CSCE-638, Programming Assignment #3
Rohan Chaudhury
UIN: 432001358

First Paper:

1. Title: "Dense Passage Retrieval for Open-Domain Question Answering"
Authors: "Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih."

Pdf: "https://aclanthology.org/2020.emnlp-main.550.pdf"

2. Task: The NLP task addressed in the paper is that of Open-Domain Question Answering. It is a critical problem in Natural Language Processing, which tries to answer questions in natural language using large-scale unstructured information databases most commonly presented in the form of documents i.e., the inherent goal is to provide a response to queries without the source domain being specified. This task can be divided into 2 stages:
(1) A contextual passage-retrieval system first picks a small subsection of passages that includes the solution to the query, and
(2) a machine reader then completely investigates the recovered context passages to find the correct response.
The DPR model acts as the contextual passage-retrieval system mentioned in the first stage for the open-domain QA task

3. Problem: In open-domain QA, retrieval is typically performed using TF-IDF or BM25, which effectively matches keywords with an inverted index and may be thought of as expressing the query and passage in high-dimensional vectors which are sparse in nature. Such a system would struggle to retrieve sentence contexts which could lead to poor retrieval performance and the **problem** of accuracy degradation of the entire QA system.

4. Solution: Instead of using such sparse representations of the passages, a dense retrieval system might be able to more accurately match and retrieve the relevant context. Dense encodings may also be learned by modifying the embedding functions, allowing for greater task-specific representations. Maximum inner product search (MIPS) techniques on the dense embeddings may be used efficiently for retrieval with particular in-memory data structures and indexing schemes. This is the proposed solution that is explored in this paper.

5. Novelty: Firstly, the authors show that using the conventional pre-trained model BERT and a dual-encoding framework, merely fine-tuning the query and passage encoders on current query-passage pairs is enough to significantly perform better than the conventional BM25 retrieval system. Their experimental findings also indicate that extra pretraining might not be required. Secondly, they confirmed that increased retrieval accuracy does indeed transfer to higher overall efficiency in the instance of open-domain question answering.

6. Evaluation: The authors' suggested Dense Passage Retriever (DPR) is extremely powerful. It surpasses BM25 by a considerable margin (65.2% for DPR vs. 42.9% for BM25 in the Top-5 accuracy), while it also produces a significant increase in end-to-end QA accuracy in comparison to ORQA (41.5% for DPR vs. 33.3% for ORQA in "open Natural Questions") ("Lee et al., 2019; Kwiatkowski et al., 2019").

7. Analysis: Several more experiments and ablation studies were carried out by the authors in order to better understand how various model training settings impact the outcomes. They demonstrated that using a generic pre-trained language model, they can train a high-quality dense retriever with only a limited amount of query-passage pairings. Providing additional training examples consistently increases retrieval accuracy. In their ablation investigation, they discovered that different similarity functions performed equally, thus they selected the simpler inner product function and enhanced the dense passage retriever by learning improved encoders. In essence, according to their empirical research and ablation experiments, more complicated model architectures or similarity functions do not always deliver extra benefits.

8. Thoughts: The authors showed that dense retrieval surpasses and can replace the conventional sparse retrieval element in open-domain question answering, which is a significant step forward in this field of study. While the authors demonstrated that a basic dual-encoder technique can be made to perform fairly well, they also demonstrated that there are several crucial aspects to properly training a dense retriever. Although some improvements can be made to this model architecture, one such improvement is discussed in the next paper that I presented in the following section.

Second Paper:

1. Title: "Dense Hierarchical Retrieval for Open-Domain Question Answering"
Authors: "Ye Liu, Kazuma Hashimoto, Yingbo Zhou, Semih Yavuz, Caiming Xiong, Philip S. Yu"
Pdf: "https://aclanthology.org/2021.findings-emnlp.19.pdf"

2. Task: The NLP task addressed in the paper is that of Open-Domain Question Answering. This work is an advancement on the aforementioned paper on DPR (Dense Passage Retrieval discussed above) for the passage-retrieval part of "open-domain question answering".

3. Problem: Currently, dense neural text retrieval techniques have shown great promise in this domain wherein dense (latent) representations of queries and documents are computed and maximum inner product search between these dense vectors gives the most probable excerpts for the queries during the retrieval process. The **problem** with the current approach lies in the fact that these dense retrieval techniques involve breaking the grounding documents into small chunks that often provide local, incomplete, and occasionally biased context and are heavily dependent on the way the documents are split which may provide erroneous and inaccurate hidden (dense) representations, thus degrading the ultimate retrieval output.

4. Solution: To solve the above-mentioned problem in Open-Domain Question Answering, the authors propose Dense Hierarchical Retrieval (DHR) which is a hierarchical paradigm

that can construct appropriate dense representations of documents by leveraging combined document-wide macroscopic semantics and passage-specific microscopic semantics. A document-level retriever (DHR-D) detects relevant documents first, after which relevant sections are retrieved with the use of a passage-level retriever (DHR-P). By looking at the document-level significance, the ranking of the obtained texts is further fine-tuned. Both DHR-D and DHR-P employ the use of BERT-based dual encoder models.

5. Novelty: In comparison to previous approaches (such as the Dense Passage Retrieval model or DPR), the following are the benefits of using the hierarchical approach and utilizing hierarchical information as discussed in this paper:

- The documents have coarse-grained information which helps in directing the passage-level retriever away from erroneous embedding function outcomes.
- The fine-grained component i.e., the passage-level retriever, will offer the critical ability to locate the relevant indications between comparable passages
- The document-level retriever removes a significant chunk of unimportant and tangential documents, resulting in a significantly quicker inference.

To summarize, they developed a novel hierarchical dense retrieval technique for "open domain question answering" consisting of a document-level retriever and a passage-level retriever that demonstrates better retrieval precision and quicker inference speed. They also employed the hierarchical information of the documents more systematically and logically, resulting in a more relevant and general passage representation that is coherent with its corresponding document.

They also used an iterative training method for DHR-D and DHR-P. In particular, researchers utilized the retriever developed during the preliminary training phase to construct hard negative instances, that could be semantically connected to the query but do not include the answer.

6. Evaluation: As demonstrated in their results, their proposed solutions outperform the existing DPR model in the Open domain question-answering task. They presented the Top-1, 20, and 100 passage-level retrieval accuracy of several techniques (BM25, BM25*, DPR, DPR*, and DHR, where * represents the same model trained on the data preprocessed using their technique for a fair comparison) on four Question answering datasets (Natural Questions, TriviaQA, WebQuestions, and CuratedTREC). The performance of DPR* is either better or comparable to DPR (better than BM25 and BM25*) which shows the benefits of their novel preprocessing technique and the performance of their model DHR is much better than both DPR and DPR* showing the obvious benefits of their model architecture.

7. Analysis: The paper also conducts an ablation study to further understand how each component of the architecture works. In their experiments, they showed that:

- Doc-level retrieval accuracy beats the BM25 results, demonstrating the effectiveness of their dense document-level retriever.
- Using the Doc-level retriever to filter the documents first to a small selection of pertinent documents would not impair final retrieval performance but will aid in filtering out certain answer-irrelevant documents. Furthermore, the rerank approach beats DHR without the reranking technique, demonstrating the importance of reranking.

8. Thoughts: I find the hierarchical structure intuitive and I believe that is how normal humans would look for relevant answers to topics. For a particular query, we would first search for the answer in the documents which have similar titles or summaries and then look for the answer in the passages of the relevant documents. This model also does something similar in the sense that it first looks for pertinent information for the queries in document summaries and then looks at the actual answer in the passages of the relevant documents. Re-ranking the retrieved passages with respect to the relevance of the document summaries also proves to be effective in the final stage of the process.

Third Paper:

1. Title: "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"
Authors: "Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova"
Pdf: "https://aclanthology.org/N19-1423.pdf"

2. Task: This paper is in the domain of language representations models and they introduced a novel language representation model that they called "BERT" ("Bidirectional Encoder Representations from Transformers").

3. Problem: One of the main reasons for BERT's success is that it utilizes a context-based embedding model, as opposed to a context-free embedding model like word2vec. Context-free models used previously provided the same embeddings for a word even when used in different contexts. This was a huge problem in the domain of Natural Language Understanding. Also, prior to BERT, standard language representation models such as OpenAI GPT were inherently unidirectional. This restricts the number of topologies available for pre-training. In the self-attention layer of the transformer in OpenAI GPT, each token only could attend to the preceding token.

4. Solution: Following are the ways in which this paper solves the above-mentioned problems:
Bert is inherently a context-based model, which means it will grasp the context before generating the embedding for the word. Furthermore, BERT was developed in order to pre-train the deep bidirectional representations from the unlabeled text data by focusing on the right and left contexts in all the layers. "BERT" addresses the above-mentioned unidirectionality restriction by pre-training with a Masked Language Model (MLM). This arbitrarily masks part of the input tokens, and the model's goal is to estimate the lexical ID of the masked word depending only on context. This allowed the representations to integrate the left and right contexts, allowing for the pre-training of a bidirectional model.

5. Novelty: BERT is offered as a method for pretraining deep bidirectional representations from unsupervised data by focusing on both the left and right context simultaneously at all levels. This bi-directionality is novel in the paper as earlier work only considered unidirectional representations. This pre-trained "BERT" model may be fine-tuned with only one extra output layer to construct cutting-edge models for a variety of tasks, including

query answering and linguistic inference, without requiring significant task-specific architectural changes. This is a type of self-supervised learning where the main task is language model learning and the particular tasks are downstream tasks that require fine-tuning.

6. Evaluation: BERT is theoretically simple but very effective. It achieved state-of-the-art results when it was published on 11 natural language processing tasks, including increasing MultiNLI precision to 86.7% (the absolute improvement amounting to 4.6%), increasing the GLUE score to 80.5% (the absolute increase amounting to 7.7%), increasing SQuAD v2.0 Test F1 score to 83.1 (the absolute improvement by 5.1 points), and increasing SQuAD v1.1 question answering Test F1 score to 93.2 (the absolute improvement by 1.5 points).

7. Analysis: They conducted different ablation studies to demonstrate the effectiveness of their architecture. From these studies, they found that removing Next Sentence Prediction and using only a left-to-right Language model (similar to OpenAI GPT) degrades the accuracy of the architecture significantly thus proving their importance. It was also the very first study to show conclusively that growing to high model sizes leads to significant gains even on very small size tasks, assuming the model has been pre-trained properly.

8. Thoughts: BERT is among the most widely utilized cutting-edge text embedding models. It has changed the landscape of NLP tasks. BERT enables the very same pre-trained model to effectively perform a wide range of NLP tasks such as text categorization, similarity detection, and so on. Almost any NLP task uses the BERT model to generate the embeddings before plugging them into the downstream task. Personally, I have also used it in several of my NLP projects and found it very effective.

Fourth Paper:

1. Title: "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension"
Authors: "Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer"
Pdf: "https://aclanthology.org/2020.acl-main.703.pdf"

2. Task: The paper discusses a method for denoising the pretraining of a sequence-to-sequence architecture for Natural Language Generation.

3. Problem: Self-supervised approaches have proven to be quite efficient in a broad variety of NLP applications. The most effective techniques have been masked language models, which are denoising auto-encoders trained to reconstitute text with a randomly selected subset of the phrases masked out. Recent research has demonstrated advantages by enhancing the allocation of masked tokens, the sequence wherein masked tokens are anticipated, and the context provided for substituting masked tokens. However, because these approaches often focus on specific sorts of end tasks (for example natural language generation, span prediction, and so on), their application is **limited**.

4. Solution: In this paper, the authors have presented BART, which essentially pre-trains a model integrating Bidirectional and Auto-Regressive Transformers. With a sequence-to-sequence model, BART is a denoising autoencoder that may be used for a highly diverse variety of end jobs. Two steps comprise pretraining: A "sequence-to-sequence" model is taught to reconstruct the original text after it has been distorted using an arbitrary noise function in step one. Despite its simplicity, BART's conventional Transformer-based neural machine translation architecture may be seen as generalizing "BERT" (because of the bidirectional encoder), "GPT" (because of the left-to-right decoder), and several other more modern pretraining approaches. The noising flexibility of this system is a crucial feature; any modifications may be made to the original text, including modifying its length.

5. Novelty: BART presented in this paper is basically almost similar to conventional sequence-to-sequence transformer architecture but with a few tweaks. They are as follows:
1. It has varied layer sizes. The basic version has 6 encoder layers, and 6 decoder layers, and the large version has 12 encoder layers and 12 decoder layers.
2. Instead of using ReLU it utilized GeLU
3. At the extreme end, there is no feed-forward network.
4. Each decoder layer also conducts cross-attention with the output from the last encoder hidden layer.

Although the model structure is straightforward, the primary contribution of this research is a thorough examination of the numerous pretraining activities.

6. Evaluation: BART performs well for comprehension tasks as well as text generation after fine-tuning. It equals RoBERTa's performance on SQuAD and GLUE while achieving state-of-the-art outcomes on a variety of summarization, dialogue, and query-answering challenges, with improvements of up to 3.5 ROUGE. "BART" also outperforms a back-translation method for machine translation by 1.1 BLEU with only target language pretraining.

7. Analysis: The authors also investigated which pretraining activities are efficient. The authors conduct this study by comparing 5 different models trained with equivalent pretraining tasks to 6 downstream tasks. The following is a summary of the author's findings:
1. BART routinely outperforms other models.
2. The masking of tokens is important
3. The pre-training goal is not the sole significant aspect. Architectural considerations such as segment-level repetition, relative position embeddings, and so on are important as well.
4. Pretraining technique performances vary greatly between tasks
5. SQuAD requires bi-directional encoders.
6. Pure language models outperform other models on the ELI5 downstream task.
7. Pretraining from left to right boosts natural language generation.

8. Thoughts: BART is very expressive and currently provides state-of-the-art performance. The reason for that is the fact that it combines the finest of both worlds by encoding the distorted texts with BERT and producing the actual document by anticipating the masked tokens using GPT. BART may be adjusted to perform well in a variety of downstream tasks

such as Token Classification, Sequence Classification, Machine Translation, and Sequence Generation.

Fifth Paper:

1. Title: "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks"
Authors: "Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, Douwe Kiela"
Pdf: "https://arxiv.org/pdf/2005.11401.pdf"

2. Task: The task of natural language generation is investigated in this study. It focuses on a general-purpose fine-tuning strategy for "retrieval-augmented generation (RAG)" models that integrate pre-trained parametric with non-parametric memory for natural language synthesis.

3. Problem: When fine-tuned on downstream NLP tasks, huge pre-trained language models have already been demonstrated to contain factual information in their parameters and attain state-of-the-art outcomes. Unfortunately, their capacity to retrieve and accurately modify knowledge remains restricted, and as a result, their effectiveness on knowledge-intensive applications falls behind task-specific models. Furthermore, giving evidence for their conclusions and upgrading their global knowledge remain unresolved research issues. These models cannot simply grow or modify their memories, cannot give easy insight into their forecasts, and may result in "hallucinations" while producing outcomes.

4. Solution: To address these aforementioned concerns, the authors developed RAG models in which the parametric memory is a pre-trained sequence-to-sequence model and the non-parametric memory is a dense vector indexing of Wikipedia retrieved using a pre-trained neural retriever. Because information can be immediately edited and expanded, and accessible knowledge can be reviewed and evaluated, hybrid systems that mix parametric memory with non-parametric memories can solve some of these previously stated challenges. Although they have only studied open-domain extractive question answering, "REALM" and "ORQA", two recently presented models that integrate masked language models with a variational retriever, have demonstrated encouraging results. So, the authors in this paper have used both parametric and non-parametric memory to improve sequence-to-sequence (seq2seq) models, the "workhorse of NLP".

5. Novelty: The authors develop a general-purpose fine-tuning method they call retrieval-augmented generation (RAG) to provide pre-trained, parametric-memory generation models with a non-parametric memory. They create RAG models using Wikipedia's dense vector index serving as the non-parametric memory, accessible by a pre-trained neural retriever, and a sequence-to-sequence transformer serving as the parametric memory. They incorporate these parts into an end-to-end trained probabilistic model. The retriever "(Dense Passage Retriever, DPR)" returns latent documents based on the input, and the seq2seq model "(BART)" subsequently takes these latent documents in addition to the input to produce the output.

6. Evaluation: The findings of the research emphasize the advantages of integrating parametric and non-parametric memory alongside natural language generation for knowledge-intensive activities, which people cannot be reasonably expected to accomplish without having access to an exterior information source. The suggested “RAG” models surpass current techniques that apply specialized pre-training goals on “TriviaQA” and attain state-of-the-art performances on “open Natural Questions”, “WebQuestions”, and “CuratedTrec”. The authors discover that unconstrained generation beats earlier extractive methods despite the fact that these are extractive challenges. The authors tested their models for knowledge-intensive generation using “MS-MARCO” and “Jeopardy question” creation, and they discovered that their models produce replies that are more accurate, detailed, and varied than a “BART” baseline. They produced results for “FEVER fact verification” within 4.3% of state-of-the-art workflow models that incorporate robust retrieval oversight.

7. Analysis: The authors discovered that “RAG” models create more detailed, diversified, and accurate language than a cutting-edge parametric-only seq2seq baseline for language generation tasks. They also show how non-parametric memory may be updated to keep the models' information up to and current as society changes. They discovered that individuals prefer “RAG” generation over solely parametric “BART” because “RAG” is more accurate and detailed. They thoroughly examined the learned retrieval component, confirming its efficacy, and they demonstrated how the retrieval index may be runtime replaced to update the model without the need for retraining.

8. Thoughts: The fact that this work is more firmly rooted in actual factual information (in this example, Wikipedia) than earlier works makes it less likely to "hallucinate" with more factual generations and provides greater control and interpretability. With immediate societal benefits, RAG might be used in a wide range of situations. Along with the benefits, there are also possible drawbacks, including the likelihood that Wikipedia and other external information sources may never be perfectly true and impartial. Considering that RAG may be used as a language model, akin worries to those for “GPT-2” are legitimate here, albeit perhaps to a smaller amount, such as the possibility that it will be used to produce offensive, false, or deceptive content in the media.