**Model Compression** is an actively pursued area of research over the last few years with the goal of deploying state-of-the-art deep networks in low-power and resource limited devices without significant drop in accuracy.

**Neural network pruning** is a method of compression that involves removing weights from a trained model. In agriculture, **pruning** is cutting off unnecessary branches or stems of a plant. In machine learning, **pruning** is removing unnecessary neurons or weights

Reference: https://analyticsindiamag.com/what-is-neural-network-pruning-and-why-is-it-important-today/

Industry-leading [neural networks](https://www.analyticsindiamag.com/will-one-shot-learning-using-hypercubes-outrank-traditional-neural-nets/) have almost unlimited compute and space at their disposal, as they are configured to run in the most powerful manner. However, developers that are creating deep learning applications for use on mobile devices don’t have this luxury.

This has led to a rise in the requirements for smaller neural networks for on-the-go applications. With the rising use of AR, facial recognition and voice assistants, it has become more accessible to use ML features on mobile devices.

These developers are looking for newer and more effective ways of reducing the size and amount of compute required for the application of neural networks. One of the most popular methods is called pruning.

What is Neural Network Pruning

Simply put, pruning is a way to reduce the size of the neural network through compression. After the network is pre-trained, it is then fine-tuned to determine the importance of connections. This is done through the ranking of the neurons from the network, with the first example being described in Yann Lecun 1990 paper ‘[Optimal Brain Damage](http://yann.lecun.com/exdb/publis/pdf/lecun-90b.pdf)‘.

The basic principles of pruning include removing unimportant weighted information using second derivative data. This results in better generalisation results, improved speed of processing the results and a reduced size as well.

Pruning is usually done in an iterative fashion, to avoid the pruning of necessary neurons. This also ensures that an important part of the network is not lost, as neural networks are a black box. The first step is to determine which neurons are important and which aren’t.

After this, the last important neuron is removed, followed by the fine-tuning of the algorithm. At this point, a decision can be made to continue the pruning process or to stop pruning,

With the rise of mobile inference and machine learning capabilities, pruning becomes more relevant than ever before. Lightweight algorithms are the need of the hour, as more and more applications find use with neural networks.

The most recent example of this comes in the form of Apple’s new products, which use neural networking to ensure a multitude of privacy and security features across products.

The easy availability of neural networks is also required due to the varied nature of their applications. Their move to mobile is also complemented by the standalone computer in flagship devices, further creating a need for an efficient program that performs the most amount of work while consuming the least amount of resources.

This is the reason pruning is more relevant today, as the applications need to get lighter and faster without sacrificing accuracy.

Dataset: <http://image-net.org/challenges/LSVRC/2012/index#introduction>

https://www.cs.toronto.edu/~kriz/cifar-100-python.tar.gz

Definition:

In computing, floating point operations per second (**FLOPS**, **flops** or **flop**/s) is a measure of computer performance, useful in fields of scientific computations that require floating-point calculations. For such cases it is a more accurate measure than measuring instructions per second.