Advanced Machine Learning Final Project

Transformer for Text Summarization on Long Documents

Munerah AlFayez

Prof.Murali Shanker

Abstract

The release of the ‘Attention Is All You Need’ [1] paper allow to reach new State-Of-Art performances in a lot of Deep Learning fields. Natural Language Processing is one the most impacted field because Transformer are now able to process much longer text sequences than with a Recurrent Neural Network. However, some problems remain. Indeed, the text length is still an issue when it is longer than thousands of tokens (pieces of words). Fortunately, some papers succeeded to modify the Transformer architecture to be able to process very long sequences at low computational costs and high speed. This work aims to present one of these recent discoveries that made State-Of-Art models on very long texts.

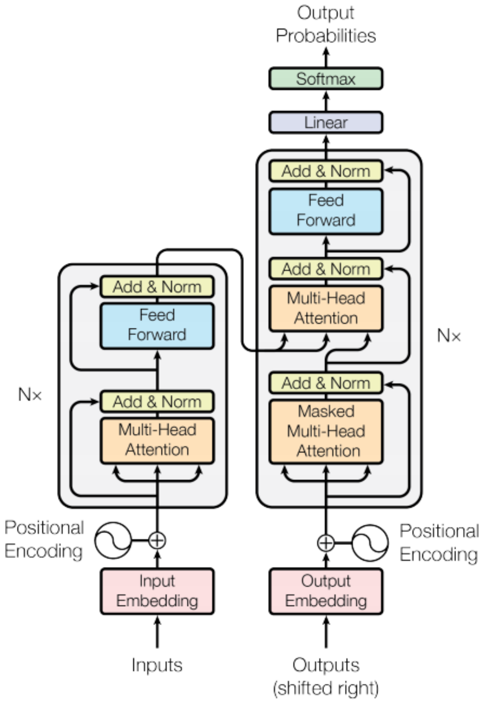
Contents

Introduction

The Transformer architecture came from the need to improve computational complexity of models that process huge data such as sequences and images. The use of the attention mechanism wasn't a new concept and thus wasn't invented by the 'Attention Is All You Need' [1] paper. This paper, however, brought a novel architecture that simplified a lot the State-Of-Art models at that time (2017).

Previously, models were based on very deep models with Recurrent layers and/or Convolutional layers depending on the task. These models took a very long time to train and thus were difficult to fine-tuned.

The Transformer architecture proposed in the paper was about text translation and used an encoder/decoder architecture without any Recurrent nor Convolutional layers. Instead, a simple feed forward network composed of dense layers was combined to the attention layers. The architecture is presented below:



However, even with this architecture, the performances (in terms of time, metrics and computational costs) were still poor on very long sequences that could contain thousands of words.

Fortunately, between 2019 and 2020, few papers about on new type of Transformer changed everything. These new architectures were very similar to the original one but with a lighter and smarter attention layer. This allowed the Transformer to process very long sequences with a linear computational cost. Compared to the exponential computational costs of the regular architecture, it is a significant improvement.

BigBird [2], Reformer [3] and Longformer [4], to name a few, are now considered as the State-Of-Art models when it comes to long sequences. In this work, I will focus on the Longformer in one of the most challenging Natural Language Processing tasks: the text abstractive summarization.

Text Summarization

Text summarization is the task of summarizing a long text into a much shorter sequence with a minimal information loss.

This task can be done in two different ways:

- Text Extraction which extracts the most significant sentences and/or group of words of the text. It is classification task similar to Named Entity Recognition.
- Text Abstractive Summarization which resumes the long document into a simple and shorter text without losing the meaning of it.

On long sequences, the Text Abstraction task performances are very weak. Often, the best method was to combined a first step of text extraction followed by a text abstractive summarization on the result (which was a shorter text) at the cost of a loss of information.

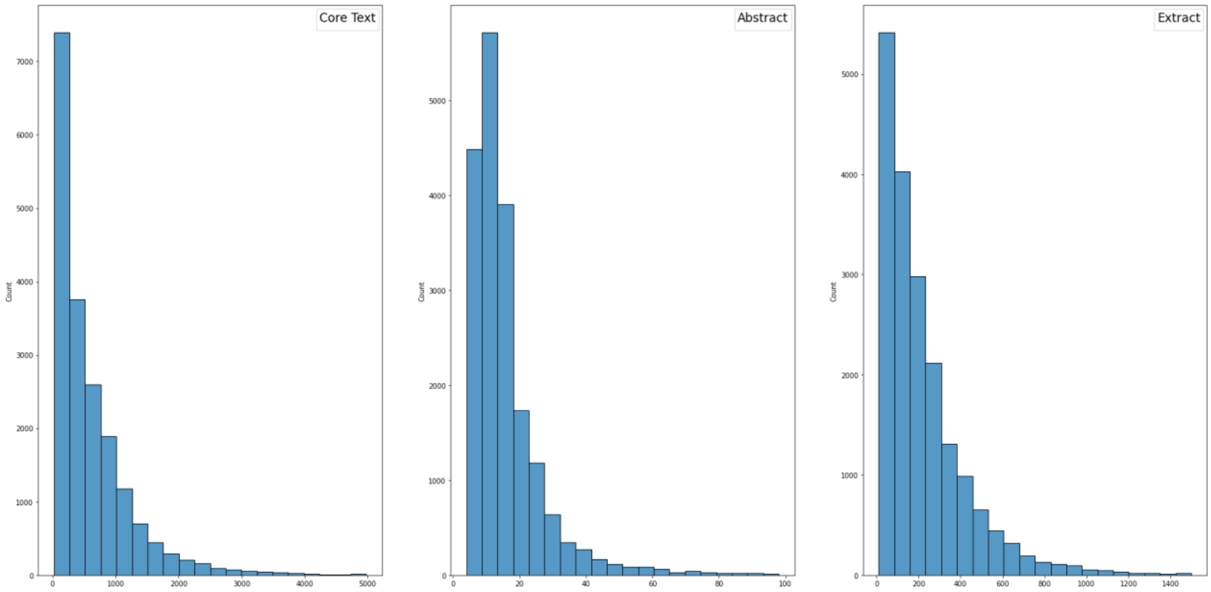
Another method was to chunked the long sequence into pieces that can be processed separately by a regular Transformer. The results were then concatenated into a bigger text on which a new step of text abstractive summarization was done.

The Debate Sum dataset

The Debate Sum Dataset is a dataset about competitive debates organized by the National Speech and Debate Association (US) since 2013. See a typical debates in the appendix (1).

The dataset consists of 187328 debates with:

- One core text which can count more than 5000 words;
- One Abstract that generally counts less than 200 words;
- One Extractive summary that counts less than 2000 words.



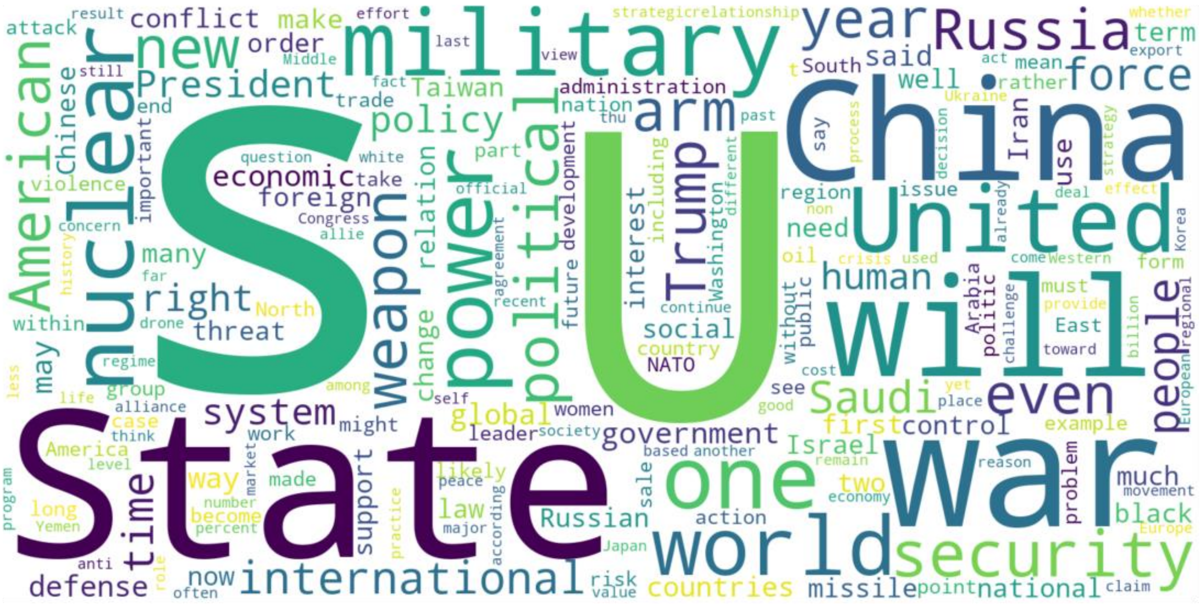
The themes of the debates vary each year but the general themes seem to be similar. During this experiment I used the 2019 debates with the following theme:

"The United States federal government should substantially reduce Direct Commercial Sales and/or Foreign Military Sales of arms from the United States."

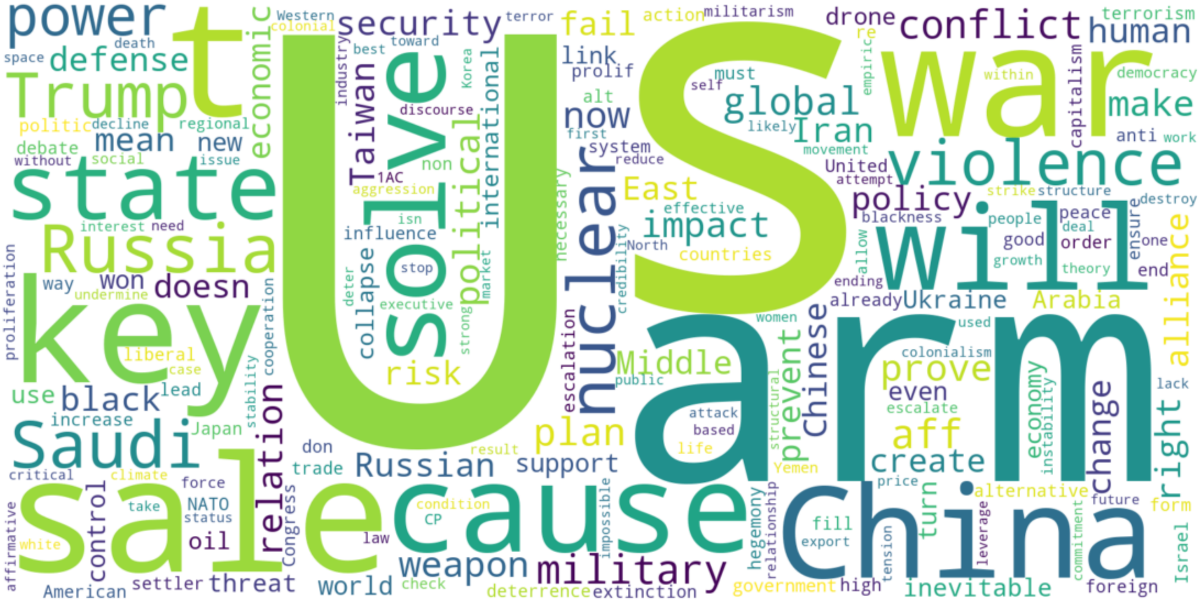
Thanks to a WordCloud analysis I can see which word is the most represented in the dataset for this year.

The same WordCloud analysis was done on the Abstracts to see if there were any difference and it seems that the abstracts are indeed covering the same topics with minor differences.

Core texts:



Abstracts:



Some words were less common in the Abstract in comparison to the Core texts. For example, **Sale**, **Key** and **Arm** are more common in the abstracts than in the core texts.

In this dataset, I used the Core texts and the abstract. The extractive summaries were discarded.

The preprocessing was done using the correct model tokenizers:

- BART Tokenizer which is smilar to the ROBERTa tokenizer, using a byte-level BPE.
- Longformer Tokenizer which is identical to the ROBERTA tokenizer.

Transformer architectures

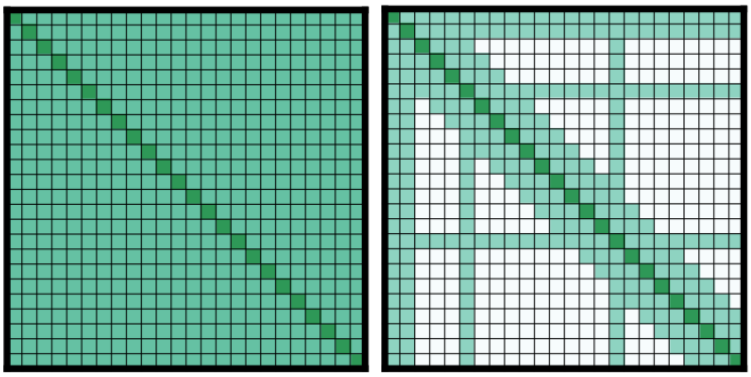
In order to compare the results, two transformers were used:

- The longformer and it new attention mechanism;
- The BART [5] transformer.

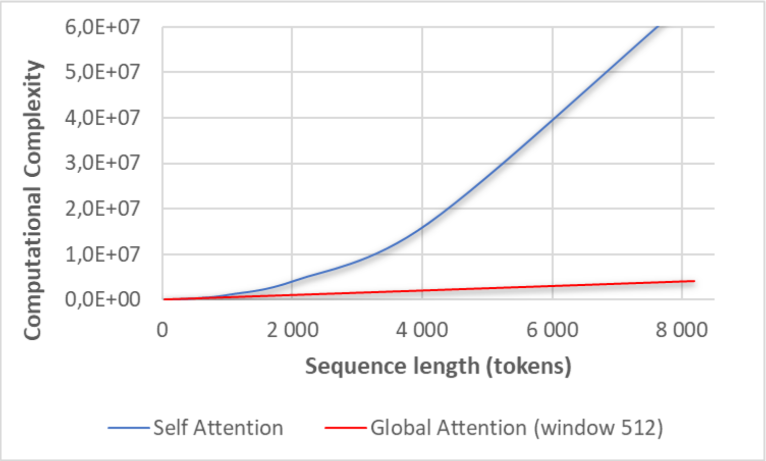
Longformer

The Longformer is a new type of Transformer that use a **Global Sliding Attention** mechanism. This mechanism is different from the regular Transformer because it doesn't compute the attention on every word. It is a 'sparse' attention that allows the model to be faster and to reduce drastically the memory costs.

To the left, the original self-attention that compute the scaled dot product to every word in the sequence. On the right, the Global Sliding attention mechanism that doesn't compute the Scaled Dot Product on every word (see the white squares). This figure comes from the Longformer paper [4]. A token with a global attention attends to all tokens across the sequence, and all tokens in the sequence attend to it.



The computational complexity of the global attention layer is O(nw) with n the sequence length and w the attention window size. It is far less expensive compared to the $O(n^2d)$ with d the embedding size of the regular transformers (regular self-attention).



The Longformer is mostly used for NLP tasks such as:

- Question Answering
- Language Modelling
- Abstractive summarization

Its usage it still very low compared to the regular transformer models such as BERT.

Longformer models can typically handle more than 1 thousand of tokens (piece of words) and can process sequences up to 16k tokens.

BART

BART uses a standard Tranformer-based neural machine translation architecture similar to the one of BERT and GPT. Thus, it uses the regular self-attention layer and its computational complexity is $O(n^2d)$.

It is one of the best models when it comes to text summarization.

BART has a sequence length limit of 512 tokens. Above this limit, the performance starts to drop significantly.

Methods

Performance comparison

This idea was to compare the validation loss on both models on different text length.

During this experiment I used the debates from the year 2019 and I kept the ones with a sequence length over 2000. I then conduct 4 different experiments with the same parameters on both models with different text lengths (512, 1024, 2048 and 4096). All of the texts come from the same text dataset and are truncated (and padded if needed) to the correct length.

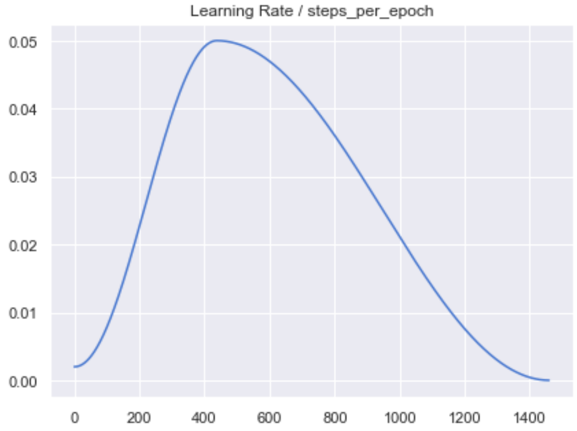
Both models were built with the help of the HuggingFace library architectures. The architectures were then wrapped into Tensorflow custom training loops.

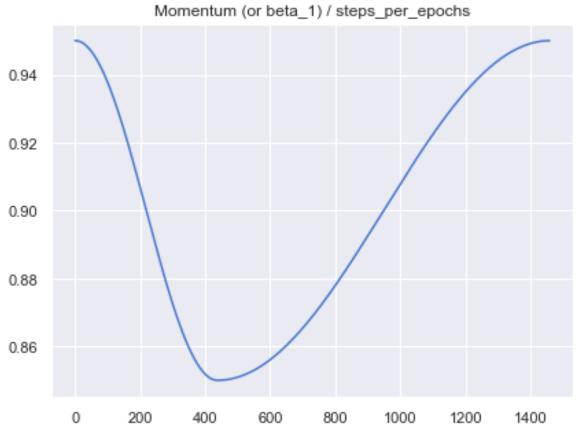
Training transformer models on very long texts can be difficult because:

- First, the training time is very long and the hyperparameter tuning is difficult;
- Second, the complexity of the model may lead toward overfitting and requests a lots of training data.

To answer both of these problems, I used pre-trained layers. The pre-trained used were the 'allenai/led-base-16384' for the longformer and the 'facebook/bart-base' for the BART.

In order to keep the pre-trained weights intact, I used a 1Cycle scheduler with a very low learning rate. The 1Cycle scheduler start with a low learning rate and then quickly increase to the maximum learning rate (1e-6). It does the opposite for the momentum which allows the model to stabilize around a good minimum.





The number of trainable parameters for each model were similar.

The BART model is composed of 6 encoders and 6 decoders with a feed forward network composed of two dense layers for a total number of trainable parameters of 139,470,681.

The Longformer model is also composed of 6 encoders and 6 decoders with a two dense layers feed forward network but for a total number of trainable parameters of 161,894,745.

Training and fine-tuning

Both models were then fine-tuned completely on 3140 texts with a length over 1000 words. The configs were identical from the previous experiment.

For the BART model, I used a sequence length of 512 tokens by truncating every text from the corpus to the correct length. For the Longformer model, the sequences were truncated at 4096 tokens and padded otherwise.

Training parameters:

- Batch size: 8 (Bart), 1 (longformer)
- Learning rate: 1e-5 (Bart), 1e-6 (longformer)
- Max_length : 512 (Bart), 4096 (Longformer)
- Epochs: 5
- Callbacks: Early Stopping + Scheduler 1Cycle

For the training a GPU P100 on google colab was used.

Note: to not harm the weights, I tried to freeze some part of the transformer (encoder, decoder, embeddings) and slowly unfreeze them during the training. However, the results weren't promising thus I skipped that part.

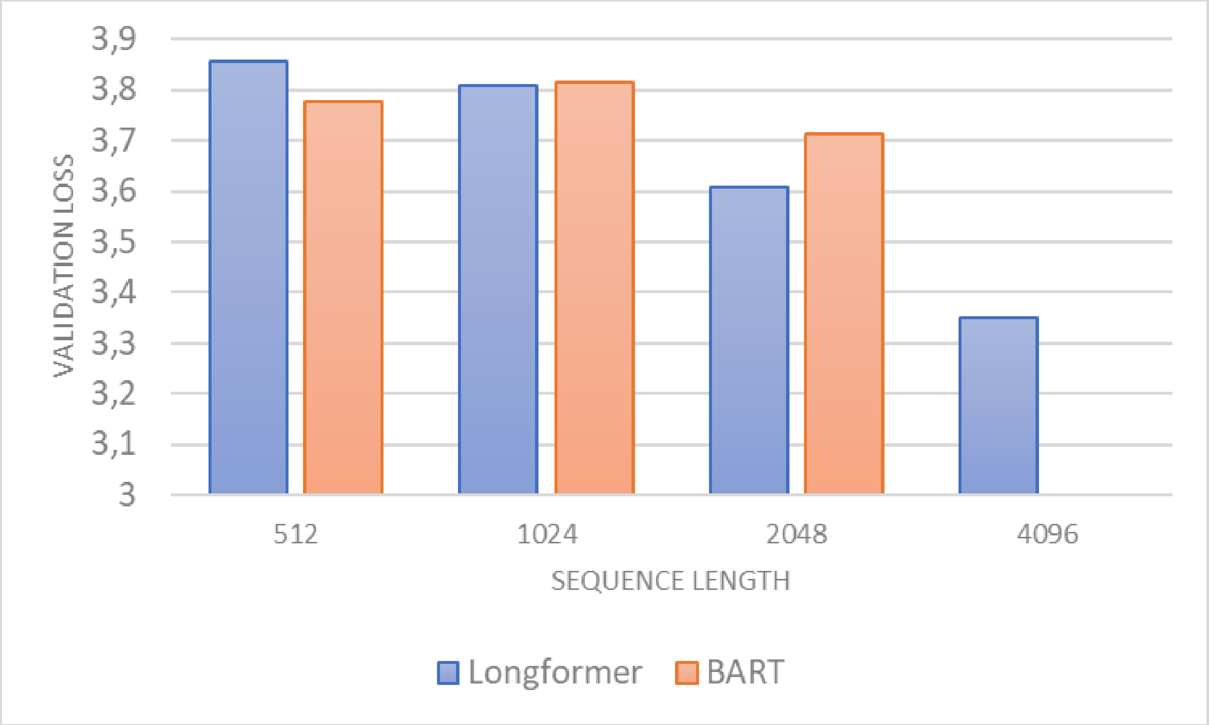
For the evaluation part, I used two different metrics:

- **BLEU** which is how much the words (and/or n-grams) in the machine generated summaries appeared in the human reference summaries. A n-gram is a sequence of words (2-gram is a sequence of two words).
- **Rouge-L** which is similar to BLEU but uses the sentence structure by keeping the longest common sequence to compute the scores. The order of the words of the sentence is the most important here.

Results

The comparison of the two models on text of different lengths showed that the Longformer effectively beats the Bart when the sequence size increase. It also showed that the BART memory consumption increases more than the Longformer one because the 4096 didn't fit into the GPU even if the BART had less trainable parameters.

The Longformer validation loss steadily decreases with the increase length whereas the BART validation loss didn't show any improvement and wasn't measurable with the 4096 tokens sequence length.



The experiment also showed that the BART outperforms the Longformer on short sequences. This can be due to the sparsity of the attention layer of the transformer.

The training was, however, always in favor of the BART model. This can be explained by the higher number of trainable parameters of the Longformer model used in this experiment.

The full training of the BART and the Longformer took 10 minutes and 1h10 per epoch respectively. This is largely due to the sequence length used for training (8 times smaller for the BART model).

Below the mean scores on the test set for the BART and the Longformer models.

	BLEU	Rouge-L
BART	15.57	13.65
Longformer	18.30	17.55

The results are a bit better for the Longformer. However, it is not enough to affirm that the Longformer is better than a regular transformer such as BART on this particular dataset.

This surprising result may be due to the fact that the abstracts of this dataset can be constructed from the beginning of the text. Hence, it seems that the end of the text doesn't bring anything new that the model doesn't already know from the beginning of the text.

Also, I used a batch size of 1 for the Longformer because of memory consumption. A very low batch size may reduce the impact of the first samples seen during each epoch (the model will gradually forget about the first samples).

Conclusion

The Longformer is an efficient tool for processing long sequences for Natural Language tasks such as text summarization. It reduces the computational costs compared to the regular transformer and improve performances thanks to its new attention mechanism.

However, as always in Deep Learning, there is no free meal. The dataset used there is very complex and it appears that the use of a regular transformer on the begin of the long texts is sufficient to keep up with the performances of the Longformer on long sequences (with 6 times less training time).

Papers & references

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[2] Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., ... & Ahmed, A. (2020). Big bird: Transformers for longer sequences. Advances in Neural Information Processing Systems, 33, 17283-17297.

[3] Kitaev, N., Kaiser, Ł., & Levskaya, A. (2020). Reformer: The efficient transformer. arXiv preprint arXiv:2001.04451.

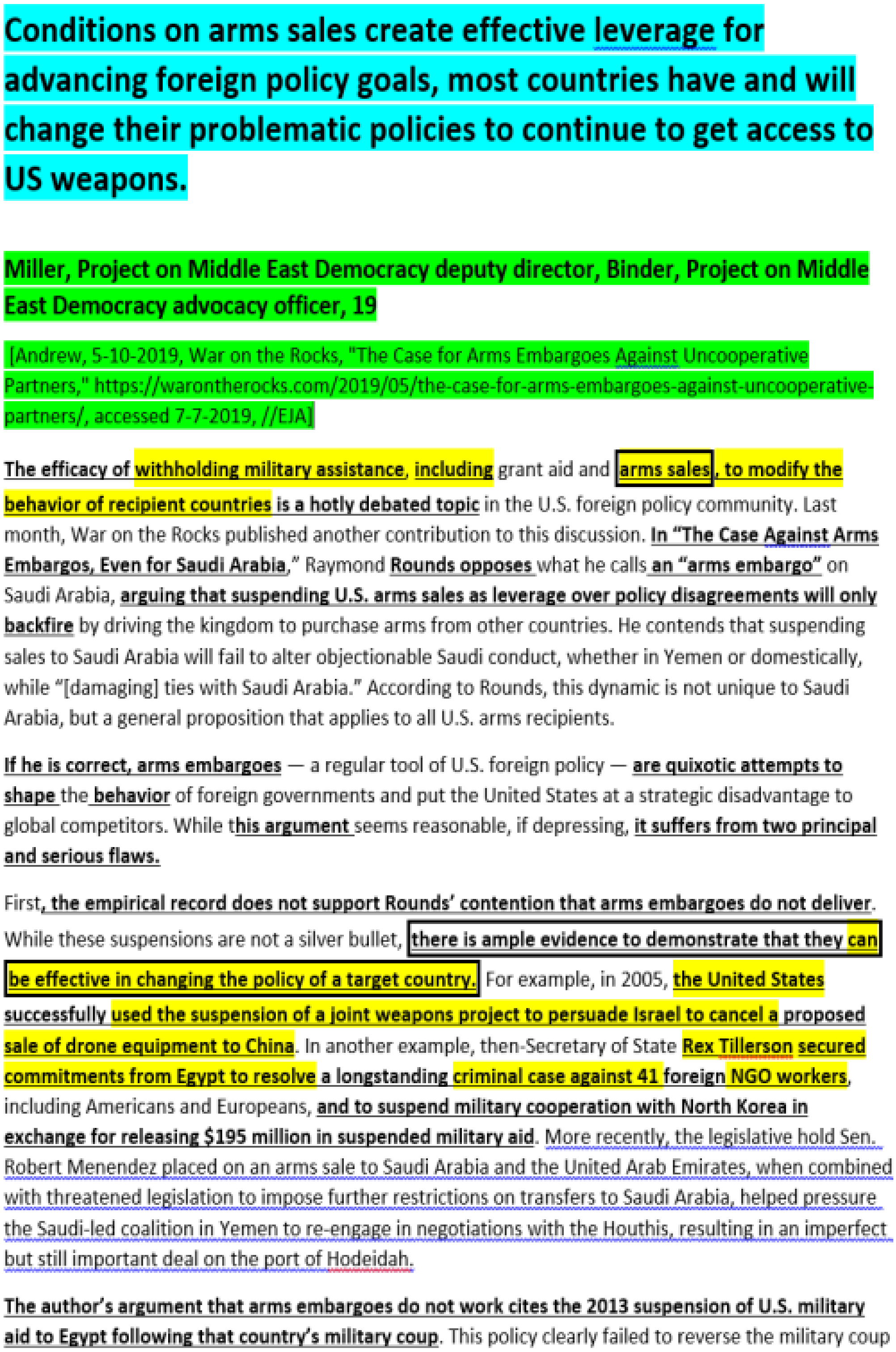
[4] Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150.

[5] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461.

[6] Smith, L. N. (2018). A disciplined approach to neural network hyper-parameters: Part 1--learning rate, batch size, momentum, and weight decay. arXiv preprint arXiv:1803.09820.

Appendix

(1) An example of debate (from the the paper of the dataset [5])



Conditions on arms sales create effective leverage for advancing foreign policy goals, most countries have and will change their problematic policies to continue to get access to US weapons.

Miller, Project on Middle East Democracy deputy director, Binder, Project on Middle East Democracy advocacy officer, 19

[Andrew, 5-10-2019, War on the Rocks, "The Case for Arms Embargoes Against Uncooperative Partners," https://warontherocks.com/2019/05/the-case-for-arms-embargoes-against-uncooperative-partners/, accessed 7-7-2019, //EJA]

The efficacy of withholding military assistance, including grant aid and **arms sales, to modify the behavior of recipient countries is a hotly debated topic** in the U.S. foreign policy community. Last month, War on the Rocks published another contribution to this discussion. **In "The Case Against Arms Embargos, Even for Saudi Arabia,"** Raymond **Rounds opposes** what he calls **an "arms embargo"** on Saudi Arabia, **arguing that suspending U.S. arms sales as leverage over policy disagreements will only backfire** by driving the kingdom to purchase arms from other countries. He contends that suspending sales to Saudi Arabia will fail to alter objectionable Saudi conduct, whether in Yemen or domestically, while "[damaging] ties with Saudi Arabia." According to Rounds, this dynamic is not unique to Saudi Arabia, but a general proposition that applies to all U.S. arms recipients.

If he is correct, arms embargoes — a regular tool of U.S. foreign policy — **are quixotic attempts to shape** the **behavior** of foreign governments and put the United States at a strategic disadvantage to global competitors. While t**his argument** seems reasonable, if depressing, **it suffers from two principal and serious flaws.**

First**, the empirical record does not support Rounds' contention that arms embargoes do not deliver**. While these suspensions are not a silver bullet, **there is ample evidence to demonstrate that they can be effective in changing the policy of a target country.** For example, in 2005, **the United States successfully used the suspension of a joint weapons project to persuade Israel to cancel a proposed sale of drone equipment to China**. In another example, then-Secretary of State **Rex Tillerson secured commitments from Egypt to resolve a longstanding criminal case against 41 foreign NGO workers**, including Americans and Europeans, **and to suspend military cooperation with North Korea in exchange for releasing $195 million in suspended military aid**. More recently, the legislative hold Sen. Robert Menendez placed on an arms sale to Saudi Arabia and the United Arab Emirates, when combined with threatened legislation to impose further restrictions on transfers to Saudi Arabia, helped pressure the Saudi-led coalition in Yemen to re-engage in negotiations with the Houthis, resulting in an imperfect but still important deal on the port of Hodeidah.

The author's argument that arms embargoes do not work cites the 2013 suspension of U.S. military aid to Egypt following that country's military coup. This policy clearly failed to reverse the military coup