Backpropagation notes

Jakub Waszczuk

December 2020

1 Addition

Let z be a sum of x and y:

$$z = x + y$$

and $\ell = g(z)$ for some differentiable function g, where ℓ represents the loss that we want to minimise. From the chain rule we have:

$$\frac{\partial \ell}{\partial x} = \frac{\partial \ell}{\partial z} \frac{\partial z}{\partial x}$$

In PyTorch, $\partial \ell / \partial z$ is something that we get as argument during the backward step. As for $\partial z / \partial x$, we can rewrite it using the sum rule as:

$$\frac{\partial z}{\partial x} = \frac{\partial (x + y)}{\partial x} = \frac{\partial x}{\partial x} + \frac{\partial y}{\partial x} = 1 + 0 = 1$$

Hence we get:

$$\frac{\partial \ell}{\partial x} = \frac{\partial \ell}{\partial z} 1 = \frac{\partial \ell}{\partial z}$$

Similarly with respect to y, $\partial \ell / \partial y = \partial \ell / \partial z$.

1.1 Element-wise addition

We can generalize this to a setting where both x and y are tensors and $+$ is element-wise. For simplicity, let's assume that x and y are vectors of length n. Then $z = x + y$ can be rewritten as:

$$\forall_{i=1..n} \; z_i = x_i + y_i$$

From the multivariate chain rule:

$$\frac{\partial \ell}{\partial x_i} = \sum_{j=1}^{n} \frac{\partial \ell}{\partial z_j} \frac{\partial z_j}{\partial x_i}$$

Due to the element-wise nature of the operation, this formula can be simplified quite significantly:

$$\frac{\partial z_j}{\partial x_i} = \frac{\partial x_j + \partial y_j}{\partial x_i} = \frac{\partial x_j}{\partial x_i} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

Therefore:
$$\left(\frac{\partial \ell}{\partial x_1}, \frac{\partial \ell}{\partial x_2}, \ldots, \frac{\partial \ell}{\partial x_n}\right) = \left(\frac{\partial \ell}{\partial z_1}, \frac{\partial \ell}{\partial z_2}, \ldots, \frac{\partial \ell}{\partial z_n}\right)$$
which can be simply expressed as:
$$\frac{\partial \ell}{\partial x} = \frac{\partial \ell}{\partial z}$$
where both $\partial \ell / \partial x$ and $\partial \ell / \partial x$ are tensors and $\partial \ell / \partial z$ is the input of the backward method.

2 Product

Let z be a product of x and y:
$$z = xy$$
The partial derivative of z w.r.t x can be calculated using the product rule:
$$\frac{\partial z}{\partial x} = \frac{\partial (xy)}{\partial x} = \frac{\partial x}{\partial x} y + \frac{\partial y}{\partial x} x = y$$
Finally using the chain rule:
$$\frac{\partial \ell}{\partial x} = \frac{\partial \ell}{\partial z}\frac{\partial z}{\partial x} = \frac{\partial \ell}{\partial z} y$$
Similarly, $\partial \ell / \partial y = (\partial \ell / \partial z) x$, and both equations generalize to arbitrarily-shaped tensors due to the element-wise nature of the operation, just as in the case of element-wise addition (see Sec. 1.1).

3 Sigmoid

Let f be the sigmoid function:
$$y = f(x) = \frac{1}{1 + \exp(-x)}$$
The derivative of sigmoid is:
$$\frac{\partial y}{\partial x} = y(1 - y)$$
Therefore, from the chain rule:
$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial y} y(1 - y)$$