16.11.22

OPENCAMPUS.sh

Application of Transformer Models

FINE-TUNING PRETRAINED MODELS

- Quiz
- Breakout Discussion
- Open Discussion
- Final Project Teams
- First Project Tasks

QUIZ

https://forms.office.com/r/gQ0fDQyVZx

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?

LOSS FUNCTION DEFINITION

```
from tensorflow.keras.losses import SparseCategoricalCrossentropy

model.compile(
    optimizer="adam",
    loss=SparseCategoricalCrossentropy(from_logits=True),
    metrics=["accuracy"],
)
model.fit(
    tf_train_dataset,
    validation_data=tf_validation_dataset,
)
```

TOKENIZING LONG SEQUENCES

```
inputs = tokenizer(
    raw_datasets["train"][2:6]["question"],
    raw_datasets["train"][2:6]["context"],
    max_length=100,
    truncation="only_second",
    stride=50,
    return_overflowing_tokens=True,
    return_offsets_mapping=True,
)
```

LEARNING RATE

- **For transformers, a good starting point to find the best value is often around 5e-5 (=0.00005).**

Adam

Adam class [source]

```
tf.keras.optimizers.Adam(
    learning_rate=0.001,
    beta_1=0.9,
    beta_2=0.999,
    epsilon=1e-07,
    amsgrad=False,
    name="Adam",
    **kwargs
)
```

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?

TOKENIZATION / IMPLEMENTATION OF FAST PREPROCESSING FUNCTIONS

```
tokenizer = AutoTokenizer.from_pretrained(checkpoint)


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


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

DYNAMIC PADDING / FORMATTING OF THE INPUT DATA

```
data_collator = DataCollatorWithPadding(tokenizer=tokenizer, return_tensors="tf")

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

- Saif/ Emmanuel/ Kristian/ Atul: Time Series Prediction Financial/Climate
- Jonathan/Julian: Arguments Mining / NER Task on data already collected
- Benjamin/ Malte/ T.-Niklas: Speech to speech models including translation
- Jeremy/ Veit/ Christian: Transcribing and summarizing Podcasts
- Laura/ Janosch/ Valentin : Training a model to produce text written in different authors' style
- Khan: Classification of activity descriptions according to keywords
- Manpreet: Unsupervised training of log data to predict user behavior
- Max: Q&A model

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

16.11.	Form project groups
23.11.	Literature review
30.11.	Dataset characteristics
04.01.	Baseline model
11.01.	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.