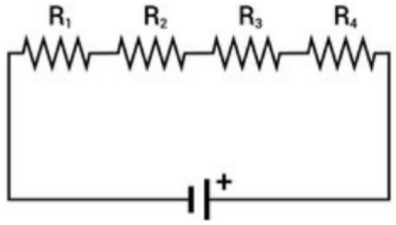
Eve Burton
Supervisor: Ben Hanson

What are Machines Learning?
An Investigation into the Knowledge Content of Machine Learning Algorithms From the Perspective of Theoretical Physics

Introduction

Can we look inside a neural network and see if it has learned the physics of an electrical circuit?
AI is becoming increasingly common, but we still do not fully understand how machine learning models learn and make predictions. This project investigates what happens inside a simple neural network when it is trained to predict the behaviour of an electrical circuit.
A four-resistor series circuit was used as a simple, controlled system where the expected physics is already known. I trained a neural network to predict the circuit behaviour and then examined its weights and biases to see whether they reflected the underlying physics.

Experimental System Circuit



Research Aim- To investigate how neural networks learn and represent the underlying physics of electrical circuits.

Methodology

1: Generating Data

1. Built a simulated 4-resistor circuit
2. Generated random resistor values and input voltages
3. Calculated the corresponding current and voltage drops using Ohm's Law
4. Used the calculated values as the model's expected outputs

2: Neural Network Training

1. Built a simple neural network with a single dense layer
2. Used resistor values and voltage drops as inputs
3. Trained the network to predict current and voltage drops
4. Examined the learned weights and biases

3: Models

1. Started with a simple four-resistor model
2. Varied the number of neurons to investigate network size
3. Analysed how network size affected the internal representations
4. Added a transistor to test a more complex physical system

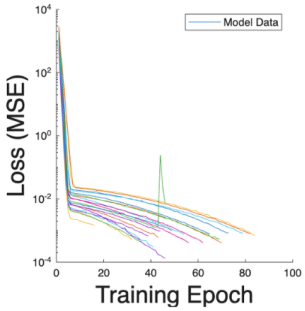
4: Physical Based Analysis

1. Extracted the learned weights and biases
2. Calculated covariance and correlation matrices
3. Identified patterns in how parameters varied across neurons
4. Calculated diagonalness to measure changes in correlation structure
5. Compared these patterns as the number of neurons increased

Results

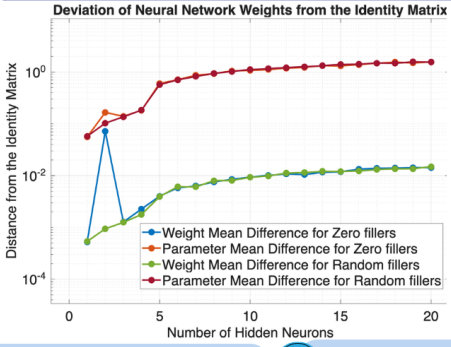
1 Model Learning

Loss decreased during training, showing that the network learned the circuit behaviour.

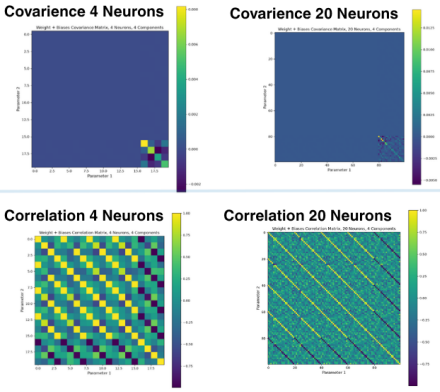


2 Physcial Representation

The learned weights showed an identity-like structure, closely matching the expected physical relationship.



3 Internal Parameter Structure

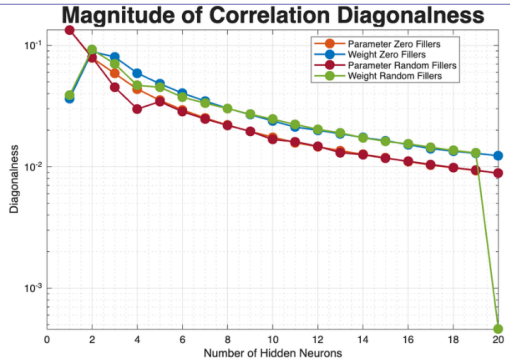


Covariance: Biases showed stronger covariance, while most weights showed relatively low covariance.

Correlation: Weights showed a strong diagonal pattern, while biases showed stronger correlations with each other.

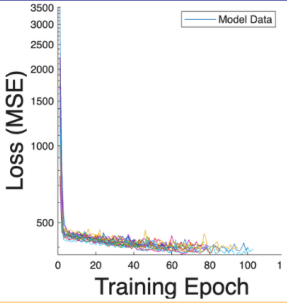
4 Effect of Network Size

Diagonalness decreased as the number of neurons increased across all parameters.



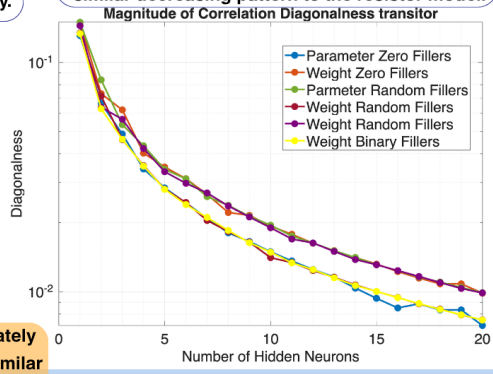
5 Transistor Model

The high loss shows that the transistor model was unable to predict the behaviour accurately.



The Transistor model could not predict accurately buts its correlation diagonalenss showed a similar decreasing patten

However, the correlation diagonalness followed a similar decreasing pattern to the resistor model.



Conclusions

- The neural network successfully learned the physics of the four-resistor circuit.
- Its internal representation changed as the number of neurons increased, moving further from the simple expected structure.
- The linear model could not accurately learn the more complex transistor behaviour.
- Similar internal patterns were still observed, suggesting that not all patterns in the network necessarily represent the underlying physics.

Scan for full paper