Batch normalization

Created	@October 12, 2021 8:34 PM
Materials	
Type	

Internal covariate shift là hiện tượng phân bố của đầu vào mỗi layer thay đổi, nguyên nhân là do sự thay đổi của W trong quá trình huấn luyện, do đó cần phải khởi tạo các parameter một cách cẩn thận và learning rates nhỏ. Điều này dẫn đến việc làm chậm quá trình huấn luyện và rất khó để huấn luyện mô hình với các hàm phi tuyến có sự bão hòa (saturating? không biết dịch đúng không nữa ☹️) (ví dụ: sigmoid và tanh)

Batch normalization là bước tiến trong việc giải quyết ***internal covariate shift***.

Batch normalization giải quyết vấn đề này thông qua 1 bước chuẩn hóa, sửa lại means và variances của input của layer.

Batch normalization cũng có lợi cho gradient flow bằng cách giảm sự phụ thuộc của gradients vào các parameter.

Batch normalization cho phép sử dụng learning rate lớn hơn mà không có nguy có bị disvergence. Hơn thế nữa, nó còn giảm sự cần thiết của Dropout và cho phép sử dụng các hàm saturating nonlinearity do tránh được việc mắc kẹt ở saturated mode.

Cho $x_i \in \mathcal{B}$, ta sẽ chuẩn hóa x theo công thức: ($\mathcal{B} \in \mathbb{R}^{m \times d}, x_i \in \mathbb{R}^{1 \times d}$)

$$
\begin{aligned}
y_i = BN(x) &= \gamma \odot \frac{x_i - \hat{\mu}_{\mathcal{B}}}{\hat{\sigma}_{\mathcal{B}}} + \beta \\
&= \gamma \odot \hat{x}_i + \beta
\end{aligned}
$$

$$
\hat{x}_i = \frac{x_i - \hat{\mu}_{\mathcal{B}}}{\hat{\sigma}_{\mathcal{B}}}
$$

γ là scale parameter và β là shift parameter, cả γ, β đều có cùng kích thước với x và là tham số mà mô hình cần phải học.

$\hat{\mu}_{\mathcal{B}}$ là sample mean và $\hat{\sigma}_{\mathcal{B}}$ là sample standard deviation tính trên mini-batch $\mathcal{B}$.

$$\hat{\mu}_{\mathcal{B}} = \frac{1}{|\mathcal{B}|}\sum_{x_i \in \mathcal{B}} x_i, \in \mathbb{R}^{1\times d}$$

$$\hat{\sigma}^2_{\mathcal{B}} = \frac{1}{|\mathcal{B}|}\sum_{x_i \in \mathcal{B}} (x_i - \hat{\mu}_{\mathcal{B}})^2 + \epsilon, \in \mathbb{R}^{1\times d}$$

ϵ > 0 để chắc chắn rằng sẽ không có phép chia cho 0.

Trong bài báo gốc, batch normalization được thêm vào trước active function(sau này cũng được sử dụng ngay sau active function).

Fully-connected layer có đầu vào x, affine transformation $Wx + b$, active function ϕ, output của layer này sẽ là:

Convolutional layer sẽ apply batch normalization sau convolution và trước active function. Batch normalization sẽ được áp dụng cho từng channel của đầu ra convolution.

Backpropagation

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

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

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

$$
\begin{aligned}
\frac{\partial \ell}{\partial \hat{\sigma}_{\mathcal{B}}^2} &= \sum_{i=1}^{m} \frac{\partial \ell}{\partial \hat{x}_i} \frac{\partial \hat{x}_i}{\partial \hat{\sigma}_{\mathcal{B}}} \frac{\partial \hat{\sigma}_{\mathcal{B}}}{\partial \hat{\sigma}_{\mathcal{B}}^2} \\
&= \sum_{i=1}^{m} \frac{\partial \ell}{\partial \hat{x}_i} (x_i - \hat{\mu}_{\mathcal{B}}) \frac{-1}{\hat{\sigma}_{\mathcal{B}}^2} \frac{1}{2} (\hat{\sigma}_{\mathcal{B}}^2)^{\frac{-1}{2}} \\
&= \sum_{i=1}^{m} \frac{\partial \ell}{\partial \hat{x}_i} (x_i - \hat{\mu}_{\mathcal{B}}) \frac{-1}{2} (\hat{\sigma}_{\mathcal{B}}^2)^{-3/2}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \ell}{\partial \hat{\mu}_{\mathcal{B}}} &= \sum_{i=1}^{m} \frac{\partial \ell}{\partial \hat{x}_i} \frac{\partial \hat{x}_i}{\partial \hat{\mu}_{\mathcal{B}}} \\
&= \sum_{i=1}^{m} \frac{\partial \ell}{\partial \hat{x}_i} (\frac{-1}{\hat{\sigma}_{\mathcal{B}}} - \frac{(x_i - \hat{\mu}_i)}{\hat{\sigma}_{\mathcal{B}}^2} \frac{\partial \hat{\sigma}_{\mathcal{B}}}{\partial \hat{\mu}_{\mathcal{B}}}) \\
&= \sum_{i=1}^{m} \frac{\partial \ell}{\partial \hat{x}_i} \frac{-1}{\hat{\sigma}_{\mathcal{B}}} + \underbrace{\sum_{i=1}^{m} \frac{\partial \ell}{\partial \hat{x}_i} (x_i - \hat{\mu}_{\mathcal{B}}) \frac{-1}{\hat{\sigma}_{\mathcal{B}}^2} \frac{\partial \hat{\sigma}_{\mathcal{B}}}{\partial \hat{\sigma}_{\mathcal{B}}^2}}_{= \frac{\partial \ell}{\partial \hat{\sigma}_{\mathcal{B}}^2}} \frac{\partial \hat{\sigma}_{\mathcal{B}}^2}{\partial \hat{\mu}_{\mathcal{B}}} \\
&= \sum_{i=1}^{m} \frac{\partial \ell}{\partial \hat{x}_i} \frac{-1}{\hat{\sigma}_{\mathcal{B}}} + \frac{\partial \ell}{\partial \hat{\sigma}_{\mathcal{B}}^2} \frac{\sum_{i=1}^{m} -2(x_i - \hat{\mu}_{\mathcal{B}})}{m} \\
\frac{\partial \ell}{\partial \hat{\sigma}_{\mathcal{B}}^2} \frac{\sum_{i=1}^{m} -2(x_i - \hat{\mu}_{\mathcal{B}})}{m} &= \frac{\partial \ell}{\partial \hat{\sigma}_{\mathcal{B}}^2} (-2) (\frac{1}{m} \sum_{i=1}^{m} x_i - \frac{1}{m} \sum_{i=1}^{m} \hat{\mu}_{\mathcal{B}}) \\
&= \frac{\partial \ell}{\partial \hat{\sigma}_{\mathcal{B}}^2} (-2) (\hat{\mu}_{\mathcal{B}} - \frac{1}{m} m \hat{\mu}_{\mathcal{B}}) \\
&= 0 \\
\Rightarrow \frac{\partial \ell}{\partial \hat{\mu}_{\mathcal{B}}} &= \sum_{i=1}^{m} \frac{\partial \ell}{\partial \hat{x}_i} \frac{-1}{\hat{\sigma}_{\mathcal{B}}}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \ell}{\partial x_i} &= \sum_{i=1}^{m} \frac{\partial \ell}{\partial \hat{x}_i} \frac{\partial \hat{x}_i}{\partial x_i} + \frac{\partial \ell}{\partial \hat{\mu}_{\mathcal{B}}} \frac{\partial \hat{\mu}_{\mathcal{B}}}{\partial x_i} + \frac{\partial \ell}{\partial \hat{\sigma}_{\mathcal{B}}^2} \frac{\partial \hat{\sigma}_{\mathcal{B}}^2}{\partial x_i} \\
&= \sum_{i=1}^{m} \frac{\partial \ell}{\partial \hat{x}_i} \frac{1}{\hat{\sigma}_{\mathcal{B}}} + \frac{\partial \ell}{\partial \hat{\mu}_{\mathcal{B}}} \frac{1}{m} + \frac{\partial \ell}{\partial \hat{\sigma}_{\mathcal{B}}^2} \frac{2(x_i - \hat{\mu}_{\mathcal{B}})}{m}
\end{aligned}
$$