

Batch Normalization, <https://arxiv.org/abs/1502.03167>

Target

Reduce the *internal covariate shift*.

Covariate shift: The input distribution to a learning system changes.

Internal covariate shift: In the course of training, the change in the distributions of internal nodes of a deep network.

Training Time

Let $\{f_1, f_2, f_3, \dots, f_n\}$ be the batched feature vector of some layer (a.k.a. that layer's output, either pre-activation or post-activation). Let's say each $f_i \in \mathbb{R}^d$. The Batch Normalization first computes

$$\begin{aligned}\mu &= \frac{1}{n} \sum_{i=1}^n f_i \\ \sigma &= \sqrt{\frac{\sum_{i=1}^n (f_i - \mu)^2}{n}} \\ \tilde{f}_i &= \frac{f_i - \mu}{\sigma}, \mu, \sigma \in \mathbb{R}^d\end{aligned}$$

second, computes the affine transform

$$\hat{f}_i = \gamma \cdot \tilde{f}_i + \beta$$

where all arithmetic are element-wise, and $\gamma, \beta \in \mathbb{R}$ are learnable.

We'll use each \hat{f}_i as next layer's input. Each new batch features $\{\hat{f}_i\}_{i=1}^n$ has mean $\mathbf{0} \in \mathbb{R}$ and standard deviation $\mathbf{1} \in \mathbb{R}$.

Note that the batch normalization is computed independently for each channel in CNN.

Testing Time

During the testing time, mostly we don't have a batched data. Hence no on-line μ, σ to compute. Instead, we use the training data to help us get them.

For each i -th batch, we have μ_i, σ_i . We can then use the Exponential Moving Average to compute the μ^-, σ^- for testing

$$\begin{aligned}\mu^- &= \alpha \mu^- + (1 - \alpha) \mu_i \\ \sigma^- &= \alpha \sigma^- + (1 - \alpha) \sigma_i\end{aligned}$$

where $\alpha \in [0, 1]$ is a pre-defined hyper parameter and μ^-, σ^- are both initialized to $\mathbf{0}$. In pytorch, α is called *momentum* and the default value is 0.1.

We use

$$\hat{f} = \frac{f - \mu^-}{\sigma^-} \cdot \gamma + \beta$$

as the testing time batch normalization.

Layer Normalization, <https://arxiv.org/abs/1607.06450>

Instance Normalization, <https://arxiv.org/abs/1607.08022>

Group Normalization, <https://arxiv.org/abs/1803.08494>

Weight Normalization, <https://arxiv.org/abs/1602.07868>

Weight Normalization, <https://arxiv.org/abs/1705.10941>