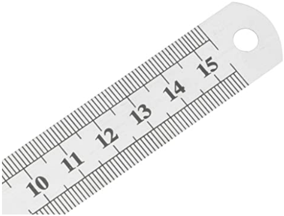

NORMALIZATION ISSUES

Vincent Barra
LIMOS, UMR 6158 CNRS, Université Clermont Auvergne

LABORATOIRE D'INFORMATIQUE,
DE MODÉLISATION ET D'OPTIMISATION DES SYSTÈMES

UNIVERSITÉ Clermont Auvergne

Why ?

Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In ICML 2015.

Why normalizing?

- Maintaining proper statistics of the activations and derivatives was a critical issue to allow the training of deep architectures.
- Batch normalization: the activation statistics during the forward pass by re-normalization.

Batch normalization can be done anywhere in a deep architecture.

During the training : shifts and rescales activations according to the mean and variance estimated on the current batch.

When testing: shifts and rescales according to the empirical moments estimated during training.

LIMOS LABORATOIRE D'INFORMATIQUE, DE MODÉLISATION ET D'OPTIMISATION DES SYSTÈMES
cnrs
UCA UNIVERSITÉ Clermont Auvergne

How ?

Batch $\mathcal{B} = \{\boldsymbol{x}_i \in \mathbb{R}^d \, i \in [\![1 \cdots B]\!]\}$

$$\textbf{Mean: } \hat{m}_{\mathcal{B}} = \frac{1}{B}\sum_{i=1}^{B} \boldsymbol{x_i} \quad \textbf{Variance: } \sigma_{\mathcal{B}}^2 = \frac{1}{B}\sum_{i=1}^{B} \|\boldsymbol{x_i} - \hat{m}_{\mathcal{B}}\|^2$$

From this

$$\boldsymbol{z}_i = \frac{\boldsymbol{x}_i - \hat{m}_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \quad \textbf{and} \quad y_{\mathcal{B}} = \gamma \odot z_{\mathcal{B}} + \beta \quad \in [\![1 \cdots B]\!]$$

$\odot$: Hadamard component-wise product
$\gamma, \beta \in \mathbb{R}^d$ parameters to optimize

How ?

During inference: shift and rescale each component of $\boldsymbol{x}$ according to statistics estimated during training.

$$y = \gamma \odot \frac{\boldsymbol{x} - \hat{m}}{\sqrt{\sigma^2 + \epsilon}} + \beta$$

$\Rightarrow$ component-wise affine transformation.

$\gamma, \beta \in \mathbb{R}^d$ need to be learned $\Rightarrow \frac{\partial \ell}{\partial \gamma}$ and $\frac{\partial \ell}{\partial \beta}$?

In the following $d = 1$ (nevermind since components are processed independently).

LABORATOIRE D'INFORMATIQUE,
DE MODÉLISATION ET D'OPTIMISATION DES SYSTÈMES

How ?

$$\frac{\partial \ell}{\partial \gamma} = \sum_{b \in \mathcal{B}} \frac{\partial \ell}{\partial y_b} \frac{\partial y_b}{\partial \gamma} = \sum_{b \in \mathcal{B}} \frac{\partial \ell}{\partial y_b} z_b$$

$$\frac{\partial \ell}{\partial \beta} = \sum_{b \in \mathcal{B}} \frac{\partial \ell}{\partial y_b} \frac{\partial y_b}{\partial \beta} = \sum_{b \in \mathcal{B}} \frac{\partial \ell}{\partial y_b}$$

Each input in the batch impacts all the outputs in the batch $\Rightarrow \frac{\partial \ell}{\partial x_b}$ not so easy.

$$\frac{\partial \ell}{\partial x_b} = \frac{\partial \ell}{\partial z_b} \frac{1}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} + \frac{2}{B} \frac{\partial \ell}{\partial \sigma_{\mathcal{B}}^2} (\boldsymbol{x}_b - \hat{m}_{\mathcal{B}}) + \frac{1}{B} \frac{\partial \ell}{\partial \hat{m}_{\mathcal{B}}}$$

Usually, $\hat{m}_{\mathcal{B}}$ and $\sigma_{\mathcal{B}}^2$ for test are estimated with a moving average during training.

LABORATOIRE D'INFORMATIQUE, DE MODÉLISATION ET D'OPTIMISATION DES SYSTÈMES

ANOTHER WAY TO NORMALIZE

L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. CoRR, abs/1607.06450, 2016.

Normalizing activations accross a layer instead of across the batch. For $\boldsymbol{x} \in \mathbb{R}^d$,

$(\bar{x}, \sigma)$ = (mean, standard deviation of the components of $\boldsymbol{x}$)

$$\text{for } i \in [\![1 \cdots d]\!] \;\; y_i = \frac{x_i - \bar{x}}{\sigma}$$

LABORATOIRE D'INFORMATIQUE, DE MODÉLISATION ET D'OPTIMISATION DES SYSTÈMES

UNIVERSITÉ Clermont Auvergne