- Internal covariate shift : Distribution of each layer's inputs changes during training, as the parameters of the previous layers change. This slows down the training by requiring lower learning rates and careful parameter initialization, and makes it hard to train models with saturating nonlinearities.
- Batch Normalization allows us to use much higher learning rates and be less careful about initialization.
- The gradient of the loss over a mini-batch is an estimate of the gradient over the training set, whose quality improves as the batch size increases.
- Due to batch norm output of an example x_i depends on all examples in the minibatch, so some noise is added to the output which acts as a regulariser.
- Batch Normalization also has a beneficial effect on the gradient flow through the network, by reducing the dependence of gradients on the scale of the parameters or of their initial values.This allows us to use much higher learning rates without the risk of divergence.
- It has been long known (LeCun et al., 1998b; Wiesler & Ney, 2011) that the network training converges faster if its inputs are whitened – i.e., linearly transformed to have zero means and unit variances, and decorrelated.
- Beta and Gamma are used to improve flexibility to revert back to original distribution or some other distribution.
- By normalizing activations throughout the network, it prevents small changes to the parameters from amplifying into larger and suboptimal changes in activations in gradients.
- Large learning rates may increase the scale of layer parameters, which then amplify the gradient during backpropagation and lead to the model explosion.But batch norm makes it possible for training to be independent of parameter scale.
- Why batchnorm only :
  - If any normalisation is applied inside gradient descent, then gradient would have to be partially given to normalisation step which reduces its effect.
  - If normalisation is outside gradient descent step => Gradient descent will not differentiate the normalisation term => Loss will be independent of bias/scale due to normalisation => bias /scale can explode which is unstable.
  - But if whitening is used in gradient descent step => expensive as $\sum$ and its derivatives have to be calculated.
  - So we find an alternative differentiable normalisation of batchnorm in which we only change variance of all features to 1 but no decorrelation.
- Each mini-batch produces estimates of the mean and variance of each activation. This way, the statistics used for normalization can fully participate in the gradient backpropagation.
- Wu + b is more likely to have a symmetric, non-sparse distribution, that is "more Gaussian" (Hyv¨arinen & Oja, 2000); normalizing it is likely to produce activations with a stable distribution.
- Batch Normalization may lead the layer Jacobians to have singular values close to 1, which is known to be beneficial for training (Saxe et al., 2013).