Project Intermediate Report 1

> Project GitHub link

> Link to Colab

Team Members

- Rafail Venediktov
- Aliaksei Korshuk
- Panov Evgenii

Current Progress

Dataset

One common dataset used for joke generation is the Reddit Jokes dataset, which contains over 200,000 jokes scraped from the popular social media site.

We managed to find Dataset Card for one-million-reddit-jokes

This corpus contains a million posts from /r/jokes. Posts are annotated with their score.

- 'type': the type of the data point. Can be 'post' or 'comment'.
- 'id': the base-36 Reddit ID of the data point. Unique when combined with type.
- 'subreddit.id': the base-36 Reddit ID of the data point's host subreddit. Unique.
- 'subreddit.name': the human-readable name of the data point's host subreddit.
- 'subreddit.nsfw': a boolean marking the data point's host subreddit as NSFW or not.
- 'created_utc': a UTC timestamp for the data point.
- 'permalink': a reference link to the data point on Reddit.
- 'score': score of the data point on Reddit.

- 'domain': the domain of the data point's link.
- 'url': the destination of the data point's link, if any.
- 'selftext': the self-text of the data point, if any.
- 'title': the title of the post data point.

Model

We will use the **Hugging Face Transformers** library to instantiate a pre-trained model.

https://arxiv.org/abs/1706.03762

Resourses

- Python documentation: **https://docs.python.org/3/tutorial/**
- Transformers library documentation: **https://huggingface.co/transformers/**
- PyTorch documentation: **https://pytorch.org/docs/stable/index.html**
- Deep Learning Book by Ian Goodfellow, Yoshua Bengio, and Aaron Courville: **https://www.deeplearningbook.org/**
- Natural Language Processing with Python book by Steven Bird, Ewan Klein, and Edward Loper: **https://www.nltk.org/book/**
- Dataset: **https://huggingface.co/datasets/SocialGrep/one-million-reddit-jokes**
- Hugging Face Transformers library: **https://huggingface.co/docs/transformers/model_doc/auto**
- GPT-4chan: This is the worst AI ever: **https://m.youtube.com/watch?v=efPrtcLdcdM**
- GPT-4chan is a language model fine-tuned from **GPT-J 6B** on 3.5 years worth of data from 4chan': **https://huggingface.co/ykilcher/gpt-4chan**

Relevant documents and articles

- **"Language Models are Unsupervised Multitask Learners"** by Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. This

paper introduces GPT-2, a large-scale language model trained on a diverse corpus of web text, and demonstrates its impressive language generation capabilities.

https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

- **"GPT-3: Language Models are Few-Shot Learners"** by Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. This paper presents GPT-3, a massive language model that achieves state-of-the-art results on a wide range of natural language processing tasks, including question answering, language translation, and text completion.

  https://arxiv.org/abs/2005.14165

- **"Decoding with Large-Scale Neural Language Models Improves Machine Translation"** by Colin Cherry, George Foster, Ankur Bapna, Orhan Firat, Wolfgang Macherey, and Naveen Arivazhagan. This paper proposes a decoder-only transformer architecture for machine translation, which achieves competitive results on the WMT14 English-German and English-French benchmarks.

  https://arxiv.org/abs/1910.03771

Each Team Member Contribution

- Rafail Venediktov - dataset search, dataset collection and data preprocessing
- Aliaksei Korshuk - search for useful information and resources to solve the problem, namely the study of GPT-2, GPT-3, read papers about Decoder Only Transformer
- Panov Evgenii - creating a model and the initial stage of its training

Each of us actively studied the above articles and learned a lot of new information about how we need to build a bot

Plan for the next three weeks

- Explore limitations
- Create a model and train it
- Read more articles on how to work with Decoder Only Transformer
- Create a complete working model that could play jokes