Training a CBOW Model: Backpropagation and Gradient Descent

- **Backpropagation**: calculate partial derivatives of cost with respect to weights and biases.

When computing the back-prop in this model, you need to compute the following:

$\frac{\partial J_{batch}}{\partial \mathbf{W}_1}, \frac{\partial J_{batch}}{\partial \mathbf{W}_2}, \frac{\partial J_{batch}}{\partial \mathbf{b}_1}, \frac{\partial J_{batch}}{\partial \mathbf{b}_2}$

- **Gradient descent:** update weights and biases

Now to update the weights you can iterate as follows:

$$\mathbf{W_1} := \mathbf{W_1} - \alpha \frac{\partial J_{\text{batch}}}{\partial \mathbf{W_1}}$$
$$\mathbf{W_2} := \mathbf{W}_2 - \alpha \frac{\partial J_{\text{batch}}}{\partial \mathbf{W_2}}$$
$$\mathbf{b}_1 := \mathbf{b}_1 - \alpha \frac{\partial J_{\text{batch}}}{\partial \mathbf{b_1}}$$
$$\mathbf{b_2} := \mathbf{b}_2 - \alpha \frac{\partial J_{\text{batch}}}{\partial \mathbf{b}_2}$$

A smaller alpha allows for more gradual updates to the weights and biases, whereas a larger number allows for a faster update of the weights. If α is too large, you might not learn anything, if it is too small, your model will take forever to train.