Capstone Final Report

Gavin Mai

May 2026

Abstract

This capstone project produces a self-contained introduction to artificial neural networks aimed at readers with no prior background in machine learning, statistics, or information theory. The accompanying paper covers the foundations of neural network training, large language models, and diffusion models, together with runnable code examples. Both the structure of the report and the strategy for explaining each topic are guided by the dual-process model of cognition popularized in Veritasium's *The Science of Thinking* [1]. The report was evaluated through a Google Form completed by 57 readers; their feedback informed targeted revisions to definitions and appendix material.

1 Introduction

Learning a new topic is hard, and while there are many reasons for this, the most fundamental is a lack of context. Complex topics often draw on a background in adjacent fields. Many machine learning principles, for example, are based on information theory: entropy, information gain, cross-entropy, and so on. The problem is that resources on machine learning and information theory each tend to teach these concepts only from their own perspective.

The clearest example of this is cross-entropy loss. In information theory, cross-entropy is the number of bits needed to encode data drawn from one distribution using a code optimized for another, and the concept is typically introduced through surprise and entropy. In computer science, cross-entropy is the loss function that measures the difference between the true distribution and the predicted distribution in multi-class classification, and the concept is typically introduced to evaluate a model's prediction. The same cross-entropy, viewed from these two fields, has different motivations and a different vocabulary, yet exactly the same mathematical formula. To make matters more confusing, in deep learning the loss is actually derived from maximum likelihood estimation, which is itself a concept from statistics rather than from computer science or information theory. The goal of this project is to introduce deep learning while supplying enough context for someone with no prior background in any of these fields to follow along.

2 Methodology

The methodology and layout of the report are inspired by Veritasium's *The Science of Thinking* [1]. The brain is modeled as having two systems. System 1 receives information

from the senses—audio, visual, tactile, and so on—and processes it quickly and reflexively. Because System 1 can hold only a small amount of information at once, it is poor at deliberate reasoning and instead handles most familiar tasks by reflex or instinct, often described as muscle memory.

System 2 is responsible for conscious thought. It is much slower and lazier than System 1, and it requires reasoning through details before producing an answer. System 2 is also responsible for the long-term retention of knowledge and for supplying System 1 with new ways to automate familiar tasks.

The concept of *chunks* explains how System 1 can improve despite its limited capacity. A chunk is any single piece of information, regardless of size. The phrase "Mona Lisa," for example, can be processed as 8 chunks letter by letter, 2 chunks word by word, or 1 chunk when recognized as the name of a single image. Learning, in this framing, is the process of getting System 2 to understand new material thoroughly enough that it can be distilled into larger chunks that System 1 can use automatically. The difficulty is that System 2 is fundamentally lazy: when information is presented as a single, easy stream, System 1 absorbs it on its own and the material never reaches System 2 for long-term processing. One of the most effective techniques in advertising exploits this directly—a brief jolt of confusion forces the viewer's attention to switch to System 2, which makes them more engaged.

With this framing in mind, the report is laid out as follows. The early sections provide an introduction in plain language, using everyday scenarios so that System 1 can follow along easily. This is largely passive learning, so the reader is not expected to retain much from these sections in detail; their purpose is to act as a hook that keeps the reader moving into the more difficult material that follows.

When the more demanding sections arrive, the sudden increase in difficulty forces the reader's attention to switch to System 2. A common obstacle in technical reading is that the search for an explanation pulls the reader away from the paper, and by the time they return they have lost momentum. To prevent this, links are provided throughout the report to short, accessible explanations of key prerequisites, written in the same approachable style as the introductory sections. Part of the editing process was asking readers explicitly whether they ever needed to consult external resources, with the goal of eliminating that need entirely.

Finally, the code sections provide a sense of purpose. Learning new information is rewarding in its own right, but without a tangible result, readers—especially those outside the field—may not be motivated to retain it. System 2 typically requires several passes over material before it is fully internalized, and the code, together with its visible outputs, gives readers a reason to come back. Customizing one of the example models inevitably requires revisiting specific sections of the report to understand how a given component should be implemented, which produces the repeated exposure that long-term retention depends on.

3 Evaluation

Evaluation was interleaved with the editing process. A Google Form was created so that readers of the draft could provide feedback. A total of 57 people responded; the large majority were computer science undergraduates. This is not an ideal sample, since computer science undergraduates already understand a substantial fraction of the material before

they begin reading. Responses prompted a few targeted edits—tightening definitions and improving several explanations in the appendix—but nothing that fundamentally changed the structure of the report.

4 Discussion

This project is significantly smaller in scope than originally intended. The initial goal was to compile all of machine learning into a single introduction; this was later narrowed to neural networks, and the final report covers only the neural network concepts most relevant to large language models and diffusion models. Whole families of architectures are deliberately left out, including time-based networks (such as RNNs and LSTMs), graph-based models, true reinforcement learning models, GANs, VAEs and many more. Aside from not having time to produce code examples, the binding constraint was narrative flow: the report is built around a single arc that begins with the basics of neural networks and ends with large language models and diffusion models, and most of the omitted topics—with the partial exception of GANs and VAEs—would have disrupted that arc. Even so, feedback from the Google Form indicates that the central goal of the project, making neural networks accessible to readers without prior background, was achieved, even if only for a biased sample.

References

[1] Veritasium. The science of thinking. YouTube video, https://www.youtube.com/watch?v=UBVV8pch1dM, 2017. Accessed: 2026-04-28.