**Common terms**

* Batch size (bs): The batch size is a hyperparameter that defines the number of samples to work through before updating the internal model parameters.
* Weight decay (wd): causes the weights to exponentially decay to zero. Prevents overgrowing of weights
* Learning rate (lr): a hyper-parameter that controls how much we are adjusting the weights of our network with respect to the loss gradient.
* Skip Connections: Skip connections are extra connections between nodes in different layers of a neural network that skip one or more layers of nonlinear processing
* ReLU (Rectifier Linear Unit): The Rectified Linear Unit is an activation function in deep learning models. The function returns 0 if it receives any negative input, but for any positive value x it returns that value back. So it can be written as f(x) = max(0, x).
* Epoch: one cycle through the full training dataset
* Loss function: helps in optimizing the parameters of the neural networks.
* Generalization: This means that a trained net could classify data from the same class as the learning data that it has never seen before

**Optimization algorithms**

* SGD (Stochastic Gradient Descent)
* Adam (A Method for Stochastic Optimization)

The name Adam is derived from adaptive moment estimation.

Adam vs SGD

Stochastic gradient descent maintains a single learning rate (termed alpha) for all weight updates and the learning rate does not change during training.

In Adam a learning rate is maintained for each network weight (parameter) and separately adapted as learning unfolds.

**Dataset**

CIFAR-10 (Canadian Institute for Advanced Research) Dataset

Contains images

<https://www.cs.toronto.edu/~kriz/cifar.html>

**Neural Networks used**

* VGG (Visual Geometry Group) (9 Layer)
* ResNet 56
* ResNet no-short 56
* DenseNet

**Types of normalization**

* **Layer Normalization:** The direction *d* is normalized in the layer level so that the direction for each layer has the same norm as the corresponding layer of *θ*
* **Filter Normalization:** The direction *d* is normalized so that the direction for each filter has the same norm as the corresponding filter in *θ*