# Optimizations

## Input Size

Varying the input size of the image to the neural network has a direct impact on inference speed and performance. Increasing the input size enables the model to capture more fine-grained details, leading to improved accuracy. It allows the network to perceive objects at different scales and enhance spatial information. However, larger input sizes result in longer inference time. Although the number of network parameters doesn’t change the amount of required operation increases as filters used for convolution and pooling have to process more pixels. Intuitively, decreasing the input size has the opposite effect.

## Quantization

## Quantization is a technique that reduces the precision of weights, biases, and activations in a neural network. Instead of using 32-bit floating-point numbers, quantization represents them with lower bit precision, such as 8-bit integers. This technique offers two main advantages at the cost of reduced accuracy:

* **Reduced Memory Footprint**: Quantization significantly reduces the memory required to store model parameters. For instance, a model using 32-bit floating-point parameters uses around 4x more memory compared to its quantized 8-bit integer version.
* **Improved Inference Speed**: Quantized models perform computations with lower precision, which can be executed faster on hardware. This acceleration can lead to improved inference speed, enabling real-time applications.