Summary: Batch Normalization

Abdul Samad Khan

May 2025

1 Overview

Ioffe and Szegedy (2015) propose batch normalization (BN) to accelerate deep network training by reducing internal covariate shift. BN normalizes layer inputs over mini-batches, adding learnable scale and shift parameters.

2 Mathematical Formulation

For a mini-batch $\mathcal{B} = \{x_1, \ldots, x_m\}$ at layer l, compute:

$$\mu_{\mathcal{B}} = \frac{1}{m}\sum_{i=1}^{m} x_i, \quad \sigma_{\mathcal{B}}^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2.$$

Normalize:

$$\hat{x}_i = \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}},$$

where ϵ prevents division by zero. Apply learnable parameters γ, β:

$$y_i = \gamma \hat{x}_i + \beta.$$

At inference, use running averages of μ and σ^2.

3 Method

BN is applied before non-linearities (e.g., ReLU). It reduces sensitivity to initialization and enables higher learning rates, improving training stability.

4 Critique

BN excels in large-batch settings but struggles with small batches, common in finance ML due to limited high-quality data. Alternatives like layer normalization may suit sequential models (e.g., transformers).