0.86 | 277 |
| Many | 0.88 | 0.98 | 0.93 | 351 |

**5. Observations & Insights**

* **Key Findings:**
  + Improved stability, lowering the learning rate to 0.0002 resulted in smoother loss curves and more stable training.
  + Better generalization, increasing the batch size to 256 improved validation accuracy and helped the model generalize better.
  + Addition of eps and betas to the optimizer:
    - Epsilon(eps=1e-6) - It is a small constant added to the denominator of the Adam update formula to prevent division by zero.
    - Betas((0.9, 0.98)) - beta1 (0.9): Controls the exponential moving average of past gradients (momentum term) while beta2 controls the exponential moving average of squared gradients (second-moment estimate).
  + The training loss remained relatively stable, but validation loss showed fluctuations, especially in the later epochs.
  + The validation accuracy peaked early and then slightly declined, indicating possible early stopping effectiveness.
* **Error Analysis:**

A collage of white circles

AI-generated content may be incorrect.

* + Confusion between ‘Medium’ and ‘Many’ Classes: The model misclassified 45 instances of ‘Many’ as ‘Medium’, suggesting some difficulty in distinguishing between high numerosity levels.
  + ‘Few’ Class is Well Learned: Precision and recall for ‘Few’ are both high meaning the model identifies this class very well.
  + Some Overfitting Still Present: Even with increased dropout and weight decay, the validation loss fluctuates in later epochs, meaning further regularization might still be needed.
* **Next Steps:**
  + Evaluate alternative dropout rates (e.g., 0.35) to balance regularization and performance.
  + Fine-tune weight decay to minimize unnecessary penalization.
  + Increase dropout to regularize the model further.

**6. Conclusion**

This experiment successfully improved generalization by reducing the learning rate, increasing batch size, and strengthening regularization techniques. The model has achieved stable training, high classification performance, and better validation trends. However, misclassification between ‘Medium’ and ‘Many’ remains a challenge, and slight overfitting persists. Future experiments will further refine generalization strategies before moving into shape-based numerosity testing.

**7. Additional Notes**

* Reproducibility was ensured by setting a fixed random seed and using pre-saved datasets.
* This run also followed the structured experiment template, making future runs easy to compare.
* Some variability in validation loss was observed, which may indicate the need for better regularization techniques.
* Early stopping was applied, preventing overfitting, but further adjustments may be needed.