Deep Learning: Backpropagation

Soon-Hyung Yook

November 14, 2023

Contents

Chapter 1

Backpropagation

1.1 Partial derivative

- Using the computational graph is a much easier way to obtain derivatives.
- However, we have to understand how we can make the computational graph.
- For this, we have to understand the chain rule first, and the relation between the computational graph and the chain rule.

Let's start with a very very simple example.

$$\begin{aligned} f(y) &= \text{a function of } y \\ y &= x_1 + x_2 \end{aligned} \tag{1.1}$$

Figure 1.1: Compuation graph for partial derivative of Eq.(1.1).

Here

$$\frac{\partial y}{\partial x_1} = \frac{\partial y}{\partial x_2} = 1 \tag{1.2}$$

Thus + node just sends the upstream input itself to the downstream branches.

1.2 Matrix Input

1.2.1 Shape of the matrices

If the input data are in the form of a matrix, then the situation becomes more complicated. Let's consider the following example.

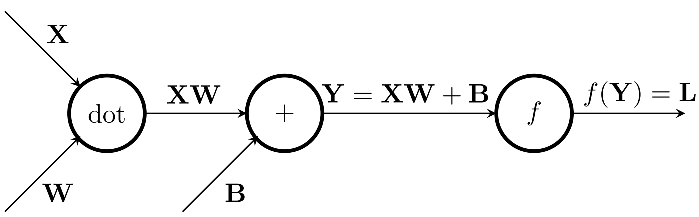

Figure 1.2: Inputs $\mathbf{X}$ and $\mathbf{W}$ are matrices. Thus $\mathbf{B}$, $\mathbf{Y}$, and $\mathbf{L}$ are also matrices.

In order to check the shape of each matrix, let's first take a walk along the forward route. For simplicity, let $\mathbf{X}$ be a (1×2) matrix (row vector) and $\mathbf{W}$ be a (2×3) matrix:

$$\mathbf{X} = (x_1 \ x_2) \tag{1.3}$$

and

$$\mathbf{W} = \begin{pmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{pmatrix}. \tag{1.4}$$

Let the bias vector $\mathbf{B}$ be a row vector,

$$\mathbf{B} = (b_1 \ b_2 \ b_3), \tag{1.5}$$

where b_1, b_2, and b_3 are some constants. The forward flow depicted in Fig.1.2 is obtained as

$$\mathbf{X}W = (x_1w_{11} + x_2w_{21} \quad x_1w_{12} + x_2w_{22} \quad x_1w_{13} + x_2w_{23}), \tag{1.6}$$

or

$$(\mathbf{X}W)_i = \sum_{j=1}^{2} x_j w_{ji}. \tag{1.7}$$

The shape of $\mathbf{X}W$ becomes (1×3) (row vector). By adding $\mathbf{B}$ to Eq.(1.7), we obtain $\mathbf{Y}$ as

$$\begin{aligned} \mathbf{Y} &= \mathbf{XW} + \mathbf{B} \\ &= (x_1w_{11} + x_2w_{21} + b_1 \quad x_1w_{12} + x_2w_{22} + b_2 \quad x_1w_{13} + x_2w_{23} + b_3) \\ &\equiv (y_1 \quad y_2 \quad y_3). \end{aligned} \tag{1.8}$$

In general,

$$(\mathbf{Y})_i \equiv y_i = (\mathbf{XW})_i + (\mathbf{B})_i = \sum_{j=1}^{2} x_j w_{ji} + b_i. \tag{1.9}$$

In Fig.1.2, we assume that $\mathbf{L} = f(\mathbf{Y})$ is a (1×3) matrix (row vector), i.e.,

$$\mathbf{L} = (f(y_1) \quad f(y_2) \quad f(y_3)) = (L_1 \quad L_2 \quad L_3). \tag{1.10}$$

1.2.2 Partial derivative $\frac{\partial \mathbf{L}}{\partial \mathbf{X}}$

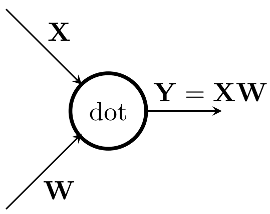

Figure 1.3: Compuation graph for a dot product.

In this example (Fig.1.2), there are two input parameters, x_1 and x_2. Thus, there are two possible cases of partial derivatives. Let's start with the simplest one:

$$\frac{\partial \mathbf{L}}{\partial \mathbf{X}} \equiv \begin{pmatrix} \frac{\partial \mathbf{L}}{\partial x_1} & \frac{\partial \mathbf{L}}{\partial x_2} \end{pmatrix} = \begin{pmatrix} \color{red}{\frac{\partial L_1}{\partial x_1}} & \frac{\partial L_1}{\partial x_2} \\ \frac{\partial L_2}{\partial x_1} & \frac{\partial L_2}{\partial x_2} \\ \frac{\partial L_3}{\partial x_1} & \frac{\partial L_3}{\partial x_2} \end{pmatrix} \tag{1.11}$$

In fact Eq.(1.11) is a Jacobian. However, sometimes $\mathbf{L}$ can be a scalar. By using the chain rule, the first element in the first column in Eq.(1.11) can be rewritten as

$$\frac{\partial L_1(\mathbf{Y})}{\partial x_1} = \frac{\partial L_1}{\partial y_1}\frac{\partial y_1}{\partial x_1} + \frac{\partial L_1}{\partial y_2}\frac{\partial y_2}{\partial x_1} + \frac{\partial L_1}{\partial y_3}\frac{\partial y_3}{\partial x_1} = \sum_{j=1}^{3} \frac{\partial L_1}{\partial y_j}\frac{\partial y_j}{\partial x_1}. \tag{1.12}$$

In general, we obtain

$$\frac{\partial L_i}{\partial x_j} = \sum_{k=1}^{3} \frac{\partial L_i}{\partial y_k}\frac{\partial y_k}{\partial x_j} \equiv \left(\frac{\partial \mathbf{L}}{\partial \mathbf{X}}\right)_{ij} \tag{1.13}$$

Since $\frac{\partial \mathbf{L}}{\partial \mathbf{Y}}$ is a (3×3) matrix,

$$\begin{aligned} \frac{\partial \mathbf{L}}{\partial \mathbf{Y}} &\equiv \begin{pmatrix} \frac{\partial \mathbf{L}}{\partial y_1} & \frac{\partial \mathbf{L}}{\partial y_2} & \frac{\partial \mathbf{L}}{\partial y_3} \end{pmatrix} \\ &= \begin{pmatrix} \frac{\partial L_1}{\partial y_1} & \frac{\partial L_1}{\partial y_2} & \frac{\partial L_1}{\partial y_3} \\ \frac{\partial L_2}{\partial y_1} & \frac{\partial L_2}{\partial y_2} & \frac{\partial L_2}{\partial y_3} \\ \frac{\partial L_3}{\partial y_1} & \frac{\partial L_3}{\partial y_2} & \frac{\partial L_3}{\partial y_3} \end{pmatrix}, \end{aligned} \tag{1.14}$$

the partial derivative $\frac{\partial \mathbf{L}}{\partial \mathbf{X}}$ becomes

$$\frac{\partial \mathbf{L}}{\partial \mathbf{X}} = \frac{\partial \mathbf{L}}{\partial \mathbf{Y}}\frac{\partial \mathbf{Y}}{\partial \mathbf{X}}, \tag{1.15}$$

where $\frac{\partial \mathbf{Y}}{\partial \mathbf{X}}$ is a (3×2) matrix,

$$\left(\frac{\partial \mathbf{Y}}{\partial \mathbf{X}}\right)_{ij} = \frac{\partial y_i}{\partial x_j}. \tag{1.16}$$

From Eq.(1.8), $y_i = \sum_{j=1}^{2} x_j w_{ji} + b_i$ and

$$\frac{\partial y_i}{\partial x_j} = w_{ji}. \tag{1.17}$$

Therefore,

$$\left(\frac{\partial \mathbf{Y}}{\partial \mathbf{X}}\right)_{ij} = w_{ji}. \tag{1.18}$$

Eq.(1.18) implies that

$$\frac{\partial \mathbf{Y}}{\partial \mathbf{X}} = \mathbf{W}^T. \tag{1.19}$$

Therefore,

$$\frac{\partial \mathbf{L}}{\partial \mathbf{X}} = \frac{\partial \mathbf{L}}{\partial \mathbf{Y}}\mathbf{W}^T. \tag{1.20}$$

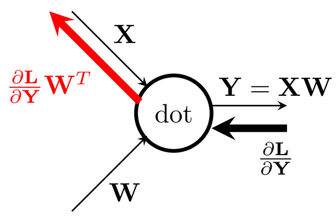

Figure 1.4: Compuation graph for the backpropagation of the dot product.

1.2.3 Partial derivative $\frac{\partial \mathbf{L}}{\partial \mathbf{W}}$

Finding an expression for partial derivative $\frac{\partial \mathbf{L}}{\partial \mathbf{X}}$ is a little bit complicated. $\frac{\partial \mathbf{L}}{\partial \mathbf{X}}$ can be expressed as

$$\frac{\partial \mathbf{L}}{\partial \mathbf{W}} = \begin{pmatrix} \frac{\partial \mathbf{L}}{\partial w_{11}} & \frac{\partial \mathbf{L}}{\partial w_{12}} & \frac{\partial \mathbf{L}}{\partial w_{13}} \\ \frac{\partial \mathbf{L}}{\partial w_{21}} & \frac{\partial \mathbf{L}}{\partial w_{22}} & \frac{\partial \mathbf{L}}{\partial w_{23}} \end{pmatrix} \tag{1.21}$$

In fact, Eq.(1.21) is not a simple matrix (It's a tensor.) As an example, let's apply the chain rule to the first element in the first column,

$$\frac{\partial \mathbf{L}}{\partial w_{11}} = \begin{pmatrix} \frac{\partial L_1}{\partial w_{11}} & \frac{\partial L_2}{\partial w_{11}} & \frac{\partial L_3}{\partial w_{11}} \end{pmatrix}. \tag{1.22}$$

The first element in Eq.(1.22) can be reexpressed by

$$\begin{aligned} \frac{\partial L_1}{\partial w_{11}} &= \frac{\partial L_1}{\partial y_1}\frac{\partial y_1}{\partial w_{11}} + \frac{\partial L_1}{\partial y_2}\frac{\partial y_2}{\partial w_{11}} + \frac{\partial L_1}{\partial y_3}\frac{\partial y_3}{\partial w_{11}} \\ &= \sum_{k=1}^{3} \frac{\partial L_1}{\partial y_k}\frac{\partial y_k}{\partial w_{11}} \end{aligned} \tag{1.23}$$

Now,

$$\frac{\partial y_1}{\partial w_{11}} = \frac{\partial}{\partial w_{11}} \sum_{k=1}^{2} x_k w_{k1} = x_1 \tag{1.24}$$

or, in general,

$$\frac{\partial y_i}{\partial w_{jk}} = \frac{\partial}{\partial w_{jk}} \sum_{n=1}^{2} x_n w_{ni} = x_j \delta_{ik}, \tag{1.25}$$

where δ_{ik} is Kronecker's delta. By using Eq.(1.25), we rewrite $\frac{\partial L_i}{\partial w_{jk}}$ as

$$\begin{aligned} \frac{\partial L_i}{\partial w_{jk}} &= \sum_{n=1}^{3} \frac{\partial L_i}{\partial y_n}\frac{\partial y_n}{\partial w_{jk}} \\ &= \sum_{n=1}^{3} \frac{\partial L_i}{\partial y_n} x_j \delta_{nk} \\ &= \frac{\partial L_i}{\partial y_k} x_j \end{aligned} \tag{1.26}$$

Therefore,

$$\begin{aligned} \frac{\partial \mathbf{L}}{\partial w_{jk}} &= \begin{pmatrix} \frac{\partial L_1}{\partial w_{jk}} & \frac{\partial L_2}{\partial w_{jk}} & \frac{\partial L_3}{\partial w_{jk}} \end{pmatrix} \\ &= \begin{pmatrix} \frac{\partial L_1}{\partial y_k} x_j & \frac{\partial L_2}{\partial y_k} x_j & \frac{\partial L_3}{\partial y_k} x_j \end{pmatrix} \\ &= x_j \begin{pmatrix} \frac{\partial L_1}{\partial y_k} & \frac{\partial L_2}{\partial y_k} & \frac{\partial L_3}{\partial y_k} \end{pmatrix} \\ &= x_j \frac{\partial \mathbf{L}}{\partial y_k} \end{aligned} \tag{1.27}$$

or

$$\left(\frac{\partial \mathbf{L}}{\partial \mathbf{W}}\right)_{jk} = x_j \frac{\partial \mathbf{L}}{\partial y_k}. \tag{1.28}$$

Since

$$\frac{\partial \mathbf{L}}{\partial \mathbf{Y}} = \begin{pmatrix} \frac{\partial \mathbf{L}}{\partial y_1} & \frac{\partial \mathbf{L}}{\partial y_2} & \frac{\partial \mathbf{L}}{\partial y_3} \end{pmatrix} \tag{1.29}$$

we obtain

$$\begin{aligned} \frac{\partial \mathbf{L}}{\partial \mathbf{W}} &= \mathbf{X}^T \frac{\partial \mathbf{L}}{\partial \mathbf{Y}} \\ &= \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \begin{pmatrix} \frac{\partial \mathbf{L}}{\partial y_1} & \frac{\partial \mathbf{L}}{\partial y_2} & \frac{\partial \mathbf{L}}{\partial y_3} \end{pmatrix} \\ &= \begin{pmatrix} x_1 \frac{\partial \mathbf{L}}{\partial y_1} & x_1 \frac{\partial \mathbf{L}}{\partial y_2} & x_1 \frac{\partial \mathbf{L}}{\partial y_3} \\ x_2 \frac{\partial \mathbf{L}}{\partial y_1} & x_2 \frac{\partial \mathbf{L}}{\partial y_2} & x_2 \frac{\partial \mathbf{L}}{\partial y_3} \end{pmatrix}. \end{aligned} \tag{1.30}$$

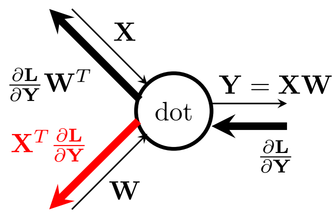

Figure 1.5: Compuation graph for the backpropagation of the dot product.

1.2.4 Partial derivative $\frac{\partial \mathbf{L}}{\partial \mathbf{B}}$

Let us first think about what $\mathbf{Y}=\mathbf{XW}+\mathbf{B}$ is. In this example, since $\mathbf{X}$ is a (1×3) matrix and $\mathbf{W}$ is a (2×3), $\mathbf{XW}$ becomes (1×3) matrix. Thus, in this example, we simply add a row vector $\mathbf{B}$ to $\mathbf{XW}$ to obtain $\mathbf{Y}$. There is no ambiguity in this example. However, if the input matrix $\mathbf{X}$ becomes (2×2), then the situation becomes not so trivial. In this case, $\mathbf{XW}$ is given by a (2×3) matrix. Thus, adding $\mathbf{B}$ to $\mathbf{XW}$ does not simply mean that a row vector $\mathbf{B}$ is added to $\mathbf{XW}$. It should be rewritten as

$$\mathbf{Y}=\mathbf{XW}+\begin{pmatrix}1\\1\end{pmatrix}\mathbf{B}=\mathbf{A}+\mathbf{DB}. \tag{1.31}$$

Here $\mathbf{A}\equiv\mathbf{XW}$ and $\mathbf{D}\equiv\begin{pmatrix}1\\1\end{pmatrix}$. Eq.(1.31) clearly shows that there is an additional dot product (or matrix multiplication) of $\begin{pmatrix}1\\1\end{pmatrix}$ and $\mathbf{B}$. The computing graph for Eq.(1.31) is depicted in Fig.1.6 Using the relation in

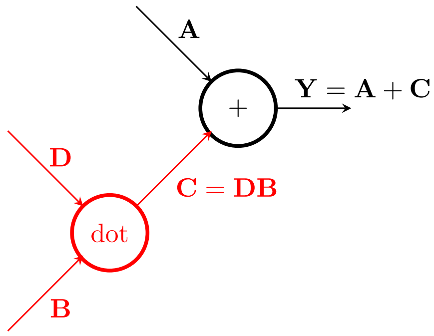

Figure 1.6: Compuation graph for the backpropagation of the dot product.

Figs.1.4 and 1.5 and addition graph, we can obtain the following computing graph. From the practical point

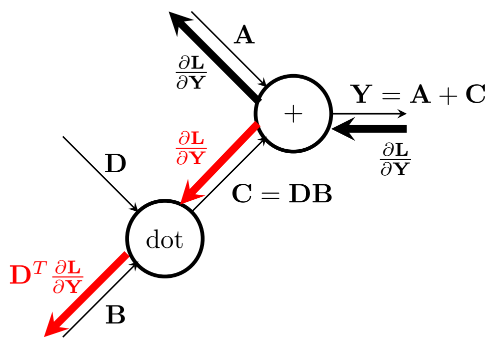

Figure 1.7: Compuation graph for a dot product.

of view, if we use python, then $\mathbf{D}^T\frac{\partial \mathbf{L}}{\partial \mathbf{Y}}$ can be expressed much simpler way. Note that $\mathbf{D}^T=\begin{pmatrix}1\\1\end{pmatrix}^T=\begin{pmatrix}1 & 1\end{pmatrix}$. Therefore, $\mathbf{D}^T\frac{\partial \mathbf{L}}{\partial \mathbf{Y}}$ is just the sum of each element of $\frac{\partial \mathbf{L}}{\partial \mathbf{Y}}$ along the `axis=0`.

1.2.5 Computing Graph with Matrix Input

Fig.1.8 shows how we can translate our analytic derivation into a computing graph. At the right of the graph, 1 is sent to the f node. Here the (local) partial derivative at f node is $\frac{\partial \mathbf{L}}{\partial \mathbf{Y}}$. The next node is the + node. So

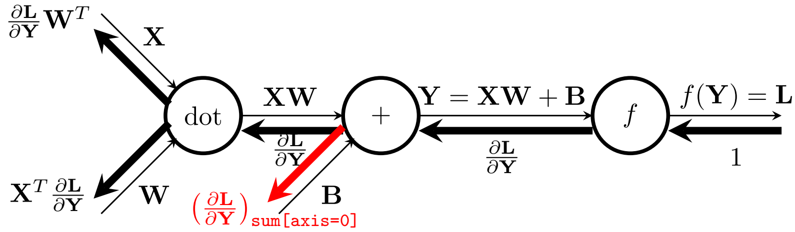

Figure 1.8: Back propagation when inputs $\mathbf{X}$ and $\mathbf{W}$ are matrices. Thus $\mathbf{B}$, $\mathbf{Y}$, and $\mathbf{L}$ are also matrices.

nothing happens when the $\frac{\partial \mathbf{L}}{\partial \mathbf{Y}}$ passes through the node. Now when it passes through the "dot" node, the $\frac{\partial \mathbf{L}}{\partial \mathbf{Y}}\mathbf{W}^T$ and $\mathbf{X}^T \frac{\partial \mathbf{L}}{\partial \mathbf{Y}}$ are assigned to each edge pointing to each input $\mathbf{X}$ and $\mathbf{W}$, respectively. Similarly, $\left(\frac{\partial \mathbf{L}}{\partial \mathbf{Y}}\right)_{\texttt{sum[axis=0]}}$ is assigned to the backward arrow to $\mathbf{B}$.