Fine-tuning of T5-BASE for Summarization

Installing necessary modules

```
!pip install torch==1.13.1+cu116 torchvision==0.14.1+cu116 torchaudio==0.13.1 torchtext==0.14.1 torchdata==0.5.1
--extra-index-url https://download.pytorch.org/whl/cu116 -U
```

```
    Looking in indexes: https://pypi.org/simple, https://download.pytorch.org/whl/cu116
    Collecting torch==1.13.1+cu116
      Downloading https://download.pytorch.org/whl/cu116/torch-1.13.1%2Bcu116-cp310-cp310-linux_x86_64.whl (1977.9 MB)
         ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 2.0/2.0 GB 419.6 kB/s eta 0:00:00
    Collecting torchvision==0.14.1+cu116
      Downloading https://download.pytorch.org/whl/cu116/torchvision-0.14.1%2Bcu116-cp310-cp310-linux_x86_64.whl (24.2 MB)
         ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 24.2/24.2 MB 9.2 MB/s eta 0:00:00
    Collecting torchaudio==0.13.1
      Downloading https://download.pytorch.org/whl/cu116/torchaudio-0.13.1%2Bcu116-cp310-cp310-linux_x86_64.whl (4.2 MB)
         ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 4.2/4.2 MB 95.9 MB/s eta 0:00:00
    Collecting torchtext==0.14.1
      Downloading torchtext-0.14.1-cp310-cp310-manylinux1_x86_64.whl (2.0 MB)
         ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 2.0/2.0 MB 23.9 MB/s eta 0:00:00
    Collecting torchdata==0.5.1
      Downloading torchdata-0.5.1-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (4.6 MB)
         ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 4.6/4.6 MB 90.1 MB/s eta 0:00:00
    Requirement already satisfied: typing-extensions in /usr/local/lib/python3.10/dist-packages (from torch==1.13.1+cu116) (4.5.0)
    Requirement already satisfied: numpy in /usr/local/lib/python3.10/dist-packages (from torchvision==0.14.1+cu116) (1.23.5)
    Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-packages (from torchvision==0.14.1+cu116) (2.31.0)
    Requirement already satisfied: pillow!=8.3.*,>=5.3.0 in /usr/local/lib/python3.10/dist-packages (from torchvision==0.14.1+cu116) (9.4
    Requirement already satisfied: tqdm in /usr/local/lib/python3.10/dist-packages (from torchtext==0.14.1) (4.66.1)
    Requirement already satisfied: urllib3>=1.25 in /usr/local/lib/python3.10/dist-packages (from torchdata==0.5.1) (2.0.7)
    Collecting portalocker>=2.0.0 (from torchdata==0.5.1)
      Downloading portalocker-2.8.2-py3-none-any.whl (17 kB)
    Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dist-packages (from requests->torchvision==0.14.
    Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (from requests->torchvision==0.14.1+cu116) (3.
    Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dist-packages (from requests->torchvision==0.14.1+cu11
    Installing collected packages: torch, portalocker, torchvision, torchtext, torchdata, torchaudio
      Attempting uninstall: torch
        Found existing installation: torch 2.1.0+cu121
        Uninstalling torch-2.1.0+cu121:
          Successfully uninstalled torch-2.1.0+cu121
      Attempting uninstall: torchvision
        Found existing installation: torchvision 0.16.0+cu121
        Uninstalling torchvision-0.16.0+cu121:
          Successfully uninstalled torchvision-0.16.0+cu121
      Attempting uninstall: torchtext
        Found existing installation: torchtext 0.16.0
        Uninstalling torchtext-0.16.0:
          Successfully uninstalled torchtext-0.16.0
      Attempting uninstall: torchdata
        Found existing installation: torchdata 0.7.0
        Uninstalling torchdata-0.7.0:
          Successfully uninstalled torchdata-0.7.0
      Attempting uninstall: torchaudio
        Found existing installation: torchaudio 2.1.0+cu121
        Uninstalling torchaudio-2.1.0+cu121:
          Successfully uninstalled torchaudio-2.1.0+cu121
    Successfully installed portalocker-2.8.2 torch-1.13.1+cu116 torchaudio-0.13.1+cu116 torchdata-0.5.1 torchtext-0.14.1 torchvision-0.14
```

```
# Installing the BLURR library and Bert-Score package

!pip install ohmeow-blurr -q
!pip install bert-score -q
!pip install wandb
!pip install sacrebleu
```

```
         ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 81.1/81.1 kB 1.7 MB/s eta 0:00:00
         ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 507.1/507.1 kB 23.0 MB/s eta 0:00:00
         ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.3/1.3 MB 49.8 MB/s eta 0:00:00
         ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 43.6/43.6 kB 6.5 MB/s eta 0:00:00
      Preparing metadata (setup.py) ... done
      Preparing metadata (setup.py) ... done
         ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 115.3/115.3 kB 15.7 MB/s eta 0:00:00
         ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 134.8/134.8 kB 19.0 MB/s eta 0:00:00
      Building wheel for rouge-score (setup.py) ... done
      Building wheel for seqeval (setup.py) ... done
         ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 61.1/61.1 kB 739.1 kB/s eta 0:00:00
    Collecting wandb
      Downloading wandb-0.16.1-py3-none-any.whl (2.1 MB)
         ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 2.1/2.1 MB 25.6 MB/s eta 0:00:00
    Requirement already satisfied: Click!=8.0.0,>=7.1 in /usr/local/lib/python3.10/dist-packages (from wandb) (8.1.7)
    Collecting GitPython!=3.1.29,>=1.0.0 (from wandb)
      Downloading GitPython-3.1.40-py3-none-any.whl (190 kB)
         ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 190.6/190.6 kB 25.2 MB/s eta 0:00:00
    Requirement already satisfied: requests<3,>=2.0.0 in /usr/local/lib/python3.10/dist-packages (from wandb) (2.31.0)
    Requirement already satisfied: psutil>=5.0.0 in /usr/local/lib/python3.10/dist-packages (from wandb) (5.9.5)
    Collecting sentry-sdk>=1.0.0 (from wandb)
      Downloading sentry_sdk-1.39.1-py2.py3-none-any.whl (254 kB)
         ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 254.1/254.1 kB 33.6 MB/s eta 0:00:00
    Collecting docker-pycreds>=0.4.0 (from wandb)
      Downloading docker_pycreds-0.4.0-py2.py3-none-any.whl (9.0 kB)
    Requirement already satisfied: PyYAML in /usr/local/lib/python3.10/dist-packages (from wandb) (6.0.1)
```

```
Collecting setproctitle (from wandb)
  Downloading setproctitle-1.3.3-cp310-cp310-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_17_x86_64.manylinux2014_x86_64.whl (3
Requirement already satisfied: setuptools in /usr/local/lib/python3.10/dist-packages (from wandb) (67.7.2)
Requirement already satisfied: appdirs>=1.4.3 in /usr/local/lib/python3.10/dist-packages (from wandb) (1.4.4)
Requirement already satisfied: protobuf!=4.21.0,<5,>=3.19.0 in /usr/local/lib/python3.10/dist-packages (from wandb) (3.20.3)
Requirement already satisfied: six>=1.4.0 in /usr/local/lib/python3.10/dist-packages (from docker-pycreds>=0.4.0->wandb) (1.16.0)
Collecting gitdb<5,>=4.0.1 (from GitPython!=3.1.29,>=1.0.0->wandb)
  Downloading gitdb-4.0.11-py3-none-any.whl (62 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 62.7/62.7 kB 10.0 MB/s eta 0:00:00
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dist-packages (from requests<3,>=2.0.0->wandb) (
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (from requests<3,>=2.0.0->wandb) (3.6)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.10/dist-packages (from requests<3,>=2.0.0->wandb) (2.0.7)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dist-packages (from requests<3,>=2.0.0->wandb) (2023.1
Collecting smmap<6,>=3.0.1 (from gitdb<5,>=4.0.1->GitPython!=3.1.29,>=1.0.0->wandb)
  Downloading smmap-5.0.1-py3-none-any.whl (24 kB)
Installing collected packages: smmap, setproctitle, sentry-sdk, docker-pycreds, gitdb, GitPython, wandb
Successfully installed GitPython-3.1.40 docker-pycreds-0.4.0 gitdb-4.0.11 sentry-sdk-1.39.1 setproctitle-1.3.3 smmap-5.0.1 wandb-0.16
Collecting sacrebleu
  Downloading sacrebleu-2.4.0-py3-none-any.whl (106 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 106.3/106.3 kB 1.5 MB/s eta 0:00:00
Requirement already satisfied: portalocker in /usr/local/lib/python3.10/dist-packages (from sacrebleu) (2.8.2)
Requirement already satisfied: regex in /usr/local/lib/python3.10/dist-packages (from sacrebleu) (2023.6.3)
Requirement already satisfied: tabulate>=0.8.9 in /usr/local/lib/python3.10/dist-packages (from sacrebleu) (0.9.0)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.10/dist-packages (from sacrebleu) (1.23.5)
Collecting colorama (from sacrebleu)
  Downloading colorama-0.4.6-py2.py3-none-any.whl (25 kB)
Requirement already satisfied: lxml in /usr/local/lib/python3.10/dist-packages (from sacrebleu) (4.9.4)
Installing collected packages: colorama, sacrebleu
Successfully installed colorama-0.4.6 sacrebleu-2.4.0
```

Import modules

```
import pandas as pd
from fastai.text.all import *
from transformers import *
from blurr.text.data.all import *
from blurr.text.modeling.all import *
from fastai.callback.wandb import *
```

```
/usr/local/lib/python3.10/dist-packages/transformers/deepspeed.py:23: FutureWarning: transformers.deepspeed module is deprecated and w
  warnings.warn(
WARNING:jax._src.xla_bridge:CUDA backend failed to initialize: Found cuDNN version 8302, but JAX was built against version 8904, whic
/usr/local/lib/python3.10/dist-packages/transformers/generation_utils.py:24: FutureWarning: Importing `GenerationMixin` from `src/tra
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/transformers/generation_tf_utils.py:24: FutureWarning: Importing `TFGenerationMixin` from `sr
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/transformers/generation_flax_utils.py:24: FutureWarning: Importing `FlaxGenerationMixin` from
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/blurr/text/modeling/question_answering.py:31: FutureWarning: load_metric is deprecated and wi
  squad_metric = load_metric("squad")
/usr/local/lib/python3.10/dist-packages/datasets/load.py:752: FutureWarning: The repository for squad contains custom code which must
You can avoid this message in future by passing the argument `trust_remote_code=True`.
Passing `trust_remote_code=True` will be mandatory to load this metric from the next major release of `datasets`.
  warnings.warn(
```

```
Downloading builder script:                              4.50k/? [00:00<00:00, 276kB/s]

Downloading extra modules:                               3.30k/? [00:00<00:00, 174kB/s]
```

Load and process data

```
df=pd.read_csv("gpt3_result.csv")
```

```
new_df =df[['cleaned_lyrics','gpt_summaries']].copy()
```

```
train_df=pd.DataFrame(columns=['lyc', 'summ'])
test_df=pd.DataFrame(columns=['lyc', 'summ'])
```

```
lyc_tr=[]
lyc_ts=[]
sum_tr=[]
sum_ts=[]
```

```
i=0
for lyrics,summary in zip(new_df.cleaned_lyrics,new_df.gpt_summaries):
  if i<5500:
    lyc_tr.append(lyrics)
    sum_tr.append(summary)
  else:
    lyc_ts.append(lyrics)
    sum_ts.append(summary)
  i=i+1
```

```
train_df['lyc']=lyc_tr
train_df['summ']=sum_tr
```

```
test_df['lyc']=lyc_ts
test_df['summ']=sum_ts
```

```
train_df["lyc"]="summarize:"+train_df["lyc"]
```

```
pip install tiktoken
```

```
    Collecting tiktoken
      Downloading tiktoken-0.5.2-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (2.0 MB)
        ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 2.0/2.0 MB 17.4 MB/s eta 0:00:00
    Requirement already satisfied: regex>=2022.1.18 in /usr/local/lib/python3.10/dist-packages (from tiktoken) (2023.6.3)
    Requirement already satisfied: requests>=2.26.0 in /usr/local/lib/python3.10/dist-packages (from tiktoken) (2.31.0)
    Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dist-packages (from requests>=2.26.0->tiktoken)
    Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (from requests>=2.26.0->tiktoken) (3.6)
    Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.10/dist-packages (from requests>=2.26.0->tiktoken) (2.0.7
    Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dist-packages (from requests>=2.26.0->tiktoken) (2023.
    Installing collected packages: tiktoken
    ERROR: pip's dependency resolver does not currently take into account all the packages that are installed. This behaviour is the sour
    llmx 0.0.15a0 requires cohere, which is not installed.
    llmx 0.0.15a0 requires openai, which is not installed.
    Successfully installed tiktoken-0.5.2
```

```
import tiktoken

def nt(string: str, encoding_name: str) -> int:
    """Returns the number of tokens in a text string."""
    encoding = tiktoken.get_encoding(encoding_name)
    num_tokens = len(encoding.encode(string))
    return num_tokens
```

Import pre-trained T5-BASE model

```
#Import the pretrained model
pretrained_model_name = "marianna13/flan-t5-base-summarization"
hf_arch, hf_config, hf_tokenizer, hf_model = get_hf_objects(pretrained_model_name,
                                                                  model_cls=T5ForConditionalGeneration)
```

```
    /usr/local/lib/python3.10/dist-packages/huggingface_hub/utils/_token.py:72: UserWarnin
    The secret `HF_TOKEN` does not exist in your Colab secrets.
    To authenticate with the Hugging Face Hub, create a token in your settings tab (https:/
    You will be able to reuse this secret in all of your notebooks.
    Please note that authentication is recommended but still optional to access public mod
      warnings.warn(
```

```
config.json: 100%                              1.53k/1.53k [00:00<00:00, 97.2kB/s]

tokenizer_config.json: 100%                    2.54k/2.54k [00:00<00:00, 71.4kB/s]

spiece.model: 100%                             792k/792k [00:01<00:00, 641kB/s]

tokenizer.json: 100%                           2.42M/2.42M [00:01<00:00, 2.24MB/s]

special_tokens_map.json: 100%                  2.20k/2.20k [00:00<00:00, 154kB/s]

pytorch_model.bin: 100%                        990M/990M [01:12<00:00, 13.5MB/s]
```

Removing **prefix**, as in Inference time, we do not need the prefix.

Note: In above cell, we manually added the prefix to the input sentences.

```
model_hf_config = hf_config.to_dict()
del model_hf_config["task_specific_params"]["summarization"]["prefix"]
hf_config = T5Config.from_dict(model_hf_config)
```

Defining preprocessor

```
preprocessor = SummarizationPreprocessor(
    hf_tokenizer,
    id_attr="id",
    text_attr="lyc",
    target_text_attr="summ",
    max_input_tok_length=1024,
    max_target_tok_length=128,
    min_summary_char_length=30
)
```

```
import warnings
warnings.filterwarnings('ignore')

proc_df = preprocessor.process_df(train_df)
```

```
text_gen_kwargs = default_text_gen_kwargs(hf_config, hf_model, task="summarization")

batch_tokenize_transform = Seq2SeqBatchTokenizeTransform(
  hf_arch, hf_config, hf_tokenizer, hf_model, text_gen_kwargs=text_gen_kwargs
)

blocks = (Seq2SeqTextBlock(batch_tokenize_tfm = batch_tokenize_transform), noop)
dblock = DataBlock(blocks=blocks, get_x=ColReader( "proc_lyc"), get_y=ColReader("proc_summ" ),
splitter=RandomSplitter())
```

```
dls = dblock.dataloaders(proc_df, bs=8)
```

```
summarization_metrics = {
  "rouge": {
      "compute_kwargs": {"rouge_types" : ["rouge1", "rouge2", "rougeL"],
          "use_stemmer": True
      },
      "returns": ["rouge1", "rouge2", "rougeL"],
  },
  "bertscore": {"compute_kwargs": {"lang" : "en"},
  "returns": ["precision", "recall", "f1"]
  },
  "bleu": {"returns": "bleu"},
  "meteor" : {"returns": "meteor"},
  "sacrebleu": {"returns" : "score"}
}
```

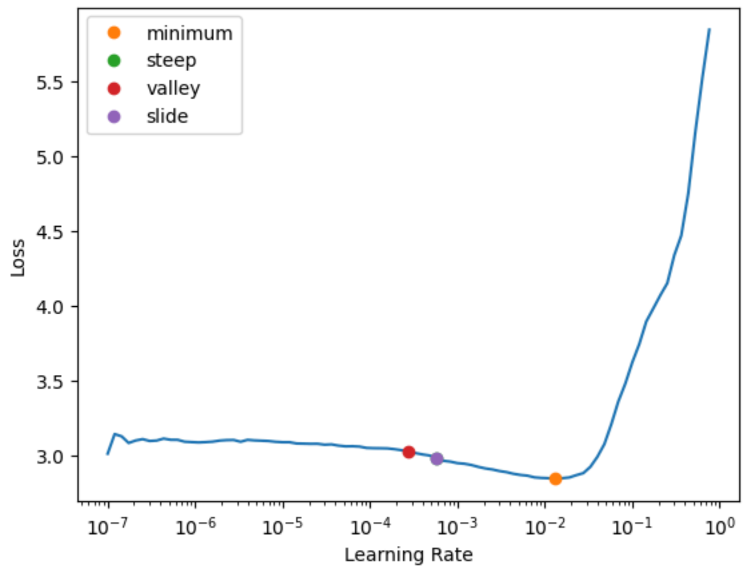
Defining learner to fine-tune the model

```
model = BaseModelWrapper(hf_model)
learn_cbs = [BaseModelCallback]

fit_cbs = [Seq2SeqMetricsCallback(custom_metrics = summarization_metrics, calc_every="epoch")]

learn = Learner(
  dls,
  model,
  opt_func=partial(Adam),
  loss_func=PreCalculatedCrossEntropyLoss(),
  cbs=learn_cbs,
  splitter=partial(blurr_seq2seq_splitter, arch=hf_arch),
)
learn.freeze()
```

```
Downloading builder script:                       5.65k/? [00:00<00:00, 132kB/s]

Downloading builder script:                       8.10k/? [00:00<00:00, 478kB/s]

Downloading builder script:                       6.06k/? [00:00<00:00, 298kB/s]

Downloading extra modules:                         4.07k/? [00:00<00:00, 312kB/s]

Downloading builder script:                       5.35k/? [00:00<00:00, 344kB/s]
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
[nltk_data] Downloading package omw-1.4 to /root/nltk_data...

Downloading builder script:                       7.65k/? [00:00<00:00, 487kB/s]
```

Optimal Learning Rate

```
# Finding Optimal Learning Rate
res = learn.lr_find(suggest_funcs=[minimum, steep, valley, slide])
```



Training the model

```
df = learn.fit_one_cycle(15, lr_max=res.valley, cbs=fit_cbs)
```

60.00% [9/15 3:29:38<2:19:45]

epoch	train_loss	valid_loss	rouge1	rouge2	rougeL	bertscore_precision	bertscor
0	2.361917	2.087417	0.417313	0.132391	0.276684	0.890513	
1	2.075847	1.855376	0.432303	0.143714	0.280130	0.892000	
2	1.925857	1.757139	0.439103	0.148422	0.287608	0.893083	
3	1.780758	1.684265	0.436357	0.149328	0.287427	0.892900	
4	1.693645	1.648512	0.440903	0.149405	0.285588	0.893161	
5	1.605728	1.623009	0.440959	0.152831	0.284772	0.893331	
6	1.524513	1.609964	0.443085	0.153957	0.289030	0.894345	
7	1.453804	1.607615	0.445127	0.154253	0.287743	0.892433	
8	1.384095	1.601479	0.447540	0.155975	0.290279	0.893908	

5.11% [7/137 00:50<15:33 1.3696]

```
config.json: 100%            482/482 [00:00<00:00, 25.8kB/s]
vocab.json: 100%             899k/899k [00:00<00:00, 1.32MB/s]
merges.txt: 100%             456k/456k [00:00<00:00, 643kB/s]
tokenizer.json: 100%         1.36M/1.36M [00:00<00:00, 1.48MB/s]
model.safetensors: 100%      1.42G/1.42G [00:06<00:00, 258MB/s]
```

Sample output

Saving the model

```
learn.metrics = None
learn = learn.to_fp32()
learn.export(fname="cnn_summary_export.pkl")
```

```
infer = load_learner(fname="cnn_summary_export.pkl")
```

```
song=test_df.lyc[5]
ground_truth=test_df.summ[5]
```

```
song
```

```
'intro rza hold stop stop goin goin goin goin holdin goin goin goin holdin goin goin g
oin holdin verse rza could soar sky like bird would disturbed see brothers curb smokin
g herb cops third precinct rush like herd charge delinquent possession greets arrest s
trip search undress make youngster come sign confession would watch cop chopper high p
owered binocular cameras attached bottom aircraft takin photographs lab pointin millim
etre shot magazine full clip gat back hovering projects spotlight shining bright movin
g targets infrared light night dark objects harassing citizens charged yet would order
flock pelicans jam propellers top see forfeit death system failure take devils eyewitn
```

```
infer.blurr_summarize(song)
```

```
[{'summary_texts': 'The song "Hold On" by RZA is a powerful and intense track that showcases the artist\'s lyrical prowess and
determination. The lyrics touch on various themes, including the power of words, the struggles of life, and the consequences of
one\'s actions. RZA\'s verses showcase his skill as a rapper, his ability to dominate the rap game, and his disdain for societal
norms. The song also touches on themes of power, control, and survival in the face of adversity. Overall, "Goin'}]
```

```
ground_truth
```

```
'\n\nThe song "What\'s Goin On" by Gravediggaz delves into various societal issues and
questions the state of the world. The lyrics touch on themes of police brutality, raci
al profiling, government surveillance, and the manipulation of power. The verses paint
a vivid picture of a corrupt system that preys on the vulnerable and perpetuates injus
tice. The song also emphasizes the need for awareness and resistance against oppressiv
```