Glen Morgenstern

DATA 37000

Prof. Edwin Lo

Dec. 14, 2025

Convolutional Neural Networks on Open Images for Classification

1. Introduction

The problem addressed in this study is the classification of images into five categories: Bicycle, Car, Cat, Dog, and Tree. While this project may have little practical application, image classification in general is a tool with applications to many fields, like medical imaging, e-commerce, and facial recognition software. The challenge lies in building models that generalize well across diverse datasets, avoiding overfitting while maintaining high accuracy. This project explores two approaches: a baseline convolutional neural network (CNN) trained from scratch, and an improved model using transfer learning with ResNet18 pretrained on ImageNet.

---

2. Data Overview

The dataset was sourced from the Open Images repository (Krasin et al.), restricted to five classes of interest. This dataset comprises high quality images and is a good source of reliable images and clean classification for image classification model training. After cleaning and removing unlabeled data, the final dataset contained 1,366 images distributed across the five categories:

- Bicycle: 156
- Car: 402
- Cat: 217
- Dog: 303
- Tree: 288

Images varied in resolution, with an average size of 992×787 pixels, minimum size of 680×576, and maximum size of 2592×1944. This variability necessitated resizing during preprocessing to ensure consistent input dimensions for the models. There were no apparent image quality issues.

---

3. Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to understand dataset characteristics:

- **Class distribution**: Car and Dog classes were more represented, while Bicycle had fewer samples. Nevertheless, each class had more than 150 images, which was plenty for the models to train on and achieve high accuracy without overfitting.
- **Image sizes**: Most images clustered around 1000 pixels in height, with some outliers at higher resolutions.
- **Visual inspection**: Example images revealed diversity in lighting, orientation, and background clutter. This helps the models learn to classify images of the same class with a variety of attributes.

Examples below:

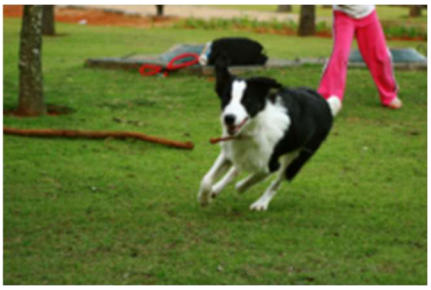
Figure 1. Bicycle example

Figure 2. Car example

Figure 3. Tree example

Figure 4. Cat example

Figure 5. Dog example



---

4. Baseline Neural Network

The baseline CNN consisted of three convolutional layers with ReLU activations and max pooling, followed by two fully connected layers with dropout regularization. Training was conducted for 10 epochs using the AdamW optimizer.

- Conv1: 3 → 32 filters, kernel size 3, stride 1, padding 1
- Conv2: 32 → 64 filters, kernel size 3, stride 1, padding 1
- Conv3: 64 → 128 filters, kernel size 3, stride 1, padding
- Pooling: MaxPool2d after each conv (kernel size 2, stride 2)
- Fully connected layers:
    - FC1: 128 × 28 × 28 → 256 units
    - FC2: 256 → 5 output classes
- Dropout: 0.5 after FC1
- Activation: ReLU throughout

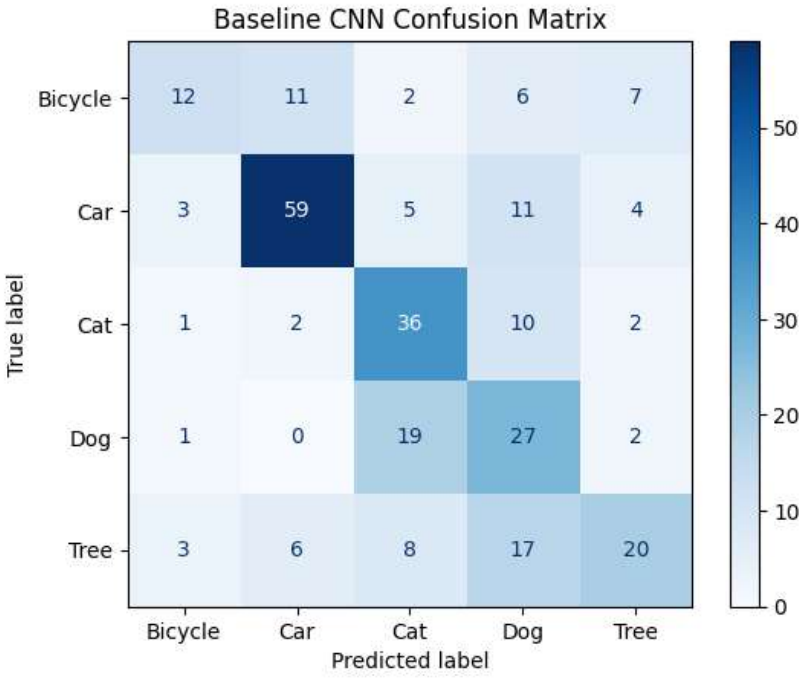
Results:

Figure 6. Baseline model training logs

Epoch	Loss	Training Accuracy	Validation Accuracy
1	60.9634	32.60%	43.43%
2	42.9659	51.10%	47.45%
3	38.0697	55.49%	45.26%
4	33.5410	60.16%	55.11%
5	28.7574	66.39%	52.19%
6	24.5109	72.99%	56.20%
7	20.4764	77.66%	55.47%
8	14.7558	83.61%	55.47%
9	10.5049	90.02%	53.65%
10	9.3753	89.93%	54.38%

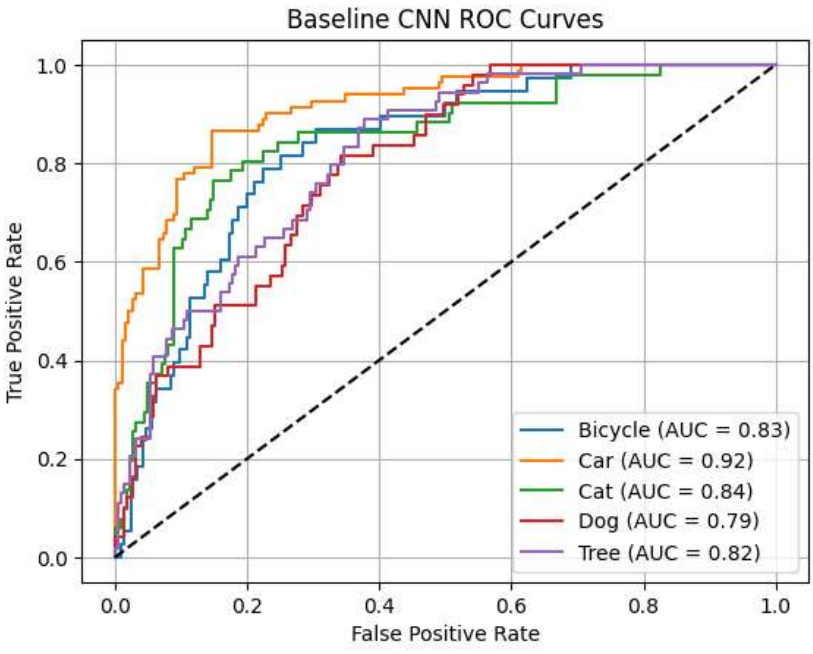
Training accuracy rose to 90.02%, showing that even this baseline model was able to learn patterns in the training set. However, validation accuracy plateaued around 52–57%, never approaching training performance. This is the hallmark of overfitting. That is, the baseline CNN model memorized training examples but failed to generalize to unseen data. After epoch 6, validation accuracy declined continuously, suggesting that the model's capacity exceeded the dataset size and diversity.

Figure 7. Baseline CNN model confusion matrix



The confusion matrix revealed misclassifications across all classes, particularly Bicycle and Dog. The Bicycle class had the lowest sample size of 156 images, which may have contributed to weaker feature learning. However, the Dog class had 303 samples. One potential explanation of poor classification performance is that there was high intra-class variability for Dog images. That is, there were too many breeds, poses, backgrounds, etc. for a shallow CNN model to capture consistent features.

Figure 8. Baseline CNN model ROC curves



The classification report showed macro F1-score of 0.53, with Car performing best (precision 0.76, recall 0.72). ROC/AUC curves indicated moderate discriminative ability, with AUC values ranging from 0.82 (Tree) to 0.92 (Car). While this baseline model did not perform poorly on the validation set (it did far better than random chance), it left much room for improvement.

5. Improved Model

The improved model used ResNet18 with pretrained ImageNet weights. Base layers were frozen, and a custom classifier head was trained for 15 epochs, with these specifications:

- Linear layer: in_features → 256
- ReLU activation
- Dropout (0.3)
- Linear layer: 256 → 5 output classes
- Activation: Softmax applied during evaluation for probability outputs.

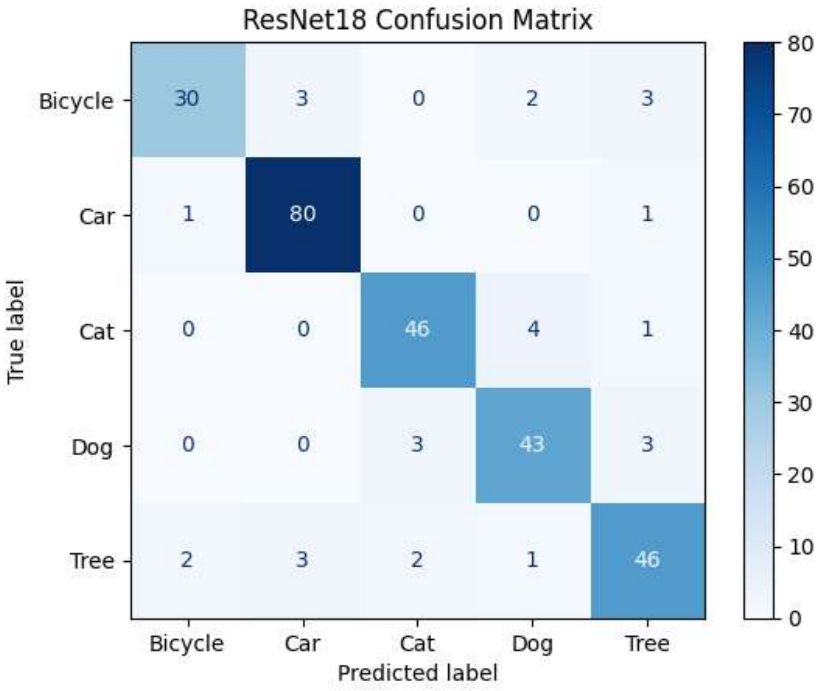
Results:

Figure 9. Improved ResNet18 model training logs

Epoch	Loss	Training Accuracy	Validation Accuracy
1	20.8699	82.42%	89.05%
2	12.5680	89.01%	85.77%
3	10.5588	90.48%	87.96%
4	8.7057	92.58%	89.05%
5	9.0611	90.93%	89.42%
6	6.8172	94.05%	87.59%
7	8.2593	91.85%	89.05%
8	7.0338	93.04%	89.42%
9	6.2759	95.05%	88.69%
10	6.1549	93.68%	88.32%
Early stopping triggered			

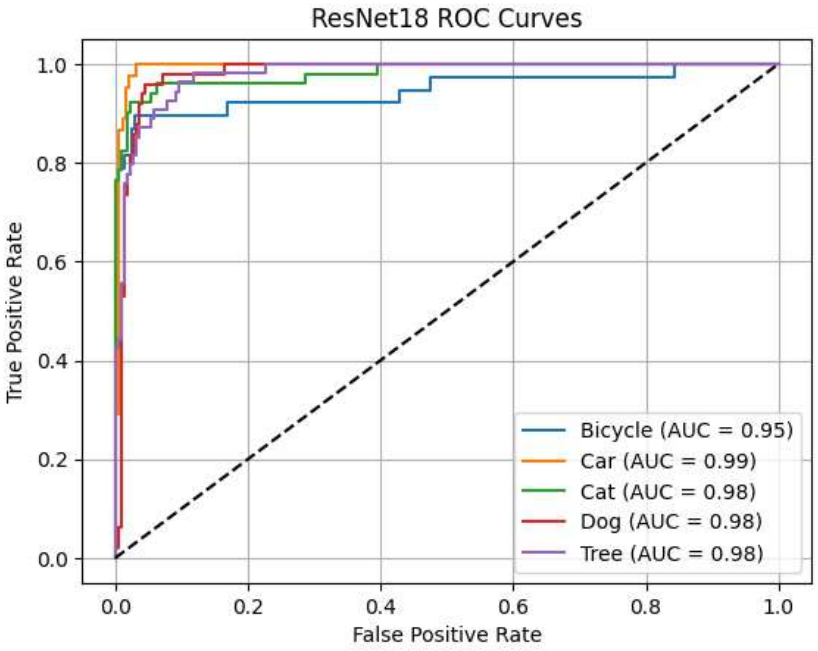
Training accuracy reached 95.05%, with validation accuracy stabilizing around 88–89%. The confusion matrix showed strong performance across all classes, with minimal misclassifications (Figure 10). Cars and Cats were classified almost perfectly, while even the hardest class (Dog) achieved strong recall. The classification report yielded a macro F1-score of 0.88, with all classes above 0.85 F1. With a macro F1-score that high, it is clear that this improved model performs consistently across all classes, not just the majority ones. Early stopping was triggered due to failure to increase validation accuracy after five epochs, but the resulting model still produced a high level of accuracy.

Figure 10. ResNet18 model confusion matrix



ROC/AUC curves demonstrated excellent discriminative ability, with AUC values between 0.95–0.99 across all classes. High performance was consistent across all five classes, speaking to ResNet18's ability to learn generalized features on a limited dataset. This model seems to be good enough for practical applications. Possibly, with a larger set of training data, this ResNet18 model could learn features even better for a classification accuracy approaching 100%.

Figure 11. ResNet18 model ROC curves



Clearly, ResNet18's transfer learning provided substantial improvements in generalization and robustness compared to the baseline CNN.

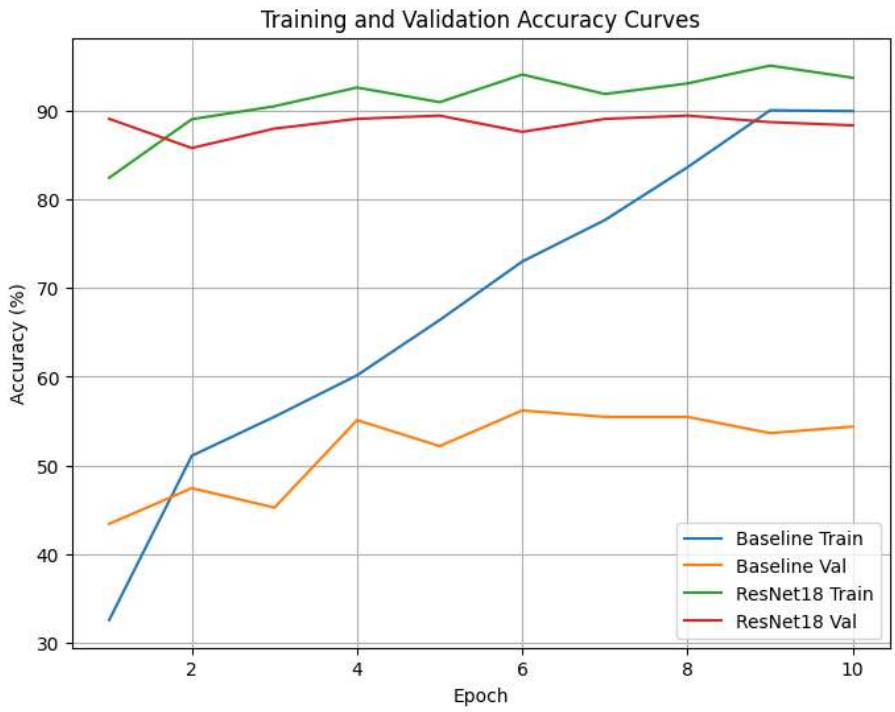
6. Comparison of Results

Figure 12. Model comparison table

Metric	Baseline CNN	ResNet18 Transfer Learning
Validation Accuracy	~56%	~89%
Macro F1-score	0.53	0.90
Best Class Performance	Car (0.74 F1)	Cat (0.92 F1), Car (0.94 F1)
ROC/AUC Range	0.82-0.92	0.97-0.99
Overfitting	High	Minimal

The baseline CNN model started from random initialization, meaning it had to learn all features (edges, textures, shapes) from scratch on a relatively small dataset (~1,300 images). This led to memorization of training data and poor generalization. Meanwhile, the ResNet18 model was pretrained on ImageNet, composed of millions of diverse images. It already "knows" how to detect low-level and mid-level features, so fine-tuning only required adapting the final layers to the five chosen classes. This shortcut dramatically improves accuracy and reduces overfitting.

Figure 13. Training and validation accuracy curves for baseline and improved models



Another feature of ResNet18 that helped improve performance was the fact that it uses skip connections (residual links) that allow gradients to flow more easily during backpropagation. This solves the vanishing gradient problem common in deeper networks, enabling ResNet18 to train effectively and capture complex patterns. Our baseline CNN, with only three convolutional layers, lacked this mechanism and struggled to represent high-level features.

As a result, our three-layer CNN model stayed “shallow” and couldn’t dig deeper to find distinguishing features of each of the five classes of interest. The 18-layer ResNet model was able to quickly pull identifying features for the five classes of interest due to its pretrained weights and skip connections. As we see in Figure 13, validation accuracy neared 90% after the first epoch of the ResNet18 model, while the relatively naïve baseline model began with a validation accuracy below 45% before climbing modestly.

---

7. Discussion

This study demonstrated the strengths of transfer learning and the limitations of training from scratch. The baseline CNN achieved high training accuracy but failed to generalize, with validation accuracy plateauing around 55% and weak per-class performance. In contrast, ResNet18 leveraged pretrained ImageNet features to achieve balanced results across all categories, with validation accuracy near 90% and consistently high F1-scores. Confusion matrices and ROC/AUC curves confirmed the superior discriminative ability of the improved model, underscoring the importance of transfer learning when data is limited.

At the same time, the baseline model revealed what does not work: shallow architectures trained on small, imbalanced datasets are prone to overfitting and poor recall, particularly for minority or visually diverse classes such as Bicycle and Dog. These results highlight practical and ethical considerations for deployment. Models trained on imbalanced data risk bias toward majority classes, and in sensitive domains such as healthcare or surveillance, issues of fairness, privacy, and interpretability must be addressed. Moreover, computational demands of deeper pretrained networks may pose challenges for deployment in resource-constrained environments.

The dataset itself imposed limitations. Class imbalance favored Cars over Bicycles, label noise may have affected training, and the relatively small sample size restricted the effectiveness of models trained from scratch. With more time, the system could be improved by expanding and balancing the dataset, fine-tuning deeper layers of ResNet18, incorporating interpretability tools such as Grad-CAM, and optimizing the model for efficiency through pruning or quantization. These steps would enhance robustness, fairness, and suitability for real-world applications.

Bibliography

Krasin I., Duerig T., Alldrin N., Ferrari V., Abu-El-Haija S., Kuznetsova A., Rom H., Uijlings J., Popov S., Kamali S., Malloci M., Pont-Tuset J., Veit A., Belongie S., Gomes V., Gupta A., Sun C., Chechik G., Cai D., Feng Z., Narayanan D., Murphy K. *OpenImages: A public dataset for large-scale multi-label and multi-class image classification*, 2017. https://storage.googleapis.com/openimages/web/index.html.