Deep Learning Technology and Application

Ge Li

Peking University

Batch Normalization

Batch Normalization

进行 Batch Normalization 的原因：

- 神经网络的训练目标是在输出层得到原始输入数据数据分布的一个映射，即在训练过程中，应该力图保证数据分布的映射关系；若数据分布在训练过程中发生了偏移，则会降低网络的泛化能力；
- 在网络训练过程中，后一层网络的输入是前一层的输出，因此，前一层网络参数的变化，将导致后一层输入数据分布的改变，且这种改变会在训练过程中向后传递并被逐步放大；这种在训练过程中，数据分布的改变称为"Internal Covariate Shift";
- 在 Batch-based Training 中，若个 Batch 的分布各不相同，网络需要在每个 Batch 的训练中适应不同的分布，从而大大降低训练速度；

所以，为了避免上述问题，可以考虑针对每层网络的每个 Batch 进行 Batch Normalization. 这是一种提高训练速度和效果的非常有效的方法.

Batch Normalization

进行 Batch Normalization 的条件：

1. 起到 Normalize 的作用：控制数据的均值与方差在一定范围内；
2. 保持 Normalize 之前的数据与 Normalize 之后数据之间的映射关系；
3. 保证 Normalize 方法 / 函数的可导性；

论文 [1] 提出了一种针对每个 Batch 进行 Normalize 的方法：

$$y^{(k)} = \gamma^{(k)}\hat{x}^{(k)} + \beta^{(k)}$$

[1]Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." arXiv preprint arXiv:1502.03167 (2015).

Batch Normalization

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1\ldots m}\}$;
Parameters to be learned: γ, β

Output: $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

$$\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma\widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

Batch Normalization

反向传播阶段的计算：

$$\frac{\partial \ell}{\partial \widehat{x}_i} = \frac{\partial \ell}{\partial y_i} \cdot \gamma$$

$$\frac{\partial \ell}{\partial \sigma_\mathcal{B}^2} = \sum_{i=1}^m \frac{\partial \ell}{\partial \widehat{x}_i} \cdot (x_i - \mu_\mathcal{B}) \cdot \frac{-1}{2}(\sigma_\mathcal{B}^2 + \epsilon)^{-3/2}$$

$$\frac{\partial \ell}{\partial \mu_\mathcal{B}} = \left(\sum_{i=1}^m \frac{\partial \ell}{\partial \widehat{x}_i} \cdot \frac{-1}{\sqrt{\sigma_\mathcal{B}^2 + \epsilon}} \right) + \frac{\partial \ell}{\partial \sigma_\mathcal{B}^2} \cdot \frac{\sum_{i=1}^m -2(x_i - \mu_\mathcal{B})}{m}$$

$$\frac{\partial \ell}{\partial x_i} = \frac{\partial \ell}{\partial \widehat{x}_i} \cdot \frac{1}{\sqrt{\sigma_\mathcal{B}^2 + \epsilon}} + \frac{\partial \ell}{\partial \sigma_\mathcal{B}^2} \cdot \frac{2(x_i - \mu_\mathcal{B})}{m} + \frac{\partial \ell}{\partial \mu_\mathcal{B}} \cdot \frac{1}{m}$$

$$\frac{\partial \ell}{\partial \gamma} = \sum_{i=1}^m \frac{\partial \ell}{\partial y_i} \cdot \widehat{x}_i$$

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^m \frac{\partial \ell}{\partial y_i}$$

Batch Normalization

数据测试阶段：

因为测试阶段输入数据可能只有一个，因此，我们使用所有 Batch 的 μ_B 的期望值代替上述公式中的 $E[x]$；使用所有 Batch 方差 δ_B^2 的无偏估计代替上述公式中的 Var$[x]$，即：

$$\mathrm{E}[x] \leftarrow \mathrm{E}_{\mathcal{B}}[\mu_{\mathcal{B}}]$$
$$\mathrm{Var}[x] \leftarrow \frac{m}{m-1}\mathrm{E}_{\mathcal{B}}[\sigma_{\mathcal{B}}^2]$$

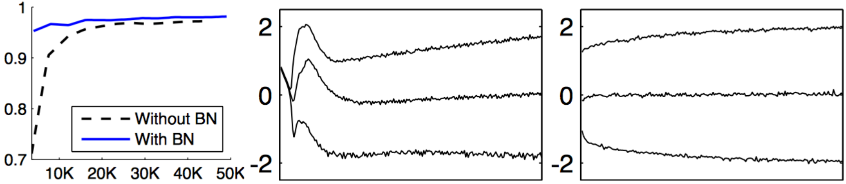
又因为，上述公式中：

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}}$$

则，Batch Normalization 层计算的公式为：

$$y = \frac{\gamma}{\sqrt{\mathrm{Var}[x]+\epsilon}} \cdot x + \left(\beta - \frac{\gamma\,\mathrm{E}[x]}{\sqrt{\mathrm{Var}[x]+\epsilon}}\right)$$

Batch Normalization 的效果：



(a) (b) Without BN (c) With BN

Thanks.