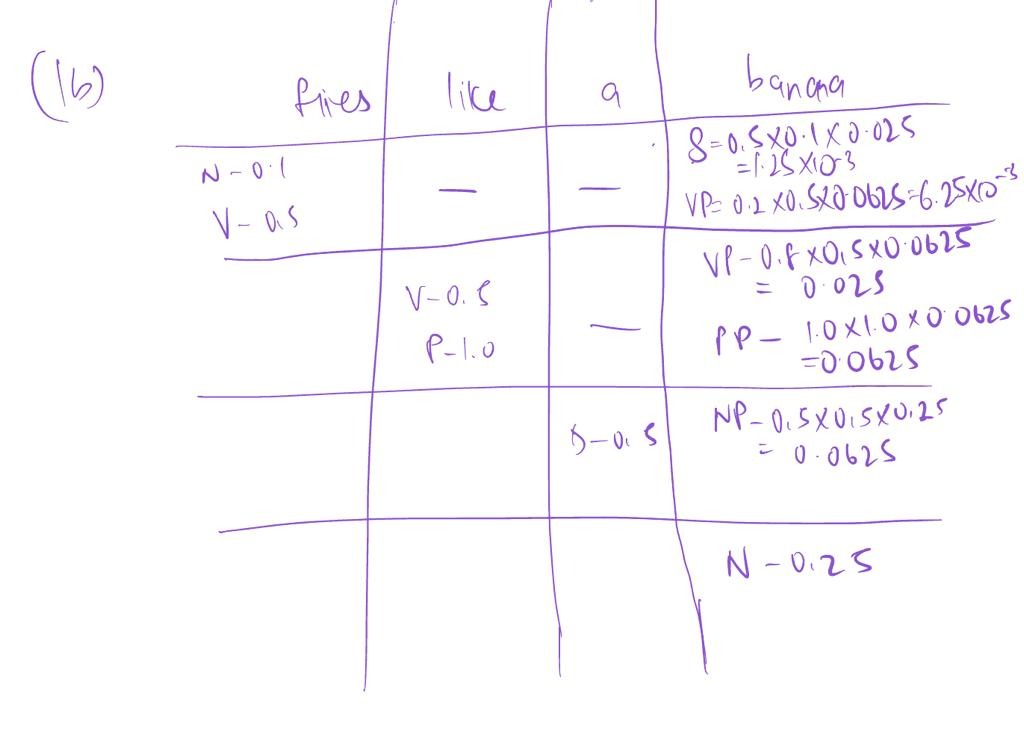
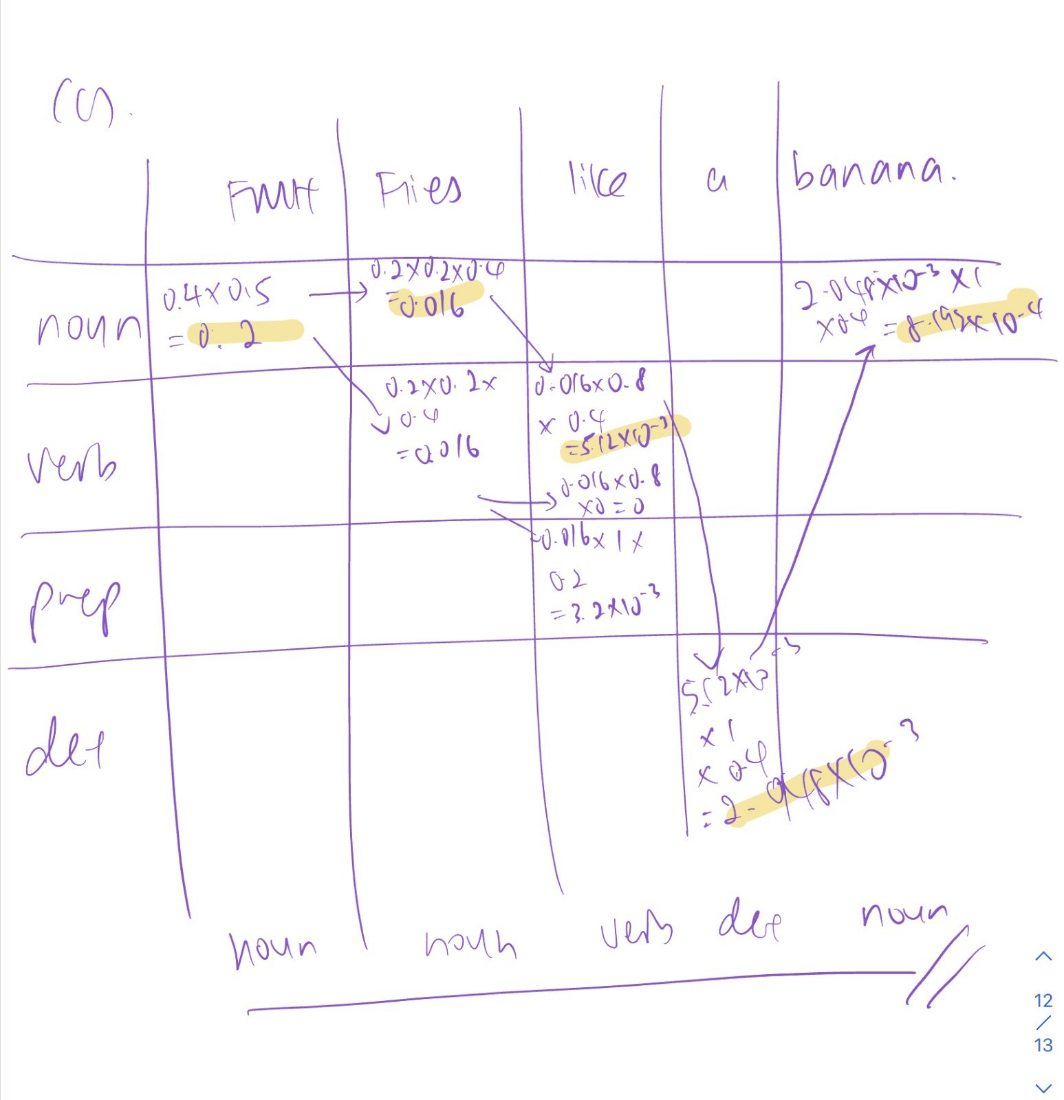
1a. Unsmoothed – 10/20, 6/20, 4/20, 0/20

Smoothed – 11/24, 7/24, 5/24, 1/24

1 b.



1 c.

T

1d. Advantages: 1) can handle rare words, spelling errors or unseen words. 2) multi-lingual capacity 3) more memory efficient (discrete space you are working space is much smaller I.e. number of possible characters vs countless word combination options)

Disadvantage: 1) cannot embody semantics/syntatics into word embeddings

2ai. Many to many. Input sentence has multiple words, and the output gives out the meaning corresponding to each word.

ii. Based on examiner’s feedback need to use the average of 4.

4 x 9000 + 1 x 1000 = 37,000 labelse

iii. Macro-average F1 score – class imbalance.

Bi) Yes. BPE is useful for low resource tasks, and can provide better representation for rare words.

ii) Character level, for the same reason as above.

c)

**For Encoders:**

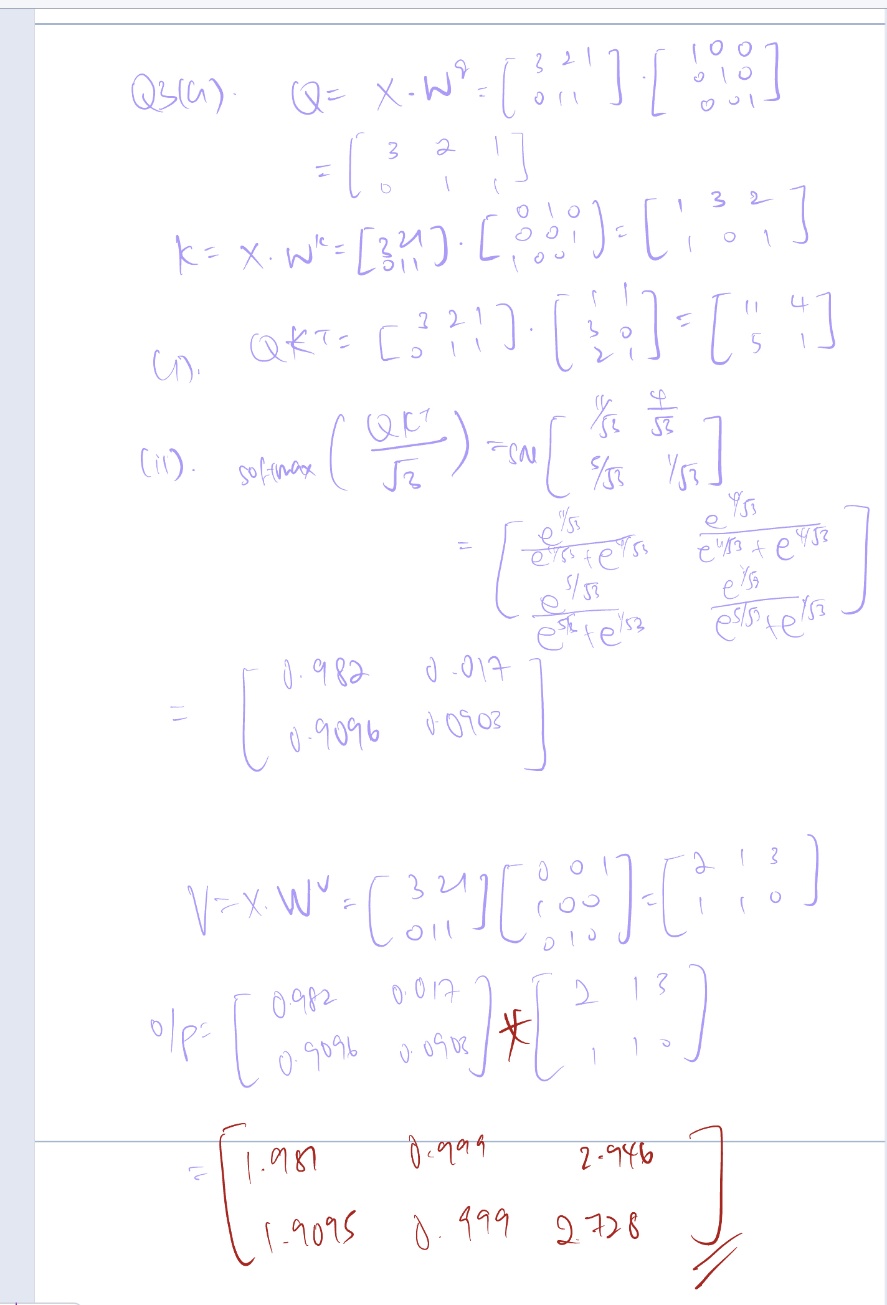
Transformer encoders process the entire input all at once (parallelisable) that’s why we need additional position encoding, whereas RNN encoders process their inputs sequentially so there is no need for position encoding.

Transformer encoder uses self-attention, but RNN encoder does not use attention.

**For Decoders:**

Transformer decoders uses masked self-attention but RNN decoder just uses cross-attention.

3a)



3b(I)

If we mask too many words, we won’t have enough context to give to the model for it to predict the mask

M2: If we only train on words that are masked, then the model learns that it only needs to construct good representations for words with mask in the input. But if we apply this model, we wouldn’t have any mask. So we want the model to construct good rep for all words (even words that are not masked).

M3: If we only trained with the M1 and M2, then the model learns that if it’s not a mask in the input, then that is an incorrect word. It learns to not trust us, it knows that if there’s a word then it needs to be a diff word in the slot. We want the model to construct good representations even if there is a correct word alrd in the slot.

(ii) d = dimension of word embeddings from BERT encoder

Input (N, d)

Output (N, 4)

Need to attached linear layers (up to max padding x d --> max padding x 4) on top of the BERT encoder, and softmax across 4 diff classes.

Loss – cross entropy loss

Ci) yes. (1) d^2 for self attention, (2) linear for position wise feed forward layer

ii) No, the weight dimensions do not depend on the Sequence length. [D x dh]. Position-wise weight matrix is not dependent on sequence length [D x dff]

iii) O(Nk) where k= window size