DLiP – Assignment 3:

Transfer Learning on Galaxy Zoo 2 Dataset

Bradley D.K.C. Spronk

February 27, 2026

Introduction

Galaxy Zoo 2 is a large-scale dataset of galaxy images labeled by morphological features. The decision tree for classification is detailed in Willett et al. [1]. Kaggle hosted a competition to predict galaxy morphologies from these images [2].

Methods

ConvNeXt Model

I used the small ConvNeXt model [3], a modern convolutional neural network with a feature extractor, fully connected layers, and an output layer (see Fig. 1). Each block consists of depthwise convolution, layer normalization, pointwise convolution, and GELU activation. The Galaxy Zoo 2 dataset was preprocessed by cropping, resizing, and normalization using ConvNeXt-specific mean and standard deviation. Training was performed in two ways: on the full dataset (normal training) and with curriculum learning, where data was staged by label confidence. In both cases, only the final layer was trained, with all other weights frozen to leverage pre-trained features. The model was trained for 20 epochs, batch size 512, using Adam optimizer with a learning rate of 2e−4.

Transfer Learning

Transfer learning is a technique where weights from a model trained on a large dataset are reused for a related task [4]. I fine-tuned a pre-trained ConvNeXt model on Galaxy Zoo 2 by updating only the final layer, efficiently adapting it to the new task.

Curriculum Learning

Curriculum learning is a strategy where a model is first trained on easier examples and gradually exposed to harder ones. [5] For this project, I split the training data into three stages based on label confidence: high (> 0.94), medium (0.85-0.94), and low (< 0.85), each about 33% of the data, or in other words: $\sim 20,000$ images.

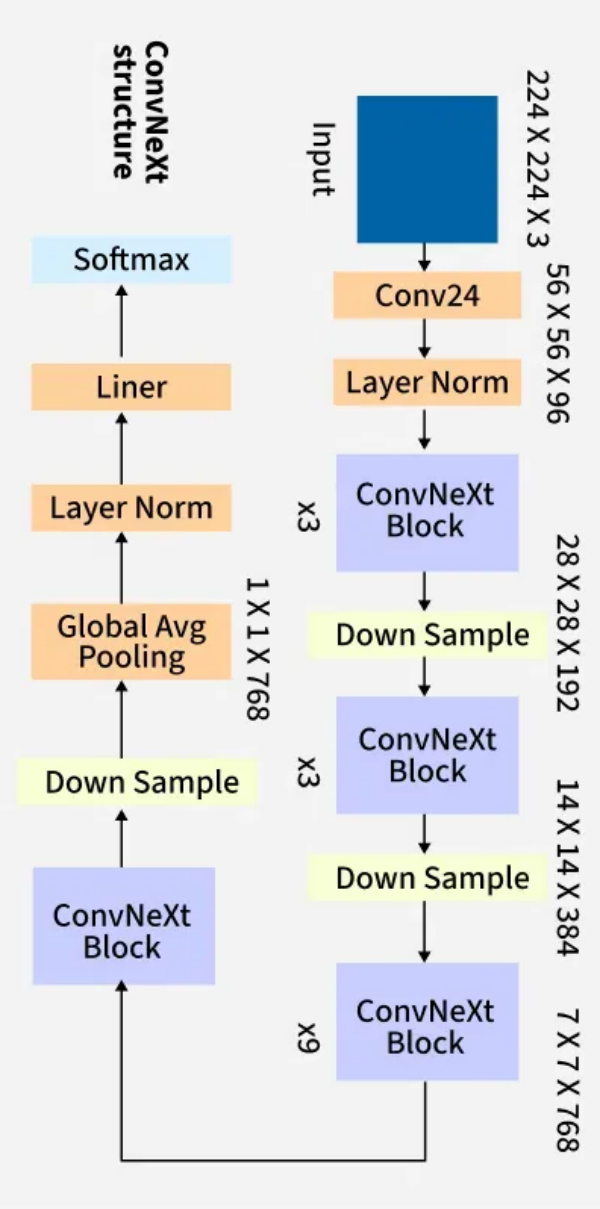

Figure 1: ConvNeXt architecture [6]: convolutional feature extractor, fully connected, and output layers. Only the final layer is trained.

Results

Figure 2 presents the training loss of the ConvNeXt model, with the results for standard training depicted in gray and those for curriculum learning illustrated in shades of green and blue. The results indicate that curriculum learning can improve classification performance by allowing the model to learn from easier examples before tackling more difficult ones.

The normal and curriculum learning models achieved final Kaggle RMSE scores of 0.XXXXX and 0.XXXXX, respectively. For comparison, the competition winner scored 0.07491.

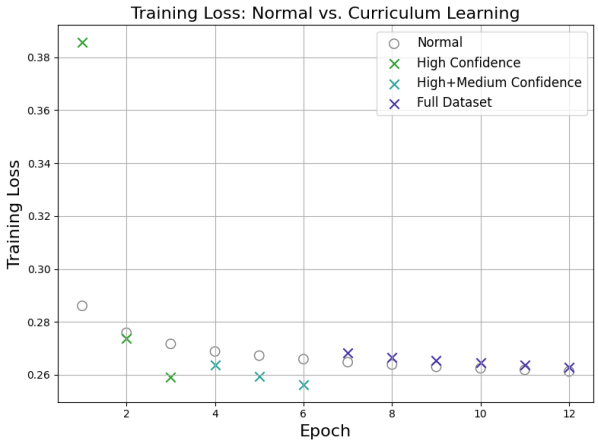

Figure 2: Training loss of the ConvNeXt model with and without curriculum learning. Note that this is not RMSE.

Discussion

My training was limited by hardware constraints, with each epoch requiring 6 minutes. Extending training with at least 10 more epochs or fine-tuning additional layers will improve the performance of both normal and curriculum learning. Curriculum learning's initial loss is higher due to a smaller dataset. The model quickly learns to classify easier examples, with loss decreasing significantly faster than normal training. One thing to note is that while it seems that both models arrive at a similar final loss in the same number of epochs, the curriculum learning model took less (real) time to get there. Curriculum learning was around 17% faster than normal learning, total training time was 1.5 hours and 1.25 hours for normal learning and curriculum learning respectively. It is unclear if this behavior will persist with different hyperparameters.

Conclusion

Transfer learning with ConvNeXt enables decent galaxy morphology classification on Galaxy Zoo 2. Further improvements could be achieved by increasing training duration, or fine-tuning more layers. Or, of course, using a bigger model and retraining it completely. Curriculum learning shows promise in reducing training time, however further investigation is needed. My results show that data selection strategies also impact results, with direct transition from high-confidence to full dataset showing promise.

References

[1] K. W. Willett *et al.*, "Galaxy zoo 2: detailed morphological classifications for 304,122 galaxies from the sloan digital sky survey." [Online]. Available: https://arxiv.org/abs/1308.3496

[2] AstroDave, "Kaggle galaxy zoo the challenge." [Online]. Available: https://www.kaggle.com/competitions/galaxy-zoo-the-galaxy-challenge

[3] Z. Liu *et al.*, "Convnext: A convnet for the 2020s." [Online]. Available: https://arxiv.org/abs/2201.03545

[4] J. Murel, "What is transfer learning?" [Online]. Available: https://www.ibm.com/think/topics/transfer-learning

[5] Y. Bengio *et al.*, "Curriculum learning." [Online]. Available: https://arxiv.org/abs/0804.0866

[6] GeeksforGeeks, "Convnext." [Online]. Available: https://www.geeksforgeeks.org/computer-vision/convnext/