Experiment Run

Experiment Run Report

Experiment Title: Numerosity-Based Categorization – Silhouettes Dataset

Date: 4/02/2025

Researcher: Karoki Evans Njogu

---

1. Experiment Details

Parameter	Value
Seed	42
Dataset Size	3000 samples
Image Size	128x128 pixels
Categories	Few (1-5), Medium (6-15), Many (>16)
Batch Size	256
Learning Rate	0.0003
Epochs	20
Optimizer	AdamW
Dropout Rate	0.4
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
Accumulation steps	2

---

2. Experiment Setup

- **Dataset:** Synthetic Dot Patterns
- **Model Architecture:** CNN-Transformer architecture
- **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

---

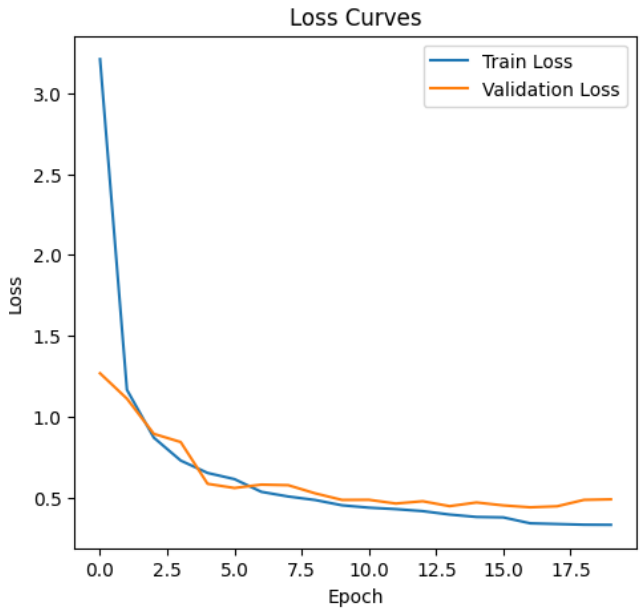
3. Training & Validation Performance

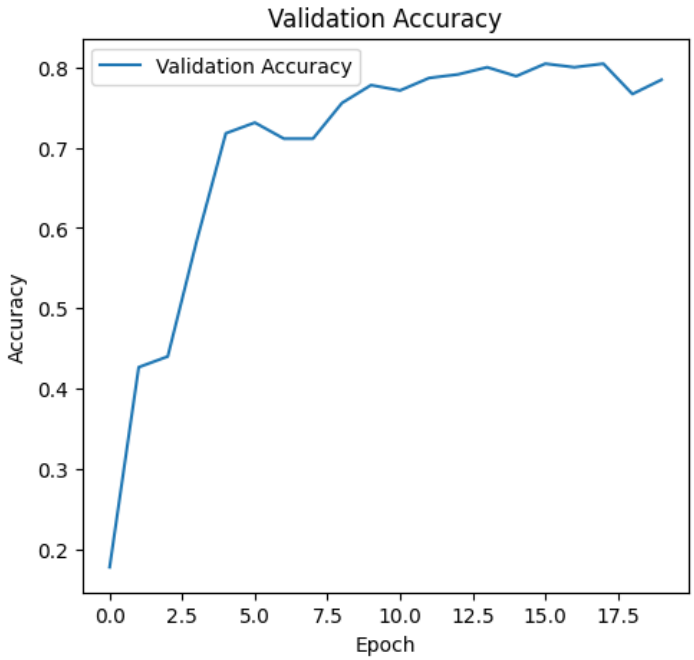
3.1 Loss and Accuracy Trends

Epoch	Train Loss	Validation Loss	Validation Accuracy (%)
1	3.2093	1.2686	17.78%
2	1.1668	1.1119	42.67%
3	0.8709	0.8944	44.00%
4	0.7294	0.8437	58.44%
5	0.6528	0.5857	71.78%
6	0.6146	0.5598	73.11%
7	0.5361	0.5807	71.11%
8	0.5072	0.5771	71.11%
9	0.4857	0.5268	75.56%
10	0.4529	0.4861	77.78%

Epoch	Train Loss	Validation Loss	Validation Accuracy (%)
11	0.4383	0.4868	77.11%
12	0.4293	0.4639	78.67%
13	0.4169	0.4780	79.11%
14	0.3958	0.4476	80.00%
15	0.3809	0.4704	78.89%
16	0.3781	0.4523	80.44%
17	0.3417	0.4405	80.00%
18	0.3373	0.4469	80.44%
19	0.3330	0.4865	76.67%
20	0.3322	0.4898	78.44%

3.2 Loss Curve & Accuracy Plot



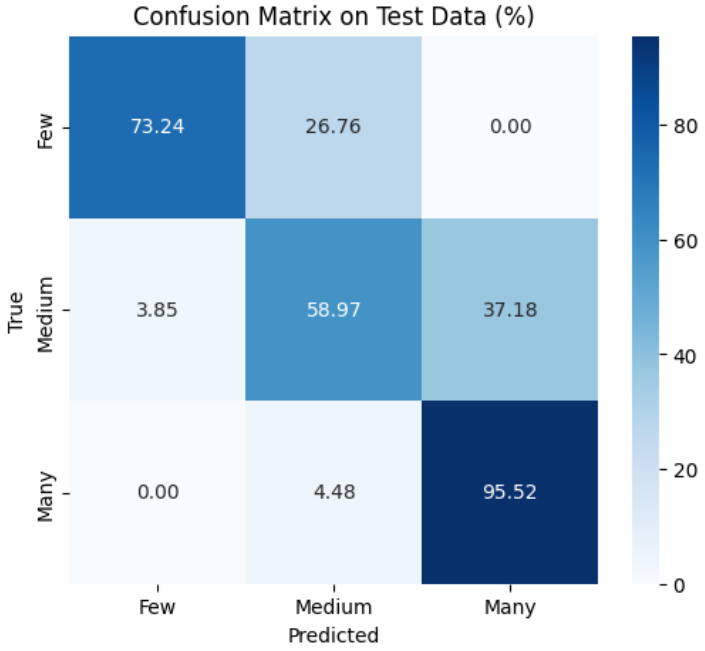


---

4. Test Set Evaluation

Final Test Accuracy: 79.33%

4.1 Confusion Matrix



4.2 Classification Report

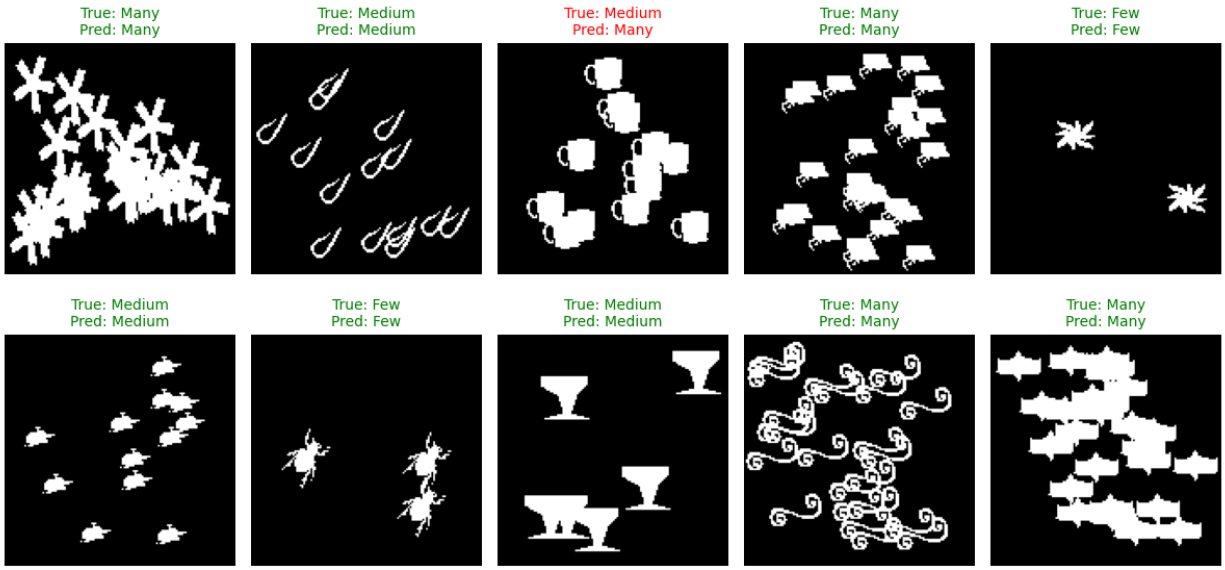
Class	Precision	Recall	F1-Score	Support
Few	0.90	0.73	0.81	71
Medium	0.76	0.59	0.66	156
Many	0.79	0.96	0.86	223

---

5. Observations & Insights

- **Key Findings:**
    - The final test accuracy was 79.33%, showing strong generalization from the training distribution.
    - Validation accuracy steadily improved, peaking at 80.44%, suggesting stable training.
    - The model performed best on 'Many' class (96% recall), which usually has more distinct spatial density.
    - The model had high precision on 'Few' (0.90), indicating strong confidence when predicting lower counts.

- **Error Analysis:**



- The 'Medium' class was the weakest, with only 59% recall, likely due to overlap with both 'Few' and 'Many'.
- Some misclassifications occurred when object scaling or spacing blurred class boundaries.
- False positives were more common for 'Medium' misclassified as 'Few' or 'Many', indicating model sensitivity to visual density and silhouette structure.

---

6. Conclusion

• The CNN-only model shows strong performance, especially in recognizing extreme classes ('Few' and 'Many').

• However, confusion around 'Medium' indicates that intermediate numerosity levels remain a challenge, possibly due to ambiguous spatial patterns.

• Compared to the CNN+Transformer hybrid, this model is slightly less accurate, but more interpretable and stable.

• Future improvements could include:

- Training on more balanced samples per class
- Adding auxiliary supervision or spatial attention
- Augmenting data with controlled overlap or crowding patterns

---