Experiment Run

Experiment Run Report

Experiment Title: Numerosity-Based Categorization - Experiment Run 1

Date: 17/02/2025

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	32
Learning Rate	0.001
Epochs	20
Optimizer	Adam
Loss Function	CrossEntropyLoss
Early Stopping	Yes (Patience = 3)
Device Used	GPU – NVIDIA L4

---

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

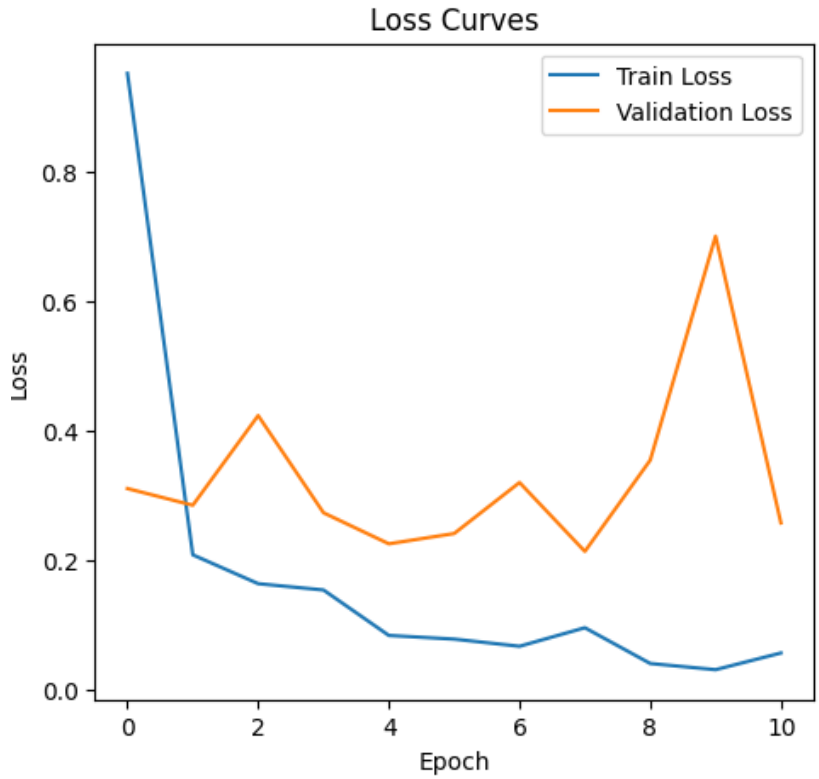
---

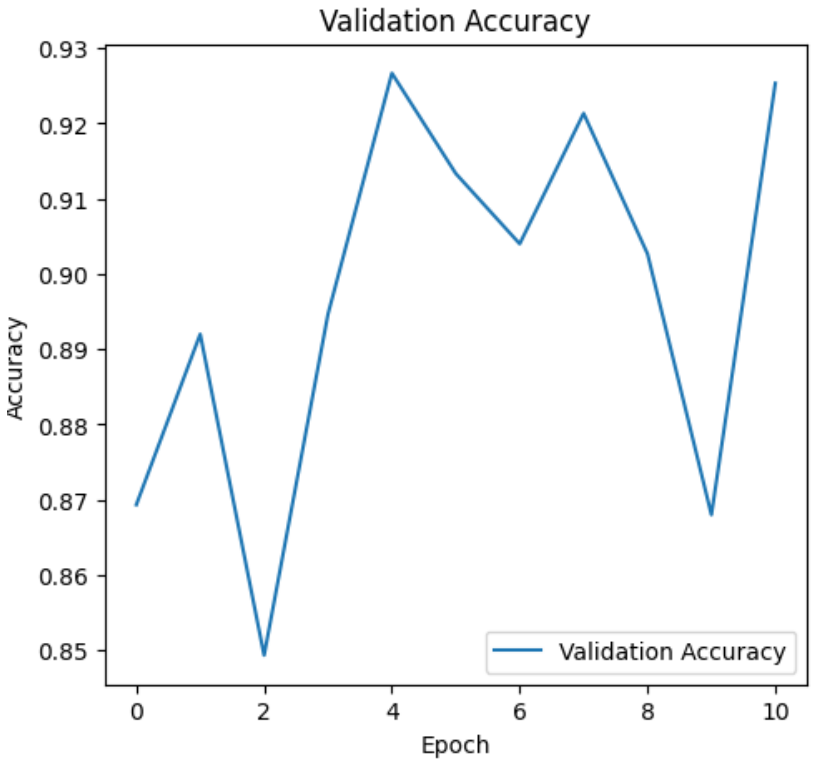
3. Training & Validation Performance

3.1 Loss and Accuracy Trends

Epoch	Train Loss	Validation Loss	Validation Accuracy (%)
1	0.9507	0.3106	86.93%
2	0.2085	0.2849	89.20%
3	0.1640	0.4232	84.93%
4	0.1543	0.2733	89.47%
5	0.0842	0.2255	92.67%
6	0.0785	0.2413	91.33%
7	0.0675	0.3200	90.40%
8	0.0961	0.2137	92.13%
9	0.0410	0.3544	90.27%
10	0.0314	0.6998	86.80%
11	0.0572	0.2576	92.53%

3.2 Loss Curve & Accuracy Plot

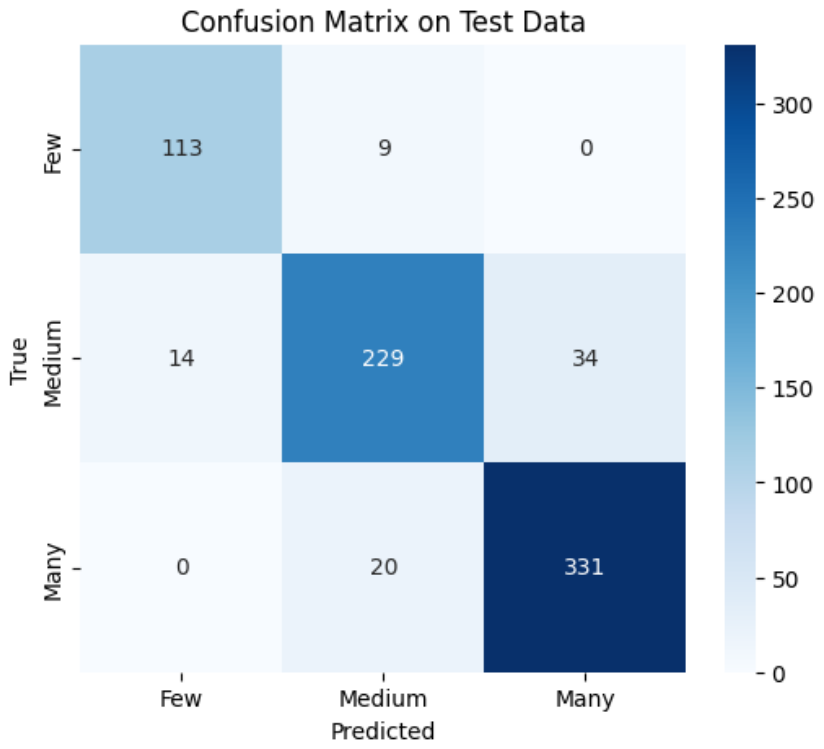




4. Test Set Evaluation

Final Test Accuracy: 89.73%

4.1 Confusion Matrix



4.2 Classification Report

Class	Precision	Recall	F1-Score	Support
Few	0.89	0.93	0.91	122
Medium	0.89	0.83	0.86	277
Many	0.91	0.94	0.92	351

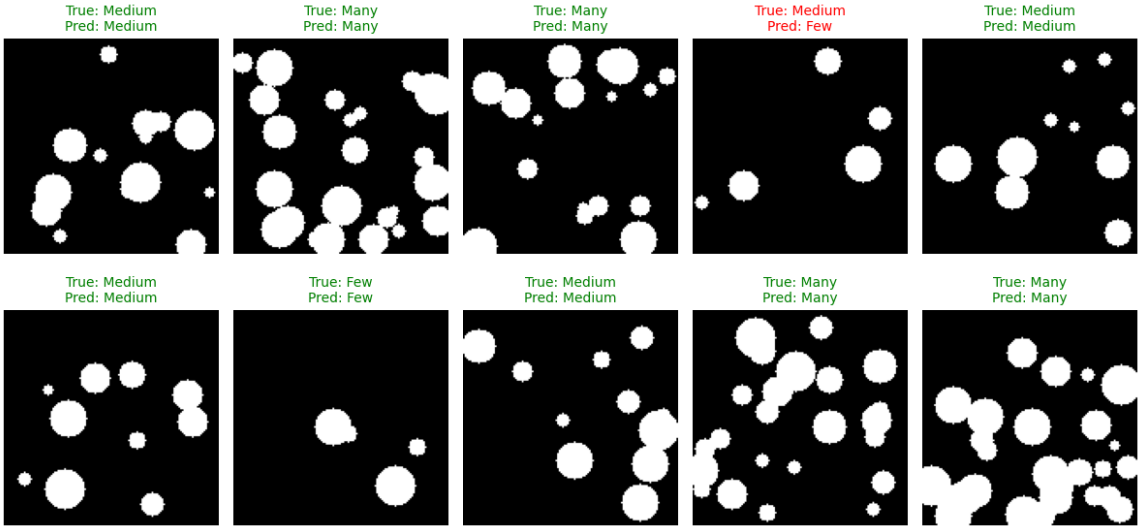
---

5. Observations & Insights

- **Key Findings:**
  - The training process showed a steady decrease in loss, but validation loss fluctuated, suggesting potential overfitting in later epochs.
  - The final test accuracy was 89.73%, which is decent but leaves room for improvement.

- The confusion matrix and classification report indicate that the "Medium" category had the most misclassifications, with some overlap between "Few" and "Medium" and "Medium" and "Many."

- **Error Analysis:**



  - Borderline cases: If an image has 5 or 6 dots, the distinction between Few and Medium is very fine, leading to possible misclassifications like the one seen in the above error analysis.
  - The model performed well in distinguishing "Few" and "Many", but had difficulty distinguishing "Medium", leading to some misclassifications.
  - The validation loss fluctuated, suggesting that the model may benefit from better regularization techniques or hyperparameter tuning.
  - Potential sources of error:
    - Similar dot distributions between "Medium" and "Few" categories.
    - Model may not have learned enough high-level features to distinguish borderline cases.
    - Refinement of thresholds.

- **Next Steps:**
  - Hyperparameter tuning: Experiment with different learning rates, batch sizes, and optimizer settings.
  - Increase epochs: Train for more than 20 epochs while monitoring loss trends.
  - Data augmentation: Introduce variations in brightness, occlusion, or slight rotations to help the model generalize better.

- Architectural improvements: Test deeper architectures or additional regularization layers to control overfitting.

---

6. Conclusion

The first experiment run on numerosity-based categorization achieved a test accuracy of 89.73%. While this is a solid starting point, analysis of the confusion matrix and classification report revealed that the Medium category had the most misclassifications, often being mistaken for Few or Many.

Improvements mentioned above will be the focus for the next run.

---

7. Additional Notes

- Reproducibility was ensured by setting a fixed random seed and using pre-saved datasets.
- The first run 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.
