Neural network

A gentle introduction: Neural network for supervised learning

Khiem Nguyen

Email	khiem.nguyen@glasgow.ac.uk
MS Teams	khiem.nguyen@glasgow.ac.uk
Whatsapp	+44 7729 532071 (Emergency only)

May 18, 2025

University of Glasgow

Chain rule in differentiation

- We review the chain rule. Assume that

$$f = f(a)$$
$$a = a(\theta)$$

- Then, the chain rule says

$$\frac{\partial f}{\partial \theta} = \frac{\partial f}{\partial a}\frac{\partial a}{\partial \theta}$$

Reasoning for back propagation

- Let us consider one neural network with single hidden layer

$$\mathbf{z}^{[1]} = \mathbf{a}^{[2]}\mathbf{W}^{[1]T} + \mathbf{b}^{[1]}$$
$$\mathbf{a}^{[1]} = g(\mathbf{z}^{[1]})$$
$$\mathbf{z}^{[2]} = \mathbf{a}^{[1]}\mathbf{W}^{[2]T} + \mathbf{b}^{[2]}$$
$$\mathbf{a}^{[2]} = g^{\mathbf{z}^{[2]}}$$

- We have the following series of mappings

$$\mathbf{a}^{[0]} \mapsto \text{Linear Mapping} \mapsto \mathbf{z}^{[1]} \mapsto \text{Activation Function} \mapsto \mathbf{a}^{[1]}$$
$$\mapsto \text{Linear Mapping} \mapsto \mathbf{z}^{[2]} \mapsto \text{Activation Function} \mapsto \mathbf{a}^{[2]} \mapsto \mathcal{L}(\mathbf{a}^{[2]}; y)$$

- Our goal: ompute the gradient of the loss function w.r.t. the learnable/model paremeters

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{[1]}} \qquad \frac{\partial \mathcal{L}}{\partial \mathbf{b}^{[1]}} \qquad \frac{\partial \mathcal{L}}{\partial \mathbf{W}^{[2]}} \qquad \frac{\partial \mathcal{L}}{\partial \mathbf{b}^{[2]}}$$

Computation of gradient of loss function

- Let us start with

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{[1]}} = \frac{\partial \mathcal{L}}{\partial \mathbf{z}^{[1]}} \frac{\partial \mathbf{z}^{[1]}}{\partial \mathbf{W}^{[1]}} = \sum_k \frac{\partial \mathcal{L}}{\partial z_k^{[1]}} \frac{\partial z_k^{[1]}}{\partial \mathbf{W}^{[1]}}$$

  The true formulation is more complex due to:

  - $\mathbf{W}^{[1]}$ is a matrix, so the derivative is done w.r.t. each component $W_{ij}^{[1]}$
  - $\mathbf{z}^{[1]}$ is a vector, so the derivative is done w.r.t. each component $z_k^{[1]}$
  - The chain rule in differentiation is actually carried out in the above formula.

$$\frac{\partial \mathcal{L}}{\partial W_{ij}^{[1]}} = \frac{\partial \mathcal{L}}{\partial \mathbf{z}^{[1]}} \frac{\partial \mathbf{z}^{[1]}}{\partial \mathbf{W}^{[1]}} = \sum_k \frac{\partial \mathcal{L}}{\partial z_k^{[1]}} \frac{\partial z_k^{[1]}}{\partial W_{ij}^{[1]}}$$

  for all i and j running in the set of parameters in $\mathbf{W}^{[1]}$

- We shall focus on the "reduced" writing notation

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{[1]}} = \underbrace{\frac{\partial \mathcal{L}}{\partial \mathbf{z}^{[1]}}}_{\text{NOT EASY}} \underbrace{\frac{\partial \mathbf{z}^{[1]}}{\partial \mathbf{W}^{[1]}}}_{\text{easy}}$$

Computation of gradient of loss function

- We shall try to compute the derivatives $\partial\mathcal{L}/\partial\mathbf{W}^{[1]}$ by chain rule

$$\frac{\partial\mathcal{L}}{\partial\mathbf{W}^{[1]}} = \underbrace{\frac{\partial\mathcal{L}}{\partial\mathbf{z}^{[1]}}}_{\text{NOT EASY}} \underbrace{\frac{\partial\mathbf{z}^{[1]}}{\partial\mathbf{W}^{[1]}}}_{\text{easy}} \qquad \frac{\partial\mathcal{L}}{\partial\mathbf{a}^{[1]}} = \underbrace{\frac{\partial\mathcal{L}}{\partial\mathbf{z}^{[2]}}}_{\text{NOT EASY}} \underbrace{\frac{\partial\mathbf{z}^{[2]}}{\partial\mathbf{a}^{[1]}}}_{\text{easy}}$$

$$\frac{\partial\mathcal{L}}{\partial\mathbf{z}^{[1]}} = \underbrace{\frac{\partial\mathcal{L}}{\partial\mathbf{a}^{[1]}}}_{\text{NOT EASY}} \underbrace{\frac{\partial\mathbf{a}^{[1]}}{\partial\mathbf{z}^{[1]}}}_{\text{easy}} \qquad \frac{\partial\mathcal{L}}{\partial\mathbf{z}^{[2]}} = \underbrace{\frac{\partial\mathcal{L}}{\partial\mathbf{a}^{[2]}}}_{\text{FINALLY EASY}} \underbrace{\frac{\partial\mathbf{a}^{[2]}}{\partial\mathbf{z}^{[2]}}}_{\text{easy}}$$

- In summary, we compute the derivative in the order

$$\frac{\partial\mathcal{L}}{\partial\mathbf{a}^{[2]}} \longrightarrow \frac{\partial\mathcal{L}}{\partial\mathbf{z}^{[2]}} \text{ using } \underbrace{\frac{\partial\mathbf{a}^{[2]}}{\partial\mathbf{z}^{[2]}}}_{\text{easy}} \longrightarrow \frac{\partial\mathcal{L}}{\partial\mathbf{a}^{[1]}} \text{ using } \underbrace{\frac{\partial\mathbf{z}^{[2]}}{\partial\mathbf{a}^{[1]}}}_{\text{easy}} \longrightarrow \frac{\partial\mathcal{L}}{\partial\mathbf{z}^{[1]}} \text{ using } \underbrace{\frac{\partial\mathbf{a}^{[1]}}{\partial\mathbf{z}^{[1]}}}_{\text{easy}}$$

$$\to \frac{\partial\mathcal{L}}{\partial\mathbf{W}^{[1]}} \text{ using } \underbrace{\frac{\partial\mathbf{z}^{[1]}}{\partial\mathbf{W}^{[1]}}}_{\text{easy}}$$

Computation of gradient of loss function

- Similarly, replacing $\dfrac{\partial \mathbf{z}^{[1]}}{\partial \mathbf{W}^{[1]}}$ with $\dfrac{\partial \mathbf{z}^{[1]}}{\partial \mathbf{b}^{[1]}}$ $\partial \mathcal{L}/\partial \mathbf{b}^{[1]}$, we arrive at

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}^{[1]}} = \underbrace{\frac{\partial \mathcal{L}}{\partial \mathbf{z}^{[1]}}}_{\text{NOT EASY}} \underbrace{\frac{\partial \mathbf{z}^{[1]}}{\partial \mathbf{b}^{[1]}}}_{\text{easy}} \qquad \frac{\partial \mathcal{L}}{\partial \mathbf{a}^{[1]}} = \underbrace{\frac{\partial \mathcal{L}}{\partial \mathbf{z}^{[2]}}}_{\text{NOT EASY}} \underbrace{\frac{\partial \mathbf{z}^{[2]}}{\partial \mathbf{a}^{[1]}}}_{\text{easy}}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{z}^{[1]}} = \underbrace{\frac{\partial \mathcal{L}}{\partial \mathbf{a}^{[1]}}}_{\text{NOT EASY}} \underbrace{\frac{\partial \mathbf{a}^{[1]}}{\partial \mathbf{z}^{[1]}}}_{\text{easy}} \qquad \frac{\partial \mathcal{L}}{\partial \mathbf{z}^{[2]}} = \underbrace{\frac{\partial \mathcal{L}}{\partial \mathbf{a}^{[2]}}}_{\text{FINALLY EASY}} \underbrace{\frac{\partial \mathbf{a}^{[2]}}{\partial \mathbf{z}^{[1]}}}_{\text{easy}}$$

- In summary, we compute the derivative in the order

$$\frac{\partial \mathcal{L}}{\partial \mathbf{a}^{[2]}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{z}^{[2]}} \text{ using } \underbrace{\frac{\partial \mathbf{a}^{[2]}}{\partial \mathbf{z}^{[2]}}}_{\text{easy}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{a}^{[1]}} \text{ using } \underbrace{\frac{\partial \mathbf{z}^{[2]}}{\partial \mathbf{a}^{[1]}}}_{\text{easy}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{z}^{[1]}} \text{ using } \underbrace{\frac{\partial \mathbf{a}^{[1]}}{\partial \mathbf{z}^{[1]}}}_{\text{easy}}$$

$$\longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{b}^{[1]}} \text{ using } \underbrace{\frac{\partial \mathbf{z}^{[1]}}{\partial \mathbf{b}^{[1]}}}_{\text{easy}}$$

Computation of gradient of loss function

- Let us now compute the derivative $\partial \mathcal{L}/\partial \mathbf{W}^{[2]}$ by chain rule

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{[2]}} = \underbrace{\frac{\partial \mathcal{L}}{\partial \mathbf{z}^{[2]}}}_{\text{NOT EASY}} \underbrace{\frac{\partial \mathbf{z}^{[2]}}{\partial \mathbf{W}^{[2]}}}_{\text{easy}}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{z}^{[2]}} = \underbrace{\frac{\partial \mathcal{L}}{\partial \mathbf{a}^{[2]}}}_{\text{ALSO EASY}} \underbrace{\frac{\partial \mathbf{a}^{[2]}}{\partial \mathbf{z}^{[2]}}}_{\text{easy}}$$

- In summary, we compute the derive in the order

$$\frac{\partial \mathcal{L}}{\partial \mathbf{a}^{[2]}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{z}^{[2]}} \text{ using } \underbrace{\frac{\partial \mathbf{a}^{[2]}}{\partial \mathbf{z}^{[2]}}}_{\text{easy}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{W}^{[2]}} \text{ using } \underbrace{\frac{\partial \mathbf{z}^{[2]}}{\partial \mathbf{W}^{[2]}}}_{\text{easy}}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{a}^{[2]}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{z}^{[2]}} \text{ using } \underbrace{\frac{\partial \mathbf{a}^{[2]}}{\partial \mathbf{z}^{[2]}}}_{\text{easy}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{b}^{[2]}} \text{ using } \underbrace{\frac{\partial \mathbf{z}^{[2]}}{\partial \mathbf{b}^{[2]}}}_{\text{easy}}$$

Computation of gradient of loss function

Back propagation process

$$\frac{\partial \mathcal{L}}{\partial \mathbf{a}^{[2]}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{z}^{[2]}} \text{ using } \underbrace{\frac{\partial \mathbf{a}^{[2]}}{\partial \mathbf{z}^{[2]}}}_{\text{easy}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{a}^{[1]}} \text{ using } \underbrace{\frac{\partial \mathbf{z}^{[2]}}{\partial \mathbf{a}^{[1]}}}_{\text{easy}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{z}^{[1]}} \text{ using } \underbrace{\frac{\partial \mathbf{a}^{[1]}}{\partial \mathbf{z}^{[1]}}}_{\text{easy}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{W}^{[1]}} \text{ using } \underbrace{\frac{\partial \mathbf{z}^{[1]}}{\partial \mathbf{W}^{[1]}}}_{\text{easy}}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{a}^{[2]}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{z}^{[2]}} \text{ using } \underbrace{\frac{\partial \mathbf{a}^{[2]}}{\partial \mathbf{z}^{[2]}}}_{\text{easy}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{a}^{[1]}} \text{ using } \underbrace{\frac{\partial \mathbf{z}^{[2]}}{\partial \mathbf{a}^{[1]}}}_{\text{easy}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{z}^{[1]}} \text{ using } \underbrace{\frac{\partial \mathbf{a}^{[1]}}{\partial \mathbf{z}^{[1]}}}_{\text{easy}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{b}^{[1]}} \text{ using } \underbrace{\frac{\partial \mathbf{z}^{[1]}}{\partial \mathbf{b}^{[1]}}}_{\text{easy}}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{a}^{[2]}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{z}^{[2]}} \text{ using } \underbrace{\frac{\partial \mathbf{a}^{[2]}}{\partial \mathbf{z}^{[2]}}}_{\text{easy}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{W}^{[2]}} \text{ using } \underbrace{\frac{\partial \mathbf{z}^{[2]}}{\partial \mathbf{W}^{[2]}}}_{\text{easy}}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{a}^{[2]}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{z}^{[2]}} \text{ using } \underbrace{\frac{\partial \mathbf{a}^{[2]}}{\partial \mathbf{z}^{[2]}}}_{\text{easy}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{b}^{[2]}} \text{ using } \underbrace{\frac{\partial \mathbf{z}^{[2]}}{\partial \mathbf{b}^{[2]}}}_{\text{easy}}$$

Back propagation process

- Repeat the formulation above

$$\frac{\partial \mathcal{L}}{\partial \mathbf{a}^{[2]}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{z}^{[2]}} \text{ using } \underbrace{\frac{\partial \mathbf{a}^{[2]}}{\partial \mathbf{z}^{[2]}}}_{\text{easy}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{a}^{[1]}} \text{ using } \underbrace{\frac{\partial \mathbf{z}^{[2]}}{\partial \mathbf{a}^{[1]}}}_{\text{easy}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{z}^{[1]}} \text{ using } \underbrace{\frac{\partial \mathbf{a}^{[1]}}{\partial \mathbf{z}^{[1]}}}_{\text{easy}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \Theta^{[1]}} \text{ using } \underbrace{\frac{\partial \mathbf{z}^{[1]}}{\partial \Theta^{[1]}}}_{\text{easy}}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{a}^{[2]}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{z}^{[2]}} \text{ using } \underbrace{\frac{\partial \mathbf{a}^{[2]}}{\partial \mathbf{z}^{[2]}}}_{\text{easy}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \Theta^{[2]}} \text{ using } \underbrace{\frac{\partial \mathbf{z}^{[2]}}{\partial \Theta^{[2]}}}_{\text{easy}}$$

where Θ can play the role of $\mathbf{W}$ or $\mathbf{b}$

- To compute the derivative $\frac{\partial \mathcal{L}}{\Theta^{[1]}}$, we need to go through the exactly same process for computing $\frac{\partial \mathcal{L}}{\Theta^{[2]}}$ but then continue one layer more.

Back propagation process

➪ **Repeat the formulation above**

$$\frac{\partial \mathcal{L}}{\partial \mathbf{a}^{[2]}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{z}^{[2]}} \text{ using } \underbrace{\frac{\partial \mathbf{a}^{[2]}}{\partial \mathbf{z}^{[2]}}}_{\text{easy}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{a}^{[1]}} \text{ using } \underbrace{\frac{\partial \mathbf{z}^{[2]}}{\partial \mathbf{a}^{[1]}}}_{\text{easy}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{z}^{[1]}} \text{ using } \underbrace{\frac{\partial \mathbf{a}^{[1]}}{\partial \mathbf{z}^{[1]}}}_{\text{easy}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \Theta^{[1]}} \text{ using } \underbrace{\frac{\partial \mathbf{z}^{[1]}}{\partial \Theta^{[1]}}}_{\text{easy}}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{a}^{[2]}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{z}^{[2]}} \text{ using } \underbrace{\frac{\partial \mathbf{a}^{[2]}}{\partial \mathbf{z}^{[2]}}}_{\text{easy}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \Theta^{[2]}} \text{ using } \underbrace{\frac{\partial \mathbf{z}^{[2]}}{\partial \Theta^{[2]}}}_{\text{easy}}$$

➪ **Backpropagation process**: Let us combinet two equations in one as follows

$$\frac{\partial \mathcal{L}}{\partial \mathbf{a}^{[2]}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{z}^{[2]}} \text{ using } \underbrace{\frac{\partial \mathbf{a}^{[2]}}{\partial \mathbf{z}^{[2]}}}_{\text{easy}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{a}^{[1]}} \text{ using } \underbrace{\frac{\partial \mathbf{z}^{[2]}}{\partial \mathbf{a}^{[1]}}}_{\text{easy}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{z}^{[1]}} \text{ using } \underbrace{\frac{\partial \mathbf{a}^{[1]}}{\partial \mathbf{z}^{[1]}}}_{\text{easy}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \Theta^{[1]}} \text{ using } \underbrace{\frac{\partial \mathbf{z}^{[1]}}{\partial \Theta^{[1]}}}_{\text{easy}}$$

$$\Downarrow \quad \Downarrow \quad \Downarrow \quad \Downarrow \quad \Downarrow$$

$$\frac{\partial \mathcal{L}}{\partial \Theta^{[2]}} \text{ using } \frac{\partial \mathbf{z}^{[2]}}{\partial \Theta^{[2]}}$$

Backpropagation process

- To compute the derivative $\frac{\partial \mathcal{L}}{\partial \Theta^{[1]}}$, we need to go through the exactly same process for computing $\frac{\partial \mathcal{L}}{\partial \Theta^{[2]}}$ but then continue one layer more.
- What to expect?
  To compute the derivative $\frac{\partial \mathcal{L}}{\partial \Theta^{[l-1]}}$, we need to go through the exactly same process for computing $\frac{\partial \mathcal{L}}{\partial \Theta^{[l]}}$ but then continue one layer more.

Backpropagaton for L-layer network

Let us repeat the backpropagation for a shallow neural network

$$\frac{\partial \mathcal{L}}{\partial \mathbf{a}^{[2]}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{z}^{[2]}} \text{ using } \underbrace{\frac{\partial \mathbf{a}^{[2]}}{\partial \mathbf{z}^{[2]}}}_{\text{easy}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{a}^{[1]}} \text{ using } \underbrace{\frac{\partial \mathbf{z}^{[2]}}{\partial \mathbf{a}^{[1]}}}_{\text{easy}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{z}^{[1]}} \text{ using } \underbrace{\frac{\partial \mathbf{a}^{[1]}}{\partial \mathbf{z}^{[1]}}}_{\text{easy}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \Theta^{[1]}} \text{ using } \underbrace{\frac{\partial \mathbf{z}^{[1]}}{\partial \Theta^{[1]}}}_{\text{easy}}$$

$$\Downarrow \quad \Downarrow \quad \Downarrow \quad \Downarrow \quad \Downarrow$$

$$\frac{\partial \mathcal{L}}{\partial \Theta^{[2]}} \text{ using } \frac{\partial \mathbf{z}^{[2]}}{\partial \Theta^{[2]}}$$

Backpropagaton for L-layer network

By computing carefully for 4-layer network, we arrive at

$$\frac{\partial \mathcal{L}}{\partial \mathbf{a}^{[4]}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{z}^{[4]}} \text{ using } \underbrace{\frac{\partial \mathbf{a}^{[4]}}{\partial \mathbf{z}^{[4]}}}_{\text{easy}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{a}^{[3]}} \text{ using } \underbrace{\frac{\partial \mathbf{z}^{[4]}}{\partial \mathbf{a}^{[3]}}}_{\text{easy}} \qquad \frac{\partial \mathcal{L}}{\partial \mathbf{z}^{[4]}} \text{ using } \underbrace{\frac{\partial \mathbf{z}^{[4]}}{\partial \Theta^{[4]}}}_{\text{easy}} \Longrightarrow \frac{\partial \mathcal{L}}{\partial \Theta^{[4]}}$$

$$\longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{z}^{[3]}} \text{ using } \underbrace{\frac{\partial \mathbf{a}^{[3]}}{\partial \mathbf{z}^{[3]}}}_{\text{easy}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{a}^{[2]}} \text{ using } \underbrace{\frac{\partial \mathbf{z}^{[3]}}{\partial \mathbf{a}^{[2]}}}_{\text{easy}} \qquad \frac{\partial \mathcal{L}}{\partial \mathbf{z}^{[3]}} \text{ using } \underbrace{\frac{\partial \mathbf{z}^{[3]}}{\partial \Theta^{[3]}}}_{\text{easy}} \Longrightarrow \frac{\partial \mathcal{L}}{\partial \Theta^{[3]}}$$

$$\longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{z}^{[2]}} \text{ using } \underbrace{\frac{\partial \mathbf{a}^{[2]}}{\partial \mathbf{z}^{[2]}}}_{\text{easy}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{a}^{[1]}} \text{ using } \underbrace{\frac{\partial \mathbf{z}^{[2]}}{\partial \mathbf{a}^{[1]}}}_{\text{easy}} \qquad \frac{\partial \mathcal{L}}{\partial \mathbf{z}^{[2]}} \text{ using } \underbrace{\frac{\partial \mathbf{z}^{[2]}}{\partial \Theta^{[2]}}}_{\text{easy}} \Longrightarrow \frac{\partial \mathcal{L}}{\partial \Theta^{[2]}}$$

$$\longrightarrow \frac{\partial \mathcal{L}}{\partial \mathbf{z}^{[1]}} \text{ using } \underbrace{\frac{\partial \mathbf{a}^{[1]}}{\partial \mathbf{z}^{[1]}}}_{\text{easy}} \longrightarrow \frac{\partial \mathcal{L}}{\partial \Theta^{[1]}} \text{ using } \underbrace{\frac{\partial \mathbf{z}^{[1]}}{\partial \Theta^{[1]}}}_{\text{easy}} \qquad \frac{\partial \mathcal{L}}{\partial \mathbf{z}^{[1]}} \text{ using } \underbrace{\frac{\partial \mathbf{z}^{[1]}}{\partial \Theta^{[1]}}}_{\text{easy}} \Longrightarrow \frac{\partial \mathcal{L}}{\partial \Theta^{[1]}}$$