Explanatory Notes for 6.390

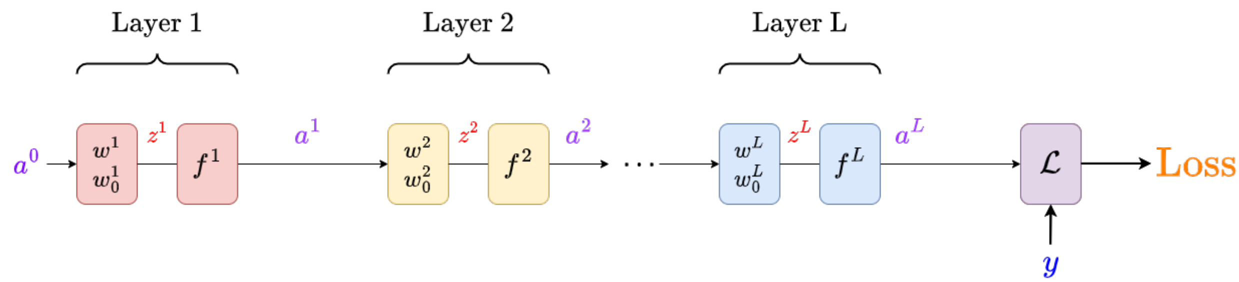
Shaunticlair Ruiz (Current TA)

Fall 2022

Many layers: Doing back-propagation

Now, we'll consider the case of many possible layers.

To make it more readable, we'll use boxes instead of circles for units.



This may look intimidating, but we already have all the tools we need to handle this problem.

Our goal is to get a **gradient** for each of our **weight** vectors w^ℓ, so we can do gradient descent and **improve** our model.

According to our above analysis in Concept 9, we need only a few steps to get all of our gradients.

Concept 1

In order to do **back-propagation**, we have to build up our **chain rule** for each weight gradient.

- We start our chain rule with one term shared by every gradient:

$$\overbrace{\frac{\partial \mathcal{L}}{\partial a^L}}^{\text{Loss unit}}$$

Then, we follow these two steps until we run out of layers:

- We're at layer ℓ. We want to get the **weight gradient** for this layer. We get this by **multiplying** our chain rule by

$$\overbrace{\frac{\partial a^\ell}{\partial z^\ell}}^{\text{Within layer}} \cdot \overbrace{\frac{\partial z^\ell}{\partial w^\ell}}^{\text{Get weight grad}}$$

  We exclude this term for any other gradients we want.

- If we aren't at layer 1, there's a previous layer we want to get the weight for. We reach layer $\ell - 1$ by multiplying our chain rule by

$$\overbrace{\frac{\partial a^\ell}{\partial z^\ell}}^{\text{Within layer}} \cdot \overbrace{\frac{\partial z^\ell}{\partial a^{\ell-1}}}^{\text{Link layers}}$$

Once we reach layer 1, we have **every single** weight vector we need! Repeat the process for w_0 gradients and then do **gradient descent**.

Let's get an idea of what this looks like in general:

$$\frac{\partial \mathcal{L}}{\partial w^\ell} = \overbrace{\left(\frac{\partial \mathcal{L}}{\partial a^L}\right)}^{\text{Loss unit}} \cdot \overbrace{\left(\frac{\partial a^L}{\partial z^L} \cdot \frac{\partial z^L}{\partial a^{L-1}}\right)}^{\text{Layer L}} \cdot \overbrace{\left(\frac{\partial a^{L-1}}{\partial z^{L-1}} \cdot \frac{\partial z^{L-1}}{\partial a^{L-2}}\right)}^{\text{Layer L}-1} \cdot \left(\cdots\right) \cdot \overbrace{\left(\frac{\partial a^\ell}{\partial z^\ell} \cdot \frac{\partial z^\ell}{\partial w^\ell}\right)}^{\text{Layer }\ell} \tag{1}$$

That's pretty ugly. If we need to hide the complexity, we can:

Notation 2

If you need to do so for **ease**, you can **compress** your **derivatives**. For example, if we want to only have the last weight term **separate**, we can do:

$$\frac{\partial \mathcal{L}}{\partial w^\ell} = \overbrace{\frac{\partial \mathcal{L}}{\partial z^\ell}}^{\text{Other}} \cdot \overbrace{\frac{\partial z^\ell}{\partial w^\ell}}^{\text{Weight term}} \tag{2}$$

But we should also explore what each of these terms *are*.