1 Introduction

Objective

There has been a Cambrian explosion in neural network architectures in recent years. Although many of the fundamental concepts underpinning these advances were developed many years ago, the size, complexity and capabilities of these new networks have advanced rapidly since approximately 2010, resulting in a field now referred to as deep learning. In this chapter we will define what we mean by deep learning, discuss some of the reasons for the rapid expansion, describe the history behind it and consider some of the applications.

What is Deep Learning?

Deep learning is a branch of machine learning involving algorithms that have many nonlinear processing stages. In most cases, deep learning refers to training neural networks that have many layers. Before approximately 2006, most neural network applications used one or two hidden layers. Since that time the size of the typical neural network has increased dramatically, creating the field of deep learning.

There is no consensus as to how many layers a neural network must have before it is considered deep. A multilayer network with one hidden layer is often referred to as a *shallow network*. A network without a hidden layer can only create linear functions, but it has been shown that shallow networks with one hidden layer can approximate any practical function or create any practical decision boundary, which is why many researchers did not originally concentrate on training deeper networks, even though the algorithms to do so have been available since the mid 1980s.

When deep networks first became popular, the numbers of layers rapidly increased, and many networks have hundreds of layers. However, as the field has matured, most recent efforts have been spent on designing more sophisticated and efficient architectures, rather than just increasing the numbers of layers.

History

The history of neural networks through the 1990s was covered in the introduction of NND2. Here we focus on the advent of deep networks starting around 2006.

Many of the fundamental concepts behind deep learning were well understood in the 1990s. Why has the field exploded in popularity more recently? People did attempt to use deeper networks in the 1990s, but they were difficult to train (as we will discuss in Chapter 3), and the results were rarely better than those achieved with shallower networks. However, in 2006 three different groups published papers ([Hinton et al., 2006], [Ranzato et al., 2006], [Bengio et al., 2006]) that introduced techniques to initialize the weights of deep networks so that the training would converge more efficiently, making deep networks finally practical. Although these initialization procedures were later found to be unnecessary, their initial introduction accelerated deep network research.

Also in 2006, Nvidia introduced the cuda language for multi-purpose programming of graphics processing units (GPUs). Prior to that time GPUs were used almost exclusively for high level 3D graphics. The introduction of cuda allowed the power of GPUs to be applied to general purpose programming. Almost immediately, several groups began to experiment with the use of cuda for neural network training. One of the first applications ([Raina et al., 2009]) demonstated the use of cuda in training deep belief networks. Their GPU implementation was up to 72 times faster than an optimized cpu implementation.

Before cuda, GPU programming was significantly more complex and limited. Programmers had to essentially "trick" the graphics pipeline into doing general-purpose computations by expressing their computations as graphics operations.

After researchers began training deep networks with GPUs, they began to win important competitions and to demonstrate that deep networks represented the state of the art in a number of application areas and could, in some cases, produce better-than-human performance. For example, in 2010 [Cireşan et al., 2010] trained a deep multilayer network to achieve the best performance (at that time) on the MNIST data set. This was followed in 2012 by [Krizhevsky et al., 2012], which won the famous Imagenet competition by a large margin.

We should note that the deep networks that won many competitions in the early 2010s were essentially just larger versions of networks that had been introduced in the 1980s – multilayer perceptrons and convolution networks.

Another factor that allowed the incredibly fast development of deep learning was the availability of larger and larger data sets. Large data sets, like Imagenet, were beginning to become available in the mid 2000s. Because deep networks have many more trainable parameters than shallow networks, more data is needed to train them adequately and to prevent overfitting.

In addition to the data sets, a number of deep learning software frameworks have been developed since 2010. Researchers no longer have to write their own cuda code to implement deep network training procedures. Deep learning frameworks, like TensorFlow (developed by Google) and PyTorch (developed by Meta) are free and open-source. They have made it possible for researchers with even a modest programming background to quickly start training and deploying deep networks.

A final factor that has fueled the growth of deep learning is the shear number of researchers working in the field. When ChatGPT was launched in November of 2022, the platform received approximately 152 million visitors in its first month. There has never before

been this level of interest in neural networks. Companies and national laboratories around the world are scrambling to find the next application area or the next breakthrough network architecture or training algorithm.

So, the increased computing power of GPUs, the extremely large data sets that have become available, the powerful deep learning frameworks, and the hundreds of thousands of practitioners who have come into the field, have supercharged deep learning progress. It is currently in a virtuous cycle – more success generates more interest, development of faster hardware and efficient software, which leads to more progress.

Applications

Many neural network applications were described in the introduction of NND2. During the 1990s and 2000s neural networks produced very successful solutions to a number of important real-world problems. Since that time, the number of applications has increased, but the principal change is that deep network performance for the most complex problems has improved tremendously. Especially in the areas of computer vision and natural language processing, deep networks are exponentially better than the neural networks (and other machine learning algorithms) that were commonly used before 2006. In many areas now, deep neural networks equal (or surpass) human level performance.

It is impossible to list all of the applications where deep neural networks are currently being used. Instead, let's just briefly review four areas in which deep learning has provided revolutionary breakthroughs. These are developments that even experts in machine learning would not have predicted in 2005.

AlphaFold

AlphaFold is a system developed by DeepMind, a subsidiary of Alphabet (Google's parent company), designed to use deep learning to predict the three-dimensional structures of proteins based on their amino acid sequences. It is one of the most significant advancements in the field of computational biology, and its designers won the Nobel prize in chemistry in 2024.

In 2018, AlphaFold debuted at the CASP13 (Critical Assessment of Techniques for Protein Structure Prediction) competition. It used deep learning techniques to predict protein structures with an accuracy that surpassed all previous state-of-the-art methods. In 2020, AlphaFold 2 [Jumper et al., 2021] was able to predict the 3D structure of a protein with a precision that rivaled that of experimental methods like X-ray crystallography and cryo-electron microscopy, but in a fraction of the time. AlphaFold has effectively solved the protein folding problem, which has been a goal of biochemistry for decades.

With the structure of proteins now more accessible, researchers are able to study disease mechanisms at a molecular level. AlphaFold has been particularly impactful in areas like:

- Understanding diseases: By predicting the structures of proteins linked to diseases like Alzheimer's, cancer, and COVID-19, AlphaFold has provided a valuable resource for understanding how these proteins might behave in the body.
- Drug development: The ability to accurately model protein structures could accelerate the development of drugs that target specific proteins, offering new avenues for treatment.

Large Language Models

Large Language Models (LLMs), enabled by the transformer [Vaswani et al., 2017], have revolutionized natural language processing and enabled human-level performance on many language tasks. This has led to tools like ChatGPT that have transformed how people interact with computers. LLMs have impacted fields from coding to education to customer service. Computer coding has been taught differently since ChatGPT, because LLMs can write code from a natural language description of the desired operation. Human iteraction in the workplace has been fundamentally changed – job applicants are using LLMs to write their resumes and cover letters, and companies are using LLMs to screen the applications. LLMs have passed the Uniform Bar Exam and the U.S. Medical Licensing Examination.

GANs and Diffusion for Image Generation

Generative Adversarial Networks (GANs) were first introduced in 2014 [Goodfellow et al., 2014]. Diffusion models were introduced in 2015 [Sohl-Dickstein et al., 2015] and were made practical for standard use in 2022 [Rombach et al., 2022]. Together, they enabled neural network generation of artificial images – what we sometimes refer to as *deepfakes*. Photo-realistic images can be created from a simple text description. This has most obviously disrupted the graphic artist profession, but there are many other applications. It can be used for molecular structure generation and materials design, style transfer, super-resolution, combined text, image, and audio generation. Short films have been generated using this technology. Actors and screenwriters have both recently gone on strike, in part, to prevent the use of this technology.

Speech Generation and Recognition

The evolution of speech technology since 2015 represents a dramatic shift from traditional methods to deep learning, and this has been revolutionary. The first neural network to generate realistic human-like speech as raw audio was WaveNet, which was developed by Google DeepMind in 2016 [Van Den Oord et al., 2016]. There have been significant yearly advances on that technology, with each step producing faster and higher fidelity results. Recently, Microsoft introduced VALL-E 2, which can synthesize high-quality personalized speech using only a three-second recording of an unseen speaker as a prompt. It is so accurate that Microsoft has not released it to the public (as this is being written), citing potential risks of misuse, such as voice spoofing and impersonation.

In 2016 Deep Speech [Amodei et al., 2016] represented a major breakthrough in speech recognition, as it was an end-to-end deep learning model that could recognize either English or Mandarin Chinese, and it was competitive with the transcription of human workers. This was further improved in 2020 with Wav2Vec 2.0 [Baevski et al., 2020], whose innovations included self-supervised learning, contrastive pretraining, and transformer-based architectures. It was a paradigm shift in speech recognition, reducing reliance on labeled data while achieving state-of-the-art performance. It broke the record for accuracy on a key speech recognition benchmark

with 100x less labeled data. The Whisper speech recognition system, introduced by OpenAI in 2023 [Radford et al., 2023], moved the state-of-the-art even further. It works on multiple languages – even on languages on which it has not been trained. It can perform speech-to-text translation, converting spoken words from one language into text in another.

Further Reading

[Jumper et al., 2021] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021

This paper introduced AlphaFold, which can accurately predict the three-dimensional structures of proteins based on their amino acid sequences. This effectively solved the protein folding problem, which had been a goal of biochemistry for decades. This work won the Nobel Prize for chemistry in 2024.

[Vaswani et al., 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017

This paper introduced the transformer architecture, on which all current large language models (LLMs) are based.

[Goodfellow et al., 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014

This paper introduced the generative adversarial network (GAN). This was the first major breakthrough in training deep networks to generate images. It works through two competing neural networks: a generator creating images and a discriminator trying to spot fakes. It led to major breakthroughs like StyleGAN (2019) which could create highly realistic human faces.

[Rombach et al., 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022

This paper introduced methods that made diffusion stable and widely accessible. It is currently the state of the art for image generation. Diffusion works by adding noise to an image in a series of stages. After enough stages, the image becomes random noise. A deep network then learns to reverse the process, removing noise from a random image in stages.

[Van Den Oord et al., 2016] Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, et al. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 12, 2016

The deep neural network introduced in this paper, WaveNet, was the first to produce raw audio.

[Radford et al., 2023] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023

This paper describes the Whisper speech recognition system of OpenAI. It works for many languages, even languages in which it was never trained.