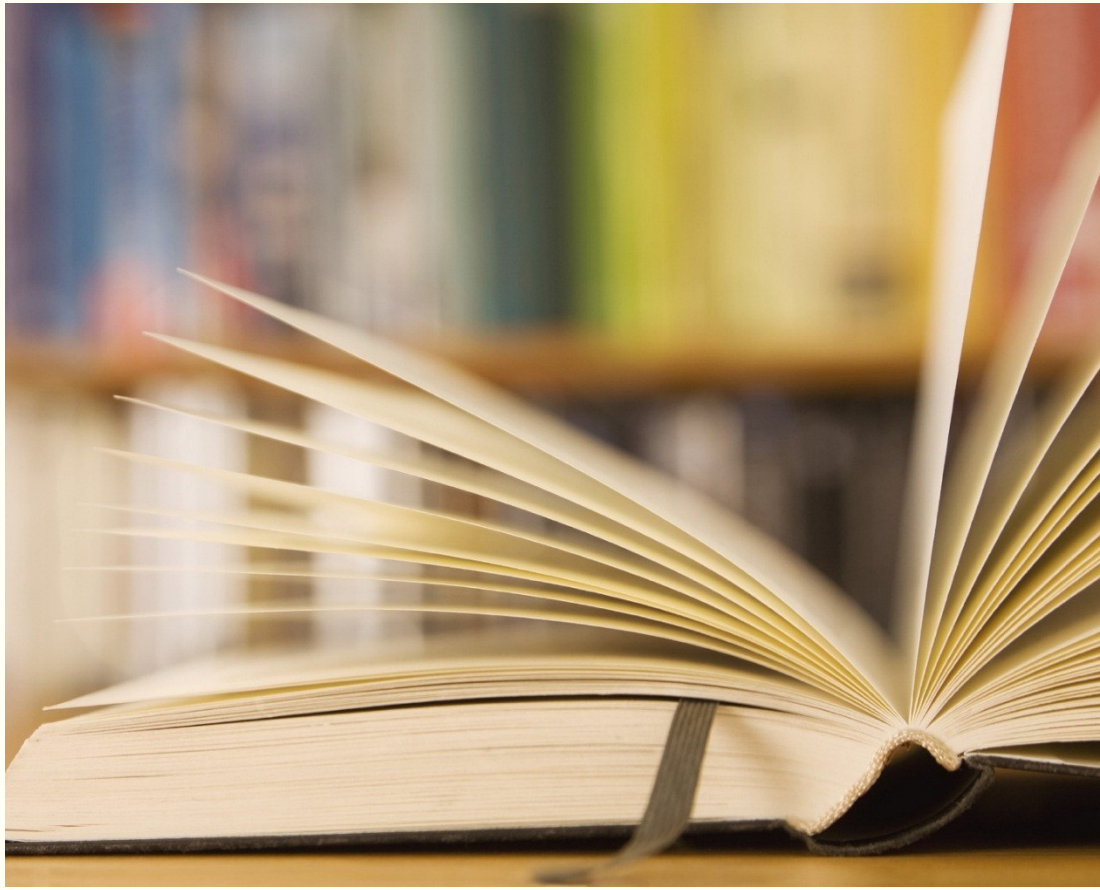
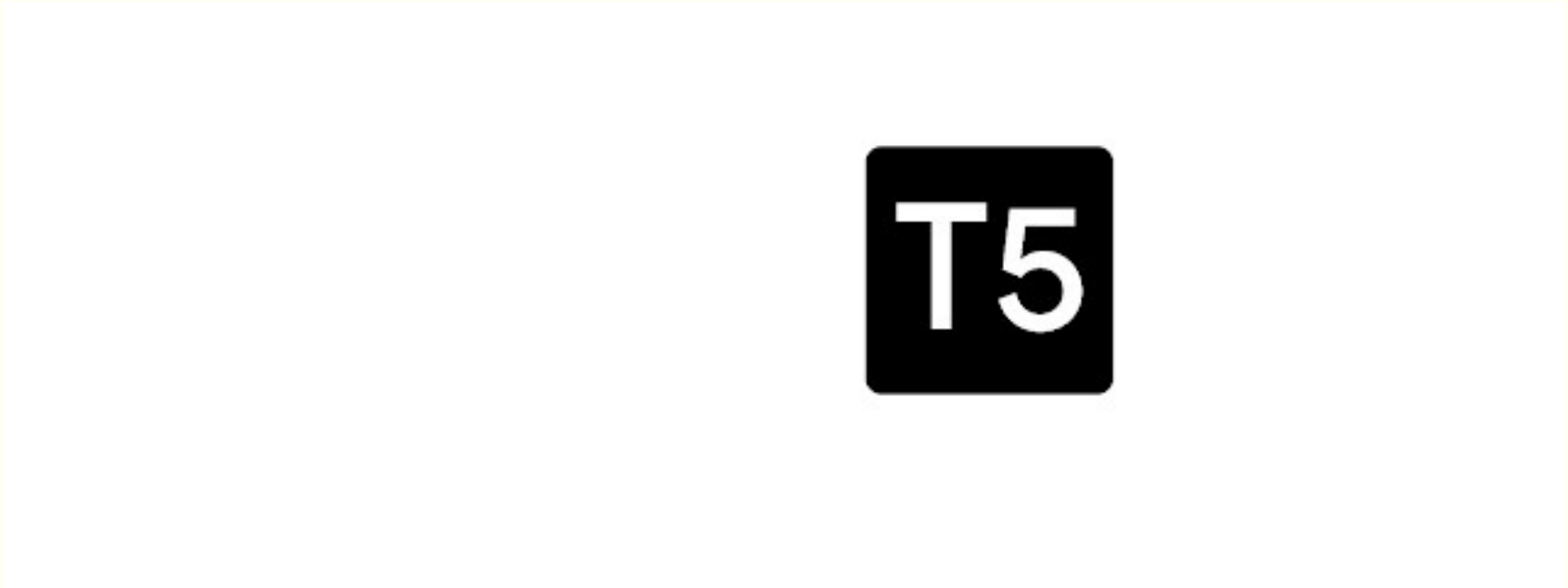
EXPLORING THE LIMITS OF TRANSFER LEARNING WITH A UNIFIED TEXT-TO-TEXT TRANSFORMER

Rohith Reddy Mada



T5 Framework



Source : https://research.google/blog/exploring-transfer-learning-with-t5-the-text-to-text-transfer-transformer/

Transfer Learning

- The model is first pretrained on a large-scale unlabeled data set via (unsupervised learning)
- It is then finetuned to specific task on a relatively small labelled dataset (supervised learning)
- Finetuning tasks used same hyperparameters and loss function (CE/MLL)
- Besides this, the paper explored Multi-task learning (in pretraining), where model is trained on multiple tasks simultaneously. This enables the model to gain task-specific insights.
- Experimented with multiple objectives in pretraining like Span Corruption, Masked Language Modelling, Autoregressive Modelling

C4 (Colossal Clean Crawled Corpus) Dataset

- Produced from an archive of "web extracted text"
- The text is processed to :
  - Retain sentences with a terminal punctuation mark ('.', '!', '?')
  - Include sentences with at least 5 words
  - Exclude pages containing "offensive words"
  - Exclude code (using keywords like "JavaScript" , "{}" , etc.)
  - Exclude duplicate data
  - Include only English sentences
- Unfiltered C4 is 6.1 TB while the processed one is 750 GB large

Downstream Tasks - Benchmarks

- GLUE and SuperGLUE – Classification tasks
  - CoLA – sentence acceptability judgement
  - SST-2 Sentiment analysis
  - WIC – Word sense disambiguation
  - Treated as a single task while finetuning ; datasets are concatenated
- CNN/Daily Mail – Text summarization
- SQuAD – Question Answering
- News Commentary v13, Common Crawl, Europarl v7 – WMT English to German
- - English to French
- - English to Romanian

Model Architecture

- Very similar to the original encoder decoder form proposed by (Vaswani et al., 2017)
- Uses Relative position embeddings
- Self-attention is used in both encoder and decoder – helps capturing both global and local context from all tokens in the sequence
- Encoder-Decoder attention – decoder can attend to different parts of encoders output / hidden representations
- Explored a variety of model sizes from 60 million to 11 billion parameters
- Tried out Encoder – Decoder, Encoder only, Decoder only, shared Encoder – Decoder architectures, Encoder – Decoder outperformed all of them

Results

- T5-Base outperformed the baseline model
- T5-11B outperformed previous models on both GLUE and SuperGLUE benchmarks
- It also performed better on other benchmarks including SQuAD (QA task) and CNN/Daily Mail (Text Summarization)
- Although it performs well in translation tasks, specialized translation models outperform in this task

Summary

Strengths

- Can be used for almost all of NLP tasks
- State of the art performance on most of the tasks
- Performance improves with scaling
- Can be finetuned for custom tasks; with smaller datasets
- Is good at both understanding(classification) and generation (QA)
- Its open source !!

Weakness

- Huge computational requirements to achieve best results (3B , 11B parameters)
- Pretraining on large scale general / internet data may cause bias. May not be used in sensitive domains like law or healthcare
- It's possible to generate a task in-specific output. There is a chance for it to output "Hamburger" for a NLI task.
- Not multilingual – limited to English, German, French, Romanian. It may not generate computer programs too
- Sometimes training for individual tasks might be easier than multi-task training

Uses in Interactive Fiction

- Scene summarization – It can be used to dynamically generate the conclusion of a scene based on command history, location, character descriptions
- Dynamic hints – It can be finetuned like a Question Answering system
- Processing user inputs – T5 can be used in parser to match player intent based on input. It can solve issues related to string matching done in general cases
- NPC dialogue generation – It can help in making actions incvolving
- Dynamic descriptions – It can be used to custom generate character, location, inventory descriptions
- Handling actions – "Go to fish pond" might be more interactive than "go up, go down, bla bla" especially if the game involves lots of locations and tasks

References

- Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." Journal of machine learning research 21.140 (2020): 1-67.
- Vaswani, Ashish, et al. "Attention Is All You Need.(Nips), 2017." arXiv preprint arXiv:1706.03762 10 (2017): S0140525X16001837.
- Mausam (2020, spring). Advanced NLP [PowerPoint slides]. Professor of Computer Science, Indian Institute of Technology Delhi. https://www.cse.iitd.ac.in/~mausam/courses/col873/spring2020/slides/08-t5.pdf
- Google Research. (2019). Exploring Transfer Learning with T5: The Text-To-Text Transfer Transformer. Retrieved from https://research.google/blog/exploring-transfer-learning-with-t5-the-text-to-text-transfer-transformer/
- Hugging Face. (n.d.). T5: The Text-To-Text Transfer Transformer. In Transformers Documentation. Retrieved from https://huggingface.co/docs/transformers/en/model_doc/t5

THANK YOU