Batch Normalization

目次

1 順伝播

学習にバッチサイズ m の入力が与えられるときの順伝播を考える。ネットワークの i 番目のニューロンへの λ 個目の入力を $x^I_{i,\lambda}$、同様に Batch Normalization layer への入力を $x_{i,\lambda}$、ネットワークの出力を $x^O_{i,\lambda}$ とする。順伝播の計算は

$$
\begin{aligned}
x_{i,\lambda} &= f_{i,\lambda}\left(x^I, \theta^I\right) \\
y_{i,\lambda} &= BN_{i,\lambda,\beta_i,\gamma_i}(x) \\
x^O_{i,\lambda} &= gi,\lambda\left(o, \theta^O\right)
\end{aligned}
$$

として計算されるとする。ここで θ^I, θ^O はそれぞれ Batch Normalization より前と後の層のパラメータとし、β_i, γ_i は Normalization 語のリスケーリングとシフトのパラメータとした。また、Normalization の計算 $BN_{i,\lambda,\beta_i,\gamma_i}(x)$ は

$$
\begin{aligned}
\mu_i &= \frac{1}{m}\sum_{\lambda=1}^{m} x_{i,\lambda} \\
\sigma_i^2 &= \frac{1}{m}\sum_{\lambda=1}^{m}(x_{i,\lambda} - \mu_i)^2 \\
\hat{x}_{i,\lambda} &= \frac{x_{i,\lambda} - \mu_i}{\sqrt{\sigma_i^2 - \epsilon}} \\
y_{i,\lambda} &= \gamma_i x_{i,\lambda} + \beta_i
\end{aligned}
$$

として計算される。

また、学習の際はミニバッチによって平均と分散が計算されるが、実際の識別の場合は学習の最後のエポックで用いた平均と分散を用いる等して計算される。

2 逆伝播

誤差関数を E と書けば、前述の計算式から各パラメータの更新量 $\Delta\theta_i^I, \Delta\theta_i^O, \Delta\beta_i, \Delta\gamma_i$ は

$$\frac{\partial y_{i,\lambda}}{\partial \beta_i} = 1, \quad \frac{\partial y_{i,\lambda}}{\partial \gamma_i} = \hat{x}_{i,\lambda}$$
$$\frac{\partial y_{i,\nu}}{\partial x_{i,\lambda}} = \frac{\gamma_i}{\sqrt{\sigma_i^2 + \epsilon}} \left(\delta_\lambda^\nu - \frac{1}{m} - \frac{1}{m} \hat{x}_{i,\nu} \hat{x}_{i,\lambda} \right)$$

と置くことで

$$\Delta\theta_i^O = -\sum_{j=1}^{n} \sum_{\lambda=1}^{m} \partial_{x_{j,\lambda}^O} E \frac{\partial x_{j,\lambda}^O}{\partial \theta_i^O}$$
$$\Delta\beta[i] = -\sum_{\lambda=1}^{m} \partial_{y_{i,\lambda}} E, \quad \Delta\gamma[i] = -\sum_{\lambda=1}^{m} \partial_{y_{i,\lambda}} E \hat{x}_{i,\lambda}$$
$$\Delta\theta_i^I = -\sum_{j=1}^{n} \sum_{\lambda=1}^{m} \sum_{\nu=1}^{m} \gamma_j \frac{1}{\sqrt{\sigma_j^2 + \epsilon}} \left(\partial_{y_{j,\nu}} E \left(\delta_\lambda^\nu - \frac{1}{m} - \frac{1}{m} \hat{x}_{j,\nu} \hat{x}_{j,\lambda} \right) \right) \frac{\partial x_{j,\lambda}}{\partial \theta_i^I}$$

として計算される。

上の式から

$$\frac{\partial E}{\partial \beta_i} = \sum_{\lambda=1}^{m} \partial_{y_{i,\lambda}} E$$
$$\frac{\partial E}{\partial \gamma} = \sum_{\lambda=1}^{m} \partial_{y_{i,\lambda}} E \hat{x}_{i,\lambda}$$
$$\frac{\partial E}{\partial x_{i,\lambda}} = \sum_{\nu=1}^{m} \gamma_i \frac{1}{\sqrt{\sigma_i^2 + \epsilon}} \left(\partial_{y_{i,\nu}} E \left(\delta_\lambda^\nu - \frac{1}{m} - \frac{1}{m} \hat{x}_{i,\nu} \hat{x}_{i,\lambda} \right) \right)$$

である。ここで論文中に示される

$$
\begin{aligned}
\frac{\partial E}{\partial \hat{x}_{i,\lambda}} &= \frac{\partial E}{\partial y_{i,\lambda}} \gamma_i \\
\frac{\partial E}{\partial \sigma_i^2} &= \sum_{\lambda=1}^{m} \frac{\partial E}{\partial \hat{x}_{i,\lambda}} (x_{i,\lambda} - \mu_i) \frac{-1}{2} (\sigma_i^2 + \epsilon)^{\frac{-3\gamma_i}{2}} \\
\frac{\partial E}{\partial \mu_i} &= \left(\sum_{\lambda=1}^{m} \frac{\partial E}{\partial \hat{x}_{i,\lambda}} \frac{-1}{\sqrt{\sigma_i^2 + \epsilon}} \right) + \frac{\partial E}{\partial \sigma_i^2} \sum_{\lambda=1}^{m} \frac{-2(x_{i,\lambda} - \mu_i)}{m} \\
\frac{\partial E}{\partial x_{i,\lambda}} &= \frac{\partial E}{\partial \hat{x}_{i,\lambda}} \frac{1}{\sqrt{\sigma_i^2 + \epsilon}} + \frac{\partial E}{\partial \sigma_i^2} \frac{2(x_{i,\lambda} - \mu_i)}{m} + \frac{\partial E}{\partial \mu_i} \frac{1}{m} \\
\frac{\partial E}{\partial \gamma_i} &= \sum_{\lambda=1}^{m} \frac{\partial E}{\partial y_{i,\lambda}} \hat{x}_{i,\lambda} \\
\frac{\partial E}{\partial \beta} &= \sum_{\lambda=1}^{m} \frac{\partial E}{\partial y_{i,\lambda}}
\end{aligned}
$$

と一致することを確認する。それぞれ代入することで

$$
\begin{aligned}
\frac{\partial E}{\partial \sigma_i^2} &= \sum_{\lambda=1}^{m} \frac{\partial E}{\partial y_{i,\lambda}} \gamma_i (x_{i,\lambda} - \mu_i) \frac{-1}{2} (\sigma_i^2 + \epsilon)^{\frac{-3}{2}} \\
&= \sum_{\lambda=1}^{m} \frac{\partial E}{\partial y_{i,\lambda}} \gamma_i (x_{i,\lambda} - \mu_i) \frac{-1}{2} (\sigma_i^2 + \epsilon)^{\frac{-3}{2}} \\
&= -\frac{1}{2} \sum_{\lambda=1}^{m} \frac{\partial E}{\partial y_{i,\lambda}} \gamma_i (\sigma_i^2 + \epsilon)^{\frac{-3}{2}} (x_{i,\lambda} - \mu_i)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial E}{\partial \mu_i} &= \left(\sum_{\lambda=1}^{m} \frac{\partial E}{\partial \hat{x}_{i,\lambda}} \frac{-1}{\sqrt{\sigma_i^2 + \epsilon}} \right) + \frac{\partial E}{\partial \sigma_i^2} \sum_{\lambda=1}^{m} \frac{-2(x_{i,\lambda} - \mu_i)}{m} \\
&= \left(\sum_{\lambda=1}^{m} \frac{\partial E}{\partial y_{i,\lambda}} \gamma_i \frac{-1}{\sqrt{\sigma_i^2 + \epsilon}} \right) + \frac{\partial E}{\partial \sigma_i^2} 0 \\
&= -\sum_{\lambda=1}^{m} \frac{\partial E}{\partial y_{i,\lambda}} \gamma_i \frac{1}{\sqrt{\sigma_i^2 + \epsilon}}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial E}{\partial x_{i,\lambda}} &= \frac{\partial E}{\partial \hat{x}_{i,\lambda}}\frac{1}{\sqrt{\sigma_i^2+\epsilon}} + \frac{\partial E}{\partial \sigma_i^2}\frac{2(x_{i,\lambda}-\mu_i)}{m} + \frac{\partial E}{\partial \mu_i}\frac{1}{m} \\
&= \frac{\partial E}{\partial y_{i,\lambda}}\gamma_i\frac{1}{\sqrt{\sigma_i^2+\epsilon}} - \frac{1}{2}\sum_{\nu=1}^{m}\frac{\partial E}{\partial y_{i,\nu}}\gamma_i(\sigma_i^2+\epsilon)^{\frac{-3}{2}}(x_{i,\nu}-\mu_i)\frac{2(x_{i,\lambda}-\mu_i)}{m} - \sum_{\nu=1}^{m}\frac{\partial E}{\partial y_{i,\nu}}\gamma_i\frac{1}{\sqrt{\sigma_i^2+\epsilon}}\frac{1}{m} \\
&= \frac{\partial E}{\partial y_{i,\lambda}}\gamma_i\frac{1}{\sqrt{\sigma_i^2+\epsilon}} - \frac{1}{m}\sum_{\nu=1}^{m}\frac{\partial E}{\partial y_{i,\nu}}\gamma_i(\sigma_i^2+\epsilon)^{\frac{-3}{2}}(x_{i,\nu}-\mu_i)(x_{i,\lambda}-\mu_i) - \frac{1}{m}\sum_{\nu=1}^{m}\frac{\partial E}{\partial y_{i,\nu}}\gamma_i\frac{1}{\sqrt{\sigma_i^2+\epsilon}} \\
&= \sum_{\nu=1}^{m}\frac{\partial E}{\partial y_{i,\nu}}\gamma_i\delta_\lambda^\nu\frac{1}{\sqrt{\sigma_i^2+\epsilon}} - \frac{1}{m}\sum_{\nu=1}^{m}\frac{\partial E}{\partial y_{i,\nu}}\gamma_i(\sigma_i^2+\epsilon)^{\frac{-3}{2}}(x_{i,\nu}-\mu_i)(x_{i,\lambda}-\mu_i) - \frac{1}{m}\sum_{\nu=1}^{m}\frac{\partial E}{\partial y_{i,\nu}}\gamma_i\frac{1}{\sqrt{\sigma_i^2+\epsilon}} \\
&= \sum_{\nu=1}^{m}\gamma_i\frac{1}{\sqrt{\sigma_i^2+\epsilon}}\left(\frac{\partial E}{\partial y_{i,\nu}}\left(\delta_\lambda^\nu - \frac{1}{m}(\sigma_i^2+\epsilon)^{-1}(x_{i,\nu}-\mu_i)(x_{i,\lambda}-\mu_i) - \frac{1}{m}\right)\right) \\
&= \sum_{\nu=1}^{m}\gamma_i\frac{1}{\sqrt{\sigma_i^2+\epsilon}}\left(\frac{\partial E}{\partial y_{i,\nu}}\left(\delta_\lambda^\nu - \frac{1}{m}\hat{x}_{i,\nu}\hat{x}_{i,\lambda} - \frac{1}{m}\right)\right) \\
&= \sum_{\nu=1}^{m}\gamma_i\frac{1}{\sqrt{\sigma_i^2+\epsilon}}\left(\frac{\partial E}{\partial y_{i,\nu}}\left(\delta_\lambda^\nu - \frac{1}{m} - \frac{1}{m}\hat{x}_{i,\nu}\hat{x}_{i,\lambda}\right)\right)
\end{aligned}
$$

となり一致する。

参考文献

[1] 斎藤康毅，“ゼロから作る Deep Learning”，オライリージャパン，(2016).

[2] Sergey loffe and Christian Szegedy (2015) : Batch Normalization : Accelerating Deep Network Training by Reducing Internal Covariate Shift. ArXiv:1502.03167 [cs] (February 2015).