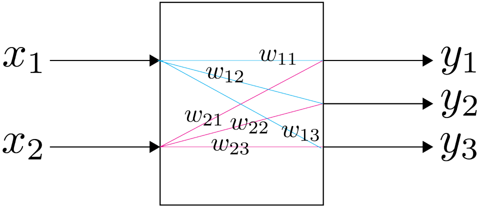
Linear layers

Computer Vision (CSCI 5520G)

Faisal Z. Qureshi

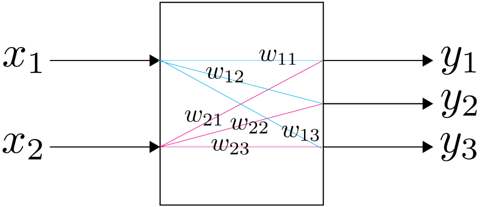
http://vclab.science.ontariotechu.ca

Linear Layer



$$y_1 = w_{11}x_1 + w_{21}x_2$$
$$y_2 = w_{12}x_1 + w_{22}x_2$$
$$y_3 = w_{13}x_1 + w_{23}x_2$$

Linear Layer



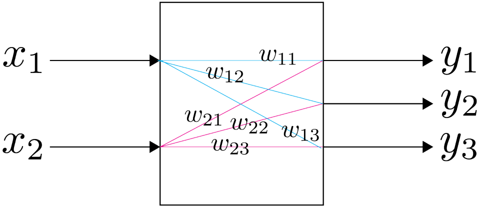
$$y_1 = w_{11}x_1 + w_{21}x_2$$

$$y_2 = w_{12}x_1 + w_{22}x_2$$

$$y_3 = w_{13}x_1 + w_{23}x_2$$

Re-writing in matrix form

Linear Layer

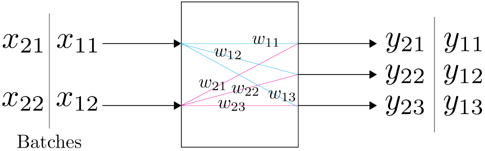


$$y_1 = w_{11}x_1 + w_{21}x_2$$
$$y_2 = w_{12}x_1 + w_{22}x_2$$
$$y_3 = w_{13}x_1 + w_{23}x_2$$

Re-writing in matrix form

$$\begin{pmatrix} y_1 & y_2 & y_3 \end{pmatrix} = \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{pmatrix}$$

Dealing with Batches



We can re-use the matrix form from before

$$\underbrace{\begin{pmatrix} y_1^{(1)} & y_2^{(1)} & y_3^{(1)} \\ y_1^{(2)} & y_2^{(2)} & y_3^{(2)} \end{pmatrix}}_{\mathbf{Y}} = \underbrace{\begin{pmatrix} x_1^{(1)} & x_2^{(1)} \\ x_1^{(2)} & x_2^{(2)} \end{pmatrix}}_{\mathbf{X}} \underbrace{\begin{pmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{pmatrix}}_{\mathbf{W}}$$

Forward Pass

The forward function for a linear layer is thus defined as

$$\mathbf{Y} = \mathbf{X}\mathbf{W}$$

where

- $\mathbf{X} \in \mathbb{R}^{B \times d_{\text{in}}}$ is the input
- $\mathbf{W} \in \mathbb{R}^{d_{\text{in}} \times d_{\text{out}}}$ is the weight matrix
- $\mathbf{Y} \in \mathbb{R}^{B \times d_{\text{out}}}$ is the output

Here B refers to the batch size.

Forward Pass

The forward function for a linear layer is thus defined as

$$\mathbf{Y} = \mathbf{X}\mathbf{W}$$

where

- $\mathbf{X} \in \mathbb{R}^{B \times d_{\mathsf{in}}}$ is the input
- $\mathbf{W} \in \mathbb{R}^{d_{\mathsf{in}} \times d_{\mathsf{out}}}$ is the weight matrix
- $\mathbf{Y} \in \mathbb{R}^{B \times d_{\mathsf{out}}}$ is the output

Here B refers to the batch size.

How would you handle the bias term?

Forward Pass

The forward function for a linear layer is thus defined as

$$\mathbf{Y} = \mathbf{X}\mathbf{W}$$

where

- $\mathbf{X} \in \mathbb{R}^{B \times d_{\text{in}}}$ is the input
- $\mathbf{W} \in \mathbb{R}^{d_{\text{in}} \times d_{\text{out}}}$ is the weight matrix
- $\mathbf{Y} \in \mathbb{R}^{B \times d_{\text{out}}}$ is the output

Here B refers to the batch size.

How would you handle the bias term? Append a 1 to each input in $\mathbf{X}$.

Backpropagation

To backpropagate, we need:

- Gradient w.r.t input: $\frac{\partial C}{\partial \mathbf{X}}$
- Gradient w.r.t weights: $\frac{\partial C}{\partial \mathbf{W}}$

We assume that we already have $\frac{\partial C}{\partial \mathbf{Y}}$. This was backpropagated by later layers. Here C denotes loss or cost that we need to minimize to "train" the network.

Backpropagation

To backpropagate, we need:

- Gradient w.r.t input: $\frac{\partial C}{\partial \mathbf{X}}$
- Gradient w.r.t weights: $\frac{\partial C}{\partial \mathbf{W}}$

We assume that we already have $\frac{\partial C}{\partial \mathbf{Y}}$. This was backpropagated by later layers. Here C denotes loss or cost that we need to minimize to "train" the network.

How to compute $\frac{\partial C}{\partial \mathbf{X}}$?

Application of chain-rule yields

$$\frac{\partial C}{\partial \mathbf{X}} = \frac{\partial C}{\partial \mathbf{Y}} \frac{\partial \mathbf{Y}}{\partial \mathbf{X}}$$

As have $\frac{\partial C}{\partial \mathbf{Y}}$ already. Lets figure out how to compute $\frac{\partial \mathbf{Y}}{\partial \mathbf{X}}$.

Structure of $\frac{\partial C}{\partial \mathbf{Y}}$

- C (loss or cost) is a scalar.
- $\frac{\partial C}{\partial \mathbf{Y}}$ as the same size as $\mathbf{Y}$, i.e. $(B \times d_{\mathsf{out}})$.

Structure of $\frac{\partial C}{\partial \mathbf{Y}}$

- C (loss or cost) is a scalar.
- $\frac{\partial C}{\partial \mathbf{Y}}$ as the same size as $\mathbf{Y}$, i.e. $(B \times d_{\text{out}})$.

For the **single input** case: $(x_1, x_2) \mapsto (y_1, y_2, y_3)$

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

Structure of $\frac{\partial C}{\partial \mathbf{Y}}$

- C (loss or cost) is a scalar.
- $\frac{\partial C}{\partial \mathbf{Y}}$ as the same size as $\mathbf{Y}$, i.e. $(B \times d_{\text{out}})$.

For the **batch input** case: $\begin{pmatrix} x_1^{(1)}, x_2^{(1)} \\ x_1^{(2)}, x_2^{(2)} \end{pmatrix} \mapsto \begin{pmatrix} y_1^{(1)}, y_2^{(1)}, y_3^{(1)} \\ y_1^{(2)}, y_2^{(2)}, y_3^{(2)} \end{pmatrix}$

$$\frac{\partial C}{\partial \mathbf{Y}} = \begin{pmatrix} \frac{\partial C}{\partial y_1^{(1)}} & \frac{\partial C}{\partial y_2^{(1)}} & \frac{\partial C}{\partial y_3^{(1)}} \\ \\ \frac{\partial C}{\partial y_1^{(2)}} & \frac{\partial C}{\partial y_2^{(2)}} & \frac{\partial C}{\partial y_3^{(2)}} \end{pmatrix}$$

Computing $\frac{\partial \mathbf{Y}}{\partial \mathbf{X}}$

Lets unpack $\frac{\partial \mathbf{Y}}{\partial \mathbf{X}}$. We begin by considering a **single input**

$$\begin{pmatrix} y_1 & y_2 & y_3 \end{pmatrix} = \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{pmatrix}$$

Expanding further

$$\begin{aligned} y_1 &= w_{11}x_1 + w_{21}x_2 \\ y_2 &= w_{12}x_1 + w_{22}x_2 \\ y_3 &= w_{13}x_1 + w_{23}x_2 \end{aligned}$$

This is a vector-valued function of two variables (x_1, x_2).

Computing $\frac{\partial \mathbf{Y}}{\partial \mathbf{X}}$

Setting up the Jacobian (the matrix of all first-order partial derivatives of a vector-valued function)

$$\begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} \\ \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} \\ \\ \frac{\partial y_3}{\partial x_1} & \frac{\partial y_3}{\partial x_2} \end{pmatrix} = \begin{pmatrix} w_{11} & w_{21} \\ \\ w_{12} & w_{22} \\ \\ w_{13} & w_{23} \end{pmatrix} = \mathbf{W}^T$$

Computing $\frac{\partial C}{\partial \mathbf{X}}$

Lets compute $\frac{\partial C}{\partial x_1}$ and $\frac{\partial C}{\partial x_2}$

Computing $\frac{\partial C}{\partial \mathbf{X}}$

Lets compute $\frac{\partial C}{\partial x_1}$ and $\frac{\partial C}{\partial x_2}$

$$\frac{\partial C}{\partial x_1} = \frac{\partial C}{\partial y_1}\frac{\partial y_1}{\partial x_1} + \frac{\partial C}{\partial y_2}\frac{\partial y_2}{\partial x_1} + \frac{\partial C}{\partial y_3}\frac{\partial y_3}{\partial x_1}$$

Computing $\frac{\partial C}{\partial \mathbf{X}}$

Lets compute $\frac{\partial C}{\partial x_1}$ and $\frac{\partial C}{\partial x_2}$

$$\frac{\partial C}{\partial x_1} = \frac{\partial C}{\partial y_1}\frac{\partial y_1}{\partial x_1} + \frac{\partial C}{\partial y_2}\frac{\partial y_2}{\partial x_1} + \frac{\partial C}{\partial y_3}\frac{\partial y_3}{\partial x_1}$$

Similarly

$$\frac{\partial C}{\partial x_2} = \frac{\partial C}{\partial y_1}\frac{\partial y_1}{\partial x_2} + \frac{\partial C}{\partial y_2}\frac{\partial y_2}{\partial x_2} + \frac{\partial C}{\partial y_3}\frac{\partial y_3}{\partial x_2}$$

Computing $\frac{\partial C}{\partial \mathbf{X}}$

Lets compute $\frac{\partial C}{\partial x_1}$ and $\frac{\partial C}{\partial x_2}$

$$\frac{\partial C}{\partial x_1} = \frac{\partial C}{\partial y_1}\frac{\partial y_1}{\partial x_1} + \frac{\partial C}{\partial y_2}\frac{\partial y_2}{\partial x_1} + \frac{\partial C}{\partial y_3}\frac{\partial y_3}{\partial x_1}$$

Similarly

$$\frac{\partial C}{\partial x_2} = \frac{\partial C}{\partial y_1}\frac{\partial y_1}{\partial x_2} + \frac{\partial C}{\partial y_2}\frac{\partial y_2}{\partial x_2} + \frac{\partial C}{\partial y_3}\frac{\partial y_3}{\partial x_2}$$

Re-writing in matrix form

$$\begin{pmatrix} \frac{\partial C}{\partial x_1} & \frac{\partial C}{\partial x_2} \end{pmatrix} = \begin{pmatrix} \frac{\partial C}{\partial y_1} & \frac{\partial C}{\partial y_2} & \frac{\partial C}{\partial y_3} \end{pmatrix} \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} \\ \frac{\partial y_3}{\partial x_1} & \frac{\partial y_3}{\partial x_2} \end{pmatrix}$$

Computing $\frac{\partial C}{\partial \mathbf{X}}$

Lets compute $\frac{\partial C}{\partial x_1}$ and $\frac{\partial C}{\partial x_2}$

$$\frac{\partial C}{\partial x_1} = \frac{\partial C}{\partial y_1}\frac{\partial y_1}{\partial x_1} + \frac{\partial C}{\partial y_2}\frac{\partial y_2}{\partial x_1} + \frac{\partial C}{\partial y_3}\frac{\partial y_3}{\partial x_1}$$

Similarly

$$\frac{\partial C}{\partial x_2} = \frac{\partial C}{\partial y_1}\frac{\partial y_1}{\partial x_2} + \frac{\partial C}{\partial y_2}\frac{\partial y_2}{\partial x_2} + \frac{\partial C}{\partial y_3}\frac{\partial y_3}{\partial x_2}$$

Re-writing in matrix form

$$\begin{pmatrix} \frac{\partial C}{\partial x_1} & \frac{\partial C}{\partial x_2} \end{pmatrix} = \begin{pmatrix} \frac{\partial C}{\partial y_1} & \frac{\partial C}{\partial y_2} & \frac{\partial C}{\partial y_3} \end{pmatrix} \underbrace{\begin{pmatrix} w_{11} & w_{21} \\ w_{12} & w_{22} \\ w_{13} & w_{23} \end{pmatrix}}_{\mathbf{W}^T}$$

Dealing with batch inputs

Lets consider **batch input**

$$\underbrace{\begin{pmatrix} y_1^{(1)} & y_2^{(1)} & y_3^{(1)} \\ y_1^{(2)} & y_2^{(2)} & y_3^{(2)} \end{pmatrix}}_{\mathbf{Y}} = \underbrace{\begin{pmatrix} x_1^{(1)} & x_2^{(1)} \\ x_1^{(2)} & x_2^{(2)} \end{pmatrix}}_{\mathbf{X}} \underbrace{\begin{pmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{pmatrix}}_{\mathbf{W}}$$

We write

$$\begin{pmatrix} \frac{\partial C}{\partial x_1^{(1)}} & \frac{\partial C}{\partial x_2^{(1)}} \\ \\ \frac{\partial C}{\partial x_1^{(2)}} & \frac{\partial C}{\partial x_2^{(2)}} \end{pmatrix} = \begin{pmatrix} \frac{\partial C}{\partial y_1^{(1)}} & \frac{\partial C}{\partial y_2^{(1)}} & \frac{\partial C}{\partial y_3^{(1)}} \\ \\ \frac{\partial C}{\partial y_1^{(2)}} & \frac{\partial C}{\partial y_2^{(2)}} & \frac{\partial C}{\partial y_3^{(2)}} \end{pmatrix} \mathbf{W}^T$$

Therefore

$$\frac{\partial C}{\partial \mathbf{X}} = \frac{\partial C}{\partial \mathbf{Y}} \mathbf{W}^T$$

Computing $\frac{\partial C}{\partial \mathbf{W}}$

Now let's turn our attention to computing $\frac{\partial C}{\partial \mathbf{W}}$.

Recall $\frac{\partial C}{\partial \mathbf{W}}$ has the same shape as $\mathbf{W}$

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

Computing $\frac{\partial C}{\partial \mathbf{W}}$

Applying chain-rule, we have

$$\frac{\partial C}{\partial \mathbf{W}} = \frac{\partial C}{\partial \mathbf{Y}} \frac{\partial \mathbf{Y}}{\partial \mathbf{W}}$$

Computing $\frac{\partial C}{\partial \mathbf{W}}$

Applying chain-rule, we have

$$\frac{\partial C}{\partial \mathbf{W}} = \underbrace{\frac{\partial C}{\partial \mathbf{Y}}}_{\checkmark} \frac{\partial \mathbf{Y}}{\partial \mathbf{W}}$$

Computing $\frac{\partial C}{\partial \mathbf{W}}$

Applying chain-rule, we have

$$\frac{\partial C}{\partial \mathbf{W}} = \underbrace{\frac{\partial C}{\partial \mathbf{Y}}}_{\checkmark} \underbrace{\frac{\partial \mathbf{Y}}{\partial \mathbf{W}}}_{?}$$

Computing $\frac{\partial C}{\partial w_{11}}, \cdots$

For **single input** case:

$y_1 = w_{11}x_1 + w_{21}x_2$, $y_2 = w_{12}x_1 + w_{22}x_2$, and $y_3 = w_{13}x_1 + w_{23}x_2$

Applying chain-rule, we have

$$
\begin{aligned}
\frac{\partial C}{\partial w_{11}} &= \frac{\partial C}{\partial y_1}\frac{\partial y_1}{\partial w_{11}} + \frac{\partial C}{\partial y_2}\frac{\partial y_2}{\partial w_{11}} + \frac{\partial C}{\partial y_3}\frac{\partial y_3}{\partial w_{11}} \\
&= \frac{\partial C}{\partial y_1} x_1
\end{aligned}
$$

Computing $\frac{\partial C}{\partial w_{11}}, \cdots$

For **single input** case:

$y_1 = w_{11}x_1 + w_{21}x_2$, $y_2 = w_{12}x_1 + w_{22}x_2$, and $y_3 = w_{13}x_1 + w_{23}x_2$

Applying chain-rule, we have

$$\begin{aligned}\frac{\partial C}{\partial w_{11}} &= \frac{\partial C}{\partial y_1}\frac{\partial y_1}{\partial w_{11}} + \frac{\partial C}{\partial y_2}\frac{\partial y_2}{\partial w_{11}} + \frac{\partial C}{\partial y_3}\frac{\partial y_3}{\partial w_{11}} \\ &= \frac{\partial C}{\partial y_1}x_1\end{aligned}$$

Simlarly

$\frac{\partial C}{\partial w_{12}} = \frac{\partial C}{\partial y_2}x_1$, $\frac{\partial C}{\partial w_{13}} = \frac{\partial C}{\partial y_3}x_1$,

$\frac{\partial C}{\partial w_{21}} = \frac{\partial C}{\partial y_1}x_2$, $\frac{\partial C}{\partial w_{22}} = \frac{\partial C}{\partial y_2}x_2$, and $\frac{\partial C}{\partial w_{23}} = \frac{\partial C}{\partial y_3}x_2$

Computing $\frac{\partial C}{\partial \mathbf{W}}$

Putting it all together for **single input** case

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

$$= \begin{pmatrix} \frac{\partial C}{\partial y_1} x_1 & \frac{\partial C}{\partial y_2} x_1 & \frac{\partial C}{\partial y_3} x_1 \\ \\ \frac{\partial C}{\partial y_1} x_2 & \frac{\partial C}{\partial y_2} x_2 & \frac{\partial C}{\partial y_3} x_2 \end{pmatrix}$$

$$= \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \begin{pmatrix} \frac{\partial C}{\partial y_1} & \frac{\partial C}{\partial y_2} & \frac{\partial C}{\partial y_3} \end{pmatrix}$$

Computing $\frac{\partial C}{\partial \mathbf{W}}$

Putting it all together for **single input** case

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

$$= \begin{pmatrix} \frac{\partial C}{\partial y_1} x_1 & \frac{\partial C}{\partial y_2} x_1 & \frac{\partial C}{\partial y_3} x_1 \\ \\ \frac{\partial C}{\partial y_1} x_2 & \frac{\partial C}{\partial y_2} x_2 & \frac{\partial C}{\partial y_3} x_2 \end{pmatrix}$$

$$= \underbrace{\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}}_{\mathbf{X}^T} \underbrace{\begin{pmatrix} \frac{\partial C}{\partial y_1} & \frac{\partial C}{\partial y_2} & \frac{\partial C}{\partial y_3} \end{pmatrix}}_{\frac{\partial C}{\partial \mathbf{Y}}}$$

Computing $\frac{\partial C}{\partial \mathbf{W}}$

For **batch input** case

$$\frac{\partial C}{\partial \mathbf{W}} = \underbrace{\begin{pmatrix} x_1^{(1)} & x_1^{(2)} \\ x_2^{(1)} & x_2^{(2)} \end{pmatrix}}_{\mathbf{X}^T} \begin{pmatrix} \frac{\partial C}{\partial y_1^{(1)}} & \frac{\partial C}{\partial y_2^{(1)}} & \frac{\partial C}{\partial y_3^{(1)}} \\ \frac{\partial C}{\partial y_1^{(2)}} & \frac{\partial C}{\partial y_2^{(2)}} & \frac{\partial C}{\partial y_3^{(2)}} \end{pmatrix}$$

Or more generally, we write

$$\frac{\partial C}{\partial \mathbf{W}} = \mathbf{X}^T \frac{\partial C}{\partial \mathbf{Y}}$$

Observations

- By using the *backpropagated* signal $\frac{\partial C}{\partial \mathbf{Y}}$, we are able to compute $\frac{\partial C}{\partial \mathbf{X}}$ and $\frac{\partial C}{\partial \mathbf{W}}$.
  - We still need to compute $\frac{\partial \mathbf{Y}}{\partial \mathbf{X}}$ and $\frac{\partial \mathbf{Y}}{\partial \mathbf{W}}$
- $\frac{\partial C}{\partial \mathbf{X}}$ is *backpropagated*.
- $\frac{\partial C}{\partial \mathbf{W}}$ is used to update weights $\mathbf{W}$ of this layer.

Useful Properties of Linear Layers

- **Global context**: Every output depends on every input.
- **Information mixing**: Inputs are combined to produce outputs.
- **Common usage**: Often the last layer in many networks.

Activation Functions

This description **ignores activation functions**, which are usually applied after linear layers.

Copyright and License