Final assignment GRA4157

Jakob Sverre Alexandersen

ja.al@fsncapital.com

November 30, 2025

Contents

1 Introduction

Click me to view the full code on Github!

This final project builds (conceptually) on my second project. A short recap of that was to go in-depth on how CNNs (Convolutional Neural Networks) are actually built. To demonstrate this, I built a relatively simple classifier (found in the `simple_bootleg_cnn_classifier.py` file). Furthermore, I ran it on the well-known MNIST dataset, which is a balanced dataset of $\sim 21,000$ images of the digits 0 to 9. It achieved both an $F1$ and accuracy score of 99%, which is considered industry standard. Since the problem is considered solved, it inspires very little motivation to do further research on the subject.

In my previous project, I very briefly mentioned Belkin et. al's theory of *double descent* [BHMM19], which challenges the traditional bias-variance tradeoff. The theory posits that test error exhibits non-monotonic relationship with model complexity and training time: performance initially worsens after the interpolation threshold (where the model perfectly fits training data), but surprisingly improves again with further increases in complexity or training epochs. However, MNIST's simplicity, with near-perfect accuracy already achieved, provides limited opportunity to empirically observe this phenomenon.

To meaningfully test double descent, I require a more challenging dataset where my model has substantial room for improvement. Therefore I will apply my CNN architecture to the CIFAR-10 dataset, which consists of $60,000$ 32×32 color images across 10 classes (airplane, car, bird, cat, deer, dog, frog, horse, ship and truck). CIFAR-10 presents significantly greater complexity due to color channels, natural image variations, and inter-class similarity, typically yielding baseline accuracies of $70 - 85\%$ without advanced techniques.

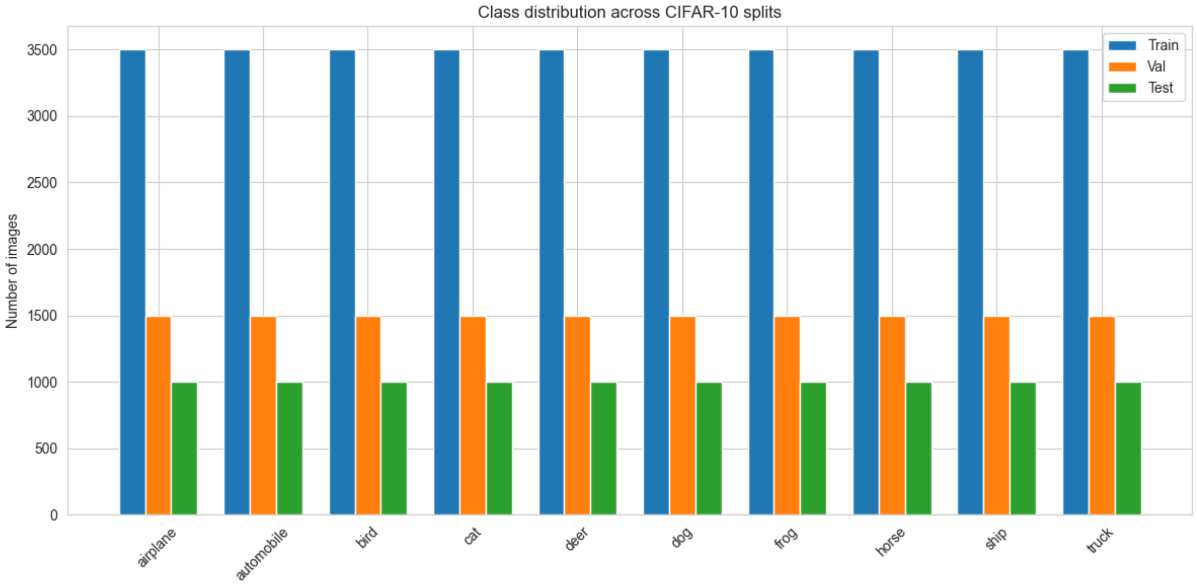
The objectives of this project are therefore twofold: **(1)** to investigate whether double descent can be empirically observed by varying model complexity (number of layers, filters, and parameters), and **(2)** to examine epoch-wise double descent by training models well beyond typical convergence points. This exploration will provide practical insights into modern deep learning phenomena that remain theoretically incomplete.

Remember to mention that I ran into an OOM issue, which is why I oved the validation through the network in batches to avoid the out-of-memory issue. the RTX is not that goated. heck ...

2 Methodology

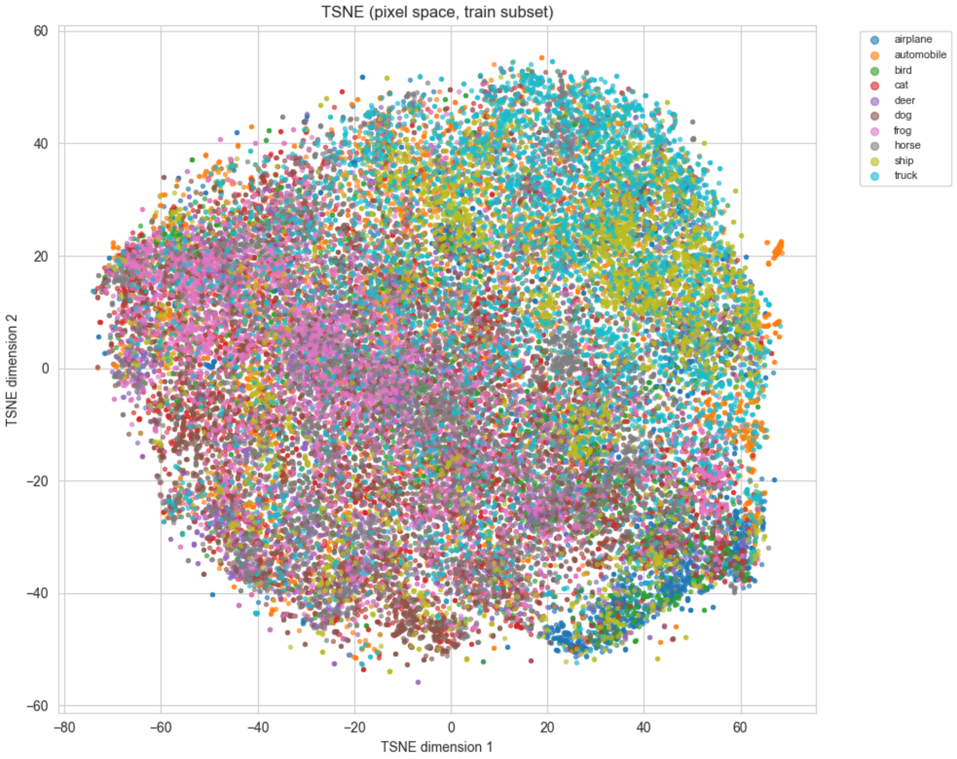
2.1 Data

Before we dive into the architectures of the different models and other techincals, it is important to achieve an understanding of the data itself. Therefore we start off by seeing how the distribution of the data looks like, to ensure that the data is indeed balanced.



The data is balanced, which means that we will not accidentally train a biased model that performs poorly on underrepresented classes. If that were the case, the model may have struggled to correctly identify the classes with fewer examples, potentially resulting in lower accuracy and recall for those classes.

If we run a tSNE analysis on the test set, we get this plot:



I can already tell from here that the model is not going spark impressive feedback, as the classes are very mixed into each other, and there is no clear separation between clusters. (TALK MORE ABOUT TSNE AND THIS IF ROOM ALSO ADD PICTURE OF IMAGES)

3 Results

4 Summary

References

[BHMM19] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, July 2019.