Ostim Technical University

SENG378 Generative AI Project

Date:21.05.2025

Name-Surname: Yusuf Berk Baytok

ID:220201023

In this report you will see the steps taken to fine tune of a pretrained model.

First, I chosen a model which is GPT2 in this instance

Then I searched for data to train the model.

To start the fine-tuning process, I found a YouTube comments data from Kaggle which contains YouTube comments and their attitude-positive, neutral, negative-.

I trained with distilGPT2 which is smaller version of GPT-2 check out the results. I tried a smaller model to achieve a smaller training time. But the results were unsatisfactory like the distinction between positive and negative comment was not observable.

So, I trained with GPT-2 to improve the outputs. The results were better since the data I trained was not big enough to get the result I desired, I decided to change the data I used.

Finally, I decided on Spotify Million dataset from Kaggle.

So, instead of comment creation, this project naturally evolved into song creation AI.

Challanges:

Noticed I did not properly setup my python environment

Downloaded wrong library so in training it used my CPU not GPU

Fixed it by properly downloading the cuda version of the library

Performance

A)
Training DistilGPT2 using YouTube Comment Data with 10 epochs took approximately took 30 minutes to train

Output Example
Sentiment: Positive | Video Title: Genarative AI | YouTube comment: thank you for this video on machine learning thank you for sharing your experience this is an invaluable video on machine learning thank you for sharing your experience and learning thank you for sharing your experience and learning thank you for sharing your experience and learning thank you for sharing

Sentiment: Negative | Video Title: Generative AI | YouTube comment: the difference between machine learning and classification is that i can predict a specific event with fewer than 1 neuron in a given neuron while training with this sort of training i still often struggle with classification problems in class because of my limited understanding of the nature of the

Sentiment: Neutral | Video Title: Generative AI | YouTube comment: i dont think it will happen anytime soon but i really hope you can watch it in action soon and its fun to watch and you can follow along with the videos for free so i can watch it asap thank you for this amazing video ৱ

B)
Training GPT-2 using Spotify Million Data with 2 epochs took around 1 hour

Example Output:

artist: The Weeknd | song: Under The Rain | lyrics:

Do you remember me

A little bit younger, baby, I mean

My body's a different shade
And my skin is a bit deeper, baby
So my hips are wider, baby

Wear like a dress, ooh
And my fingers are long
Cause I can't move, baby, oh
And my toes are long

Sometimes I keep my feet on my ass
On the street (I think)
If I move, I'm gonna be

(Under the rain)
Under the rain

Sometimes I keep my feet on my ass
On the street (I think)
If I move, I'm gonna be

Oh yeah

You think I'm like you, oh yeah
A little bit younger (Oh yeah, baby)
Baby I'm a little bit older

I'm a little bit younger, baby
Baby you make me cry, baby
Like you make me cry

In the streets
They used to dance for me
A little bit longer, baby
They never wanted me back
And now I'm the same way
They used to dance, yeah
And they used to dance

(Under the rain)
Under the rain

Under the rain

They used to walk for me
Yeah, yeah (Oh yeah, baby)
Like a dog walking, yeah (Baby)

(Under the rain)
Under the rain

References:

Links:

Spotify Million Song Dataset
https://www.kaggle.com/datasets/notshrirang/spotify-million-song-dataset

Youtube Comments Dataset

https://www.kaggle.com/datasets/atifaliak/youtube-comments-dataset

DistilGPT2

https://huggingface.co/distilbert/distilgpt2

GPT-2

https://huggingface.co/openai-community/gpt2