AI and Deep Learning

3-3 Batch Normalization

Zhonglei Wang

WISE, SOE, CHOW

XMU

(Version v1)

Contents

1. Introduction

2. Forward propagation

3. Backpropagation

4. Disucssion for dropout and batch normalization

Introduction

1. Before 2015, there existed optimization issues for training deep neural networks
   - Instability for backpropagation
   - Severe sensitivity to parameter initialization
2. For example, there is an issued, called internal covariate shift
   - Minor modification of parameters in shallower layers may change the distribution of inputs for deep layers
   - Parameters of deeper layers focus on “a moving target”
3. Such instability may lead to the following problems
   - Gradient vanishing
   - Gradient exploding

Introduction

1. Whitening is populator in classical statistical learning
   - Use linear transformation to achieve mean 0 and identity covariance matrix
   - Algorithm may converge fast
2. Such whitening is impossible for deep learning
   - Size of the training dataset is usually large
   - We cannot guarantee whitening for each layer
   - Computation cost for whitening is high

Introduction

1. Ioffe and Szegedy (2015) proposed batch normalization (BN)
   - Implemented independently for each feature (associated with a certain neuron): Does not consider covariance among features
   - Mean and standard error is obtained by the current mini-batch
2. Normalize values after linear transformation but before activation for EACH neuron
3. BN has the following properties
   - Tries to solves the interval covariate shift problem
   - Introduce two more parameters to allow for heterogeneity

Forward propagation

1. For a mini-batch with m training examples, the forward propagation for the lth layer is

$$\boldsymbol{Z}^{[l]} = \boldsymbol{A}^{[l-1]}(\boldsymbol{W}^{[l]})^{\mathrm{T}} + (\boldsymbol{b}^{[l]})^{\mathrm{T}} \in \mathbb{R}^{m \times d^{[l]}}, \quad \boldsymbol{A}^{[l]} = \sigma^{[l]}(\boldsymbol{Z}^{[l]}) \in \mathbb{R}^{m \times d^{[l]}}$$

2. Denote $\boldsymbol{Z}^{[l]} = (\boldsymbol{z}_1^{[l]}, \ldots, \boldsymbol{z}_m^{[l]})^{\mathrm{T}}$

3. After linear transformation, we consider the following new calculations:

$$\begin{aligned}
\boldsymbol{\mu}^{[l]} &= m^{-1}(\boldsymbol{Z}^{[l]})^{\mathrm{T}}\mathbf{1} \in \mathbb{R}^{d^{[l]} \times 1}, \\
\boldsymbol{\sigma}^{2[l]} &= m^{-1}\sum_{i=1}^{m}\left(\boldsymbol{z}_i^{[l]} - \boldsymbol{\mu}^{[l]}\right) \circ \left(\boldsymbol{z}_i^{[l]} - \boldsymbol{\mu}^{[l]}\right) \in \mathbb{R}^{d^{[l]} \times 1}, \\
\tilde{\boldsymbol{z}}_{i,norm}^{[l]} &= \frac{\boldsymbol{z}_i^{[l]} - \boldsymbol{\mu}^{[l]}}{\sqrt{\boldsymbol{\sigma}^{2[l]} + \epsilon}} \in \mathbb{R}^{d^{[l]} \times 1}, \\
\tilde{\boldsymbol{z}}_i^{[l]} &= \boldsymbol{\gamma}^{[l]} \circ \tilde{\boldsymbol{z}}_{i,norm}^{[l]} + \boldsymbol{\beta}^{[l]} \in \mathbb{R}^{d^{[l]} \times 1}, \\
\tilde{\boldsymbol{Z}}^{[l]} &= (\tilde{\boldsymbol{z}}_1^{[l]}, \ldots, \tilde{\boldsymbol{z}}_m^{[l]})^{\mathrm{T}} \in \mathbb{R}^{m \times d^{[l]}}
\end{aligned}$$

Forward propagation

1. Vectorization for the red calculations, but leave the blue parts alone

$$\boldsymbol{Z}^{[l]} = \boldsymbol{A}^{[l-1]}(\boldsymbol{W}^{[l]})^{\mathrm{T}} \in \mathbb{R}^{m\times d^{[l]}},$$

$$\boldsymbol{\mu}^{[l]} = m^{-1}(\boldsymbol{Z}^{[l]})^{\mathrm{T}}\mathbf{1} \in \mathbb{R}^{d^{[l]}\times 1},$$

$$\check{\boldsymbol{Z}}^{[l]} = \boldsymbol{Z}^{[l]} - \mathbf{1}(\boldsymbol{\mu}^{[l]})^{\mathrm{T}} \in \mathbb{R}^{m\times d^{[l]}},$$

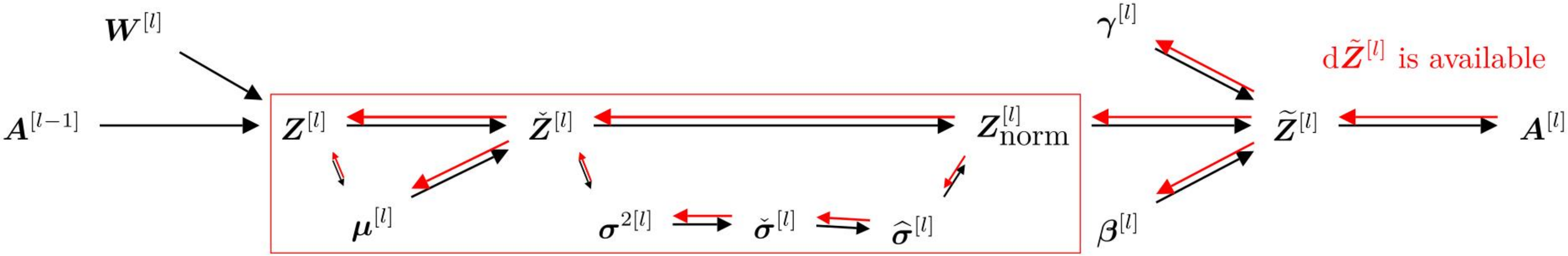
$$\boldsymbol{\sigma}^{2[l]} = m^{-1}\sum_{i=1}^{m} \check{\boldsymbol{z}}_i^{[l]} \circ \check{\boldsymbol{z}}_i^{[l]} \in \mathbb{R}^{d^{[l]}\times 1},$$

$$\check{\boldsymbol{\sigma}}^{[l]} = \sqrt{\boldsymbol{\sigma}^{2[l]} + \epsilon} \in \mathbb{R}^{d^{[l]}\times 1},$$

$$\widehat{\boldsymbol{\sigma}}^{[l]} = (\check{\boldsymbol{\sigma}}^{[l]})^{-1} \in \mathbb{R}^{d^{[l]}\times 1},$$

$$\boldsymbol{Z}_{\mathrm{norm}}^{[l]} = \check{\boldsymbol{Z}}^{[l]} \circ \{\mathbf{1}(\widehat{\boldsymbol{\sigma}}^{[l]})^{\mathrm{T}}\} \in \mathbb{R}^{m\times d^{[l]}}$$

2. Notice that the previous bias term $\boldsymbol{b}^{[l]}$ is useless for batch normalization



Backpropagation for the red parts:

$$\mathrm{d}\boldsymbol{Z}^{[l]} = \mathrm{d}\boldsymbol{Z}_1^{[l]} + \mathrm{d}\boldsymbol{Z}_2^{[l]} \qquad \mathrm{d}\boldsymbol{\beta}^{[l]} = \left(\mathrm{d}\tilde{\boldsymbol{Z}}^{[l]}\right)^{\mathrm{T}} \mathbf{1} \qquad \mathrm{d}\boldsymbol{\gamma}^{[l]} = \left(\mathrm{d}\tilde{\boldsymbol{Z}}^{[l]} \circ \boldsymbol{Z}^{[l]}_{\text{norm}}\right)^{\mathrm{T}} \mathbf{1}$$

$$\mathrm{d}\check{\boldsymbol{Z}}^{[l]} = \mathrm{d}\check{\boldsymbol{Z}}_1^{[l]} + \mathrm{d}\check{\boldsymbol{Z}}_2^{[l]} \qquad \mathrm{d}\check{\boldsymbol{\sigma}}^{[l]} = -\mathrm{d}\widehat{\boldsymbol{\sigma}}^{[l]} \circ \widehat{\boldsymbol{\sigma}}^{[l]} \circ \widehat{\boldsymbol{\sigma}}^{[l]} \qquad \mathrm{d}\boldsymbol{Z}^{[l]}_{\text{norm}} = \mathrm{d}\tilde{\boldsymbol{Z}}^{[l]} \circ \left\{\mathbf{1}\left(\boldsymbol{\gamma}^{[l]}\right)^{\mathrm{T}}\right\}$$

$$\mathrm{d}\boldsymbol{Z}_1^{[l]} = \mathrm{d}\check{\boldsymbol{Z}}^{[l]} \qquad \mathrm{d}\boldsymbol{\sigma}^{2[l]} = \mathrm{d}\check{\boldsymbol{\sigma}}^{[l]} \circ \widehat{\boldsymbol{\sigma}}^{[l]}/2 \qquad \mathrm{d}\check{\boldsymbol{Z}}_1^{[l]} = \mathrm{d}\boldsymbol{Z}^{[l]}_{\text{norm}} \circ \left\{\mathbf{1}\left(\widehat{\boldsymbol{\sigma}}^{[l]}\right)^{\mathrm{T}}\right\}$$

$$\mathrm{d}\boldsymbol{\mu}^{[l]} = -\left(\mathrm{d}\check{\boldsymbol{Z}}^{[l]}\right)^{\mathrm{T}} \mathbf{1} \qquad \mathrm{d}\check{\boldsymbol{Z}}_2^{[l]} = 2m^{-1}\check{\boldsymbol{Z}}^{[l]} \circ \left\{\mathbf{1}\left(\mathrm{d}\boldsymbol{\sigma}^{2[l]}\right)^{\mathrm{T}}\right\} \qquad \mathrm{d}\widehat{\boldsymbol{\sigma}}^{[l]} = \left(\mathrm{d}\boldsymbol{Z}^{[l]}_{\text{norm}} \circ \check{\boldsymbol{Z}}^{[l]}\right)^{\mathrm{T}} \mathbf{1}$$

$$\mathrm{d}\boldsymbol{Z}_2^{[l]} = m^{-1}\mathbf{1}\left(\mathrm{d}\boldsymbol{\mu}^{[l]}\right)^{\mathrm{T}}$$

Remarks about BN

1. Disadvantage
   - Since we normalize "inputs" before activation, many different "inputs" may result in same "activations"
   - Besides, batch normalization also introduces more model parameters
2. Advantage
   - Stablize forward propogation (The variance can be controlled by $\boldsymbol{\gamma}$'s)
   - Higher learning rates (Batch normalization can make loss and its gradients more smooth)
   - Regularization
     - ▷ We injects noises from other training examples throught mean and variance
     - ▷ Thus, batch normalization may improve the generality of the network

Remarks about BN

1. When applied to validation or test, we usually use the sample mean and variance from the training dataset, instead
   - The sample size may be limited
2. Check Santurkar (2018) and others for more discussion
3. We may introduce other normalization methods in the future sections
4. Why not do batch normalization after activation?

Remarks about dropout and BN

1. BN normalizes each feature to achieve better convergence rate and generality
2. Dropout improves robustness through randomly removing neurons
3. If BN is applied after dropout, we may have problems (Pitfall 1)
   - During training, dropout may change the variance of neurons
   - However, BN stabilizes variance for each mini-batch
   - Moreover, dropout is not conducted for the validation and test data
   - It may cause the shift of statistics, leading to instability (Li et al., 2019)
4. Recommended order:

$$\text{Linear} \rightarrow \text{BN} \rightarrow \text{Activation} \rightarrow \text{Dropout}$$

Remarks about dropout and BN

1. The implementation of both BN and dropout may lead to over regularization (Pitfall 2)
   - BN is a regularization method
   - Implementation of both BN and dropout may result in underfitting
2. Recommendation
   - Use BN + dropout with smaller removal probability
   - Use BN without dropout

Summary

1. One obstacle for training deep neural networks is internal covariate shift
2. Based on the idea of whitening, BN normalizes each feature individually
3. BN stabilizes training, enables larger learning rates, and provides mild regularization