Deep Learning Optimization

Calculating the Inference Time

Vocabulary

- **FLOPs: Floating Point Operations**
  - The number of computations the model will have to perform
  - Tells us how complicated our model is
- **FLOPS: Floating Point Operations per Second**
  - The number of operations per second our model can do
  - Tells us how good our hardware is
- **MACs: Multiply–Accumulate Computations**
  - An operation that includes an addition and a multiplication
  - 1 MAC = 2 FLOPs

Calculating the FLOPs

- **Convolutions**: MACs =. Number of Kernel x Kernel Shape x Output Shape
- **Fully Connected**: MACs = Input Size x Output Size
- **Pooling Layers**: FLOPs = Height x Width x Depth
  - With Strides: FLOPs = (Height/Stride)x(Widht/Stride)x Depth

Calculating the Inference Time

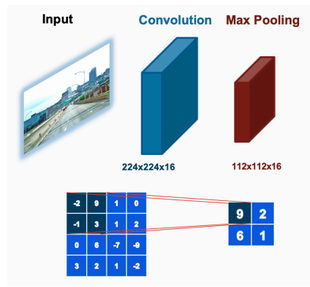
TIme = FLOPs / FLOPS

Optimizing a Deep Learning Model

Reducing the Number of Operations

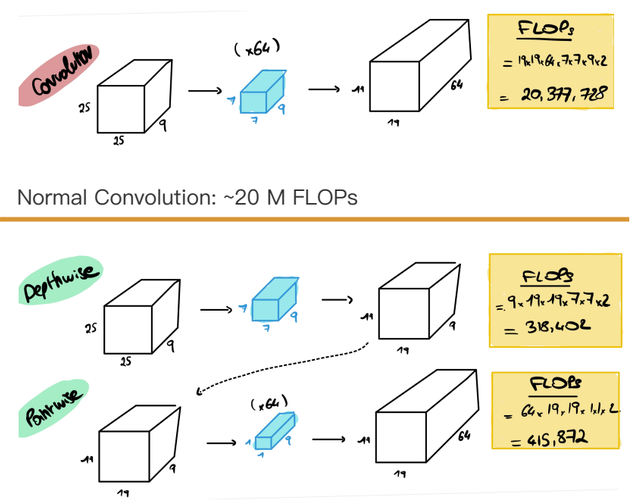
Pooling

Reduce the amount of parameters being passed from one layer to another



Separable Convolutions

A depthwise followed by a pointwise convolution

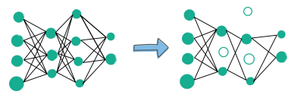


Normal Convolution: ~20 M FLOPs

Separable Convolution: ~734K FLOPs

Model Pruning

Remove the weights that are redundant or that don't matter by setting them to 0



Reducing the Size of a Model

Compression

- **Quantization**
  - Map values from a large set to a smaller one
  - Example: 2.9089002 —> 2.9
- **Weight Sharing**
  - Share weights between layers to reduce the amount of weights we store
  - Using the K–Means algorithm for example

Knowledge Distillation

Transfer the knowledge learned by a large and accurate model (teacher) to a smaller and lighter one (student)