RADHIKA PATWARI

18CS10062

30/2/20

Class Test-4 [CS60010]

1. (a)(i) Complexity of (q.k) = 36×36

Addition of v

$X = n \times 100$

$Q = n \times 36$

$K = n \times 36$

$E = n \times n$

$A = n \times n$

$V = n \times 25$

Complexity $= n^2 \times 36 \times 36$ (dot product) / alignment score

$+ (n \times n)(n \times 25)$ (AV)

$= n^2 1296 + 25n^3$

(ii) for 1 word:

$W_q = 100 \times 36$ (Query weight matrix)

$W_k = 100 \times 36$ (key vector weight matrix)

$W_v = 100 \times 25$ (value weight matrix)

for n words total learnable parameters

$= (n \times 100 \times 36) + (n \times 100 \times 36) + n(100 \times 25)$

$= 7200n + 2500n$

$= 9700n$

(iii) $\alpha_{5,30} = \frac{(q_5)(k_{30})}{\sqrt{36}}$

8. (a) The giant [MASK] had been wedged diagonally across Egypt's [MASK] Canal since March 23

(b) During training, 80% of the time, replace unknown words with [MASK], 10% of time replace random word and 10% of time, keep same.

80% keep (MASK)

eg. The giant [MASK] had been wedged diagonally across Egypt's [MASK] Canal since March 23

10% of time, replace with random word,

eg. the giant box had been wedged diagonally across Egypt's Dump Canal since March 23

10% of time, keep original word,

eg. The giant vessel had been wedged diagonally across Egypt's Suez Canal since March 23.

(c) A vector representing masked word is returned which is passed through a linear multiclass classifier of vocabulary size to find word of maximum probability for each [MASK]

This gives the predicted words for each of the masked word of input sentence.

(d) BERT has 2 components while computing loss.

→ the model must recover MASKED words correctly

→ the model must predict sentence consecutiveness.

---

4. (a) ~~RNN has sequential training~~

RNN has problem of long range dependecies due to vanishing gradient during backpropagation.

But in transformers, information of all vectors are aggregated in self attention layers during prediction of vector representation of each word.

(b) Transformer requires 8 attention heads and various encoder-decoder layers stacked together. there it is computationally more expensive & require more parameters than RNN.

(c) (i) 80000 distinct words have different context. Representing them using vectors and putting all into vocabulary size, will increase the confidence of

predicting this word during NLP task. Hence model can learn effectively.

(ii) This is effective in tagging words into various classes like verbs, nouns, preposition as we are tokenising the set of word pieces.

3. (a)

(b) GPT is a unidirectional contextual pretrained Model but BERT is bidirectional.

BERT is more suitable as machine translation, the words used depends on both the left and right context of the sentence. But GPT will considers only the left or right part. This may result in false predictions. BERT on the other hand, considers the entire sentence and hence is more effective and better for machine translation

(c)