FPT Education
FPT UNIVERSITY

AIP391 Project

Text Summarization with Seq2seq Model

Hoang Tuan Minh, Nguyen Kieu Mai
Advisor: Nguyen Quoc Trung

ABSTRACT: Text summarization is an important task in natural language processing, as it can help users quickly and accurately understand the main points of a long text. There are various approaches to developing an abstractive summarization model, including deep learning methods such as LSTM and Transformer-based models. Many studies have evaluated the effectiveness of these models using metrics such as ROUGE and BLEU scores. These studies have found that abstractive summarization models are generally effective in generating summaries that capture the key information in a text. However, there are still some challenges that need to be addressed, such as demonstrating coherence and consistency in the generated summaries, and reducing the risk of generating biased or inaccurate information. Overall, the findings suggest that abstractive text summarization models have the potential to greatly enhance our ability to process and understand large amounts of text, but more research is needed.

1. INTRODUCTION:

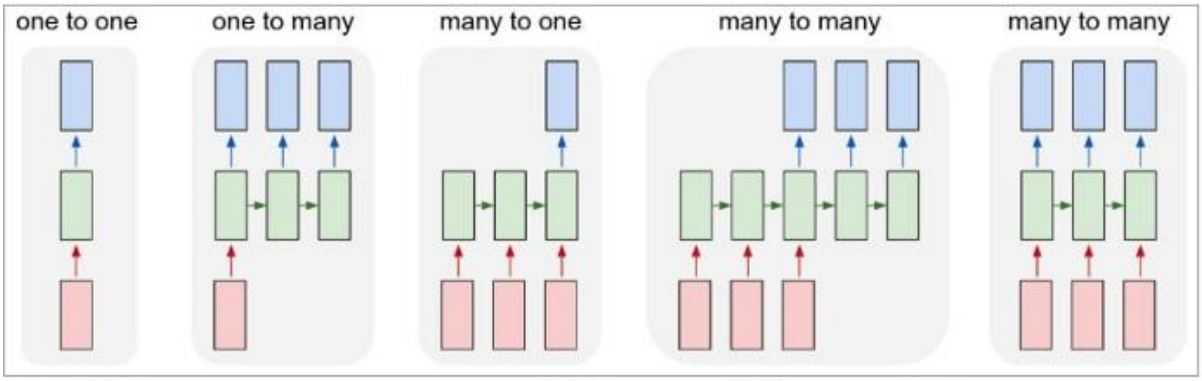
Text summarization is the process of condensing lengthy texts while retaining their main points. This helps to summarize a lengthy text and cut down on the amount of time needed to read it. In this situation, we can use a deep learning model created using an encoder-decoder sequence-to-sequence model to create a text summarizer rather thandepending on manual summarizing.

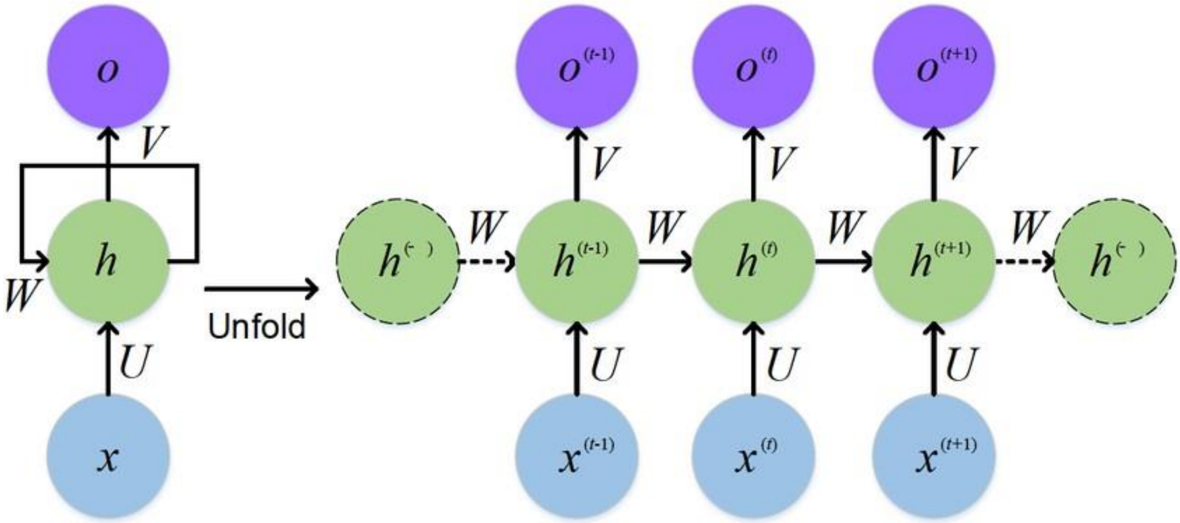
An RNN normally has input and output vectors that are fixed in size, meaning that their lengths are predetermined. Nonetheless, in use scenarios where the input and output sequences are not required to be fixed and the same length, such as speech recognition, machine translation, etc., this is not desirable.

The aforementioned issue can be resolved with the aid of sequence-to-sequence (seq2seq) models. When given an input, the encoder-decoder seq2seq model first generates an encoded representation of the model, which is then transferred to the decoder to generate the required output. The size of the input and output vectors in this situation does not haveto be fixed.

2. RELATED WORK:

- RNN: A Recurrent Neural Network is a type of neural network that contains loops, allowing information to be stored within the network. In short, Recurrent Neural Networks use their reasoning from previous experiences to inform the upcoming events. Recurrent models are valuable in their ability to sequence vectors, which opens up the API to performing more complicated tasks.





Recurrent Neural Networks can be thought of as a series of networks linked together. They often have a chain-like architecture, making them applicable for tasks such as speech recognition, language translation, etc. An RNN can be designed to operate across sequences of vectors in the input, output, or both. For example, a sequenced input may take a sentence as an input and output a positive or negative sentiment value. Alternatively, a sequenced output may take an image as an input, and produce a sentence as an output.

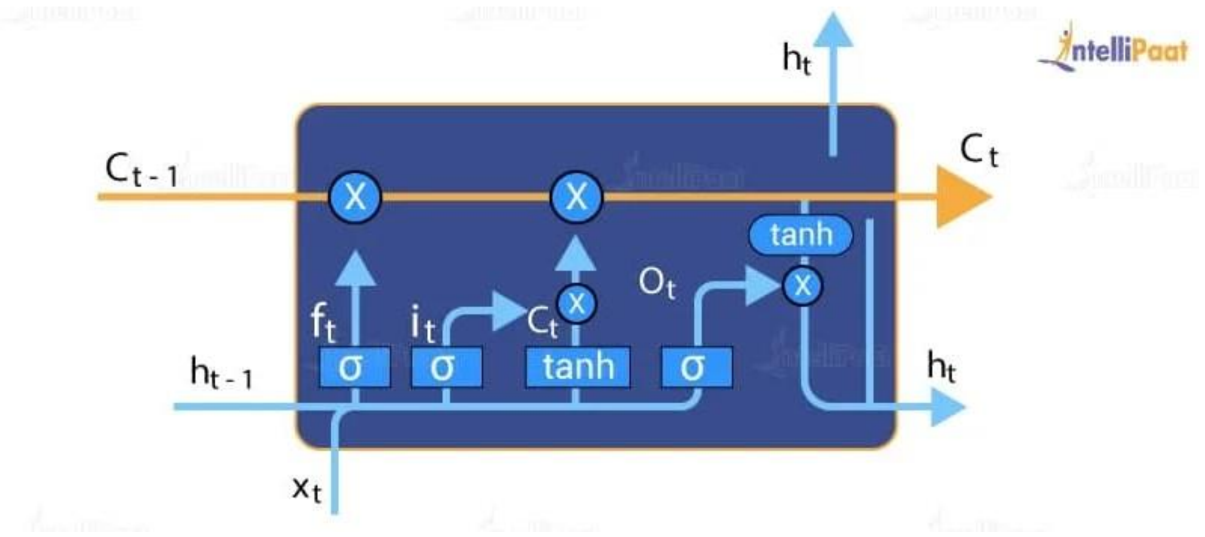
3. DATA PREPARATION:

Here we're going to use the News Summary Dataset. It consists of two CSV files: one contains information about the author, headlines, source URL, short article, and complete article, and another which only contains headlines and text. In the current application, you will extract the headlines and text from the two CSV files to train the model.

	text	summary
0	Saurav Kant, an alumnus of upGrad and IIIT-B's...	upGrad learner switches to career in ML & Al w...
1	Kunal Shah's credit card bill payment platform...	Delhi techie wins free food from Swiggy for on...
2	New Zealand defeated India by 8 wickets in the...	New Zealand end Rohit Sharma-led India's 12-ma...
3	With Aegon Life iTerm Insurance plan, customer...	Aegon life iTerm insurance plan helps customer...
4	Speaking about the sexual harassment allegatio...	Have known Hirani for yrs, what if MeToo claim...
...	...	...
102910	Mansha Mahajan 24 Feb 2017,Friday http://india...	Rasna seeking ?250 cr revenue from snack categ...
102911	Dishant Sharma 03 Aug 2017,Thursday http://ind...	Sachin attends Rajya Sabha after questions on ...
102912	Tanya Dhingra 03 Aug 2017,Thursday http://www....	Shouldn't rob their childhood: Aamir on kids r...
102913	Pragya Swastik 07 Dec 2016,Wednesday http://in...	Asha Bhosle gets ?53,000 power bill for unused...
102914	Chhavi Tyagi 03 Aug 2017,Thursday http://india...	More than half of India's languages may die in...

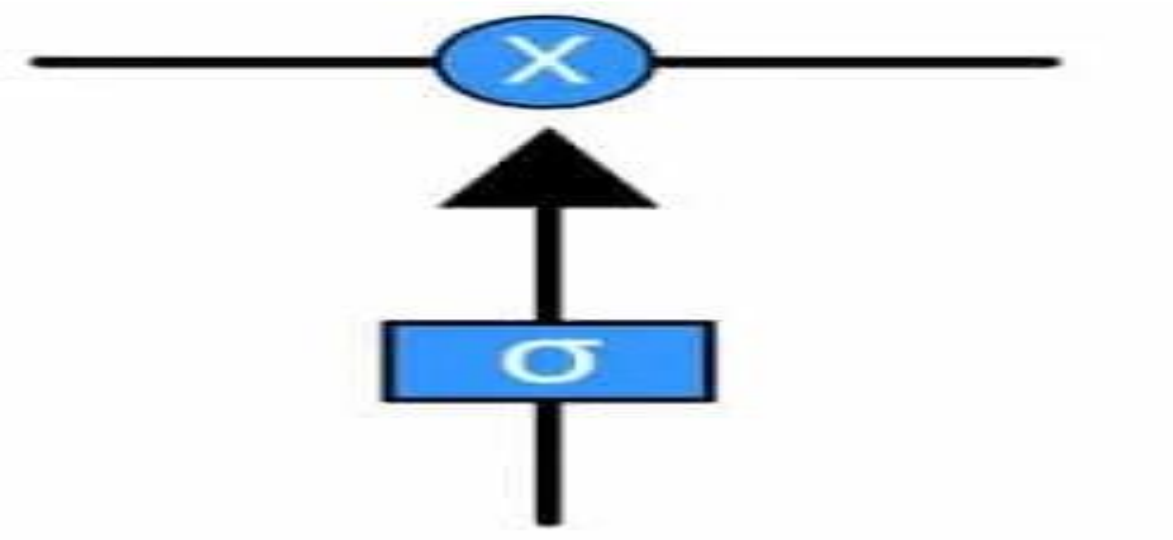
4. METHODS:

- LSTM: The central role of an LSTM model is held by a memory cell known as a 'cell state' that maintains its state over time. The cell state is the horizontal line that runs through the top of the below diagram. It can be visualized as a conveyor belt through which information just flows, unchanged.



Information can be added to or removed from the cell state in LSTM and is regulated by gates. These gates optionally let the information flow in and out of the cell. It contains a pointwise multiplication operation and a sigmoid neural net layer that assist the mechanism.

The sigmoid layer gives out numbers between zero and one, where zero means ‘nothing should be let through,’ and one means ‘everything should be let through.’



(IMPROVE RNN disadvantage of vanishing gradient.)

A Shared Text-To-Text Framework

With T5, we propose reframing all NLP tasks into a unified text-to-text-format where the input and output are always text strings, in contrast to BERT-style models that can only output either a class label or a span of the input. Our text-to-text framework allows us to use the same model, loss function, and hyperparameters on any NLP task, including machine translation, document summarization, question answering, and classification tasks (e.g., sentiment analysis). We can even apply T5 to regression tasks by training it to predict the string representation of a number instead of the number itself.

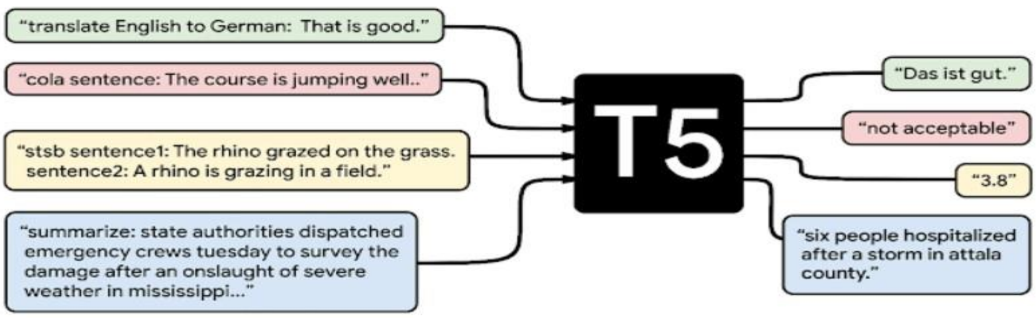


Diagram of our text-to-text framework. Every task we consider uses text as input to the model, which is trained to generate some target text. This allows us to use the same model, loss function, and hyperparameters across our diverse set of tasks including translation (green), linguistic acceptability (red), sentence similarity (yellow), and document summarization (blue). It also provides a standard testbed for the methods included in our empirical survey.

A Systematic Study of Transfer Learning Methodology

With the T5 text-to-text framework and the new pre-training dataset (C4), we surveyed the vast landscape of ideas and methods introduced for NLP transfer learning over the past few years. The full details of the investigation can be found in our paper, including experiments on:

- ***model architectures***, where we found that encoder-decoder models generally outperformed "decoder-only" language models;
- ***pre-training objectives***, where we confirmed that fill-in-the-blank-style denoising objectives (where the model is trained to recover missing words in the input) worked best and that the most important factor was the computational cost;
- ***unlabeled datasets***, where we showed that training on in-domain data can be beneficial but that pre-training on smaller datasets can lead to detrimental overfitting;
- ***training strategies***, where we found that multitask learning could be close to competitive with a pre-train-then-fine-tune approach but requires carefully choosing how often the model is trained on each task;
- and ***scale***, where we compare scaling up the model size, the training time, and the number of ensembled models to determine how to make the best use of fixed compute power.

Insights + Scale = State-of-the-Art

To explore the current limits of transfer learning for NLP, we ran a final set of experiments where we combined all of the best methods from our systematic study and scaled up our approach with Google Cloud TPU accelerators. Our largest model had 11 billion parameters and achieved state-of-the-art on the GLUE, SuperGLUE, SQuAD, and CNN/Daily Mail benchmarks. One particularly exciting result was that we achieved a near-human score on the SuperGLUE natural language understanding benchmark, which was specifically designed to be difficult for machine learning models but easy for humans.

Extensions

T5 is flexible enough to be easily modified for application to many tasks beyond those considered in our paper, often with great success. Below, we apply T5 to two novel tasks: closed-book question answering and fill-in-the-blank text generation with variable-sized blanks.

5. RESULT:

T5 – model

[19680/19680 1:11:43, Epoch 4/4]

Epoch	Training Loss	Validation Loss	Rouge1	Rouge2	Rougel	Rougelsum	Gen Len
1	1.790700	1.545628	0.511400	0.276400	0.469500	0.469500	15.449300
2	1.678100	1.477610	0.521100	0.284300	0.478300	0.478300	15.564700
3	1.649100	1.451922	0.524000	0.287400	0.481400	0.481400	15.569800
4	1.638700	1.442738	0.525700	0.288500	0.483000	0.483000	15.588800

LSTM - model

```
Epoch 47/50
88517/88517 [==============================] - 353s 4ms/sample - loss: 2.8391 - val_loss: 3.2304
Epoch 48/50
88517/88517 [==============================] - 353s 4ms/sample - loss: 2.8298 - val_loss: 3.2272
Epoch 49/50
88517/88517 [==============================] - 355s 4ms/sample - loss: 2.8196 - val_loss: 3.2276
Epoch 50/50
88517/88517 [==============================] - 353s 4ms/sample - loss: 2.8119 - val_loss: 3.2409
```

```
Review: taking dig at bjp for its proposed rath yatra west bengal cm mamata banerjee on friday said that rath yatras are not car
ried out to kill people those who carry out yatras to kill common people indulge in yatras she added this comes after calcutta h
igh court recently put stay on bjp rath yatra in the state
Original summary: start rath yatras are not carried out to kill people wb cm end
Predicted summary:  start no one dead people will die in wb cm mamata end
ROUGE-1: 0.14
ROUGE-2: 0.03
ROUGE-L: 0.07
ROUGE-Lsum: 0.00


Review: a picture of russian mp natalya leaning against wall ahead of vladimir putin inauguration has gone viral reacting to the
picture twitter user wrote current mood natalya other users tweeted am natalya at every party and maybe she was ordered to open
and close the door
Original summary: start pic of mp leaning against wall before putin oath goes viral end
Predicted summary:  start mp cm meets twitter after russia open defecation end
ROUGE-1: 0.12
ROUGE-2: 0.00
ROUGE-L: 0.12
ROUGE-Lsum: 0.00


Review: actor purab kohli has said that the release of films on fridays is like the appraisal period for actors just like employ
ees in corporate offices have their annual appraisal period where they get nervous about what will happen with their salaries we
actors too feel the first friday said purab he added that ultimately audience is the best judge
Original summary: start friday releases are like period for actors end
Predicted summary:  start film industry is like to make film industry end
ROUGE-1: 0.07
ROUGE-2: 0.03
ROUGE-L: 0.07
ROUGE-Lsum: 0.00
```

6. CONCLUSION:

Honestly, summarizing tasks can't be automatically evaluated by a computer fairly. Because ROUGE score and BLEU score evaluate the summarizing tasks by checking how many words matched from the original sentences to the summarized sentences. The only way we know whether our model performs well on the dataset is to let it be checked by human experts.

In our scenario, we can just evaluate the two methods by looking at their VAL_LOSS. Where T5 performs very well while LSTM performs not that well at all, LSTM's VAL_LOSS is larger, about 2-3 times more than T5's VAL_LOSS.

It's easy to understand as T5-small has 60 million parameters, it's a pretrained model studied by Google research. And T5 itself is studied based on the pretrained model BERT with many finetunes.

References

[1] Tahmid Hasan , Abhik Bhattacharjee , Md Saiful Islam, Kazi Samin, Yuan-Fang Li,

Yong-Bin Kang, M. Sohel Rahman, Rifat Shahriyar, XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Language
[2] Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In Text Summarization Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
[3] Hsu, W.T., Lin, C.K., Lee, M.Y., Min, K., Tang, J., Sun, M.: A unified model for extractive and abstractive summarization using inconsistency loss. In: Computational Linguistics Association, pp. 132–141. Australia(2018). URL: https://doi.org/10.18653/v1/P18-1013
[4] Dalal, V., Malik, L.: A survey of extractive and abstractive text summarization techniques. In: International Conference on Emerging Trends in Engineering and Technology, pp. 109–110. USA (2013) . URL: https://doi.org/10.1109/ICETET.2013.31
[5] Wei, R., Huang, H., Gao, Y.: Sharing pre-trained BERT decoder for a hybrid summarization. In: Chinese Computational Linguistics, pp. 169–180 (2019). URL: https://doi.org/10.1007/978-3-030-32381-3_14
[6] Paulus, R., Xiong, C., Socher, R.: A deep reinforced model for abstractive summarization. In: International Conference on Learning Representations (2018)
[7] Song, K., Wang, B., Feng, Z., Liu, R., Liu, F.: Controlling the amount of verbatim copying in abstractive summarization. In: Conference on Artificial Intelligence, pp. 8902–8909 (2020). URL: https://doi.org/10.1609/aaai.v34i05.6420
[8] Nallapati, R., Zhou, B., Dos Santos, C., Gulcehre, C., Xiang, B.: Abstractive text summarization using sequence-to-sequence RNNs and beyond. In: Conference on Computational Natural Language Learning, pp. 280–290 (2016) . URL: https://doi.org/10.18653/v1/K16-1028
[9] Mozer, M.: BERT Does Business: Implementing the BERT Model for Natural Language Processing at Wayfair (2019)
[10] Dat Quoc Nguyen and Anh Tuan Nguyen, VinAI Research, Vietnam; NVIDIA, USA, PhoBERT: Pre-trained language models for Vietnamese. URL: https://arxiv.org/pdf/2003.00744.pdf
[11] Annie Louis, Joshua Maynez, OPINESUM: Entailment-based self-training for abstractive opinion summarization