NLP HW3

Ido Shapira Yoav Rabinovich  
319021044 208275735

**Question 1**

1.

a. can be interpreted as a categorical probability distribution because if we view it as a categorical probability distribution with n categories such that each category is I and the probability is then we can see it has all the needed properties:   
 , .

b. The categorical distribution α puts almost all of its weight on some αj when the dot product between the query q and a specific key kj is significantly larger than the dot products between the query and all other keys, indicating a strong similarity or match between q and kj compared to the other keys.

c. The output C will be very close to for the j from the last question.

d. In intuitive terms, this means that when the query q closely matches or aligns with a specific key kj compared to the other keys, the resulting output c will be strongly influenced by the vector vj associated with that key. It implies that the model focuses its attention and assigns a higher weight to the key that exhibits a stronger similarity or relevance to the query, resulting in a more pronounced impact on the final output.

2.

a. We will observe M = .

b. We will observe .

3.

a)

We will define .

Since we can say that .

The fact that says that approximately and we already solve this case in question 2b:

All in all, .

b)

Now we have

As in last question,

We also know that , to make the description simpler we will sample and define

The dot product between q and will be zero just like in previous question so we need to calculate the following SoftMax participants:

Which means where .

4.

a)

We will define q1 and q2 as follows:

As before

With similar calculations, we infer .

b)

As in previous question we will sample

All in all,

**Question 2**

4. When we evaluate our model that did not do pretrain we get for the dev set:  
Correct: 9.0 out of 500.0: 1.7999999999999998%

When we predict London every time we get for the dev set:  
Correct: 25.0 out of 500.0: 5.0%

We can see that the model now is very bad. Even worse that a model that give the same prediction for every input.

6. When we evaluate our model that did do pretrain we get for the dev set:  
Correct: 90.0 out of 500.0: 18.0%

7. The pretrained vanilla model achieved high accuracy because it had been trained on relevant data, acquiring knowledge of language patterns, grammar, and semantics. This prior training provided a solid foundation for better predictions and responses. In contrast, the non-pretrained model probably did not gain enough understanding of the language semantics and structure from the fine tune data to successfully infer the correct answers from the data.