Batch Normalization Notes

Justin Johnson

February 23, 2016

Assume a set of scalar inputs $x_1, \ldots, x_N$. Batch normalization uses the sample mean and variance to output a set of scalar outputs $y_1, \ldots, y_N$ that are zero-mean and unit variance. The transform is defined as follows:

$$\mu = \frac{1}{N}\sum_{k=1}^{N} x_k \qquad\qquad v = \frac{1}{N}\sum_{k=1}^{N}(x_k - \mu)^2$$

$$\sigma = \sqrt{v + \epsilon} \qquad\qquad y_i = \frac{x_i - \mu}{\sigma}$$

where ϵ is a small positive constant used to avoid dividing by zero.

To use batch normalization as a component of a neural network, we need to derive an expression to be used in back-propagation.

We use $\mathcal{I}[p]$ to denote the indicator function which is 1 if the predicate p is true and 0 otherwise. We then compute

$$\frac{\partial \mu}{\partial x_i} = \frac{1}{N} \tag{1}$$

$$\frac{\partial v}{\partial x_i} = \frac{1}{N}\sum_{k=1}^{N}\frac{\partial}{\partial x_i}\left[(x_k - \mu)^2\right] = \frac{2}{N}\sum_{k=1}^{N}(x_k - \mu)\frac{\partial}{\partial x_i}\left[x_k - \mu\right] \tag{2}$$

$$= \frac{2}{N}\sum_{k=1}^{N}(x_k - \mu)\left(\mathcal{I}[i = k] - \frac{1}{N}\right) \tag{3}$$

$$= \frac{2}{N}\sum_{k=1}^{N}\left(x_k\mathcal{I}[i = k] - \frac{x_k}{N} - \mu\mathcal{I}[i = k] + \frac{\mu}{N}\right) \tag{4}$$

$$= \frac{2}{N}\big(x_i - \mu\big) \tag{5}$$

$$\frac{\partial \sigma}{\partial x_i} = \frac{1}{2\sqrt{v + \epsilon}}\frac{\partial}{\partial x_i}\big[v + \epsilon\big] = \frac{1}{2\sigma}\frac{2}{N}\big(x_i - \mu\big) = \frac{x_i - \mu}{N\sigma} = \frac{y_i}{N} \tag{6}$$

$$
\begin{aligned}
\frac{\partial y_j}{\partial x_i} &= \frac{\partial}{\partial x_i}\left[\frac{x_j-\mu}{\sigma}\right] && (7)\\
&= \frac{1}{\sigma}\frac{\partial}{\partial x_i}\Big[x_j-\mu\Big]\sigma - \frac{1}{\sigma^2}\frac{\partial \sigma}{\partial x_i}\big(x_j-\mu\big) && (8)\\
&= \frac{1}{\sigma}\left(\mathcal{I}[i=j]-\frac{1}{N}\right) - \frac{x_j-\mu}{\sigma^2}\frac{y_i}{N} && (9)\\
&= \frac{\mathcal{I}[i=j]}{\sigma} - \frac{1}{\sigma N} - \frac{y_i y_j}{\sigma N} && (10)
\end{aligned}
$$

For back-propagation, we assume that ℓ is the scalar-valued loss of the entire network. By the chain rule we know that

$$
\begin{aligned}
\frac{\partial \ell}{\partial x_i} &= \sum_{j=1}^{N}\frac{\partial \ell}{\partial y_j}\frac{\partial y_j}{\partial x_i} && (11)\\
&= \frac{1}{\sigma}\sum_{j=1}^{N}\frac{\partial \ell}{\partial y_j}\left(\mathcal{I}[i=j]-\frac{1}{N}-\frac{y_i y_j}{N}\right) && (12)\\
&= \frac{1}{\sigma}\frac{\partial \ell}{\partial y_i} - \frac{1}{\sigma N}\sum_{j=1}^{N}\frac{\partial \ell}{\partial y_j} - \frac{y_i}{\sigma N}\sum_{j=1}^{N} y_j\frac{\partial \ell}{\partial y_j} && (13)
\end{aligned}
$$

If we define

$$
\frac{\partial \ell}{\partial x} = \left(\frac{\partial \ell}{\partial x_i},\ldots,\frac{\partial \ell}{\partial x_N}\right) \qquad\qquad \frac{\partial \ell}{\partial y} = \left(\frac{\partial \ell}{\partial y_i},\ldots,\frac{\partial \ell}{\partial y_N}\right)
$$

then we have a simple vectorized expression to use in backpropagation:

$$
\frac{\partial \ell}{\partial x} = \frac{1}{\sigma}\left(\frac{\partial \ell}{\partial y} - \overline{\frac{\partial \ell}{\partial y}}\,\mathbf{1} - y\left(\overline{y\odot\frac{\partial \ell}{\partial y}}\right)\right) \tag{14}
$$

where $\overline{z}$ is the mean of a vector z, $\mathbf{1}$ is a constant vector of ones, and $\odot$ denotes elementwise multiplication of vectors.