BatchNorm Excerpt

Batch normalization is a regularization technique. **It works by taking an average of the mean and standard deviations of the activations of a layer** and using those to normalize the activations.

However, this can cause problems because the network might want some activations to be really high in order to make accurate predictions. To fix this, there are two learnable parameters (they will be updated in the SGD step), usually called gamma and beta. After normalizing the activations to get some new activation vector y, a batchnorm layer returns gamma\*y + beta.

The activations can have a mean or variance which is independent from the mean and standard deviation of the results of the previous layer.

Those statistics are learned separately, making training easier on a model.

The behavior is different during training and validation –

During training the mean and standard deviation of the batch is used to normalize the data, whereas during validation it uses a running mean of the statistics calculated during training.