WEEKLY PARTICIPATION 9

Recall that batch normalization layers are used as follows[1]:

$$\begin{aligned}
\boldsymbol{a}^\ell &= \mathrm{BN}_{\boldsymbol{\gamma}^\ell,\boldsymbol{\beta}^\ell}(\boldsymbol{W}^\ell \boldsymbol{o}^{\ell-1}) = \boldsymbol{\gamma}^\ell \odot \left(\frac{\boldsymbol{W}^\ell \boldsymbol{o}^{\ell-1} - \boldsymbol{\mu}_B}{\sqrt{\boldsymbol{\sigma}_B^2 + \epsilon}} \right) + \boldsymbol{\beta}^\ell \\
&= \boldsymbol{\gamma}^\ell \odot \overline{\boldsymbol{o}^{\ell-1}} + \boldsymbol{\beta}^\ell, \\
\boldsymbol{o}^\ell &= \sigma(\boldsymbol{a}^\ell).
\end{aligned}$$

For convenience, we have denoted the affinely transformed output of layer $\ell - 1$ by

$$\overline{\boldsymbol{o}^{\ell-1}} = \frac{\boldsymbol{W}^\ell \boldsymbol{o}^{\ell-1} - \boldsymbol{\mu}_B}{\sqrt{\boldsymbol{\sigma}_B^2 + \epsilon}}.$$

In the above expressions, ϵ is a small positive number to guard against division by zero, $\odot$ denotes element-wise multiplication, and $\boldsymbol{\mu}_B, \boldsymbol{\sigma}_B^2 \in \mathbb{R}^{n_{\ell-1}}$ are the vectors of minibatch means and minibatch variances for each neuron in layer $\ell - 1$:

$$\boldsymbol{\mu}_B = \frac{1}{m} \sum_{i=1}^m (\boldsymbol{o}^{\ell-1})_i, \qquad \boldsymbol{\sigma}_B = \frac{1}{m} \sum_{i=1}^m ((\boldsymbol{o}^{\ell-1})_i - \boldsymbol{\mu}_B)^2.$$

The scale and shift vectors $\boldsymbol{\gamma}^\ell, \boldsymbol{\beta}^\ell \in \mathbb{R}^{n_{\ell-1}}$ are parameters that must be learned during training, using backpropagation. Let f be the training objective.

(1) Verify that $\boldsymbol{J}_{\boldsymbol{o}^\ell}(\boldsymbol{a}^\ell) = \mathrm{diag}(\sigma'(\boldsymbol{a}^\ell))$, and use this fact to give an expression for $\nabla_{\boldsymbol{a}^\ell} f$, assuming that $\nabla_{\boldsymbol{o}^\ell} f$ is known.

(2) Verify that $\boldsymbol{J}_{\boldsymbol{a}^\ell}(\boldsymbol{\gamma}) = \mathrm{diag}\left(\overline{\boldsymbol{o}^{\ell-1}}\right)$, and use this fact to give an expression for $\nabla_{\boldsymbol{\gamma}^\ell} f$ in terms of $\nabla_{\boldsymbol{a}^\ell} f$.

(3) Verify that $\boldsymbol{J}_{\boldsymbol{a}^\ell}(\boldsymbol{\beta}^\ell) = \boldsymbol{I}$, and use this fact to give an expression for $\nabla_{\boldsymbol{\beta}^\ell} f$ in terms of $\nabla_{\boldsymbol{a}^\ell} f$.

(4) Observe (no need to verify this) that for any $i \in [n_\ell]$,

$$\begin{aligned}
\left[\boldsymbol{J}_{\boldsymbol{a}^\ell}(\boldsymbol{W}^\ell)\right]_{i,:} &= \left[\frac{\partial (\boldsymbol{a}^\ell)_i}{\partial (\boldsymbol{W}^\ell)_{p,q}} \right]_{p,q=1}^{n_\ell, n_{\ell-1}} \\
&= \mathrm{diag}\left(\frac{\boldsymbol{\gamma}^\ell}{\sqrt{\boldsymbol{\sigma}_B^2 + \epsilon}} \right) \boldsymbol{e}_i (\boldsymbol{o}^{\ell-1})^T,
\end{aligned}$$

so for any vector $\boldsymbol{v} \in \mathbb{R}^{n_\ell}$,

$$\boldsymbol{J}_{\boldsymbol{a}^\ell}(\boldsymbol{W}^\ell)^T \boldsymbol{v} = \mathrm{diag}\left(\frac{\boldsymbol{\gamma}^\ell}{\sqrt{\boldsymbol{\sigma}_B^2 + \epsilon}} \right) \boldsymbol{v} (\boldsymbol{o}^{\ell-1})^T.$$

Use this fact to give an expression for $\nabla_{\boldsymbol{W}^\ell} f$ in terms of $\nabla_{\boldsymbol{a}^\ell} f$.

[1] Notice that we do not include a bias $\boldsymbol{b}^\ell$ in the call to the BN primitive because $\boldsymbol{\beta}^\ell$ is our bias.