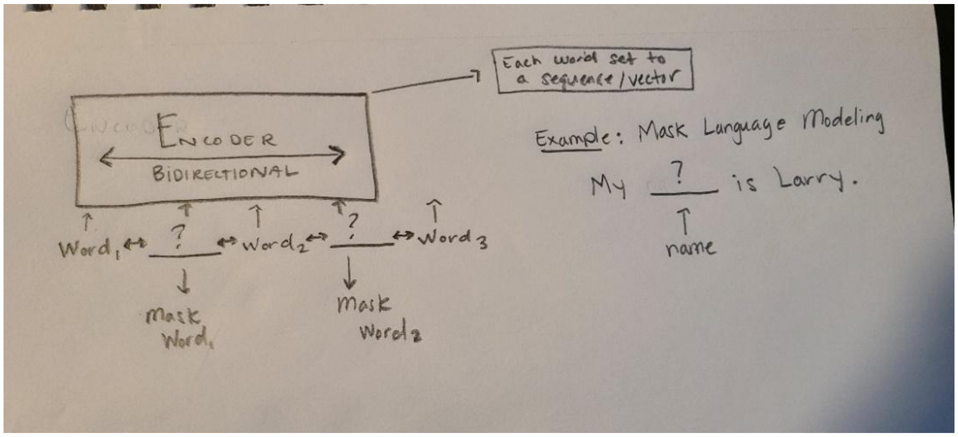
Larry Vue – DS776 Homework 9

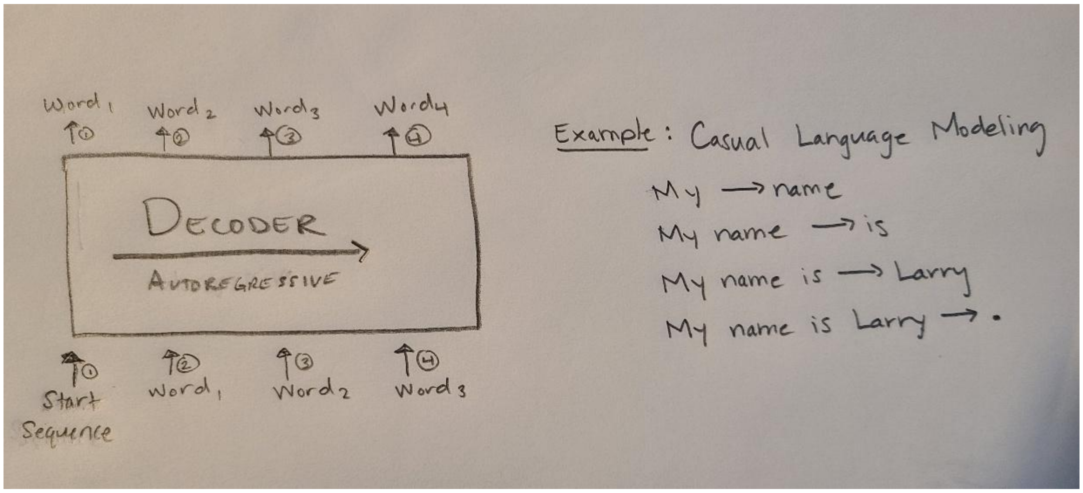
Introduction:

Natural Language Processing (NLP) popularity has grown recently, and chatbots like ChatGPT have risen prominently. What truly is a chatbot, and what more can NLP do? NLP is a transformer-based model with three primary architectural types: encoder-only, decoder-only, and encoder-decoder. They are transformer-based models work with tokens, small text from a context broken into words or letters. To do this, the model uses self-attention to understand each word or token. Self-attention also allows the token to look at other tokens in the sentence and decide which one is most important to the meaning of the context. This helps the model understand how words relate to each other in context.

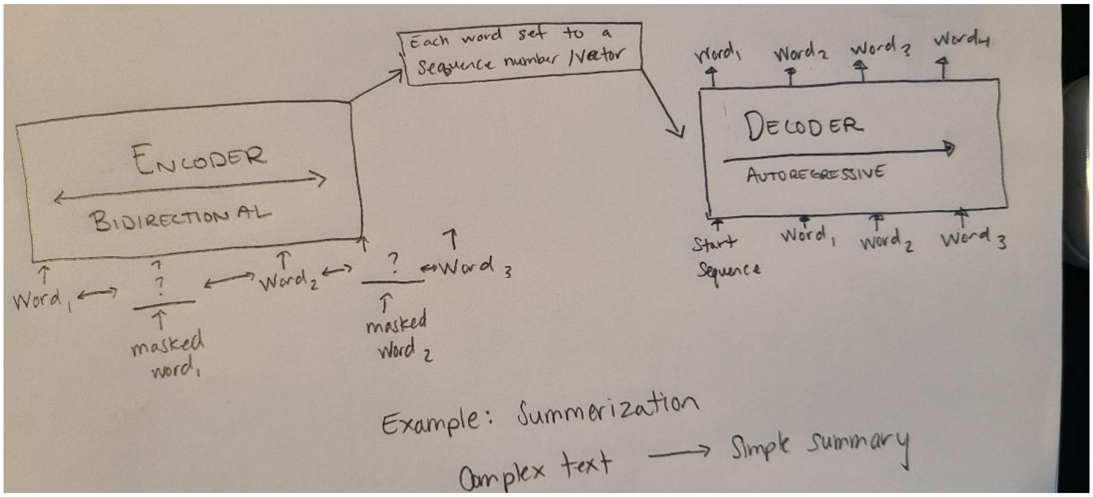
- Encoder-only models like BERT are helpful for information extraction and classification tasks. They use bidirectional self-attention, which means the model looks at both the words before and after a given word to understand the context. Encoder models are trained by hiding some words and asking the model to figure them out or fix the sentence in the example image below. An example of what this model can do is a classification called named entity recognition (NER), which is good at answering extractive questions (Raschka, 2023). Another modern-day use of this is text classification in emails for spam detection. (*Encoder Models - Hugging Face NLP Course*)



- Decoder-only models use just the decoder portion of a transformer. They can generate text one word/token at a time using the words/token that came before it, as shown below in the image. Autoregressive means they can only predict the next word based on what they have already seen, making it a left-to-right process, also called Casual Self-Attention. These models excel at summarization, question-answering, and dialogue systems. Decoders are good at generating language based on prior information they learned, usually one token at a time. Today's examples are the GPT series (GPT2, GPT-4), where decoder-only transformers use text generation like humans do, such as summarization or creative writing. (*Encoder Models - Hugging Face NLP Course*)



- Encoder-decoder models use the transformer's encoder and decoder parts, also known as sequence-to-sequence models like BART and T5. The encoder reads the entire input sentence to understand the meaning. Second, the decoder generates the output sentence one word at a time. To do this, the model uses casual self-attention, which only looks at earlier words it has generated. It also uses Cross-attention, which looks at the encoder's understanding of the input. These models are more advanced as they can learn to replace a random span of text with a token. The model also can learn to fill in missing parts. Today, usage includes machine translation, such as English into French, summarization, and question answering. (*Sequence-to-sequence Models Sequence-to-sequence-models - Hugging Face NLP Course*, n.d.)



Model Architectures and Attention Mechanisms:

BERT, GPT, and BART are transformer architectures that are designed for different tasks. Below, we will compare them and examine their different attention features.

- BERT—Bidirectional Encoder Representations from Transformers only use the encoder portion. It is an NLP model introduced by Devlin et al. (2019) that improved earlier language models and enabled bidirectional pretraining. BERT can look at each word's left and right context during pretraining. Bert uses masked language modeling (MLM), which randomly masks out words in a sentence and asks the model to predict the missing words.
- GPT- Generative Pretrained Transformer (GPT) is an Autoregressive architecture with a decoder-only architecture. GPT is pre-trained in casual language modeling (CLM), which is trained to predict the next word in a sequence using only the words to the left. This is called casual self-attention, which enables natural and coherent language generation. This makes the model suitable for generating text that flows logically. (Brown et al., 2020)
- BART—Bidirectional and Auto-Regressive Transformers is a sequence-to-sequence pretraining framework. It combines the Encoder and Decoder for summarization, translation, and question-answering tasks. BART is trained by corrupting text with various noisy schemes to reconstruct the original text. This will use infilling, sentence shuffling, and token deletion/rotation (Lewis et al., 2019). In the real world, BART is used in scientific research or healthcare to get a quick summary of the core insight from the complex text. (https://www.width.ai/post/bart-text-summarization)

Pretraining Objectives and Key Terminology:

1. *Masked Language Modeling (MLM)—BERT*: In MLM, random words in the sentence are replaced with a mask token. The model is trained to predict the original word by looking at the words before and after it. An example would be that in health care, sentences such as 'diabetes' are masked. BERT can infer it from surrounding words in medical records.

2. *Casual Language Modeling (CLM) -GPT*: CLM trains the model to predict the next word in a sentence using only previous words. This is a left-to-right approach. For example, GPT can help with healthcare patient support chatbots such as "Please arrive tomorrow at" with a time based on earlier words.

3. *Autoregression:* This is in the decoder architecture and refers to predicting the next token using previously generated tokens in a sequence. Autoregression is one word at a time, a text generation in GPT and BART's decoder. An example of autoregression is that I can personalize prescriptions by refilling previous prescriptions needed in their electronic health records.

4. *Denoising Autoencoder (BART):* BART is trained by corrupting text and learning to reconstruct it back to the original text. It combines BERT's encoder with GPT's decoder. (*Encoder Models - Hugging Face NLP Course*) For example, doctors can use BART to summarize long technical reports into easy-to-read summary.

5. *Self-Attention vs. Casual Attention:*

   a) Self-attention, using the BERT/BART encoder, occurs when each word attends to all others to capture the full context. For example, a model could generate a refill reminder based on a patient's electronic health record's passed prescription.

   b) Casual Attention is used in the GPT and BART decoders. In this case, a model will only look back to learn from its past token (words) to predict the next. This is a left-to-right generation of text that shows a coherent generation.

6. *Fine-Tuning:* Fine-tuning is when a pre-trained model is trained for a specific task using a smaller dataset. For example, fine-tuning BERT on a mental health forum can help indicate high-risk behavior and help get better treatment.

7. *Transfer Learning:* Transfer Learning refers to reusing a model trained on one task and applying it to a different task with minimal extra training. For example, a BERT model trained on general English can be fine-tuned to identify patient diagnoses in medical notes.

Fine Tuning Approaches:

BERT, GPT, and BART are models that can be fine-tuned for specific tasks such as text classification, summarization, and answer questioning.

- BERT is an encoder-only model that excels in understanding tasks. For Text Classification, it adds a SoftMax classifier to classify text into categories. BERT uses a named entity recognition (NER) to identify and label specific terms in a text for summarization. This text or entity could be a person's name, location, organization, and more. Lastly, BERT needs to be trained to answer questions by finding an answer in a passage. These answer questions will be trained on the provided text. BERT can be more specialized when fine-tuned on labeled data, which can be based on the need for a specific task.

- GPT is a decoder-only model that is designed to generate text by predicting the next word based on what it has already seen. Fine-tuning for GPT in Text Generation: GPT is trained well to write a structured text based on a prompt. GPT is also good at summarization, as it uses chatbots and dialogues. GPT can hold task-specific conversations to provide advice and summarization. For question answering, fine-tuning GPT can answer questions such as providing clinical answers on drug dosages.
- BART is a powerful model that combines both BERT and GPT. The model transforms one type of text into another. BART's fine-tuning for summarization creates concise summaries from long reports or papers. It can generate strong, coherent outputs with its flexibility in using an encoder-decoder. To answer questions, fine-tune BART and the information found in the given text or document to provide an answer based on the question. Barte uses sequence-to-sequence, which extracts answers and generates answers based on context. Bart can take in complex text such as medical or research text and output a simplified version of the original text. It would preserve the important information but make it easier and faster to understand for everyone.

My field of interest is healthcare and scientific NLP. I would focus on fine-tune BART to help generate patient-friendly summaries of complex medical reports, which usually contain technical language that is difficult for patients to understand. BART's encoder would be used to read and understand the input text, while the decoder generates a simple summary text version. This makes BART a strong choice for such technical text from the healthcare and science fields.

Real-World Application & My Professional Interest:

I am interested in the potential of NLP models to be applied to healthcare and science vocations. Below are ideas for how these models, BERT, GPT, or BART, can be used to solve real-world challenges.

- Applications: Summarizing Radiology and Diagnostic Reports for Patients.
    - This paper focuses on deep learning based on medical imaging report generation and explores different CNN- RNN, attention-based, and reinforced learning-based models. Reading images for radiologists is very time-consuming. Fine-tuning BART can help by making detailed and accurate reports using diagnostic images to train on. This approach uses BART's encoder-decoder architecture to convert features from medical images into descriptive text that copies the quality of human-written reports. BART would be best for this task because it can take complex input, such as CNN models, and generate coherent output to form diagnostic reports (Pang et al., 2023).
- Fine-tuning BERT for Patient/Clinical Use of Medical Question Answering

    - BERT would be best used to create an efficient medical question-answering system. Stanford did this by using a large-scale dataset of over 100,000 human question-answer pairs from Wikipedia articles called SQuAD (Stanford Question Answering Dataset), (Rajpurkar et al., 2016). In healthcare, answering clinical questions while reviewing patient records is time-consuming and error-prone. Fine-tuning BERT with the clinical version of the SQuAD dataset can automate accurate answers. BERT would be ideal for this task because of its bidirectional self-attention mechanism because it can look at both sides of the token to make an accurate prediction.

- AI for Personalized Health Coach
    - People with conditions of obesity, diabetes, or hypertension require more ongoing support and need lifestyle coaching. Fine-tuning GPT to create a conversational AI system that provides personalized health advice and answers patient questions based on medical guidelines. GPT would be best for this task because it can generate easy-to-understand and appropriate responses. GPT can be used to fine-tune and personalize doctor-patient interaction datasets.

Conclusion:

This report covered three transformer-based NLP models: BERT, GPT, and BART. I explored their model architectures, fine-tuning approaches, and strengths/weaknesses. Here are the key takeaways for each application.

- BERT is best for deep understanding of input text, like classification and question answering.
- GPT is good at generating human-like text and having conversations like humans.
- BART is most flexible by combining BERT and GPT for tasks like transformation and summarization.

I am most interested in healthcare and science NLP applications, and BART is the most promising model for achieving my goals. BART is ideal for complex reports in the healthcare/science field and can output simple summaries for everyday people who need them. Another reason is that BART can be adapted to transform complex science procedures or clinical guidelines into step-by-step instructions to make them easy to follow. Integrating these models into clinical and scientific fields will improve decision-making and enhance work and everyday uses of challenging contexts as NLP advances. As these NLPs evolve, they will drive transformative change in many industries.

Work Cited

1. *BART Text Summarization Vs. GPT-3 Vs. BERT: An In-Depth Comparison | Width.ai*. www.width.ai/post/bart-text-summarization.
2. Devlin, Jacob, et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” *arXiv.org*, 11 Oct. 2018, arxiv.org/abs/1810.04805.
3. *Encoder Models - Hugging Face NLP Course*. huggingface.co/learn/nlp-course/en/chapter1/5?fw=pt.
4. Lewis, Mike, et al. “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension.” *arXiv.org*, 29 Oct. 2019, arxiv.org/abs/1910.13461.
5. Pang, Ting, et al. “A Survey on Automatic Generation of Medical Imaging Reports Based on Deep Learning.” *BioMedical Engineering OnLine*, vol. 22, no. 1, May 2023, https://doi.org/10.1186/s12938-023-01113-y.
6. Rajpurkar, Pranav, et al. “SQuAD: 100,000+ Questions for Machine Comprehension of Text.” *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Jan. 2016, https://doi.org/10.18653/v1/d16-1264.
7. Raschka, Sebastian, PhD. “Understanding Encoder and Decoder LLMs.” *Ahead of AI*, 17 June 2023, magazine.sebastianraschka.com/p/understanding-encoder-and-decoder.