11.05.22

OPENCAMPUS.sh

Transformers for Natural Language Processing and Beyond

FINE-TUNING PRETRAINED MODELS

- Quiz
- Breakout Discussion
- Discussion of Some Special Issues
- Project Teams
- First Project Tasks

QUIZ

https://forms.office.com/r/if3nDTQNuk

BREAKOUT DISCUSSION

Suppose you want to predict bakery sales. They depend on the weather, the holidays, weekdays, and many more variables.

- What is best: an encoder model, a decoder model, or an encoder-decoder model?

OPEN DISCUSSION

- What is trained during fine-tuning: Only the head or the full model?
- What are the pros and cons of training just the head and training the full model?
- Are there other alternatives?

IMPLEMENTATION OF FAST PREPROCESSING FUNCTIONS

```
def tokenize_function(example):
    return tokenizer(example["sentence1"],
                     example["sentence2"], truncation=True)

tokenized_datasets = raw_datasets.map(tokenize_function, batched=True)
```

DYNAMIC PADDING

```
tf_train_dataset = tokenized_datasets["train"].to_tf_dataset(
    columns=["attention_mask", "input_ids", "token_type_ids"],
    label_cols=["labels"],
    shuffle=True,
    collate_fn=data_collator,
    batch_size=8,
)
```

PROJECTS

- Chris/Dariush: Classification of student emails to predict if an argument is correctly included
- Friedrich/Nicolas/Dustin/Wang: Detection of Transposable Elements in Genome Sequences
- Prosper/Julien/Alwin: Generating Marketing Content for NFTs
- Dieter/Desmond: Time Series Prediction for Electric Motors

FIRST PROJECT TASKS

(1) Setup a project channel in the Chat.

(2) Define a common repository or GoogleDrive to exchange the program code.

(3) Decide on times for regular project meetings.

(4) Schedule a meeting with the Instructor

PROJECT MILESTONES

- 11.05. Form project groups
- 18.05. Literature review
- 25.05. Dataset characteristics
- 01.06. Baseline model
- 08.06. Model & model evaluation (Joint Coding)
- 15.06. Project presentations

LITERATURE REVIEW

- Search for transformer models applied to similar problems
- Focus on the structure of the input and of the output
- Are there pretrained models that you can use?
- Which type of model is best suited?
- Do you need tokenization?
- Do you need a type of embedding layer?

TODOS UNTIL NEXT WEEK

- Complete chapter 4 (Sharing Models and Tokenizers) and chapter 5 (The Datasets Library) of the Hugging Face course

- Literature Review:
  Each team should review current publications and answer the questions from the slide before.