CS224n Winter 2019 Homework 4

SUNet ID: 05794739
Name: Luis Perez
Collaborators:

By turning in this assignment, I agree by the Stanford honor code and declare that all of this is my own work.

Problem 1

(a) In “utils.py”.

(b) In “model_embeddings.py”.

(c) In “nmt_model.py”.

(d) In “nmt_model.py”.

(e) In “nmt_model.py”.

(f) In “nmt_model.py”.

(g) As per the code, the masks end-up setting the $e_{t,i} = -\infty$ at all positions corresponding to ‘pad’ tokens. This corresponds to assigning $-\infty$ energy to the “annotations” (encoder hidden states) corresponding to our padded sequences. After running through a Softmax, this leads to a zero-probability for these states, which means they don’t contribute to our overall attention vector a_t since this is the result of a weighed (by the softmax probability) average of the encoder hidden states.

It is necessary to use the masks in this way since the ‘pad’ tokens are artificial additions (mainly for performance) which should not be used by our decoder to generate translations. This information is not only useless, but could harm the overall ability of the model to translate sentences since many tranlations could be mis-matched in length, leading to ‘pad’ tokens being alinged with real worlds.

(h) Trained on Azure VM.

(i) TODO

(j) i Dot Product Attention

- **Advantage**: The biggest advantage is computational, as well as the intepretability of this mechanism. The computational complexity of dot product is $O(n)$. Furthermore, this mechanism is very intuitive, where values with high cosine-similarity to the query will receive the most attention.

- **Disadvantage**: One big disadvantage of this method is a practical one. In order to use this attention mechanism, the values and the query must be of the same dimension. Generally speaking, this is an unnecessary restriction. Additionally, there are not explicit parameters to learn for this attention step, therefore the model is more restricted.

ii Multiplicative Attention:

- **Advantage**: With the introduction of the weight matrix, there is no longer a restriction that the dimension of the query and the values must match. Furthermore, the weight matrix itself can be learned, allowing for more expressivity in the model (ie, a value vector that's close to the query vector can be given a low-score due to the linear transformation it first undergoes).
- **Disadvantage**: Computational more complex than dot product attention, and some level of interpretability is lost since now there is an additional linear transformation of the value vectors.

iii Additive Attention

- **Advantage**: Explicitly allows for tuning the "attention dimensionality" thereby allow for techniques which try to bottleneck or expand the capacitor of the attention mechanism.
- **Disadvantage**: The most complex of all of the mechanisms, with the least level of interpretability.

Problem 1

(a) i 1

ii 2

iii 3

iv 4

v 5

vi 1 1