ESSENCE

INTERNSHIP : WEEKLY REPORT

(RENU BOJJA - 1RVU22CSE129)

COLAB : Renu_Essence.ipynb

WEEK: 10Jan - 25Jan

Get the list of the most popular language models, think about pros-cons of different models if there is research on it. Big challenge: how computationally-intensive are those models, can you run them on Google Colab? Maybe try and code them.

What I did:

- Tried 4 language models (GPT-2, BERT, T5, XLNet)
- Compared speed and creative generation of models.
- Computed tokenization and inference time of the models

Next Steps:

- Compare LM's and LLM's with word embedding models
- Build reproducible pipeline

Challenges:

- Need an API key to work on GPT-3 / GPT-4 models

WEEK: 29Jan - 16Feb

Fine-tune the pre trained BERT Model

What I did:

- Fine-tuned BERT model.
- Was working on fine-tuning the T5 model; will continue after completing the fine-tuning of the TinyBERT model
- Research on TinyBERT and BERT models, exploring their differences.
- Computed the computation time for both; TinyBERT takes more time than BERT mode which is an issue and I'm currently addressing it.
- Will fine-tune the model to achieve better results.

WEEK: 16Feb - 28Feb

Work on how smaller versions of models perform like TinyBERT (smaller size models to preserve our limited resources with acceptable (not ideal!) level of accuracy).

What I did:

- Implemented TinyBert model and fine tuned it.
- Compared Bert and TinyBert model's computation time.

- Observations + code are enclosed in the colab attached in this doc.

Challenges:
Tried fine-tuning from Hugging Face but needed an authentication token. Logged into my Hugging Face account and got a free token, but when using it, it said invalid.

WEEK: 6Mar - 15Mar

T5, GPT, Bert, XLNet, TinyBert:
1) Compute the average tokenization time of a task title across the dataset. I.e., an average speed of sentence tokenization

What I did:

- Randomly selected a sentence from the column task name of dataset_20240111.csv and calculated its tokenization time.
- Computed average tokenization time.

Observation:

- On real data, average tokenization time of a model:
- T5: 0.00011271800634995946 sec
  GPT: 0.00017567250713612298 sec
  BERT: 0.09176518707155432 sec
  XLNet: 0.00022623898848047797 sec
  TinyBERT: 0.05756535844982795 sec
- Increasing order: T5 < GPT < XLNet < TinyBERT < BERT

WEEK: 18Mar - 3Apr

2) Fine-tune the models on the dataset take the "task type" classes. train-val split should include all classes, like StratifiedSplit from sklearn.