NNproject_

April 20, 2023

[ ]:
```
pip install transformers datasets evaluate
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-
wheels/public/simple/
Requirement already satisfied: transformers in /usr/local/lib/python3.9/dist-
packages (4.28.1)
Requirement already satisfied: datasets in /usr/local/lib/python3.9/dist-
packages (2.11.0)
Requirement already satisfied: evaluate in /usr/local/lib/python3.9/dist-
packages (0.4.0)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.9/dist-
packages (from transformers) (23.1)
Requirement already satisfied: regex!=2019.12.17 in
/usr/local/lib/python3.9/dist-packages (from transformers) (2022.10.31)
Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.9/dist-
packages (from transformers) (4.65.0)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.9/dist-
packages (from transformers) (1.22.4)
Requirement already satisfied: tokenizers!=0.11.3,<0.14,>=0.11.1 in
/usr/local/lib/python3.9/dist-packages (from transformers) (0.13.3)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.9/dist-
packages (from transformers) (6.0)
Requirement already satisfied: requests in /usr/local/lib/python3.9/dist-
packages (from transformers) (2.27.1)
Requirement already satisfied: filelock in /usr/local/lib/python3.9/dist-
packages (from transformers) (3.11.0)
Requirement already satisfied: huggingface-hub<1.0,>=0.11.0 in
/usr/local/lib/python3.9/dist-packages (from transformers) (0.13.4)
Requirement already satisfied: pandas in /usr/local/lib/python3.9/dist-packages
(from datasets) (1.5.3)
Requirement already satisfied: responses<0.19 in /usr/local/lib/python3.9/dist-
packages (from datasets) (0.18.0)
Requirement already satisfied: dill<0.3.7,>=0.3.0 in
/usr/local/lib/python3.9/dist-packages (from datasets) (0.3.6)
Requirement already satisfied: fsspec[http]>=2021.11.1 in
/usr/local/lib/python3.9/dist-packages (from datasets) (2023.4.0)
Requirement already satisfied: pyarrow>=8.0.0 in /usr/local/lib/python3.9/dist-
packages (from datasets) (9.0.0)
```

```
Requirement already satisfied: xxhash in /usr/local/lib/python3.9/dist-packages
(from datasets) (3.2.0)
Requirement already satisfied: aiohttp in /usr/local/lib/python3.9/dist-packages
(from datasets) (3.8.4)
Requirement already satisfied: multiprocess in /usr/local/lib/python3.9/dist-
packages (from datasets) (0.70.14)
Requirement already satisfied: async-timeout<5.0,>=4.0.0a3 in
/usr/local/lib/python3.9/dist-packages (from aiohttp->datasets) (4.0.2)
Requirement already satisfied: multidict<7.0,>=4.5 in
/usr/local/lib/python3.9/dist-packages (from aiohttp->datasets) (6.0.4)
Requirement already satisfied: charset-normalizer<4.0,>=2.0 in
/usr/local/lib/python3.9/dist-packages (from aiohttp->datasets) (2.0.12)
Requirement already satisfied: aiosignal>=1.1.2 in
/usr/local/lib/python3.9/dist-packages (from aiohttp->datasets) (1.3.1)
Requirement already satisfied: yarl<2.0,>=1.0 in /usr/local/lib/python3.9/dist-
packages (from aiohttp->datasets) (1.8.2)
Requirement already satisfied: frozenlist>=1.1.1 in
/usr/local/lib/python3.9/dist-packages (from aiohttp->datasets) (1.3.3)
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.9/dist-
packages (from aiohttp->datasets) (23.1.0)
Requirement already satisfied: typing-extensions>=3.7.4.3 in
/usr/local/lib/python3.9/dist-packages (from huggingface-
hub<1.0,>=0.11.0->transformers) (4.5.0)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.9/dist-
packages (from requests->transformers) (3.4)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.9/dist-packages (from requests->transformers) (2022.12.7)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in
/usr/local/lib/python3.9/dist-packages (from requests->transformers) (1.26.15)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.9/dist-
packages (from pandas->datasets) (2022.7.1)
Requirement already satisfied: python-dateutil>=2.8.1 in
/usr/local/lib/python3.9/dist-packages (from pandas->datasets) (2.8.2)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.9/dist-
packages (from python-dateutil>=2.8.1->pandas->datasets) (1.16.0)
```

[ ]:
```
from huggingface_hub import notebook_login

notebook_login()
```

```
VBox(children=(HTML(value='<center> <img\nsrc=https://huggingface.co/front/
↪assets/huggingface_logo-noborder.sv…
```

[ ]:
```
from datasets import load_dataset

squad = load_dataset("squad", split="train[:5000]")
```

```
squad = squad.train_test_split(test_size=0.2)

squad["train"][0]
```

```
{'answers': {'answer_start': [515], 'text': ['Saint Bernadette Soubirous']},
 'context': 'Architecturally, the school has a Catholic character. Atop the␣
 ↪Main Building\'s gold dome is a golden statue of the Virgin Mary.␣
 ↪Immediately in front of the Main Building and facing it, is a copper statue␣
 ↪of Christ with arms upraised with the legend "Venite Ad Me Omnes". Next to␣
 ↪the Main Building is the Basilica of the Sacred Heart. Immediately behind␣
 ↪the basilica is the Grotto, a Marian place of prayer and reflection. It is a␣
 ↪replica of the grotto at Lourdes, France where the Virgin Mary reputedly␣
 ↪appeared to Saint Bernadette Soubirous in 1858. At the end of the main drive␣
 ↪(and in a direct line that connects through 3 statues and the Gold Dome), is␣
 ↪a simple, modern stone statue of Mary.',
 'id': '5733be284776f41900661182',
 'question': 'To whom did the Virgin Mary allegedly appear in 1858 in Lourdes␣
 ↪France?',
 'title': 'University_of_Notre_Dame'
}
```

```
WARNING:datasets.builder:Found cached dataset squad (/root/.cache/huggingface/da
tasets/squad/plain_text/1.0.0/d6ec3ceb99ca480ce37cdd35555d6cb2511d223b9150cce08a
837ef62ffea453)
```

[ ]:
```
{'answers': {'answer_start': [515], 'text': ['Saint Bernadette Soubirous']},
 'context': 'Architecturally, the school has a Catholic character. Atop the Main
Building\'s gold dome is a golden statue of the Virgin Mary. Immediately in
front of the Main Building and facing it, is a copper statue of Christ with arms
upraised with the legend "Venite Ad Me Omnes". Next to the Main Building is the
Basilica of the Sacred Heart. Immediately behind the basilica is the Grotto, a
Marian place of prayer and reflection. It is a replica of the grotto at Lourdes,
France where the Virgin Mary reputedly appeared to Saint Bernadette Soubirous in
1858. At the end of the main drive (and in a direct line that connects through 3
statues and the Gold Dome), is a simple, modern stone statue of Mary.',
 'id': '5733be284776f41900661182',
 'question': 'To whom did the Virgin Mary allegedly appear in 1858 in Lourdes
France?',
 'title': 'University_of_Notre_Dame'}
```

[ ]:
```
from transformers import AutoTokenizer

tokenizer = AutoTokenizer.from_pretrained("distilbert-base-uncased")

def preprocess_function(examples):
    questions = [q.strip() for q in examples["question"]]
    inputs = tokenizer(
        questions,
```

```
        examples["context"],
        max_length=384,
        truncation="only_second",
        return_offsets_mapping=True,
        padding="max_length",
    )

    offset_mapping = inputs.pop("offset_mapping")
    answers = examples["answers"]
    start_positions = []
    end_positions = []

    for i, offset in enumerate(offset_mapping):
        answer = answers[i]
        start_char = answer["answer_start"][0]
        end_char = answer["answer_start"][0] + len(answer["text"][0])
        sequence_ids = inputs.sequence_ids(i)

        # Find the start and end of the context
        idx = 0
        while sequence_ids[idx] != 1:
            idx += 1
        context_start = idx
        while sequence_ids[idx] == 1:
            idx += 1
        context_end = idx - 1

        # If the answer is not fully inside the context, label it (0, 0)
        if offset[context_start][0] > end_char or offset[context_end][1] <␣
↪start_char:
            start_positions.append(0)
            end_positions.append(0)
        else:
            # Otherwise it's the start and end token positions
            idx = context_start
            while idx <= context_end and offset[idx][0] <= start_char:
                idx += 1
            start_positions.append(idx - 1)

            idx = context_end
            while idx >= context_start and offset[idx][1] >= end_char:
                idx -= 1
            end_positions.append(idx + 1)

    inputs["start_positions"] = start_positions
    inputs["end_positions"] = end_positions
    return inputs
```

[ ]:
```
tokenized_squad = squad.map(preprocess_function, batched=True,␣
↪remove_columns=squad["train"].column_names)
```

```
Map:   0%|          | 0/4000 [00:00<?, ? examples/s]
Map:   0%|          | 0/1000 [00:00<?, ? examples/s]
```

[ ]:
```
from transformers import DefaultDataCollator

data_collator = DefaultDataCollator()
```

[ ]:
```
from transformers import AutoModelForQuestionAnswering, TrainingArguments,␣
↪Trainer

model = AutoModelForQuestionAnswering.from_pretrained("distilbert-base-uncased")

training_args = TrainingArguments(
    output_dir="my_awesome_qa_model",
    evaluation_strategy="epoch",
    learning_rate=2e-5,
    per_device_train_batch_size=16,
    per_device_eval_batch_size=16,
    num_train_epochs=3,
    weight_decay=0.01,
    push_to_hub=True,
)

trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=tokenized_squad["train"],
    eval_dataset=tokenized_squad["test"],
    tokenizer=tokenizer,
    data_collator=data_collator,
)

trainer.train()
```

```
Some weights of the model checkpoint at distilbert-base-uncased were not used
when initializing DistilBertForQuestionAnswering: ['vocab_transform.weight',
'vocab_layer_norm.weight', 'vocab_transform.bias', 'vocab_projector.bias',
'vocab_projector.weight', 'vocab_layer_norm.bias']
- This IS expected if you are initializing DistilBertForQuestionAnswering from
the checkpoint of a model trained on another task or with another architecture
(e.g. initializing a BertForSequenceClassification model from a
BertForPreTraining model).
- This IS NOT expected if you are initializing DistilBertForQuestionAnswering
from the checkpoint of a model that you expect to be exactly identical
```

```
(initializing a BertForSequenceClassification model from a
BertForSequenceClassification model).
Some weights of DistilBertForQuestionAnswering were not initialized from the
model checkpoint at distilbert-base-uncased and are newly initialized:
['qa_outputs.weight', 'qa_outputs.bias']
You should probably TRAIN this model on a down-stream task to be able to use it
for predictions and inference.
Cloning https://huggingface.co/Rekhni/my_awesome_qa_model into local empty
directory.
WARNING:huggingface_hub.repository:Cloning
https://huggingface.co/Rekhni/my_awesome_qa_model into local empty directory.
/usr/local/lib/python3.9/dist-packages/transformers/optimization.py:391:
FutureWarning: This implementation of AdamW is deprecated and will be removed in
a future version. Use the PyTorch implementation torch.optim.AdamW instead, or
set `no_deprecation_warning=True` to disable this warning
  warnings.warn(
```

```
<IPython.core.display.HTML object>
```

```
<IPython.core.display.HTML object>
```

[ ]:
```
TrainOutput(global_step=750, training_loss=2.3414713134765623,
metrics={'train_runtime': 18821.7159, 'train_samples_per_second': 0.638,
'train_steps_per_second': 0.04, 'total_flos': 1175877900288000.0, 'train_loss':
2.3414713134765623, 'epoch': 3.0})
```

[ ]:
```
trainer.push_to_hub()
```

```
Upload file pytorch_model.bin:   0%|          | 1.00/253M [00:00<?, ?B/s]
```

```
Upload file runs/Apr20_04-15-21_24ce5b36a57e/events.out.tfevents.1681964126.
  ↪24ce5b36a57e.472.0:   0%|          …
```

```
To https://huggingface.co/Rekhni/my_awesome_qa_model
   e795dd7..2ba0c11  main -> main

WARNING:huggingface_hub.repository:To
https://huggingface.co/Rekhni/my_awesome_qa_model
   e795dd7..2ba0c11  main -> main

To https://huggingface.co/Rekhni/my_awesome_qa_model
   2ba0c11..c9a4316  main -> main

WARNING:huggingface_hub.repository:To
https://huggingface.co/Rekhni/my_awesome_qa_model
   2ba0c11..c9a4316  main -> main
```

[ ]:
```
'https://huggingface.co/Rekhni/my_awesome_qa_model/commit/2ba0c11f074912fbfbd18d2d767215e6dd2d6788'
```

[ ]:
```
question = "How many programming languages does BLOOM support?"
context = "BLOOM has 176 billion parameters and can generate text in 46␣
 ↪languages natural languages and 13 programming languages."
```

[ ]:
```
from transformers import pipeline

question_answerer = pipeline("question-answering", model="my_awesome_qa_model")
question_answerer(question=question, context=context)
```

[ ]:
```
{'score': 0.17916885018348694,
 'start': 10,
 'end': 95,
 'answer': '176 billion parameters and can generate text in 46 languages natural languages and 13'}
```