COLX581 Team Project “CoverGeni” - A Project Proposal

[Draft by Meiyu, review by Hari, Jaya and Sara]

Introduction

The goal of our project is to build a fine-grained mini-ChatGPT (named “CoverGeni”) , which is designed to generate resumes and cover letters based on job descriptions from the tech field. By nature, it is a language generation task, and it takes the job description as input to a sequence of text and turns it into a structured, certain style of resumes and cover letters. This might involve parameter efficient finetuning, reinforcement learning and prompting engineering to some extent.

Data

We will be working with job descriptions from LinkedIn jobs and Indeed in the tech field (data analyst, data scientist, machine learning engineer roles etc).

Scope of Corpus

- Genre: job description
- Source: LinkedIn and Indeed job postings
- Size: a typical job description consists of around 1000 tokens, we are aiming to collect at least 10000 examples
- Language: English
- Data format: The corpus will be stored as a .csv file

Engineering

- computing infrastructure: Google Colab
- deep learning method: language generation with T5/GPT2
- framework: PyTorch

Previous Works (minimal)

- We have reviewed a similar project which was carried out using T5. The model is a fine-tuned version of t5-base on cover letter samples scraped from Indeed and JobHero. The author has explored the following hyperparameters such as

learning rate, train batch, eval batch size, seed and optimizer etc., which is a great example of taking multiple hyperparameters into consideration into our project. For more information, please refer to:
https://huggingface.co/nouamanetazi/cover-letter-t5-base

- There's another project related to fine-tuned job resumes based on gpt2, which could be a reference as well.
https://huggingface.co/czw/gpt2-base-chinese-finetuned-job-resume

- This one doesn't contain a lot of information but it is based on gpt2 as well.
https://huggingface.co/BigSalmon/CoverLetter

Evaluation

Our main metrics would be BLEU score and perplexity. We want to also include some visualizations of statistics as a vehicle of serving insights from the project. We will see and modify components of our project once we start working on engineering.