

```
#Installation
!pip install datasets
!pip install transformers
!rm seq2seq_trainer.py
!wget https://raw.githubusercontent.com/huggingface/transformers/master/examples/seq2seq
!pip install rouge_score
```

```
import datasets
import transformers
import pandas as pd
from datasets import Dataset
```

```
#Tokenizer
from transformers import RobertaTokenizerFast
```

```
#Encoder-Decoder Model
from transformers import EncoderDecoderModel
```

```
#Training
from transformers.trainer_seq2seq import Seq2SeqTrainer
from transformers.training_args_seq2seq import Seq2SeqTrainingArguments
from transformers import TrainingArguments
from dataclasses import dataclass, field
from typing import Optional
```

```
Attempting uninstall: urllib3
Found existing installation: urllib3 1.24.3
Uninstalling urllib3-1.24.3:
Successfully uninstalled urllib3-1.24.3
Attempting uninstall: dill
Found existing installation: dill 0.3.6
Uninstalling dill-0.3.6:
Successfully uninstalled dill-0.3.6
Successfully installed datasets-2.6.1 dill-0.3.5.1 huggingface-hub-0.10.1 multi
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheelhouse/pypi
Collecting transformers
  Downloading transformers-4.24.0-py3-none-any.whl (5.5 MB)
    |████████████████████████████████████████| 5.5 MB 14.4 MB/s
Requirement already satisfied: importlib-metadata in /usr/local/lib/python3.7/dist-packages (4.6.4)
Requirement already satisfied: filelock in /usr/local/lib/python3.7/dist-packages (3.10.0)
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.7/dist-packages (2022.10.31)
Collecting tokenizers!=0.11.3,<0.14,>=0.11.1
  Downloading tokenizers-0.13.1-cp37-cp37m-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (7.6 MB)
    |████████████████████████████████████████| 7.6 MB 54.7 MB/s
Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.7/dist-packages (4.64.1)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.7/dist-packages (6.0.1)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.7/dist-packages (23.0)
Requirement already satisfied: requests in /usr/local/lib/python3.7/dist-packages (2.28.1)
Requirement already satisfied: huggingface-hub<1.0,>=0.10.0 in /usr/local/lib/python3.7/dist-packages (0.10.1)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.7/dist-packages (1.24.2)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.7/dist-packages (4.4.0)
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in /usr/local/lib/python3.7/dist-packages (3.1.0)
Requirement already satisfied: zipp>=0.5 in /usr/local/lib/python3.7/dist-packages (3.15.0)
Requirement already satisfied: chardet<4,>=3.0.2 in /usr/local/lib/python3.7/dist-packages (3.7.4)
Requirement already satisfied: urllib3<2,>=1.25 in /usr/local/lib/python3.7/dist-packages (1.26.15)
```

```
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in /usr/
Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.7/dist-pack

Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.7/d
Installing collected packages: tokenizers, transformers
Successfully installed tokenizers-0.13.1 transformers-4.24.0
rm: cannot remove 'seq2seq_trainer.py': No such file or directory
--2022-11-07 01:35:27-- https://raw.githubusercontent.com/huggingface/transformers
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.108.1
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.108.1
HTTP request sent, awaiting response... 404 Not Found
2022-11-07 01:35:27 ERROR 404: Not Found.
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheelhouse/pypi
Collecting rouge_score
  Downloading rouge_score-0.1.2.tar.gz (17 kB)
Requirement already satisfied: absl-py in /usr/local/lib/python3.7/dist-packages
Requirement already satisfied: nltk in /usr/local/lib/python3.7/dist-packages (1
Requirement already satisfied: numpy in /usr/local/lib/python3.7/dist-packages (1
Requirement already satisfied: six>=1.14.0 in /usr/local/lib/python3.7/dist-pack
Requirement already satisfied: joblib in /usr/local/lib/python3.7/dist-packages
Requirement already satisfied: tqdm in /usr/local/lib/python3.7/dist-packages (1
Requirement already satisfied: click in /usr/local/lib/python3.7/dist-packages (1
Requirement already satisfied: regex>=2021.8.3 in /usr/local/lib/python3.7/dist-
Building wheels for collected packages: rouge-score
  Building wheel for rouge-score (setup.py) ... done
  Created wheel for rouge-score: filename=rouge_score-0.1.2-py3-none-any.whl siz
  Stored in directory: /root/.cache/pip/wheels/84/ac/6b/38096e3c5bf1dc87911e358f
Successfully built rouge-score
```

```
import datasets
from datasets import load_dataset, load_metric
from transformers import AutoTokenizer
from transformers import DataCollatorForSeq2Seq
from transformers import AutoModelForSeq2SeqLM, Seq2SeqTrainingArguments, Seq2SeqTrainer
from transformers import AutoModelForSeq2SeqLM, DataCollatorForSeq2Seq, Seq2SeqTrainingArguments
from transformers import create_optimizer, AdamWeightDecay
from transformers import RobertaTokenizerFast
from transformers import EncoderDecoderModel
from transformers import TrainingArguments
from transformers.trainer_seq2seq import Seq2SeqTrainer
from transformers.training_args_seq2seq import Seq2SeqTrainingArguments
from dataclasses import dataclass, field as dataclassfield
from typing import Optional
import nltk
import numpy as np
from transformers import RobertaConfig, RobertaModel
from transformers import DataCollatorForSeq2Seq
nltk.download('punkt')

metric = load_metric("rouge")
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
```

```
[nltk_data]   Unzipping tokenizers/punkt.zip.
```

```
/usr/local/lib/python3.7/dist-packages/jupyter_kernel_launcher.py:21: FutureWarning: 1.
```

```
from huggingface_hub import notebook_login
notebook_login()
```

```
Login successful
```

```
Your token has been saved to /root/.huggingface/token
```

```
pip install git+https://github.com/huggingface/datasets.git
```

Collecting git+<https://github.com/huggingface/datasets.git>

Cloning <https://github.com/huggingface/datasets.git> to /tmp/pip-req-build-5a9

Running command git clone -q <https://github.com/huggingface/datasets.git> /tmp/

Installing build dependencies ... done

Getting requirements to build wheel ... done

Preparing wheel metadata ... done

Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.7/dist-pack

Requirement already satisfied: fsspec[http]>=2021.11.1 in /usr/local/lib/python3.

Requirement already satisfied: importlib-metadata in /usr/local/lib/python3.7/d

Requirement already satisfied: pyarrow>=6.0.0 in /usr/local/lib/python3.7/dist-p

Requirement already satisfied: dill<0.3.7 in /usr/local/lib/python3.7/dist-packa

Requirement already satisfied: multiprocessing in /usr/local/lib/python3.7/dist-pac

Requirement already satisfied: aiohttp in /usr/local/lib/python3.7/dist-packages

```
df = load_dataset('scientific_papers', 'arxiv', download_mode="force_redownload")
```

Requirement already satisfied: huggingface-hub<1.0.0,>=0.2.0 in /usr/local/lib/p

```
from transformers.data import data_collator
```

```
tokenizer = RobertaTokenizerFast.from_pretrained("roberta-base")
```

```
tokenizer.bos_token = tokenizer.cls_token
```

```
tokenizer.eos_token = tokenizer.sep_token
```

```
#parameter setting
```

```
batch_size=256 #
```

```
encoder_max_length=40
```

```
decoder_max_length=8
```

```
def process_data_to_model_inputs(batch):
```

```
    # tokenize the inputs and labels
```

```
    inputs = tokenizer(batch["Text"], padding="max_length", truncation=True, max_length=
```

```
    outputs = tokenizer(batch["Summary"], padding="max_length", truncation=True, max_leng
```

```
    batch["input_ids"] = inputs.input_ids
```

```
    batch["attention_mask"] = inputs.attention_mask
```

```
    batch["decoder_input_ids"] = outputs.input_ids
```

```
    batch["decoder_attention_mask"] = outputs.attention_mask
```

```
    batch["labels"] = outputs.input_ids.copy()
```

```
    # because RoBERTa automatically shifts the labels, the labels correspond exactly to
```

```
    # We have to make sure that the PAD token is ignored
```

```
    batch["labels"] = [[-100 if token == tokenizer.pad_token_id else token for token in ]
```

```
    return batch
```

```
def robertaTokenize(model_checkpoint, dataset, subset):
```

```
    global tokenizer
```

```
    sum=df
```

```
    tokenizer=RobertaTokenizerFast.from_pretrained("roberta-base")
```

```
    tokenized_sum=sum.map(process_data_to_model_inputs, batched=True)
```

```
    return(tokenizer, tokenized_sum)
```

```
def weightless_model(model_checkpoint, tokenizer):
```

```
    #make a model that is not pretrained
```

```
    config=RobertaConfig()
```

```

model=RobertaModel(config)
model.init_weights()

data_collator=DataCollatorForSeq2Seq(tokenizer=tokenizer,model=model)
return(model,data_collator)

def get_my_hyper_params(model_checkpoint, my_epochs, floating_point):
    batch_size = 256
    model_name = model_checkpoint
    args = Seq2SeqTrainingArguments(
        f"{model_name}-science-papers",
        per_device_train_batch_size=batch_size,
        per_device_eval_batch_size=batch_size,
        predict_with_generate=True,
        evaluation_strategy='epoch',
        do_train=True,
        do_eval=True,
        logging_steps=4,
        save_steps=16,
        eval_steps=500,
        warmup_steps=500,
        overwrite_output_dir=True,
        save_total_limit=1,
        fp16=floating_point,
        push_to_hub=True,
    )
    return args

```

```

from transformers import EncoderDecoderModel

roberta_shared = EncoderDecoderModel.from_encoder_decoder_pretrained("roberta-base", "1

```

```

# set special tokens
roberta_shared.config.decoder_start_token_id = tokenizer.bos_token_id
roberta_shared.config.eos_token_id = tokenizer.eos_token_id

# sensible parameters for beam search
# set decoding params
roberta_shared.config.max_length = 40
roberta_shared.config.early_stopping = True
roberta_shared.config.no_repeat_ngram_size = 3
roberta_shared.config.length_penalty = 2.0
roberta_shared.config.num_beams = 4
roberta_shared.config.vocab_size = roberta_shared.config.encoder.vocab_size

```

```

training_args = Seq2SeqTrainingArguments(
    output_dir="./",
    per_device_train_batch_size=batch_size,
    per_device_eval_batch_size=batch_size,

```

```

    predict_with_generate=True,
    #evaluate_during_training=True,
    do_train=True,
    do_eval=True,
    logging_steps=2,
    save_steps=16,
    eval_steps=500,
    warmup_steps=500,
    overwrite_output_dir=True,
    save_total_limit=1,
    fp16=True,
)

# instantiate trainer
def get_trainer(model,tokenizer,tokenized_sum,data_collator,training_args):
    trainer= Seq2SeqTrainer(
        model=roberta_shared,
        args=training_args,
        compute_metrics=compute_metrics,
        tokenizer=tokenizer,
        data_collator=data_collator,
        train_dataset=tokenized_sum["train"],
        eval_dataset=tokenized_sum["test"],
    )

```

```

def my_train_model():
    model_name = "roberta_shared"
    dataset = "scientific_papers"
    subset = "arxiv"
    epochs = 5
    floating_point = True

    token_tuple = robertaTokenize(model_name, dataset, subset)

    model_tuple = weightless_model(model_name, token_tuple[0])

    params = get_my_hyper_params(model_name, epochs, floating_point)

    trainer = get_trainer(model_tuple[0], token_tuple[0], token_tuple[1], model_tuple[1],
                           params)

    trainer.train()

    return trainer

```

```

trainer=my_train_model()

```

```

from transformers import EncoderDecoderModel

roberta_shared = EncoderDecoderModel.from_encoder_decoder_pretrained("roberta-base", "1

```

Some weights of the model checkpoint at roberta-base were not used when initializing.
 - This IS expected if you are initializing RobertaModel from the checkpoint of a
 - This IS NOT expected if you are initializing RobertaModel from the checkpoint of
 Some weights of RobertaForCausalLM were not initialized from the model checkpoint
 You should probably TRAIN this model on a down-stream task to be able to use it for
 The following encoder weights were not tied to the decoder ['roberta/pooler']

```
# load rouge for validation
rouge = datasets.load_metric("rouge")

def compute_metrics(pred):
    labels_ids = pred.label_ids
    pred_ids = pred.predictions

    # all unnecessary tokens are removed
    pred_str = tokenizer.batch_decode(pred_ids, skip_special_tokens=True)
    labels_ids[labels_ids == -100] = tokenizer.pad_token_id
    label_str = tokenizer.batch_decode(labels_ids, skip_special_tokens=True)

    rouge_output = rouge.compute(predictions=pred_str, references=label_str, rouge_type=rouge.get_supported_metrics())

    return {
        "rouge2_precision": round(rouge_output.precision, 4),
        "rouge2_recall": round(rouge_output.recall, 4),
        "rouge2_fmeasure": round(rouge_output.fmeasure, 4),
    }
```

```
training_args = Seq2SeqTrainingArguments(
    output_dir=".",
    per_device_train_batch_size=batch_size,
    per_device_eval_batch_size=batch_size,
    predict_with_generate=True,
    #evaluate_during_training=True,
    do_train=True,
    do_eval=True,
    logging_steps=2,
    save_steps=16,
    eval_steps=500,
    warmup_steps=500,
    overwrite_output_dir=True,
    save_total_limit=1,
    fp16=True,
)

# instantiate trainer
def get_trainer(model, tokenizer, tokenized_sum, data_collator, training_args):
    trainer = Seq2SeqTrainer(
        model=roberta_shared,
```


```
args=training_args,  
compute_metrics=compute_metrics,  
train_dataset=train_data,  
eval_dataset=val_data,  
)  
trainer.train()
```



```

Using cuda_amp half precision backend
The following columns in the training set don't have a corresponding argument in
/usr/local/lib/python3.7/dist-packages/transformers/optimization.py:310: FutureWarning,
FutureWarning,
***** Running training *****
    Num examples = 550000
    Num Epochs = 3
    Instantaneous batch size per device = 256
    Total train batch size (w. parallel, distributed & accumulation) = 256
    Gradient Accumulation steps = 1
    Total optimization steps = 6447
    Number of trainable parameters = 153654873
/usr/local/lib/python3.7/dist-packages/transformers/models/encoder_decoder/modeling_encoder_decoder.py:100:
warnings.warn(DEPRECATION_WARNING, FutureWarning)

```



Step Training Loss

2	12.172600
4	12.127800
6	12.250600
8	12.076300
10	12.069600
12	12.169000
14	12.000000
16	12.012300
18	11.573900
20	11.395800
22	11.115300
24	10.694600
26	10.511600
28	10.261700
30	9.967600
32	9.663800
34	9.373600
36	9.023400
38	8.839300
40	8.578600
42	8.382000
44	8.112800
46	7.982400

48	7.817200
50	7.646000
52	7.554400
54	7.456600
56	7.189300
58	7.226500
60	7.139200
62	7.120800
64	6.976500
66	6.933900
68	6.706700
70	6.749800
72	6.796400
74	6.669100
76	6.618300
78	6.686000
80	6.503600
82	6.482500
84	6.522900
86	6.464800
88	6.438300
90	6.350400
92	6.385400
94	6.207300
96	6.329800
98	6.219300
100	6.248000
102	6.150500
104	6.196600
106	6.048800
108	6.053200
110	6.217800