DLP - Week-3

Graded Assignment - 3

1) Download the Yelp review dataset "Yelp/yelp_review_full". Split each sample by calling the string method ".split()" and choose the correct statements about the dataset. **2 points**

A. The dataset contains close to 99 million words
B. There are more than 300 samples that contain a single word
C. There are less than 300 samples that contain only a single word
D. "Cheesy-melty-roasted-cauliflower-with-fresh-bread-crumbs-on top.\\nTo-die-for." is one of the single words in the dataset
E. The average length of a sample is 134.1
F. The distribution of the length of the samples is right skewed

- [x] A
- [x] B
- [ ] C
- [x] D
- [x] E
- [x] F

} NB

Partially Correct.
Score: 1.2

2) Load the "bert-base-uncased" pre-trained tokenizer and choose the correct statements about the tokenizer. **2 points**

A. The tokenizer is used for the BERT model with the context length of 512
B. The tokenizer has 5 special tokens
C. Tokenizing a sample that contains more than 512 words would result in truncation of all tokens beyond the length 512
D. Tokenizer inserts all the special tokens when it processes a single sample as an input
E. Tokenizer inserts [CLS] and [SEP] special tokens when it processes a single sample as an input
F. Tokenizer inserts only [CLS]special token when it processes a single sample as an input

→ go, cls, sep, eot and 1 more

→ it gets added automatically

- [x] A
- [x] B
- [ ] C
- [ ] D
- [x] E
- [ ] F

3) Use "BertConfig" and "BertForMaskedLM" to construct the default (original) BERT model. Choose the correct statements **1 point**

A. The model has 12 Bert layers
B. The model has 6 Bert layers
C. The model uses absolute position embeddings
D. The word embedding (token embedding) layer has about 23 million learnable parameters
E. The total number of parameters in the model is close to 110 million

→ GPT 2 has 12 self attn heads (12 bert layers)

→ Yes, from the calculations in week-4 (L2)

→ 116.435 to be exact

- [x] A
- [ ] B
- [x] C
- [x] D
- [x] E

4) Double the context length from 512 to 1024 (you can change it in the configuration). Count the number of parameters and enter the change in the number of parameters (in millions) compared to the default configuration.

0.39

Yes, the answer is correct.
Score: 1

Accepted Answers:

(Type: Range) 0.38,0.41

→ even after doubling the context length, the no. of parameters doesn't change much.

1 point

5) Pack (chunk) the samples such that the length of all the samples in the dataset is 512 (for efficient training). Define a mapping function that implements the following procedure
1. Take a batch of 1000 samples
2. Tokenize it to get input IDs and attention mask
3. Concatenate all the input IDs
4. Chunk the concatenated IDs into a size of 512
5. Drop the last chunk if its length is less than 512
6. Pack all the chunks
7. Iterate over all the batches in the dataset

Store the resulting dataset in the variable "ds_chunked". Enter the total number of samples in the new dataset.
Note: the batch size should be kept at 1000 while calling "ds.map()" for theanswer to match.

246695

} some confusion in this ques. check NB

6) Split the new dataset into training and test sets with the test_size=0.05 and seed=42. Use the appropriate data collator function for the MLM objective and set the masking probability to 0.2. Use the data loader from PyTorch to load a batch of samples, and enter the token ID corresponding to the unmasked token

-100

Yes, the answer is correct.
Score: 1

Accepted Answers:

(Type: Numeric) -100

1 point

7) Create a small BERT model by changing the following hyper-parameters and keeping the other hyper-parameters as is
- num_hidden_layers = 6
- hidden size: 384
- intermediate_size: 1536

and start training the model with a batch of size 8 for an epoch. What is the loss value at the end of the training?
Note: You may optionally save the checkpoints for every N-th step.

5.848