Text Summarization: A Comparative Analysis

Derek Parker and Andrew Downing

Problem Description and Motivation

- We address a problem of **automatic text summarization**
- Goal: generate concise, coherent summaries from longer bodies of text using a decoder and encoder-decoder model.
- Given an *article*, our models produce a shortened version that preserves the main ideas and key information.
- While our initial problem was more focused towards intaking new articles, it can be further expanded out to include other domains.
    - With the amount of content being posted online, whether it be journalistic articles, research papers, social media essays, and more, it can be overwhelming to intake vast amounts of information, making automatic summarization a useful tool in instances where a summary has not already been provided.

Ethical or Societal Impacts

- Summaries may leave out critical context, may lead to misleading interpretations with the original article.
    - Neural Text Summarization: A Critical Evaluation (Kryściński et al, 2019)
        - News articles frequently present key points in the beginning, study found that 25% of the samples were difficult to interpret, even by humans.
- Summaries may not reflect the original tone, intent, and emphasis of the original article. Summaries are based on initial input and are reflective of the original tone and language present in the article.
- If summarization model is trained on biased data, may reinforce biases.
    - Identifying and Reducing Gender Bias in Word-Level Language Models (Bordia and Bowman (2019)
        - CNN/Daily Mail dataset had fewer gender bias compared to other datasets.

Proposed Solution

Problem Context

- Existing summarization models vary in architecture and performance
- Want to compare how a decoder-only vs an encoder-decoder model affects the summary quality and fidelity

Proposed Solution

- Compare GPT-2 (decoder-only) and T5-small (encoder-decoder) on a shared summarization dataset
- Analyze how the architecture affects:
  - Similarity between original article and the articles highlights
  - The fluency of the output
  - Factual accuracy
- Using prompt-based generation for GPT-2 and task-specific for T5

Background and Related Work

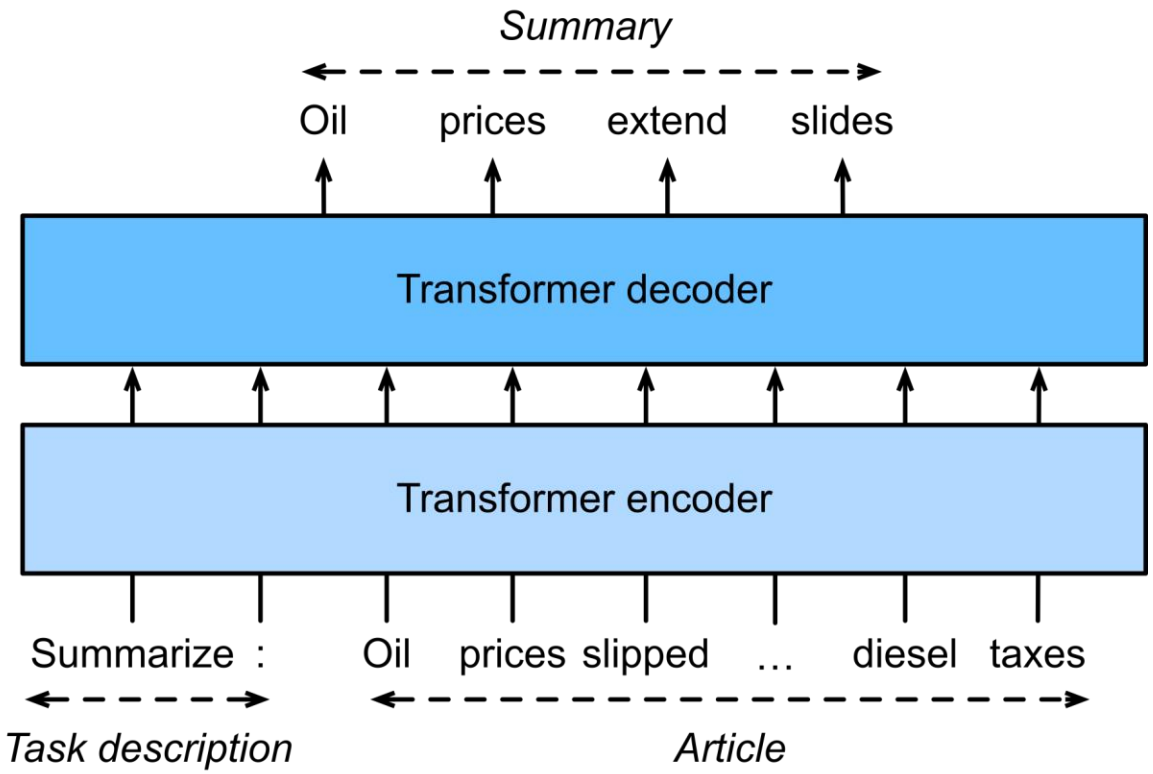
The Dataset

- The CNN/Daily Mail dataset is popular for summarization research and finetuning LLMs.
  - Long input articles
  - Includes article highlights (key bullet point take aways from an article)

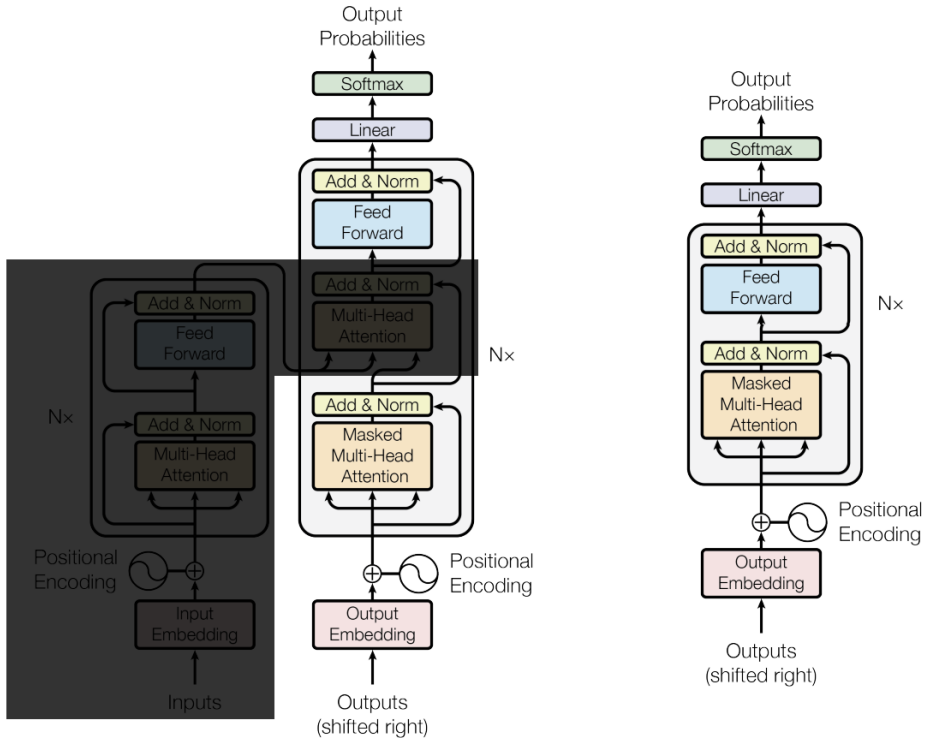
Past Architecture Work

- GPT-2 (Radford et al., 2019)
  - Introduced GPT-2 which was trained on OpenAI's internal WebText dataset (massive web corpus)
  - Discuses how GPT-2 can perform a wide range of tasks with no task specific training, via zero-shot
  - Shows GPT-2 performs decently on summarizing, better when paired with prompt engineering or minimal fine-tuning
- T5 (Raffel et al., 2020)
  - Introduced T5 model, reformulated every NLP task as a text-to-text problem
  - Pretrained on text summarization tasks, enables strong zero-shot performance

Model Diagrams



Adapted from Zhang et al., Dive into Deep Learning, 2023, https://D2L.ai, CC BY-SA 4.0



"Summarize the following article in 2 sentences: [article]"

Implementation Progress

- **Dataset** was retrieved from *Kaggle*, contained 300k *training*, 30k *dev* and *validation* observations.
- 10% of the original *dataset*, leaving roughly 30k training, 10k dev/val observations.
  - Dataset was collected by using archived versions of CNN articles from April 2007 to April 2015 and the Daily Mail articles were obtained from June 2010 and April 2015.
- Use **T5-Small** model, an *encoder-decoder* model, that has 77 *million parameters*.
- Utilize the **GPT-2** model, a *decoder* only model with *1.5B parameters*, to compare the differences in performance and their outputs.
- Used *HuggingFace* transformers library with PyTorch to run the T5-Small and GPT-2 models locally

00083697263e215e5e7ed a753070f08aa374dd45	There are a number of job descriptions waiting for Darren Fletcher when he settles in at West Brom b...	Tony Pulis believes Saido Berahino should look up to Darren Fletcher . Pulis insists Berahino has be...

Preliminary Results

Measures the unigram overlap (per-word similarity)

Measures the bigram overlap (word-pair similarity)

LCS that's useful for structural similarity over a sentence

LCS over multiple sentences. Tuned for summarization

Model	Rouge-1	Rouge-2	Rouge-L	Rouge-L Sum
T5-Small	39.23	17.88	27.87	34.14
GPT-2	20.04	10.28	14.36	18.13

Scores have been scaled by 100x

LCS (Longest Common Subsequence) - captures in-sequence matches.

Preliminary Results

Article: A middle-school teacher in China has inked hundreds of sketches that are beyond be-leaf. Politics teacher Wang Lian, 35, has created 1000 stunning ink drawings covering subjects as varied as cartoon characters and landscapes to animals, birds according to the People's Daily Online. The intricate...

T5: the intricate scribbles on leaves feature Wang's favorite sites across the city of Nanjing. she uses an ink pen to gently draw the local scenes and buildings.

GPT2: A middle-school teacher in China has inked hundreds of sketches that are beyond be-leaf. Politics teacher Wang Lian, 35, has created 1000 stunning ink drawings covering subjects as varied as cartoon characters and landscapes to animals, birds according to the People's Daily Online. The intricate scribbles on leaves feature Wang's favourite sites across the city of Nanjing, which include the Presidential Palace, Yangtze River Bridge, the ancient Jiming Temple...

Refence Summary: Works include pictures of Presidential Palace and Yangtze River Bridge. Has inked 1,000 pieces of art on leaves in last two years. Gives work away to students in form of bookmarks and postcards.

Challenges and Insights

- **GPT-2:** Frequently had issues repeating the same summarized output. Needs to be fine tuned extensively to produce coherent outputs
- **T5-Small:** Maximum input token length of 512, average *mean token length* of all articles is 781
- Training for both models required a fair amount of compute power and time
- *Encoder-decoder* performs well with the specific task prompted for, *decoder* can work with generating text, has trouble producing *abstractive* summaries
- *Evaluation metrics* may not always reflect the true quality of a summary, highlights the importance of using *human evaluation* to look into a model's performance.
- GPT-2 was trained to predict the next word based on the past tokens
- T5 was trained with a "text-to-text" objective (translation, summarization, question answering)

References

- Kryściński, W., Keskar, N. S., McCann, B., Xiong, C., & Socher, R. (2019). Neural text summarization: A critical evaluation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 540–551). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1051
- Bordia, S., & Bowman, S. R. (2019). *Identifying and reducing gender bias in word-level language models*. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 629–634. https://doi.org/10.18653/v1/N19-1063
- Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2023). *Dive into Deep Learning*. Cambridge University Press. https://D2L.ai
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). *Exploring the limits of transfer learning with a unified text-to-text transformer*. Journal of Machine Learning Research, 21(140), 1-67. http://jmlr.org/papers/v21/20-074.html
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners*. OpenAI Blog. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Thank you

Questions