Sri Ram Eswaran (srirame2@illinois.edu)

CS410 – Text Information Systems – Technology Review (For Fourth Credit)

Text-To-Text Transfer Transformer T5 Google

Motivation:

There are a number of recent developments in deep learning, text mining and natural language processing. Great improvements in these domains were accelerated by models like BERT (Bidirectional Encoder Representation), GPT (Generative Pre-trained Transformer) series and the latest, being T5 transformer. Thus, I would like to provide a summary of the above models and talk more about the Text-To-Text Transformer T5 model by Google. This can be a one-stop for people who intend to get some basic grasp on industrial deep learning models.

Pre-Trained Models and Transfer Learning:

Deep learning models tend to be super-large, that account for their ability to have higher accuracies. The number of parameters is of the order of millions. Training such models would be time and resource consuming. Thus, on-the-fly training for real-time operations is not feasible. Thus, pre-trained models come with models that are 'pre'-trained on a huge model and then, their parameters are fine-tuned to suit a specific task. Such models start to define the future of deep learning models. Transformer models are the most sought-out kind of such techniques. This transfer of knowledge from a huge learning set to its application on a fine-grained task is the primary motivation behind the emerging domain of transfer learning.

BERT and the GPT series:

Pre-Trained models can come with context-free or context-oriented representations. Word2vec model is a sample context-free model that embeds words in a corpus vocabulary. The consequence is that, a term like "bat" which has two meanings (one meaning refers to a sports equipment and the other is a bird) will have one embedding irrespective of its context. But, the BERT model would represent the two meanings separately and therefore, allow for more contextual directionality. BERT uses the technique of masking certain tokens and condition every other token to predict those masked terms in a bidirectional manner.

Generative Pre-trained Transformers are language models that produces text that resembles human language generation. Once the model has an initial trigger, the model shall resume a meaningful continuation of the trigger. A classic example that emphasizes the power of such a model is the following: Given the trigger, "Really liked this movie!", the GPT-2 generated the following text [4]: "Loved the character's emotions at being in constant danger, and how his inner fears were slowly overcome by these events. Also loved that he is so focused on surviving; even for a while it felt like something out of Batman v Superman was showing up every now again because you always knew your enemy would show up eventually anyways :) The ending theme really stuck with me too… I mean yeah, they did

have to…". As we can see, in spite of the trigger not having any explicit token as a review, the model understood that and generated a suitable continuation. There are variations in GPT, including GPT-2, GPT-3 where GPT-3 is more robust and can cover and handle more subtle topics of text generation.

T5: The Text-To-Text Transfer Transformer

While BERT consisted of encoder blocks and GPT-2 consisted of decoder blocks, the Text-To-Text Transfer Transformer (T5) consists of a combination of both. The model uses the technique of converting any language problem into a text-to-text problem.

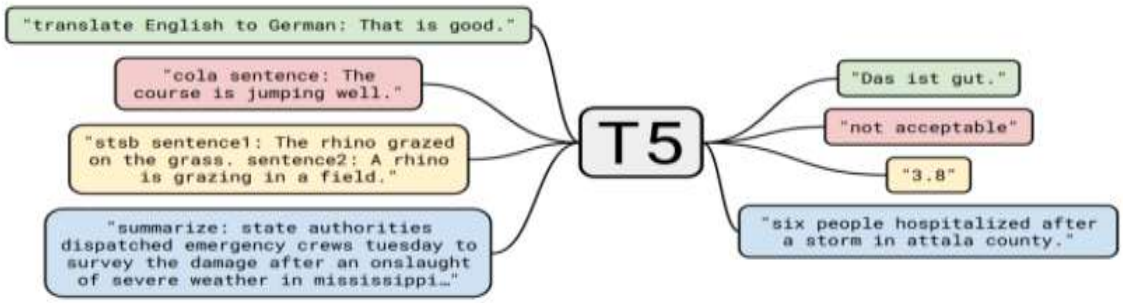

Fig 1: T5 Transformer Model Architecture [3]

The above figure portrays how T5 works to our needs. Multifarious tasks like language translation, summarization, classification can be dealt by the model depending on the input. Thus, almost tasks of deep learning are attempted to be converted to a text-to-text problem.

The training happens as follows: the model is trained to generate text output, for all kinds of tasks including classification and regression. To fulfil the requirement of a uniform dataset that could serve multiple tasks, a Colossal Clean Crawled Corpus (C4) was created. It is significant to note that it is twice the order of magnitude of Wikipedia dataset.

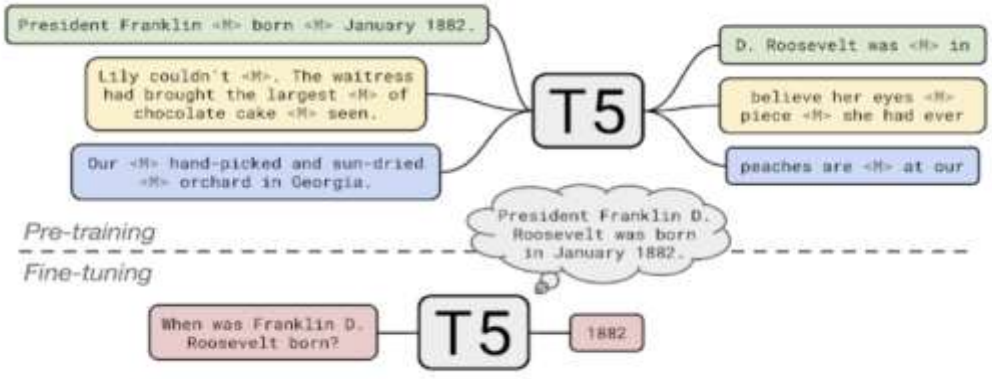

Fig 2: T5 Distinction b/w Pre-Training and Fine Tuning [2]

The above figure depicts the working of T5 to Question-Answering. Given a sentence, T5 employs blank-masking technique to predict other tokens in the sentence. No context is thus lost during the fine tuning and the internalization of the knowledge from the C4 sentence helps it answer the question. The following are the operations of a T5 model [4].

1. CoLA – Classify grammatical correctness, given a sequence.
2. RTE – Classify inference of a sentence
3. MNLI – Classify the contradiction among hypotheses
4. MRPC – Classify semantic equivalence of sentences
5. QNLI – Classify an answer to a question
6. QQP – Classify semantic equivalence of questions
7. SST2 – Classify sentiment of a sentence
8. STSB – Classify sentiment on a scale
9. CB – Classify contradiction between a premise and hypothesis
10. COPA – Classify the correct choice for a question
11. MuliRc – Classify an answer to a question
12. WiC – Classify word sense disambiguation
13. WSC – Predict ambiguous pronoun
14. Summarization – Summarize a paragraph
15. SQuAD – Answer a question
16. WMT1 – Machine Translation (English to German)
17. WMT2 – Machine Translation (English to French)
18. NQ – Closed Book Answering

Conclusion:

T5 is a very flexible model that can be extended to diverse tasks with appropriate fine-tuning. However, the size of the model restricts this research to resourceful organizations and is highly costly for students and scholars outside. Therefore, programs which allow scholars some direct access to models at a nominal cost can be beneficial for faster improvements in this direction.

Acknowledgements and References:

I would like to thank Prof. ChengXiang Zhai for allowing this opportunity to explore relevant technology to the course CS410 Text Information Systems of the University of Illinois Urbana Champaign

I would also like to acknowledge the following references that have helped me understand the models and for the resources I've used in this review.

[1] https://arxiv.org/pdf/1706.03762.pdf

[2] https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

[3] https://medium.com/syncedreview/google-t5-explores-the-limits-of-transfer-learning-a87afbf2615b#:~:text=T5%20is%20an%20extremely%20large,they%20authors%20aim%20to%20solve.

[4] https://www.toptal.com/deep-learning/exploring-pre-trained-models

[5] https://towardsdatascience.com/hands-on-googles-text-to-text-transfer-transformer-t5-with-spark-nlp-6f7db75cecff

[6] https://arxiv.org/abs/1910.10683

[7] https://huggingface.co/docs/transformers/model_doc/t5

[8] https://www.youtube.com/watch?v=91iLu6OOrwk

[9] https://wandb.ai/mukilan/T5_transformer/reports/Exploring-Google-s-T5-Text-To-Text-Transformer-Model--VmlldzoyNjkzOTE2

[10] https://arxiv.org/abs/1810.04805

[11] https://openai.com/blog/tags/gpt-2/

[12] https://openai.com/api/

[13] https://becominghuman.ai/attention-is-all-you-need-16bf481d8b5c

[14] https://en.wikipedia.org/wiki/GPT-2

[15] https://en.wikipedia.org/wiki/BERT_(language_model)

[16] https://en.wikipedia.org/wiki/GPT-3