5. (16 points) Suppose you want to embed bigrams instead of words. Recall the skip-gram model:

$$P(\text{context} = y \mid \text{word} = x) = \frac{\exp(\mathbf{v}_x^\top \mathbf{c}_y)}{\sum_{y' \in \text{vocab}} \exp(\mathbf{v}_x^\top \mathbf{c}_{y'})}$$

a. (3 points) In your first attempt, you are going to learn a bigram vector for every bigram. That is, for a sentence like *the cat saw*, you will form word-context pairs (b(the cat), u(saw)), (b(cat saw), u(the)), where b and u denote bigram and unigram respectively (for maximal clarity). **Note that contexts are still unigrams.**

With **v** and **c** denoting the word and context vectors respectively, write out the model for skip-gram with word bigrams and context unigrams: $P(\text{context} = y \mid \text{word} = b)$ (b denotes a bigram).

V_b will be vector size of all bigram "words"

$$\boxed{P(\text{context} = y \mid \text{word} = b) = \frac{\exp(V_b^T C_y)}{\sum_{y' \in \text{vocab}} \exp(V_b^T C_{y'})}}$$

b. (2 points) What is the big-O runtime of computing the probability for a single bigram-context pair? Express this answer in terms of $|V|$, the vocabulary size, $|B|$, the number of bigrams attested in the data, and d, the dimension of the vectors.

Big-O notation involves giving the runtime independent of small constants: for example, taking the dot product of two d-length vectors is an $O(d)$ operation because the time it takes is proportional to the length of the vectors, as each pair of numbers has to be multiplied and then summed together. Assume you are describing the total number of operations and there is no parallelization available.

~~$O(V \cdot d + B \cdot d)$~~

$O(Vd + Bd)$

dimension of vocab embeddings and bigram embeddings

c. (2 points) Assume that making a single gradient step on a single example takes the same time as what you reported in part (b). What is the big-O runtime of model training for M (bigram, context) examples on N epochs? Express this in terms of your answer to part (b), which you can denote as E. (You may also use the other quantities defined in the part (b) question as needed.)

$O(E M N)$

N epoch loops through M example predictions/updates

d. (2 points) How many parameters are in the model? Express this in terms of the quantities in part (b). Your answer to this part should be exact.

$(V + B) \cdot d$ parameters