Abstractive Summarization with BigPatent

Ziling Huang
University of California Berkeley

TABLE OF CONTENTS

01

Summarization

Motivation, Research Goal, Literature Review

How many minutes does an average reader take?

Time Taken to Read

arXIV paper

10,000 words in 30 minutes

Privacy Agreement

2,500 words in 8 minutes

Patent document

3,600 words in 12 minutes

Summary

100 words in 30 seconds

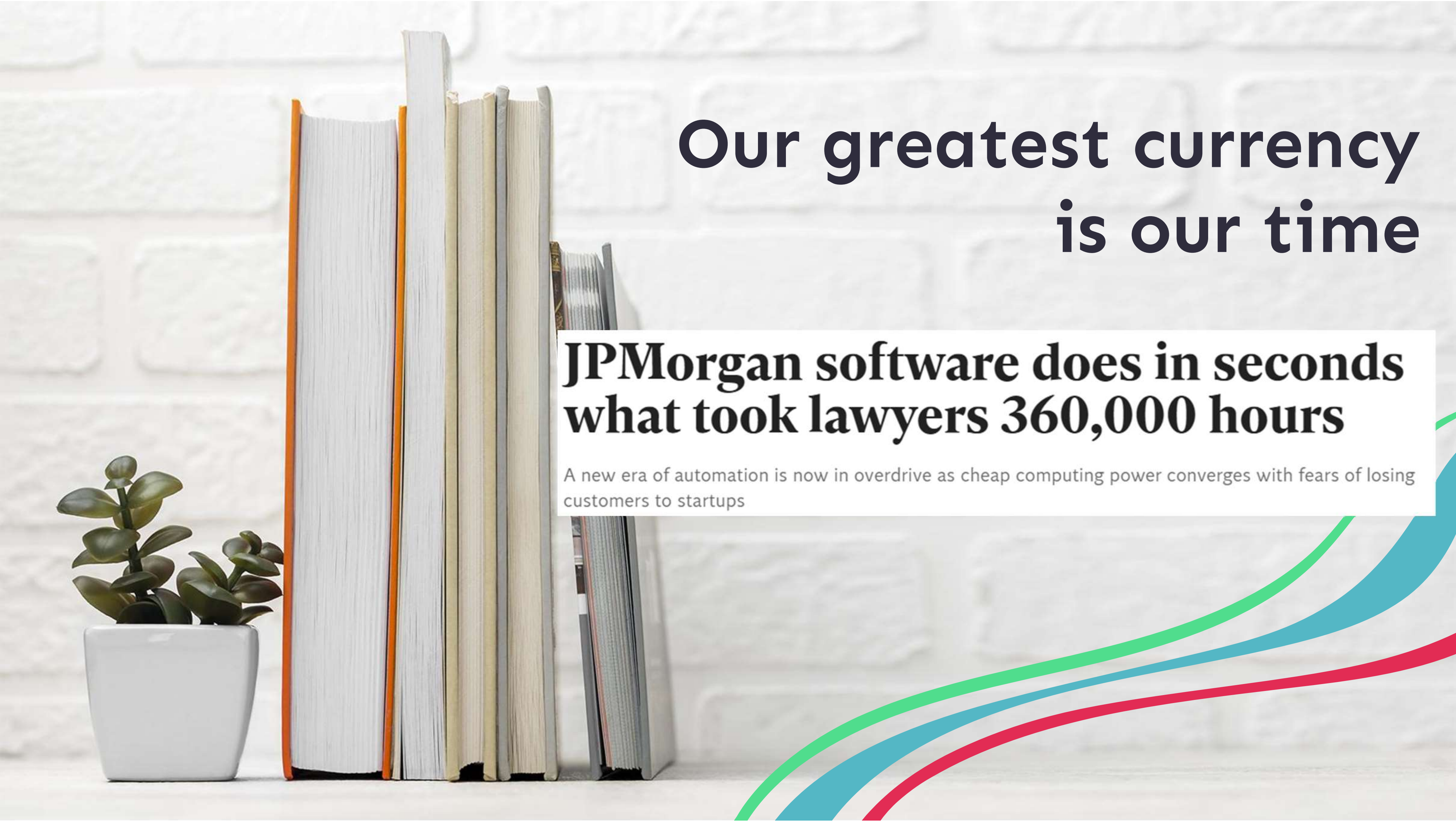
How about 1.3 million Patent Documents?

74 years

How about their summaries?

1 year

Our greatest currency is our time



JPMorgan software does in seconds what took lawyers 360,000 hours

A new era of automation is now in overdrive as cheap computing power converges with fears of losing customers to startups

RESEARCH GOAL

Can we use extractive-then-abstractive approaches to outperform sparse attention mechanism (Bigbird) in Long Document Summarization?

Metrics : ROUGE and QuestEval (Nov 2021)

LITERATURE REVIEW

Extract-then-Generate

Roberta extractor and BART generator with dynamic extract scoring

Sparse Attention

BigBird: Long sequences, random, window and global attention

Divide-and-Conquer

Identify a document structure, extract from each section, then consolidate

02

BigPatent Dataset

Quick Overview

BigPatent Dataset (Sharma et al., 2019)

01
Abstractive
More novel n-grams in summary versus input

02
Structure
Salient input found in 80% of text. Part by part how to make and use invention.

03
Stats
1.3 mil Cases; 9 categories
Mean Text Length: 3,572
Mean Summary Length: 116
Mean Compression Ratio: 36.4

04
Global Content

03

Methods & Analysis

Experiment, Qualitative and Quantitative

THE EXPERIMENT

Baseline

BigBird-Pegasus	4,096 input length, 150 output length, 5 beams, length penalty 0.8
T5	512 input length, 150 output length, 2 beams, length penalty 1.0

Variants

TF-IDF sentence scoring then T5	Top 15 sentences, Best : Minimum >8 words per sentence + Baseline T5
TF-IDF-Vectors Pairwise Cosine Similarity then T5	Best : Square Root (input), 30 character minimum per sentence + Baseline T5

Pretraining and Training

Baseline

BigBird-Pegasus	BB: Books, CC-News, Stories, Wikipedia Pegasus: C4, HugeNews 1.2mil BigPatent training
T5	C4 6,400 BigPatent training (1,600 val, 2000 test)

Variants

TF-IDF sentence scoring then T5	C4 6,400 BigPatent training (1,600 val, 2000 test)
TF-IDF-Vectors Pairwise Cosine Similarity then T5	C4 6,400 BigPatent training (1,600 val, 2000 test)

Evaluation Metrics

Use both!

QuestEval (Scialom et al., Nov 2021)

- Higher correlation with human judgment over 4 dimensions versus ROUGE:
  - consistency, coherence, fluency and relevance
- Reference-less Metric
- Unifies prior QA approaches by using both precision & recall under one framework

Classic ROUGE 1,2, L

- F-Score
- Overlapping n-grams in the generated summary versus the gold summary

Criticism

- More than one valid summary can exist for a document
- Factual consistency is not measured by n-gram metrics

Results Table

	BB-P	T5	TFIDF-T5	TFIDF-COSSIM-T5
ROUGE-1	38.5% ✓	33.9%	31.2% ✗	34.4% ✓
ROUGE-2	15.4% ✓	10.8%	8.8% ✗	10.8% ✓
ROUGE-L	26.3% ✓	22.9%	21.6% ✗	23.3% ✓
QuestEval	33.5% ✓	30.5%	29.2% ✗	30.2%

RESULTS ANALYSIS

Input length

Bigbird-Pegasus: 4,096
BigPatent: 80% of mean 3,572 document has salient input

Post-analysis

>4,096 QuestEval : 31.5%
<4,096 QuestEval : 34.2%

Pretraining

BigBird-Pegasus was pre-trained on more types of datasets

Training and Test

T5 was trained on less training samples (computing power limitations).
A smaller test sample than the original BigBird paper

RESULTS ANALYSIS

Restrictiveness

TF-IDF-COS-SIM extraction params less restrictive : square root of text length versus TF-IDF top 15

Approach

TF-IDF-COS-SIM learns from the document.
TF-IDF learns from all documents across patent categories

Validation loss pattern

Manual Review

Difficult to achieve good pre-processing over long text and variants
Repetitive, incoherent statements, lower informativeness

Conclusion

Longer input lengths using sparse attention produce significantly better results than extractive-then-generative

Future exploration to level the transfer learning playing field for the models.
Experiment with other types of extractive techniques.

Thank you to the faculty advisors Daniel, Mark, Sandip

Thank you to all of you for being a great audience!

REFERENCES

- Sharma et al. (2019). BIGPATENT: A Large-Scale Dataset for Abstractive and Coherent Summarization, Page Number (1-10). https://arxiv.org/pdf/1906.03741.pdf

- Zhang et al. (2019). PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization, Page Number (1-55). https://arxiv.org/pdf/1912.08777.pdf

- Raffel et al. (2019). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, Page Number (1-67). https://arxiv.org/pdf/1910.10683v3.pdf

- Chen and Bansal (2018). Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting, Page Number (1-4). https://arxiv.org/pdf/1805.11080.pdf

REFERENCES

- Hsu et al. (2018). A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss, Page Number (1-5). https://arxiv.org/pdf/1805.06266.pdf

- Mao et al. (2021). DYLE : Dynamic Latent Extraction for Abstractive Long-Input Summarization, Page Number (1-6). https://arxiv.org/pdf/2110.08168.pdf#page=10&zoom=100,88,94

- Zaheer et al. (2020). Big Bird: Transformers for Longer Sequences, Page Number (1-4). https://arxiv.org/pdf/2007.14062.pdf

- Beltagy et al. (2020). Longformer: The Long-Document Transformer, Page Number (1-5). https://arxiv.org/pdf/2004.05150.pdf

REFERENCES

- Gidiotis and Tsoumakas. (2020). A Divide-and Conquer Approach to the Summarization of Long Documents, Page Number (1-6). https://arxiv.org/pdf/2004.06190.pdf

- Scialom et. Al. (2021). QuestEval : Summarization asks for Fact-based Evaluation, Page Number (6594-6601) https://aclanthology.org/2021.emnlp-main.529.pdf

- Transformers Documentation: https://huggingface.co/transformers/

- Tutorial and Code on using Cosine Similarity: https://medium.com/@krause60/using-cosine-similarity-to-build-a-python-text-summarization-tool-d3c8228549bf

REFERENCES

- Tutorial on Pre-training T5: https://www.youtube.com/watch?v=KMyZUIraHio&t=499s
- Tutorial on Evaluation with BigBird-Pegasus using PubMed Data: https://colab.research.google.com/github/vasudevgupta7/bigbird/blob/main/notebooks/bigbird_pegasus_evaluation.ipynb
- Article Explaining TF-IDF : https://towardsdatascience.com/text-summarization-using-tf-idf-e64a0644ace3
- BigPatent Dataset : https://evasharma.github.io/bigpatent/

Appendix

Example: Publication number US-2012183074-A1

Gold Reference Summary
in video encoding it is common to encode the image data to remove redundancies in the information to be transmitted . while the pixel data is usually arranged in pixel blocks , the blocks can be arranged in one or more groups of n — m blocks called tiles . the tiles avoid the need to send header information on a tile - by - tile or block - by - block basis , and simplifies parallel processing of the tiles . bits from respective tiles may then be reformatted to recreate bits according to a raster - scan direction . this enables the decoder to receive the bits in a regular raster - scan format , but also have the ability to decode the tiles once the bits are reformatted . by partitioning an image into tiles of size n— m , it is possible to further exploit the intra - frame correspondence of images in a vertical direction as well as horizontal direction since the tiles need not destroy as many dependencies between blocks in a tile as if the blocks where organized in slices or slice groups .

Appendix

Example: Publication number US-2012183074-A1

T5 only predicted Summary
the present invention relates to a video encoder for use in encoding images. the encoder includes a plurality of tiled blocks, each tiled block having a vertical dimension and a horizontal dimension. the encoder also includes a plurality of processors capable of simultaneously performing encoding operations on multiple tiles. the encoder further includes a plurality of processors capable of simultaneously performing encoding operations on multiple tiles.

Appendix

Example: Publication number US-2012183074-A1

BigBird Pegasus Predicted Summary
The present embodiment introduces the notion of tiles to exploit the two dimensional dependencies between blocks while also supporting the exploitation of multiple processors, if available in the encoder, to simultaneously perform encoding operations on multiple tiles. The partitioning of a frame into tiles is completely specified by the numbers n and m, eliminating the need for a slice header, which is a basic requirement in conventional slice processing. here, n and m are the height and width of a tile measured in number of blocks. typically, the values of n and m are conveyed to the decoder in the sequence header or picture header resulting in negligible transmission bandwidth overhead. in addition to unilaterally transmitting the n and m numbers to the decoder in the sequence or picture header, an alternative is to have

Appendix

Example: Publication number US-2012183074-A1

TF-IDF-Cos-Sim-then-T5 Predicted Summary
a multicore decoder is provided with a plurality of processing cores each capable of processing at least one tile. the encoder and decoder are configured to perform a handshaking operation in which the encoder and decoder operate in a tile partitioning mode. the encoder and decoder are configured to perform a handhaking operation on the encoder and decode the tiles in a tile partitioning mode. the encoder and decoder are configured to perform a handhaking operation on the encoder so as to place the bits in a tile partitioning mode.

Appendix

Example: Publication number US-2012183074-A1

TF-IDF-then-T5 Predicted Summary
a computer system and method for providing data communication to a plurality of devices is disclosed. the computer system includes : a first network link that provides data communication to a plurality of devices ; a second network link that provides data communication to a plurality of devices ; and a third network link which provides data communication to a plurality of devices.

THANKS

Do you have any questions?

zilingh@berkeley.edu