Architecture for the CBOW Model
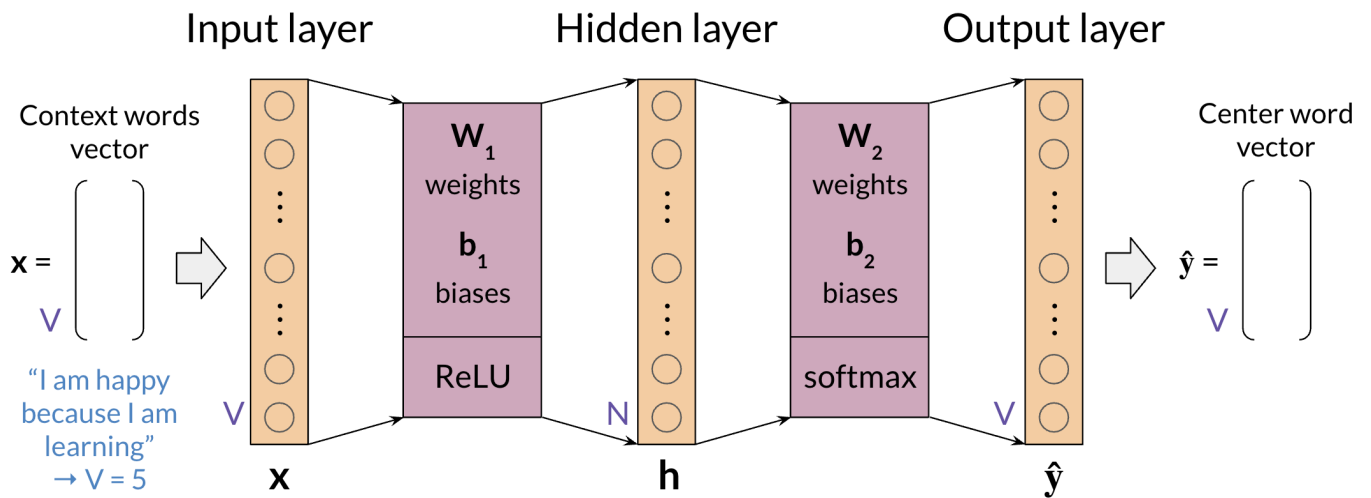
The architecture for the CBOW model could be described as follows

You have an input, *X*, which is the average of all context vectors. You then multiply it by W_1 and add b_1. The result goes through a ReLU function to give you your hidden layer. That layer is then multiplied by W_2 and you add b_2. The result goes through a softmax which gives you a distribution over *V*, vocabulary words. You pick the vocabulary word that corresponds to the arg-max of the output.