2024 Fall NLP - Individual Project Report

News Text Generation using Different Advanced Model -Classical and Deep learning Approaches: A Comparative Study

George Washington University

Master of Science in Data Science Program

Course: DATS_6312_10 Natural Language Processing

Professor Name : Dr. Amir Jafari

Group 4:

Keerthana Aravindhan [G34700275]

GitHub Repository: GitHub Link

Introduction:

"In the digital age, the ability to generate creative and coherent text has become increasingly important for engaging news consumers. This project addresses the challenge of generating diverse and contextually relevant text from a few words or sentences as input, leveraging advanced Natural Language Processing (NLP) models.

Our focus is on implementing and evaluating text generation methods, ranging from classical Markov Chains to modern architectures such as Long Short-Term Memory (LSTM) networks and Transformer-based models like GPT-2 and T5. These techniques will be applied to news datasets to explore how well they balance coherence and diversity in the generated outputs.

The insights from this project aim to advance automated text generation, with applications in news writing, personalized content delivery using few seed words or sentences as input.

The report outlines the datasets, methodologies, and evaluation metrics such as perplexity. It also discusses key findings, challenges faced, and potential future enhancements to improve the quality of automated news text generation.

Individual Contribution:

Focused on Data preprocessing to input text according to Transformer Models. Developed a generic GPT-2 model and pretrained the GPT-2 model from scratch. Additionally, I experimented with the BART model to analyze its performance in generating text, comparing its behavior with that of the GPT models.

Development of the Algorithm:

Generic GPT-2:

A pre-trained GPT-2 model was selected to generate text without domain-specific adaptations. Text input was tokenized and passed through the model's transformer-based architecture to generate sequences iteratively. Various decoding strategies, including greedy decoding, beam search, and nucleus sampling, were explored to balance coherence and diversity.

Training GPT-2 from scratch:

Utilizing the GPT-2 model as the foundation for generating domain-specific text. GPT-2 outperformed other models, delivering superior results in terms of fluency and coherence. However, its generic nature necessitated fine-tuning to achieve domain specificity for tasks. A new tokenizer was trained on the entire dataset to create vocabulary specific to the domain. To enhance the model's understanding and representation of specific contexts, special tokens were added. Data was processed in blocks to preserve sequence coherence.

The GPT-2 model was configured to train, bypassing pretrained weights. This approach provided a blank slate for embedding domain-specific features directly into the model. GPT-2 was trained with a causal language modeling (CLM) objective and is therefore powerful at predicting the next token in a sequence. Language Model (LM) head on top of the GPT-2 transformer architecture is added. LMHead essentially means to equip a linear layer to transform the transformer's hidden states into vocabulary-sized predictions, making it perfect for language modeling tasks. The model's vocabulary size was dynamically adjusted to match the tokenizer. Hyper parameters were fine-tuned to optimize training.

The choice of decoding strategy is critical in language model generation. Traditional methods like beam search and top-k sampling can suffer from over-representation of low-probability candidates. To address this, Nucleus Sampling (top-p sampling) was employed. This method samples from a dynamically determined "nucleus," retaining only the most relevant tokens within a cumulative probability threshold (p). This approach enhances the reliability and diversity of the generated text. The fine-tuned model's performance was evaluated using metrics like perplexity and validation loss, indicating the quality of the generated text.

Hyper params are fine tuned while training, to make the model understand the news structure. Learning rate (LR) schedulers were used to increase the LR gradually, since providing large values of LR initially destabilized the model. As an optimizer, AdamW was used, as it provided the weight decay which helped to reduce the overfitting while training.

Generic BART:

The BART (Bidirectional and Auto-Regressive Transformer) model was chosen for its architecture, combination of bidirectional encoding (similar to BERT) and autoregressive decoding (like GPT). The encoder is a transformer model trained to process the input in a bidirectional fashion, understanding context from both left and right of the tokens. The decoder, on the other hand, generates tokens in an autoregressive manner, predicting the next token based on previously generated tokens. To implement the generic BART model:

The pre-trained BART model (facebook/bart-large-cnn) and its associated tokenizer are loaded from the transformers library by Hugging Face. The input prompt is first tokenized, encodes it into input IDs, and passes it to the model for generation. Hyper parameters are used to tune the generation of the model. The generated content is decoded back into human-readable text and adjusted to ensure it ends with a complete sentence.

Fine tuned BART:

To enhance the relevance and coherence of generated texts specific to news contexts, the BART model was fine-tuned on domain-specific data:

News articles from various sources were compiled and preprocessed. Input-output pairs were constructed, where the input was a starting phrase and the output was the corresponding article content.Tokenization was performed using BART’s tokenizer to ensure compatibility with its

vocabulary, and inputs were padded or truncated to maintain a consistent sequence length. Hyperparameters such as learning rate, batch size, and number of epochs were optimized through trial and error. The model was fine-tuned using cross-entropy loss to minimize the difference between predicted and actual token sequences.

Results:

Evaluating the quality and accuracy of the texts generated by transformers is not an easy task, as there is no single metric or criterion to measure them. Human evaluation, which involves asking experts or users to rate the texts on various aspects, is more reliable and comprehensive, but also more costly and time-consuming. Continually improving the quality and accuracy of texts generated by transformers is an active research area.

1. **Perplexity**: Measures how well the model predicts a sequence of words. A lower perplexity indicates better performance.

	Generic BART	Fine tuning BART	Generic GPT-2	Fine tuning GPT-2
Perplexity	1.23	4	7.1	1.69

BART Generated Text:

Covid *is a mobile phone app that allows users to send text messages to friends and family. The app is available in the U.S. and Canada. For more information on Covid, visit www.covid.com. For confidential support on suicide matters call the Samaritans on 08457 90 90 90 or visit a local Samaritans branch, see www.samaritans.org for details. In the UK, call the National Suicide Prevention Line on 0800-273-8255 or visit http://www.suicideprevention.org.*

Fine Tuned BART Generated Text:

covid *news update and the news update is a news update but is the news.Covidnews update to messenger news update could be an arrest for money.comits are the news series. covid business move more than the world of money.*

Fine tuned GPT2 Generated Text:

Covid*-19: More than 500,000 Americans are living with COVID-19. The US Centers for Disease Control and Prevention says the number of Americans with COVID-19 is at its highest level in more than a year. This includes people who have been exposed to certain types or conditions like vaping — which can lead... [Read more...] In August, the CDC announced that it had identified a new group on coronavirus*

with high levels of COVID-19. "As we continue our work with our community partners around the country to protect our citizens from this outbreak," the agency said,. "We continue working together as a community to prevent further spread of this virus.". The CDC said the number of new cases in the United States has risen by more than 6 times since the start of the pandemic, with the number now at a record high. "Despite all efforts being made to reduce the number….we remain committed to keeping Americans safe, and to doing everything we

Artificial Intelligence*: New York Times. A new technology called AI, developed by Google and Microsoft to help us understand the world around us has emerged in the form of AI-powered machines that can make predictions about our future. It's a huge step forward for the future of artificial intelligence... but it will take years of trial and error before we know how to use it. Read the full story..... The New York Times's "The Art Of The Artificial Mind" is on sale now. For more information, visit www.-nytimes.com/artoftheificial-mind/, and follow @nytimes_on..... The New York Times is a global news organization with an international reach and a global reach of over 2 million readers across more than 4.1 billion countries (US).. Its website is www://www2npo.org/, and its news content is available at www.nytimes.com/news/. For more information, visit www.*

Summary:

BART, an encoder-decoder model, is primarily designed for tasks like summarization, translation, and text reconstruction. However, it demonstrated limitations in generating coherent text from an initial prompt or a few words. Due to its architecture, when provided with a prompt such as "Covid," BART encodes the input and generates output from the decoder in a manner that often provides repitation. Its bidirectional encoder-decoder structure makes it more suitable for tasks where the context is well-defined, such as providing content based on a title or heading. This characteristic renders BART less effective for the project's goal of generating coherent and diverse news articles from minimal input.

In contrast, GPT-2 exhibited superior performance for text generation tasks. As an autoregressive transformer model, GPT-2 naturally excels at generating coherent and contextually relevant text based on initial prompts. Even the generic pre-trained GPT-2 model produced better results compared to BART, though its outputs were more generic and less aligned with the stylistic structure of news articles.

Fine-tuning GPT-2 on domain-specific datasets, such as news articles, significantly improved its ability to produce structured, and contextually accurate content resembling actual news pieces. Fine-tuning also reduced issues such as repetition, particularly when sampling strategies like top-k sampling or nucleus sampling were employed.

The perplexity scores of text generated by both BART and GPT-2 were approximately similar, reflecting comparable capabilities in predicting the next word in a sequence. But, BART's outputs often began with a relevant context but deviated significantly as the text progressed. Conversely,

GPT-2 maintained contextual relevance throughout the generated content, producing data closely aligned with the initial prompt and project objectives.

Conclusion:

GPT-2 proved to be a more suitable model for generating coherent, contextually appropriate, and stylistically accurate news articles from minimal input compared to BART. While BART's architecture aligns better with structured input-output tasks like summarization or title-based content generation, GPT-2's autoregressive nature and fine-tuned capabilities make it ideal for open-ended text generation.

Further improvement:

- Use carefully crafted prompts to provide clearer guidance to the model.
- Preprocess Data in such a way to provide X and Y where X will be headlines or title of news article and Y will be the news article. This would help the model to generate fake articles from news titles, where this can be more used as an application.
- Narrow the training dataset's focus to specific topics, such as sports news, rather than diverse news categories. This ensures more domain-specific content generation.
- Increase the number of attention heads or layers in GPT-2 to improve its contextual understanding and generation capabilities.
- Consider leveraging larger versions of GPT-2, such as GPT-2 Large or GPT-2 XL, for more nuanced text generation.

Code Percentage:

Total Lines: 700
Lines from internet : 500
Lines modified : 200
Lines added: 200
Percentage = (500-200)/(500+200) * 100 = 42.857

References:

1. 🤗 *Transformers*. (n.d.). https://huggingface.co/transformers/
2. Schmid, P. (2021, December 15). Fine-tune a non-English GPT-2 Model with Huggingface. *Medium*. https://towardsdatascience.com/fine-tune-a-non-english-gpt-2-model-with-huggingface-9acc2dc7635b

3. Hugging Face. *BART Model Documentation*. Retrieved from https://huggingface.co/docs/transformers/en/model_doc/bart.
4. ProjectPro. *Transformers BART Model Explained*. Retrieved from https://www.projectpro.io/article/transformers-bart-model-explained/553.
5. Illustrated GPT-2 Jalammar, J. (n.d.). *Illustrated GPT-2*. Retrieved from https://jalammar.github.io/illustrated-gpt2/.
6. Yuqian Tan. *Progressive Generation*. GitHub repository. Available at: https://github.com/tanyuqian/progressive-generation. Accessed: December 2024.