Early Stopping is **a regularization technique for deep neural networks that stops training when parameter updates no longer begin to yield improves on a validation set**.

A **validation set** is **a set of data used to train artificial intelligence (AI) with the goal of finding and optimizing the best model to solve a given problem**.

By **preprocessing data**, we **make it easier to interpret and use**. This process eliminates inconsistencies or duplicates in data, which can otherwise negatively affect a model's accuracy. Data preprocessing also ensures that there aren't any incorrect or missing values due to human error or bugs.

transfer learning is **a machine learning method where we reuse a pre-trained model as the starting point for a model on a new task**. To put it simply—a model trained on one task is repurposed on a second, related task as an optimization that allows rapid progress when modeling the second task.

VGG16 is **object detection and classification algorithm which is able to classify 1000 images of 1000 different categories with 92.7% accuracy**. It is one of the popular algorithms for image classification and is easy to use with transfer learning.

The VGG model stands for the Visual Geometry Group

What is a **Neural Network Activation Function**? An Activation Function **decides whether a neuron should be activated or not**. This means that it will decide whether the neuron's input to the network is important or not in the process of prediction using simpler mathematical operations.

the usage of ReLU **helps to prevent the exponential growth in the computation**

we use **ReLU in hidden layer to avoid vanishing gradient problem and better computation performance , and Softmax function use in last output layer** .

ReLU activation function, is perhaps the most common function used for hidden layers. It is common because **it is both simple to implement and effective at overcoming the limitations of other previously popular activation functions**, such as Sigmoid and Tanh.

The **softmax activation function** **transforms the raw outputs of the neural network into a vector of probabilities**.

This function **takes any real value as input and outputs values in the range of 0 to 1**. The larger the input (more positive), the closer the output value will be to 1.0, whereas the smaller the input (more negative), the closer the output will be to 0.0,

Why Relu over sigmoid :

However, when *n* hidden layers use an activation like the sigmoid function, *n* small derivatives are multiplied together. Thus, the gradient decreases exponentially as we propagate down to the initial layers.

A small gradient means that the weights and biases of the initial layers will not be updated effectively with each training session. Since these initial layers are often crucial to recognizing the core elements of the input data, it can lead to overall inaccuracy of the whole network.

Solutions:

The simplest solution is to use other activation functions, such as ReLU.

A CNN can be instantiated as a **Sequential model** because **each layer has exactly one input and output and is stacked together to form the entire network**.

What is the difference between sequential model and functional model?

For example, **in sequential model you can only stack one layer after another, while in functional model you can connect a layer to literally any other layer**.

Cross-entropy is commonly used in **machine learning** as a loss function.
What is cross entropy loss in ML?

Cross entropy loss is a metric used to measure how well a classification model in machine learning performs. The loss (or error) is measured as a number between 0 and 1, with 0 being a perfect model. The goal is generally to get your model as close to 0 as possible.

The **loss function** is **a method of evaluating how well your machine learning algorithm models your featured data set**.

Optimizers are **algorithms or methods used to change the attributes of your neural network such as weights and learning rate in order to reduce the losses**. Optimizers help to get results faster.

The name is derived from adaptive moment estimation. The optimizer is called **Adam** because **uses estimations of the first and second moments of the gradient to adapt the learning rate for each weight of the neural network**.

Stochastic gradient descent (**SGD**)
Stochastic gradient descent (often abbreviated SGD) is an iterative method for **optimizing an objective function with suitable smoothness properties** (e.g. differentiable or subdifferentiable).
Memory requirement is less compared to the GD algorithm as the derivative is computed taking only 1 point at once.