How do machines learn: can a neural network be called artificial intelligence?

James Mann

Abstract

In this project I aim to explain the functioning of a neural network and decide if it is truly intelligent or just complex pattern analysis. My finished product is a convolutional neural network trained on a dataset of 30000 labelled images to differentiate between pictures of 15 species of animal; it achieved an 82% accuracy on validation data. Alongside this, the project delves into the epistemological and metaphysical implications arising from advancements in modern AI and broader philosophical questions about the nature of intelligence and the rise of machine learning.

I Introduction

We stand on the precipice of an Artificial Intelligence revolution. Systems employing AI have pervaded every department of modern life. This EPQ aims to dissect the machinations of neural networks, seeking to understand their function and the nature of their ‘intelligence’, if it is deserving of that label. The driving question "How do machines learn: can a neural network be called artificial intelligence?” provokes a nuanced exploration of what it is to be intelligent, and the grey boundaries between that, and complex pattern recognition.

To ground this exploration, I have built a convolutional neural network, in this project I will explain at a high level each step in its development and the part it plays in creating a final system capable of appearing intelligence-like. This shall run in tandem to investigations into the broader epistemological and metaphysical implications of machines that can think.

II What is Intelligence?

In my research for this project, a recurring theme I encountered was the reliance upon analogies and thought experiments acting as proxies to more abstract questions. The first, perhaps most notable, incident of this was Alan Turing proposing his eponymous Turing Test (Turing, 1950). He argues against the absurdity inherent in questions like "can machines think" and how they hinge upon the consensus fallacy that interpretations of words like "think" are universally ubiquitous. That is not to say Turing considered his test infallible, he doesn't, rather he considers it informative in a different way. Largely, the outcome of the test is irrelevant; Turing's true intention is to force you to re-examine your own understanding of intelligence.

Similarly, the Chinese Room thought experiment (Searle, 1980), takes an abstract concept, strong AI, and designs a contradictory gedankenexperiment. Strong AI is the belief that the computer is not merely a tool in the study of the mind; rather, the appropriately programmed computer is a mind. By contrast, weak AI holds that a computer can only ever be a tool, never itself possess a mind. The Chinese Room goes as follows:

> "Imagine a native English speaker who knows no Chinese locked in a room full of boxes of Chinese symbols (a data base) together with a book of instructions for manipulating the symbols (the program). Imagine that people outside the room send in other Chinese symbols which, unknown to the person in the room, are questions in Chinese (the input). And imagine that by following the instructions in the program the man in the room is able to pass out Chinese symbols which are correct answers to the questions (the output). The program enables the person in the room to pass the Turing Test for understanding Chinese but he does not understand a word of Chinese." – (Searle's summary, 1999)

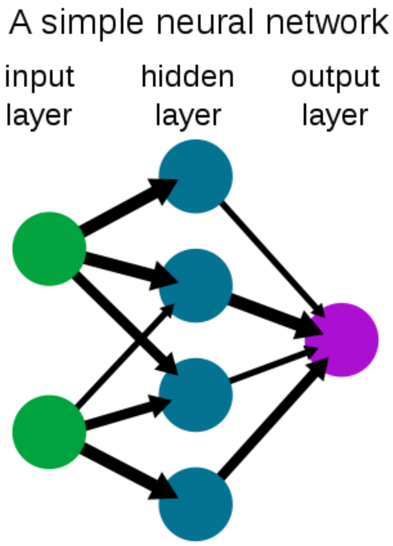
In his paper, Searle proposes something called intentionality

Searle believes that partisans of Strong AI would consider the system to possess a cognitive state, an understanding, and that this is demonstrably false. I disagree with Searle. I do not find it readily apparent that the system is unintelligent, Searle's dismissal is, to me, an act of mind chauvinism. From our omniscient god's-eye-view we can, with high confidence, say that the system is not intelligent; however, could any onlooker from within the system not also say, with equally high confidence, that the system they have before them is intelligent. The system certainly exhibits intelligence-like abilities. Searle's argument hinges upon his assumption that both process and outcome are necessary to prove intelligence; and that no computerised process will ever be sufficient. I align more with the functionalist view of being hardware ambivalent, preferring the computational theory of mind that the human brain is simply highly specialised hardware. This would place decision-making, reasoning, and perception as the high-level consequences of emulations within the brain.

The significance of this to our investigation is that before being able to classify something 'artificial intelligence', it is first prudent to have a rigid definition for artificial intelligence. The purpose of this section is to show that there are many such definitions.

III Neural Networks

Recent successes in AI have been, almost entirely, due to the advent of the neural network. The neural network is heavily inspired by the structure of the human brain, specifically, the interconnectedness of neurons. A neural network consists of an input layer, one or more hidden layers, and an output layer. A layer is just a stack of neurons. The input layer receives raw data, which then makes a pass through the hidden layers, and finally the output layer provides a prediction or decision based on the input. Each neuron in a layer is connected to all neurons in the following layer and these connections each carry an associated weight. The strength of the connections, i.e., the weights, are what the network adjusts to learn from the data.



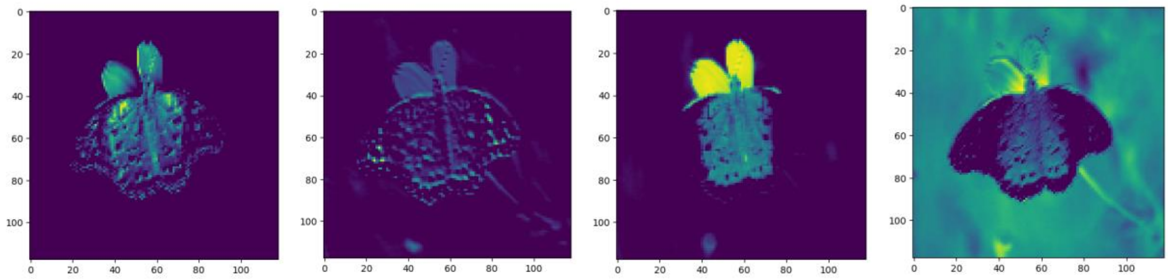
The fundamental operation in a neural network is the weighted sum of inputs to each neuron, followed by the application of an activation function. The activation function introduces non-linearity into the model, enabling it to learn more complex patterns. It is activation functions that make neural networks universal approximators – capable of learning any pattern. In a fashion, this is analogous to the function of the Chinese Room. The configuration and values of the weights would be the set of instructions, both static and input invariant. And it is the job of the driver (either the man or the network's algorithm) to take inputs and transform them into outputs.

The learning in a neural network occurs as it is trained on a dataset. In supervised learning the data comes in pairs: features, and a target. Training in a neural network refers to a process called backpropagation. In this process the loss function computes the error of the current weights by comparing the true targets with the network's predictions. The error is propagated backwards across all the weights, which are then adjusted accordingly, hence the term backpropagation. This process is repeated until the network hits some kind of minima in the loss function and it has learned the dataset.

My product is a variation of the neural network called the convolutional neural network (LeCun, Bottou, Bengio, Haffner, 1998). The difference is some new types of layers: convolutional layers, max pooling layers, and a flatten layer. I will go over each, illustrating the effects they have on a starting image of a butterfly.

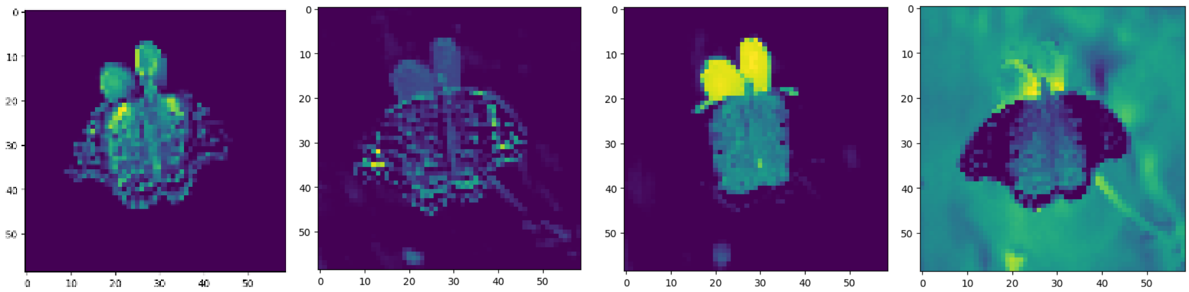
The Convolutional Layers

A convolutional layer applies numerous filters to the input data. These filters are capable of detecting patterns - such as edges, colours, and textures - in small sections of the input data. The output of this process, known as a feature map, represents the locations and strengths of these detected patterns. This facilitates the network's ability to recognize and differentiate key features, thereby contributing to its overall understanding of the input data. A feature map is a collection of images that correspond to each filter.



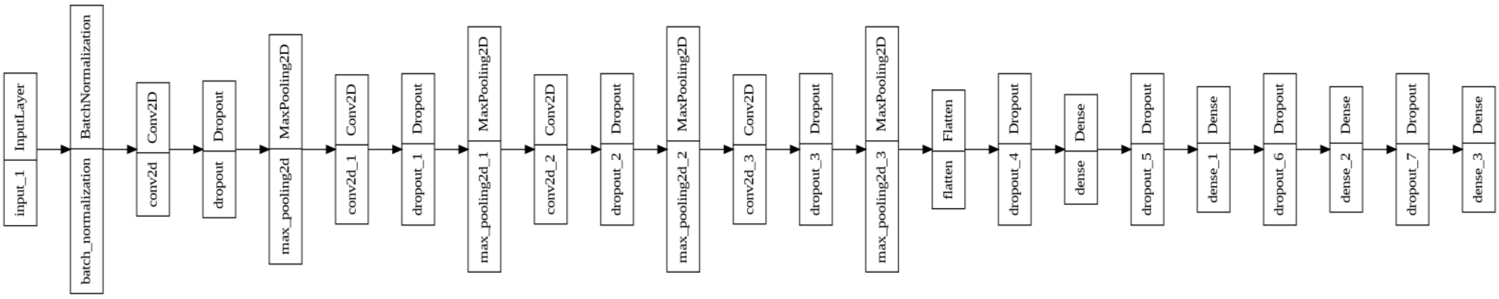
Max Pooling Layers

Max Pooling layers work by reducing the dimensionality of the data, that is the number of pixels it contains. By sliding a window around the image and averaging the pixels values in that window, it effectively downsamples the data without compromising on the inferences made by the convolutional layers.



The Flatten Layer

In both convolutional and max pooling layers the shape of the data remains constant, still incompatible with the typical neural network. The flatten layer solves this problem by condensing the data down into one dimension. Visually this can be thought of as peeling off the layers of pixels, one by one, and stacking them on top of each other vertically. This transforms the data into a format the network can use.



Put all together, the full architecture for my network, my model, looks like this.

The batch normalisation and dropout layers are something added to aid the training process. They limit a problem called overfitting. When training a model, it is best practice to keep some of your dataset to one side, as something to *validate* the model's performance. This is because neural networks, especially large ones, will memorise instead of generalise. This hurts the performance of the model on data it hasn't seen before. Overfitting is when the model learns the *noise* of the data (that is the slight anomalies) instead of focusing on the signal, what we are actually trying to teach it. In my case, I could fairly easily create a model that would achieve >99% accuracy on the training data; that would then perform poorly on validation data.

Now the basics of neural networks have been established we can combine this with the first section and return to our broader question "How do machines learn: can a neural network be called artificial intelligence?". Just as Searle did, and Turing before him, I will ground this in tangibility by rephrasing the question. Can ***my*** neural network be called artificial intelligence.

A weak AI proponent may argue that the neural network bears no intelligence: lacking agency, totally not self-aware, incapable of formulating or pursuing goals. Furthermore, it is unable to explain its decisions or to reason beyond the data it was trained on. This perspective is reminiscent of the argument posed by Searle, implying that a system cannot be considered truly intelligent simply because it follows predefined procedures to manipulate symbols and generate outputs.

However, adopting a functionalist perspective, the outlook is markedly different. Materially agnostic functionalism holds that mental states, including intelligence, are not determined by the inner workings of the mind, but rather by the roles these states play and the behaviours they realise. To a functionalist, my neural network could be described as intelligent. It, just as we do, extrapolates from patterns to make inference and even adapts its internal model based on feedback from its environment (in the form of backpropagation during training).

Taken so holistically, this bears a striking resemblance to aspects of our own, human intelligence. Generalising on data and extending that learning to new, unseen scenarios is fundamentally human. The neural network achieves this without any recognisable spark of life. If we remain process ambivalent, then the demonstration of such capabilities is sufficient evidence to consider the network artificial intelligence.

That said, it's important to remember that intelligence is not a binary characteristic, but rather a spectrum. The intelligence demonstrated by my neural network is narrow and specific; it's adept at categorizing animal species but would falter outside of this particular task. This differs from human intelligence, which is typically characterized as a general intelligence: the ability to understand, learn, adapt, and apply knowledge across a variety of tasks.

Therefore, while the neural network exhibits a form of intelligence, it's paradigmatically dissimilar the kind of intelligence a human or even a simple animal might possess. In this sense, we might consider it a form of artificial narrow intelligence (ANI) – a manifestation of weak AI.

IV Broader Implications

We stand on the precipice of an Artificial Intelligence (AI) revolution. Revolution, here, is intended as abstractly as possible. If we view history as an attempt to hedonically optimise civilisation within certain confines afforded by our environment then revolution, as I use it, refers to those moments we punch through frontiers; and there is a sudden decoupling of the space we are afforded, and our ability to fill it. In the Agricultural Revolution humans went from hunter-nomads to settled farmers. The consequential decouple of resource supply from resource demand cast back the frontiers as the grip of Darwinian economics loosened. It was the ensuing vacuum that became art, architecture, and human culture’s general flourishing. In the Industrial Revolution there was a decouple between economic output and the labour required to produce it. This led to unprecedented wealth creation, but also new forms of inequality and social tension as the disparity between those capable of capitalising upon new technologies and those who weren’t became stark.

Underlying each of these revolutions is a broader paradigm shift as two historically intertwined qualities, or quantities, diverge. In the coming AI revolution, I predict the decouple to be between the amount of intelligence residing in the universe, and the amount of that intelligence which resides in humans. It is prudent to note here that although history does tend to repeat itself, there are some key aspects of the AI revolution that will have no mirror in anteriority. I use the metaphor of the space between a civilisation and its environmental barriers to symbolize opportunity for growth. Technologies emerge, further the boundaries of the human condition and humanity occupies the gap. I make this analogy because it illuminates a significant danger of the AI revolution. Presently, the limits of the environment are largely driven by the civilisation that occupies it.