Symbol Manipulation in Neural Networks

Jonas Peters, Philipp Haack

August 1, 2025

Contents

1 Introduction

The human brain is a remarkably complex and efficient system, capable of performing an extraordinary range of cognitive tasks with ease. Its ability to learn from experience, adapt to new environments, and process vast amounts of sensory information has long fascinated scientists and engineers alike.

The study of its functionality is partially in the domain of computer science: We try to model the brain with computer algorithms, which is inspired by the idea that the human brain consists of elementary computational units, neurons, which itself can be sufficiently modeled with computers, and whose interplay creates the emergent complex behavior we see. This leads to the following premise:

Premise 1.1. *In theory, the human brain can be sufficiently modelled by a computer program.*

Over many decades scientists have been trying to make progress towards this monumental task of modelling the human brain. As a result, two paradigms on how to implement intelligent models emerged, they are called *connectionism* and *symbolism*. While connectionist models rely on training data, symbolist models have all their inference rules and algorithms directly coded into them. As a result, many connectionist models are black box ... associated models...

What up?

> See [**?**] for an in-depth discussion.

The mere fact that the human brain consists of neurons does not prove or disprove one of the paradigms.