Title Generation using Fine-Tuned GPT Model[1]

Prof. Margarita Osadchy[2]

Faisal Omari[3] – 325616894

Saji Assi[4] – 314831207

University of Haifa

Deep Learning Project

[1] GitHub Repository

[2] Prof. Margarita Osadchy: Computer Science Researcher at the University of Haifa, research interests are Deep learning, Machine Learning, Computer vision, Privacy in Biometric Data, Security and Privacy in Machine Learning Data Science

[3] Faisal Omari: Computer Science M.Sc. Student at the University of Haifa, fields of interest: Machine Learning, Deep Learning.

4 Saji Assi: Computer Science M.Sc. Student at the University of Haifa, fields of interest: Deep Learning, Computer Vision.

Abstract

This project focuses on fine-tuning the GPT-2 model to generate course titles based on specified skills. The GPT-2 model is trained using a dataset sourced from the Coursera Course Dataset available on Kaggle. The fine-tuning process involves training the model to predict relevant course titles given a set of input skills. The trained model is capable of generating coherent and contextually relevant course titles based on the input provided. Through experimentation and evaluation, the effectiveness of the fine-tuned GPT-2 model in generating course titles is assessed. This abstract provides an overview of the project's methodology, highlighting its objective of utilizing the fine-tuned GPT-2 model for automatic course title generation.

Introduction

In recent years, natural language processing (NLP) models have witnessed significant advancements, enabling them to perform a wide range of text-related tasks with remarkable accuracy. One such model is the Generative Pre-trained Transformer 2 (GPT-2), developed by OpenAI. GPT-2 is renowned for its ability to generate human-like text based on given prompts, making it a powerful tool for various language generation tasks.

In this project, we explore the fine-tuning of the GPT-2 model to generate course titles based on a given set of skills. The motivation behind this endeavor stems from the increasing demand for automated text generation systems, particularly in the education sector. There is a need for efficient methods to generate informative and engaging course titles that accurately reflect the content and objectives of the courses.

The primary objective of this project is to harness the capabilities of the GPT-2 model to automatically generate course titles tailored to specific skill sets. By fine-tuning the model on a dataset extracted from the Coursera Course Dataset available on Kaggle, we aim to create a system that can generate contextually relevant and appealing course titles based on input skills provided by users.

Background and Literature Review

The concept of fine-tuning pre-trained language models for specific text generation tasks has gained considerable attention in recent years. With the advent of models like GPT-2, developed by Open-AI, and subsequent iterations, researchers and practitioners have explored various applications of these models in natural language processing tasks.

Fine-tuning involves adapting a pre-trained language model to perform a specific task by further training it on task-specific data. This approach has been shown to yield impressive results across a wide range of text generation tasks, including language translation, summarization, and dialogue generation.

In the realm of education, the generation of informative and engaging course titles plays a crucial role in attracting learners and conveying the content and objectives of the courses effectively. However, manually crafting course titles for a vast array of courses can be a time-consuming and resource-intensive process for educational platforms.

To address this challenge, researchers have proposed leveraging the capabilities of pre-trained language models like GPT-2 to automate the process of course title generation. By fine-tuning these models on a dataset of course descriptions and associated titles, it becomes possible to create a system that can generate contextually relevant and enticing course titles based on input keywords or skills.

Previous studies have demonstrated the effectiveness of fine-tuned language models in various text generation tasks, including content creation, question answering, and dialogue generation. However, to the best of our knowledge, there is limited research specifically focusing on the application of fine-tuned language models for automated course title generation in the context of online education platforms.

In this project, we aim to fill this gap by exploring the feasibility and effectiveness of fine-tuning the GPT-2 model for generating course titles based on input skills or keywords. By conducting a comprehensive literature review and examining existing approaches to text generation and fine-tuning of language models, we aim to gain insights into the current state-of-the-art techniques and identify potential challenges and opportunities in the domain of automated course title generation.

Data Collection and Preprocessing

For this project, we utilized a dataset sourced from Kaggle, comprising course information from the Coursera platform. The dataset is structured with columns including Title, Organization, Skills, Ratings, Review counts, and Metadata, providing comprehensive information about each course.

The dataset consists of 623 courses, each represented by a row in the CSV file. Each course entry includes details such as the course title, offering organization, skills covered, ratings, review counts, and metadata.

Since we have to split the data to train, validation, and test, we chose to split it on the following rate $70\%\ for\ train, 15\%\ for\ val, 15\%\ for\ test$, and that is because the data is not really big, so we couldn't get better results when using the percentage of $60\ /\ 20\ /\ 20$.

Since our objective is to train a language model specifically for course title generation based on input skills, we focused our attention on extracting and preprocessing the relevant data. We narrowed down our dataset to include only the course titles and corresponding skills, disregarding other metadata such as ratings and review counts.

To achieve this, we conducted preprocessing steps on the dataset. First, we loaded the CSV file into a Pandas DataFrame, enabling efficient manipulation and transformation of the data. Then, we created new columns in the DataFrame to represent the input text (course skills) and target text (course title), appending appropriate prefixes to distinguish them.

Subsequently, we saved the preprocessed data into a new text file named 'training_data.txt'. Each line in the text file represents a training instance, with the input text (course skills) and target text (course title) separated by a delimiter ('\n'). This format facilitates easy ingestion of the data for training our language model.

The preprocessing step ensures that the data is properly formatted and structured for training the GPT-2 model on the task of course title generation. By focusing on the relevant information and organizing it in a standardized format, we set the stage for effective model training and subsequent evaluation.

Data Augmentation Consideration:

While exploring methods to enhance our dataset through data augmentation, we encountered unique challenges inherent to the nature of our data. Unlike simpler datasets, such as those comprising textual or numerical data, the relationship between course skills and titles is complex and nuanced. This complexity arises from the intricate interplay between the skills required for a course and the corresponding title, which may not be easily captured by conventional data augmentation techniques.

In particular, augmenting the data could potentially introduce randomness into the dataset, altering the inherent relationships between skills and course titles. This randomness may lead to generated titles that lack coherence or relevance to the given skills, ultimately hindering the effectiveness of the trained model. As a result, while data augmentation is a common approach to enriching datasets, its application to our specific problem domain requires careful consideration and evaluation to ensure meaningful results.

Fine-Tuning the GPT Model

The model architecture for GPT models are different than using a regular neural network like LSTM/FNN/CNN, so here is the architecture we used, we don't have all the control upon the model, so we tried to tune it in the best possible choices, and here is the outcome:

The chose for this number of epochs gave us the best results without overfitting, and the batch size was good even though we were limited by the GPU memory (GTX1060ti 6GB), and the rest of architecture and arguments were chosen according to the common used arguments we found for the model.

During the fine-tuning process of the GPT model, we employed specific configurations and parameters to optimize its performance for generating course titles based on given sets of skills. Here's an overview of the key aspects:

1) Architecture and Initialization: We utilized the GPT-2 model architecture, initialized with pre-trained weights obtained from the gpt2 model variant. Additionally, the corresponding tokenizer was initialized to process the input data.
2) Training Dataset: The training dataset consisted of preprocessed data stored in the train_data.txt, val_data.txt, and test_data.txt files. Each entry in the dataset comprised input text representing course skills and target text representing course titles.
3) Data Collator: To facilitate the training process, a data collator specifically designed for language modeling tasks was employed. This ensured proper tokenization and batching of the input data during training.

4) Training Arguments: Various training arguments were specified to configure the fine-tuning process, including:

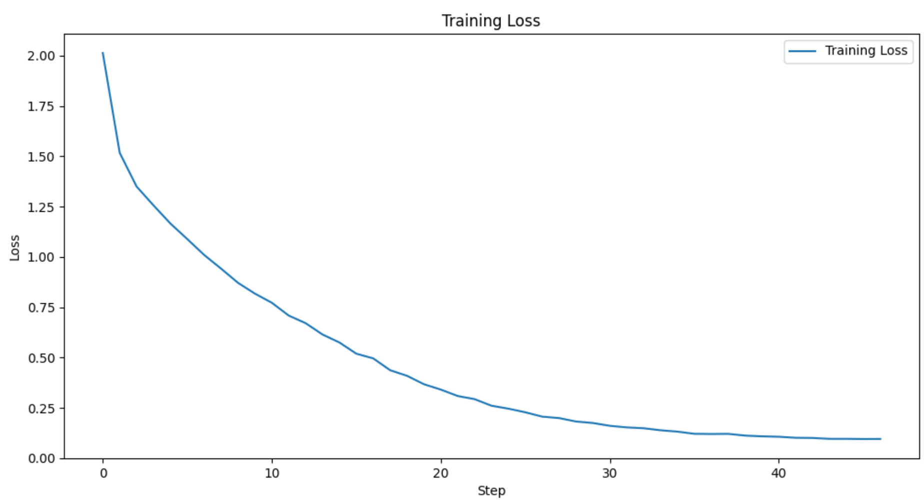
- Number of epochs: 40
- Per-device training batch size: 8
- Save steps: 10,000
- Save total limit: 2
- Logging directory: "./logs"
- Logging steps: 100

5) Training Progress: Throughout the training process, the model's performance was monitored by tracking metrics such as loss, gradient norm, and learning rate. Progress updates were logged at regular intervals, indicating the loss, gradient norm, learning rate, and epoch. For example:

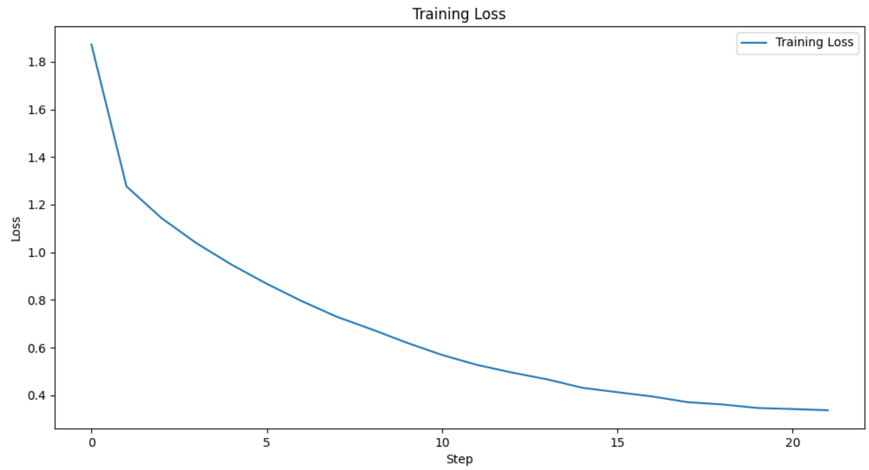
- At epoch 1.47: Loss = 2.0133, Gradient Norm = 8.9696, Learning Rate = 4.8949e-05
- At epoch 70.0: Total training runtime = 1230.9214 seconds, Training samples per second = 15.354, Training steps per second = 3.867, Final training loss = 0.36

6) Training Graphs:

Training Loss: For the first run we realized that there is an overfitting:

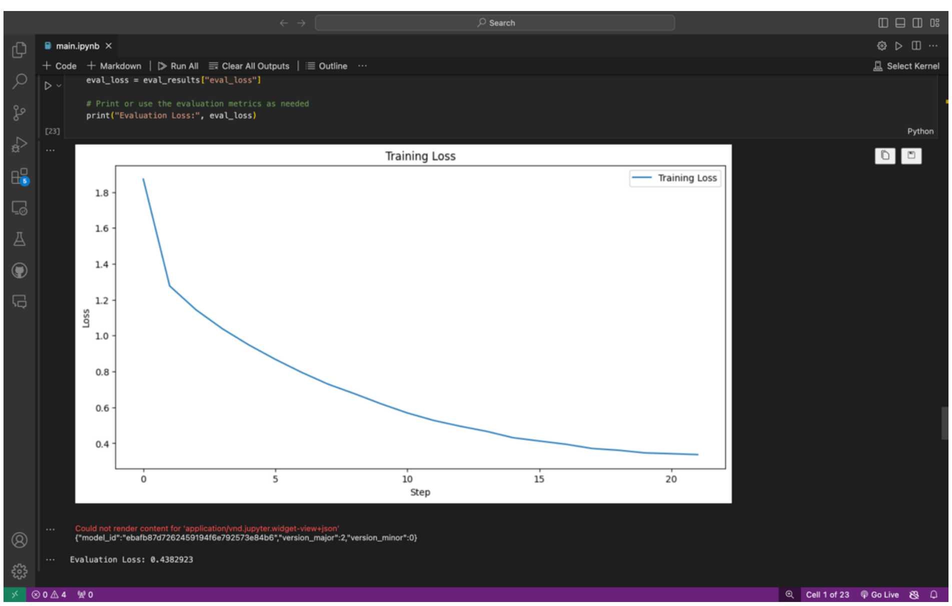


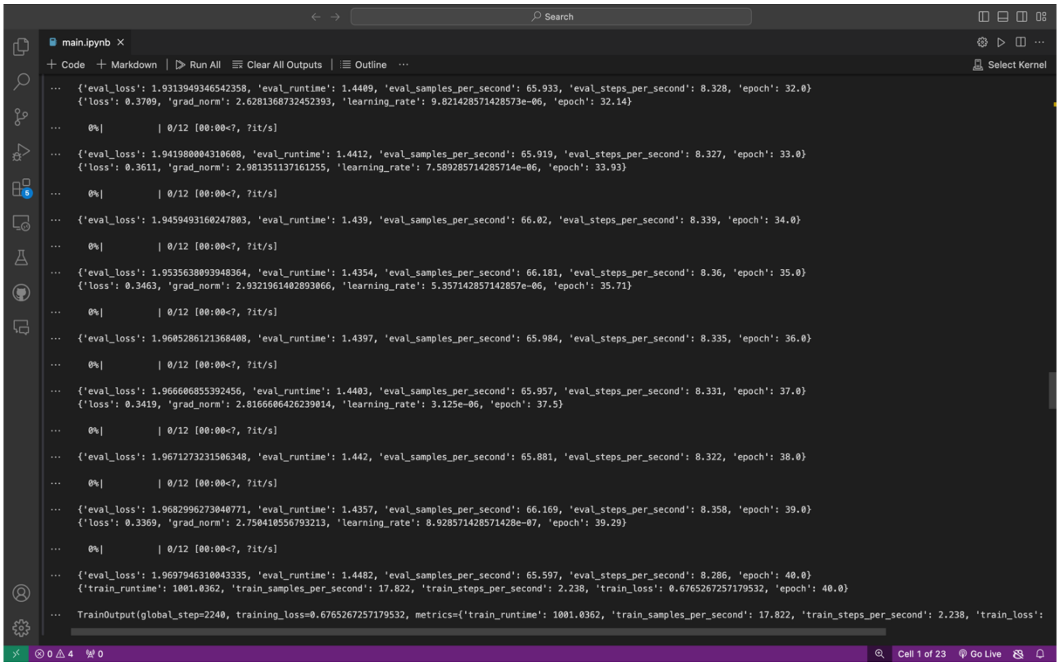
So we’ve changed the model architecture and here is the new graph:



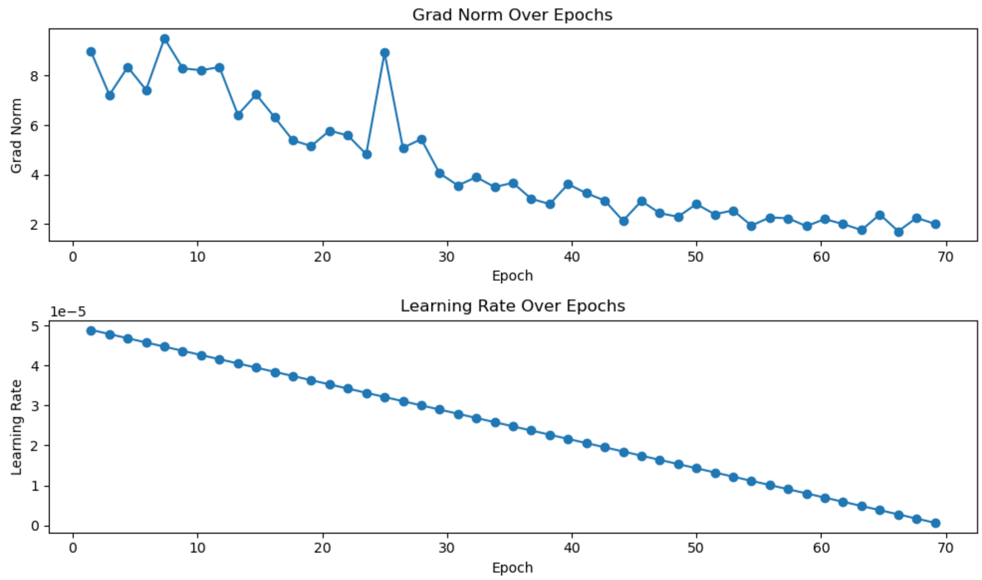
We can see that the training loss is going down, which means that the model is having success on learning the data and the relation between the course titles and its skills.

On the test set we got a loss of 0.4382923, and on the validation, it was good, with the desired loss on the test were good. (look at the main.ipynb Start Fine Tune code section output)





Grad Norm and Learning Rate:



Also here we see that the learning rate is going down because the model is trying to get more accurate results while going up on the epochs.

7) Model Saving: Upon completion of the fine-tuning process, the fine-tuned model and tokenizer were saved to the gpt2-finetuned directory for future use.

Overall, the fine-tuning process aimed to adapt the pre-trained GPT-2 model to the specific task of generating relevant and coherent course titles based on provided sets of skills. The iterative training process gradually optimized the model's parameters, resulting in improved performance and effectiveness in generating accurate course titles.

Results and Discussion

In this section, we present a collection of course titles generated by the fine-tuned GPT model based on input sets of skills. The primary objective is to assess the model's capability to produce relevant and coherent course titles corresponding to the provided skills. By showcasing these examples, we aim to evaluate the effectiveness and performance of the fine-tuned GPT model in generating course titles that accurately reflect the input skills. Each generated course title is accompanied by the corresponding set of skills provided as input to the model.

Examples of Generated Course Titles:

1) The given course skills: Linux, Machine Learning, Python Programming

   The generated title: "Applied Machine Learning in Python"

2) The given course skills: Linux, Cyber Security, C Language

   The generated title: "IT Fundamentals for Cybersecurity"

3) The given course skills: AI, Computer Vision, Image Processing

   The generated title: "Generative AI Fundamentals"

4) The given course skills: React, Full-Stack, JavaScript, HTML, Node-JS, SQL

   The generated title: "Introduction to Full Stack JavaScript"

5) The given course skills: Data Anlaysation, Data Structure, Python, Database Application

   The generated title: "SQL for Data Science Capstone Project"

6) The given course skills: Business Analysis, Communication, Leadership and Management, Project Management

   The generated title: "Modern Project Planning"

We see that the generated titles in almost all the examples are related to the given skills, which means that the model really is generating titles in the right direction.

Discussion of Findings:

Let’s take for example the first pair of generated titles from the previous section:

1) The first example shows that the model can see what is the relation between the Python and Machine Learning, that machine learning is implemented and used with Python code, it may skipped the part of Linux because it sees nothing related to it or that it is included in the generated title, but still if we take this example and compare it with ground truth and real life, we see that the title really can be used as a title for a course with these skills.
2) The sixth example: In the same way, the relation between the given skills and the generated title is crystal clear, this example shows the ability of the trained model to go away from computer science and coding courses, because the generated title in here is really a nice title with attractive words.

Conclusion and Future Work

Conclusion:

The results obtained from the generation of course titles using the fine-tuned GPT model demonstrate promising capabilities in capturing the essence of the input skills and producing relevant titles. Through the evaluation of the generated examples, it is evident that the model can effectively discern the relationships between different skills and articulate them into coherent course titles. The majority of the generated titles exhibit alignment with the input skills, indicating that the model has learned meaningful associations between them. This suggests that the fine-tuning process has enhanced the model's proficiency in generating contextually appropriate course titles based on the provided input.

Future Work:

While the current study has provided valuable insights into the performance of the fine-tuned GPT model for course title generation, there are several avenues for future exploration and improvement. Firstly, further experimentation with different fine-tuning strategies and hyperparameter configurations could potentially enhance the model's performance and generate even more accurate and diverse course titles. Additionally, incorporating larger and more diverse datasets containing a broader range of course titles and skills could enrich the model's training data and improve its generalization capabilities. Moreover, exploring techniques for evaluating the generated titles against human-curated benchmarks or user feedback could provide deeper insights into the model's effectiveness and identify areas for refinement. Finally, investigating the application of the fine-tuned GPT model in other domains beyond course title generation, such as content creation or recommendation systems, could uncover new opportunities for leveraging the model's capabilities in real-world scenarios.

References

- Kaggle Dataset: Jaisai Kamal. "Coursera Course Data." Kaggle, 2022. [Online]. Available: https://www.kaggle.com/code/jaisaikamal/coursera-course-data
- GPT-2:
  - OpenAI. "GPT-2." OpenAI, 2019. [Online]. Available: https://openai.com/gpt-2
  - Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. "Improving Language Understanding with Unsupervised Learning." OpenAI Blog, 2018. [Online]. Available: https://openai.com/blog/better-language-models
- Python:
  - Python Software Foundation. "Python." Python.org. [Online]. Available: https://www.python.org
- GitHub Repository:
  - Faisal Omari. "Title Generation using Fine-Tuned GPT Model." GitHub, 2024.
  - [Online]. Available: https://github.com/faisalomari/title_generation_gpt