CS 4395.001
Motalib Rahim
Mxr170012
Dr. Karen Mazidi
Portfolio: ACL Paper - Summary

Training Data is More Valuable than you Think:

A Simple and Effective Method by Retrieving from Training Data

The paper is a comprehensive overview of the formidable research of: Shuohang Wang, Yichong Xu, Yuwei Fang, Yang Liu, Siqi Sun, Ruochen Xu, Chenguang Zhu and Michael Zeng. The research was under a well-known company called Microsoft for their Azure Cognitive Service. Training data to this day is already a very powerful technique that satisfies many impressive business models. In the paper the team plans to utilize the training data with some external knowledge. They discovered that retrieval-based methods to be great in natural language processing tasks. With the goal of achieving a more effective, efficient and a better performance. This was made possible through the creation of **REINA**, aka **RE**trieving from the train**IN**g dat**A**.

REINA has shown that it can benefit natural language generation and natural language understanding significantly. To be specific the focus was on tasks such as summarization, machine translation, language modeling and question answering. Integrating REINA and utilizing 11 viable datasets in total resulted in significantly better performance compared to similar pre-trained models. The 11 datasets for each of the tasks (respectively) used were: 1) Multi-News, WikiHow, Xsum, NEWSROOM and BigPatent 2) WikiText2, WikiText103 3) WMT16 (en-tr), WMT16 (en-de) 3) CQSA, PIQA, aNLI. For summarization tasks with REINA got state of the art performance for datasets Xsum, CommonsenseQA (leaderboard No.1) and Big Patent. Other 3 summarization results showed that BART-base matches with that of BART-large which is impressive as there are more parameters. For fair comparison similar pre-trained model was used and REINA outperformed all. Furthermore, the retrieval process of REINA can simply scale up through leveraging labeled data from different datasets. This also outperforms baselines trained on exact dataset.

The team's method or process falls under supervised learning of machine learning where it makes use of labeled datasets to train algorithms. Where then it produces prediction and classification of data. The team emphasize on the effectiveness of their approach that greatly supports supervised learning. With leveraging training data there can generate high quality results. Their perspective was as there could be many parameters, it is not feasible to for a model to capture everything from a training data. The resources required for that would be massive. To mitigate the hurdle of the training data, they approach it by recapturing related training data. As a result, it

gathers necessary information that enhances model performance at inference. Therefore, the model is created. Firstly, most matched instances with labels are retrieved. The retrieved data is concatenated with input sequence. Then it is fed into the model for output generation.

Based upon BM25, the retrieval engine was built for its speed. The methods are designed as a mathematical function for all the tasks or even processing. For the retrievable, the corpus in indexed into a list of key-value pairs. Then when an input is taken the retrieval engine epsilon matches with all keys and return top K most similar keys. The input is the fed into the model to generate output. The terms/data inside function can be concatenated but data for different tasks would be different format design. So, for every task they make different combinations of function f for each of the tasks: 1) Summarization - the purpose is to generate a summary given a document. So, they build an index for the pairs document and summary where document is the key and summary the value. 2) Language modeling - they use Seq2Seq where given a chunk predict next chunk of text. Given an input the most similar keys are looked for in the index and prepend corresponding chunk. 3) Machine translation - they translate text from to source to target language. While only concatenating the retrieved text in target for faster processing. 4) Question Answering – This is a bit complex as it requires the machine to generate answer which is done by commonsense knowledge. The concatenation is indexed for question and ground truth choice. So, for an input, model is given choices which is then concatenated with every choice as a query and related training instances get retrieved. The function concatenates both retrieved question and answers with input. Then model predicts and output.

With the above REINA is then evaluated through the datasets which are well known. REINA can handle long document retrieval without training unlike dense passage retrieval-based method. What makes it more efficient is that REINA only needs retrieval once per chunk. For optimization their information retrieval is based on Lucene Index and model training on transformers library. The experiments conducted on an 8-GPU machines. For different tasks tuning was required whether it be for hyper-parameters, baseline, method, length and appending. As they conducted experiments, their metric was based on Rouge-1/2/L scores. As mentioned, they achieved state of the art and could beat larger models such as REINA (BART-base) beats PEGASUS-LARGE and BART-large on BigPatent and WikiHow data. They find that while REINA being trained on only target task it can reach utmost effectiveness. After the experiment they can see drastic

improvements and more room for improvement via different models of sequence to sequence.

Overall, the authors have conducted a great job which led to valuable outcome. The creation and implementation of REINA has shown with proven results, significant gains in terms of performance, effectiveness, and efficiency. This outcome of was their main motive behind the research. The best part is as mentioned REINA can be integrated in any different model for various different tasks with ease. The authors paper based on the impressive work has contributed to 31 citations with Yang Liu having the most personal citation of 1181.