IIT KHARAGPUR AI4ICPS I HUB FOUNDATION

Hands-on **A**pproach to **AI, Cohort-2, July – October 2024**

Assignment 10: Transformers

Due date: Friday 20th September 2024, EOD – IST.

Important Instructions for submitting solutions

1. Submit the solution to all questions in the assignment should be submitted in a **single PDF file with not more than 500 words**.
2. Any plagiarism if detected will automatically attract **zero marks** for that assignment.
3. It is preferable if the **text of PDF file can be extracted** through a PDF extractor e.g. PyPDF. For example, pictures of handwritten text are not extractable, whereas PDF generated by MS Word, Latex, etc., are.
4. Exceptionally good solutions with extractable text may receive **special appreciation** from the teachers.

Q1. Suppose you are pretraining a BERT model with 8 layers, 768-dim hidden states, 8 attention heads, and a sub-word vocabulary of size 40k. Also, your feed-forward hidden layer is of dimension 3072. What will be the number of parameters of the model? You can ignore the bias terms, and other parameters used corresponding to the final loss computation from the final encoder representation. The BERT model can take at most 512 tokens in the input.

Q2. Suppose, you give the following input to your transformer encoder: {flying, arrows}. The input embeddings for these two words are **[0,1,1,1,1,0] and [1,1,0,-1,-1,1]**, respectively. Suppose you are trying to represent the first word ‘flying’ with the help of self-attention in the first encoder. For the first attention head, the query, key and value matrices just take the 2 dimensions from the input each. Thus, the first 2 dimensions define the query vector, and so on. What will be the self-attention output for the word ‘flying’ corresponding to this attention head. You are using the scaled dot vector. **Note: use appropriate scaling.**

Q3. Suppose you are using BERT-base for topic classification of documents with 5 classes. How many task-specific parameters will you need to use? On the other hand, what if you were using BERT-base for language identification in code-switched dataset, how many task-specific parameters will need to be used? Assume that there are two languages, English and Hindi, using the same Roman script, and each word belongs to exactly one of these languages.