The Best Joke Generator

Author: Alyssa Rodriguez

Problem Overview

The problem I am addressing is a problem of generating funny and coherent jokes using a machine learning approach. The goal is to build a joke generator using a Recurrent Neural Network trained on a dataset of existing jokes from Kaggle. The system learns patterns in joke structures and punchlines to generate new jokes. This project can be used for entertainment or inspiration.

Data

I am using a dataset from Kaggle. This is a raw dataset that comes in the form of a csv file. It provides 15,324 jokes for me to work with. I've gone through some of them and they seem like funny puns. The following link I will provide is the source of my dataset.

https://www.kaggle.com/datasets/usamabuttar/dad-jokes?resource=download

Method

I am using an existing library for the Recurrent Neural Network. I am using PyTorch to do this. I have not introduced any new variations or methods for my data just yet but I do plan to in the future. As of now, I will evaluate my results by if they make sense or not. As of now, my model seems to still need a lot of training, but it is learning. I will most likely then compare the jokes it generated to existing one-liner jokes on the internet.

Preliminary Experiments

So far, I have a joke generator that has the spirit, but isn't quite ready to tell a joke. I've cleaned my data of any special characters (there was a lot) and most of the punctuation and unnecessary spaces in the vocabulary.

My loss is good (~0.7) however, the jokes are still pretty much gibberish. This may be because I only trained on 5,000 of the jokes in the dataset, instead of the 15,000 jokes that were provided in the dataset. I did this to shorten the amount of time it takes to train, but will experiment with using more of the jokes in the dataset.

```
print("\n--- Sample Joke ---\n")
print(generate_joke("why did the", length=250, temperature=0.8))
```

```
2.6.0+cu124
CUDA available: True
Epoch 1, Loss: 0.7268
Epoch 2, Loss: 0.7467

--- Sample Joke ---

why did the toilet pair of tick joke? inflation mugger-its computerman leet if thesher houses clean on a pried a factor was to blowd, 'no one candle of car, as a guine. and the pains and himsel
```

As you can see, my model understands the format of jokes, but as far as the content is concerned, it is random.

I believe that more training will lead to better results. I will test with more epochs or possibly more jokes.

Related Work

https://hdsr.mitpress.mit.edu/pub/wi9yky5c/release/3 This paper researches computational techniques for both detecting and generating humor, highlighting how these systems can enhance our understanding of this uniquely human trait. It emphasizes the broader implications of machines grasping humor, setting the stage for advancements in human-computer interaction. This is similar to my project because it explores the implications of machines creating humor, even when machines can't necessarily understand it.

https://aclanthology.org/2020.latechclfl-1.4/ This paper offers an overview of existing systems designed for generating humor, such as jokes and short humorous texts. This also identifies their strengths and weaknesses. The paper also proposes evaluation criteria like humorousness and complexity, providing a structured approach to assess humor generation systems. This differentiates it from other works that may not systematically evaluate computational humor methods. This paper is similar to my work because it discusses using RNN for joke generation.

https://www.sciencedirect.com/science/article/pii/S0306457321000297 This paper introduces a framework for generating jokes that align with moral values. The proposed joke generator aims to produce humor that is both entertaining and ethically appropriate. This focus on the moral dimensions of humor generation sets it apart from other research that may not address the ethical implications of computational humor. While my project isn't focused on the ethical implications of joke generation, I do want to make sure that the jokes that are generated are appropriate.

https://arxiv.org/abs/2405.07280 This paper from Cornell explores the creation of one-liner jokes through multistep reasoning, which results in reconstructing the process behind crafting humorous content. The emphasis on multistep reasoning and empirical evaluation distinguishes this work from others that might not delve into the cognitive processes involved in humor creation.

https://arxiv.org/abs/2409.01232 This paper presents an interpretable framework for humor classification and included multiple humor theories, aiming to bridge the gap between theoretical research and computational detection. By creating proxy features that reflect different aspects of humor theories, the framework achieves a high F1 score

in humor detection tasks. Its approach offers a unique perspective compared to other studies that may rely more heavily on data-driven methods. In contrast to my work, this paper is focused more on theory than actual data.

Division Of Labor

I have undertaken all the research and responsibilities of this project.

Timeline

I plan to do more training and explore more options as much as possible before the deadline. As far as remaining steps go, I just need to modify my code to clean my data better and train my model on more data, all of which can be accomplished in a little over a week.

References

https://arxiv.org/abs/2409.01232

https://arxiv.org/abs/2405.07280

https://www.sciencedirect.com/science/article/pii/S0306457321000297

https://aclanthology.org/2020.latechclfl-1.4/

https://hdsr.mitpress.mit.edu/pub/wi9yky5c/release/3

https://www.kaggle.com/datasets/usamabuttar/dad-jokes?resource=download