Making Words Less Wordy

Applying the Transformer to Abstractive Sentence Compression

Khyati Agrawal

Advisor: Karthik Narasimhan

PRINCETON UNIVERSITY

Motivation

- Length of sentences can be reduced simply by reordering, substituting or deleting redundant words
- Summarization" or "deletion" has been the focus of most existing research to date
- Can human-like editing be mimicked using a ML model?

- Applications:
  1. Better news reporting by enhancing sentence readability
  2. Shortening papers and technical documents
  3. Learning tool for beginners in English language

Approach

- The field is "Abstractive" summarization is relatively unexplored
- The Transformer model has been successful for other applications such as Image generation and multi-document summarization

1. Is a Transformer model suitable for "Abstractive" sentence compression?
2. If yes, how can the model be fine-tuned for the task?

- **Experiment variables:**

1. Quality of dataset: "Best" or "Best and Average" target compressions
2. Encoder hyperparameters: Discussed in detail on the right
3. Decoder parameters: Tested various values for "beam-size" and "alpha"
4. Training time: Tested 20K, 30K, 50K or 125K training steps
5. Types of regularization: Tried attention dropout and layer prepost process dropout

Hyperparameter Tuning

Hyperparameter	Initial	Tuned	Tried values:
Hidden Layers	6	3	Integer values in [2, 4]
Hidden Size	512	256	Values in: {128, 256, 512}
Learning rate	0.05	0.1925	Float values in [0.05, 0.25]
Attention dropout	0.1	0.6	Float values in [0.4, 0.7]

- All hyperparameters moved towards a simpler model
- Best "Beam Size" was 4 and "Alpha" was 0.6

Training steps? Either 20K or 30K

Results

Model	CR	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
OperationNet	0.655	0.362	0.174	0.337	.2630
TunedTransformer	0.777	0.757	0.568	0.755	0.440

Table 6: Comparison with (Yu et 2018)

Input: Two of these studies provide the basis to form ratios of the WTP of different age cohorts to a base age cohort of 40 years. These ratios can be used to provide Alternative age-adjusted estimates of the value of avoided premature mortalities.

Prediction: Two studies provide the basis to form of the WTP of different age cohorts to a base age cohort of 40 years. These ratios can be used to provide Alternative age-adjusted estimates of avoided premature mortties.

Lot of room syntactic improvement!
Post-processing to correct grammar and spelling

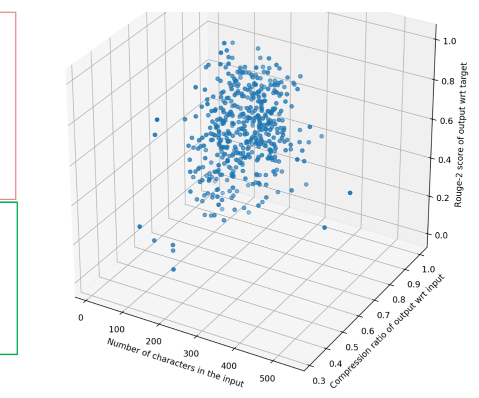
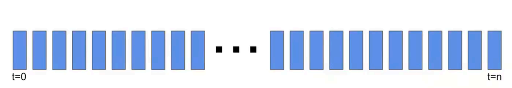


Figure 1: Variation of metrics with input length

Prior Approaches

- **Task: Text to text sequences**

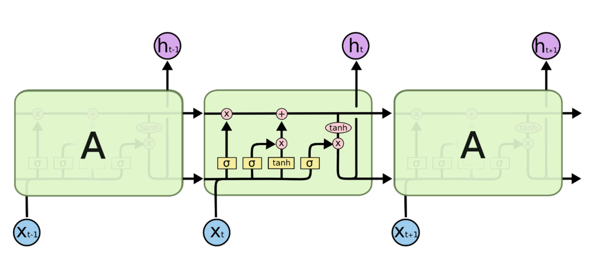


"Abstractive" compression summary:

- Encoder-Decoder (use of Two LSTMs) architecture used in Machine Translation applied in 2014
- Breakthrough when "Attention" mechanism was introduced by Bahdanau et 2016
- Most work focused on "Abstractive" summarization, i.e. conversion of a document a summary of few lines

"Extractive" sentence compression summary:

- Started off with Rule-based models using ILP
- RNNs marked the shift from Rule-based to learning models
- LSTM (Gated RNN) solved the vanishing gradient problem beating the state of the Art



The repeating module in an LSTM contains four interacting layers.

A simple illustration of attention:

- The context vector is simply a vector of weights that represents the importance of each input for translating a particular input.
- For example: To translate "I" to "Je", the model learns to attend to "I" the most, and learns to give high importance to "am" because the translation for "I" in French is either "Je" or "J' " based on the succeeding verb.

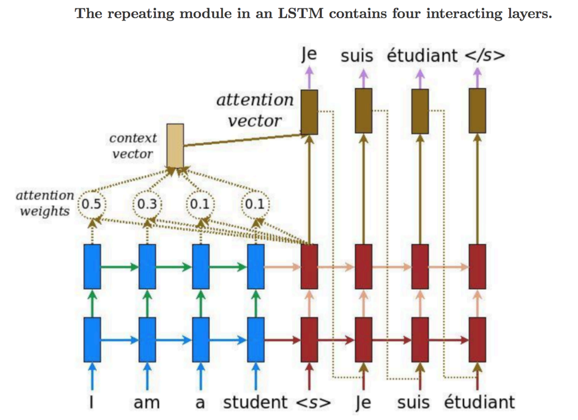


Figure 8: Illustration of the "Attention" mechanism for Machine Translation tasks

Lot of work on summarizing documents but what about sentences?

- "Operation Network" by Yu et 2018 : ROUGE-2 recall score of .174 and a compression ratio of 0.65 on the MSR Abstractive sentence compression dataset. (I compare my results to this model as a benchmark)
- Very little work on the sentence level

Implementation- Model

The Transformer model and mechanism from Vaswani et 2018

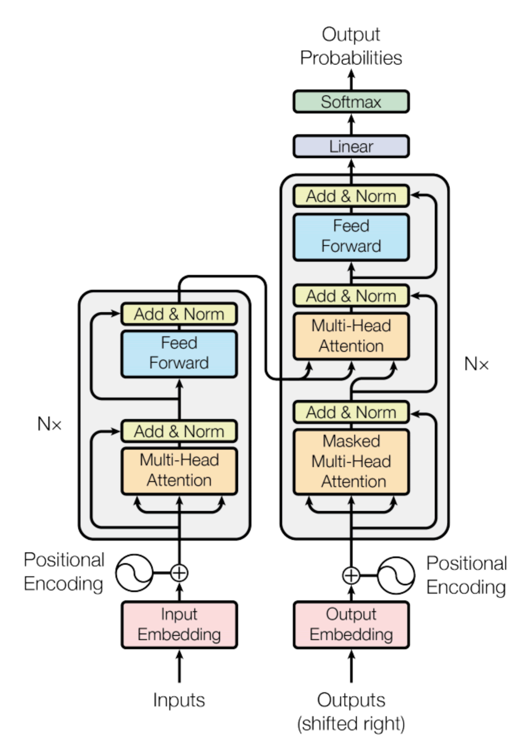
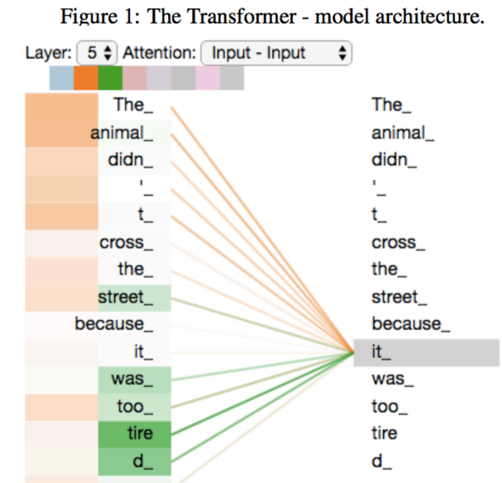


Figure 1: The Transformer - model architecture.



- Encoder-Decoder architecture
- The Encoder has any number N of layers, where each has 2 sub-layers:
  - **Multi-head self attention** layer
  - Fully connected feed forward layer.
- Residual connections: output from each sub-layer becomes LayerNorm(x + Sublayer(x))
- Attention in the Transformer:

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

Salient features:

- No convolutions, no recurrence
  - Attention is all you need!
  - Fast-parallelizable training:
  - Sequential operations reduced from O(n) in RNNs to O(1)
  - No vanishing gradient problem
  - Maximum path length between two time steps reduced from O(n) to O(1)
- Multi-head attention for multiple contexts

Benefits from using Multi-head attention:

- Notice in the figure on the left
- **"it"** can be encoded to independently attend to the **"the animal"**, i.e. the noun it replaces, and to **"tired"** i.e. the state of that animal.
- Distant and finer dependencies captured better

Improvement from default

Input	Old Prediction	New Prediction
Meantime, she has graduated from Coppin State College, works as a drug and alcohol counselor, left subsidized housing and plans to start work on a master's degree this year.	Members has graduated from Coppin State College, works and alcohol **counselzed** and plans to start on a master this year.	Meantime, she has graduated from Coppin State College, works as drug and alcohol counsel plans to work on a master's degree
We'll also award more small grants through our Species Action Fund to help a wide range of animals, including whooping cranes, giant river otters, and Tibetan antelopes.	We'll also award more small grants through our Species **ion** Fund, including **whoing** cranes, giant river and Tibetan **pes**	We'll award more small grants through our Species Fund to help a wide range of animals, including **woping** cranes, giant river ters, and Tibetan antelopes.

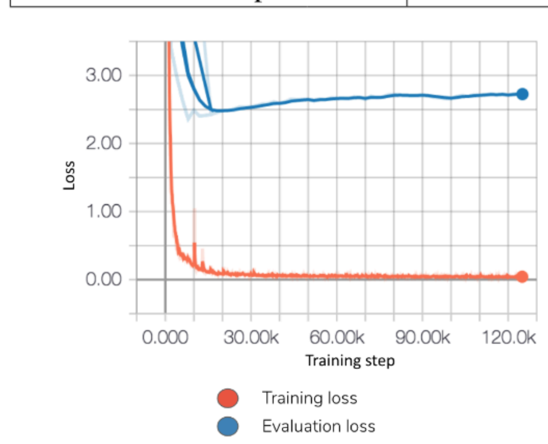


Figure 2: Transformer Default

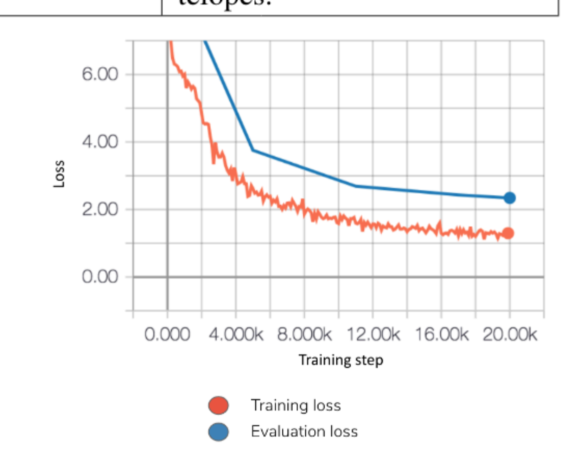


Figure 3: Transformer Tuned

Model	ROUGE-2 F1	ROUGE-L
Base Transformer	0.508	0.610
Hyperparameter Tuned	0.530	0.613

- **Reduced overfitting**
- **Better qualitative predictions**

Analysis

More from the Figure 1:

1. Trade off between compression quality and compression ratio: High compression tends to come at the cost of low ROUGE-2 recall scores

From the Figure 2:

1. Target CR is around 0.7
2. Model's CR around 0.77
3. Model's performance to comparable to targets

Caveats:

1. Inconsistent CR
2. Huge dispersion in CR for small inputs
3. Low ROUGE-2 scores for outlier sentences: i.e. input length <100
4. Low compression ratios for input sentences with length > 300

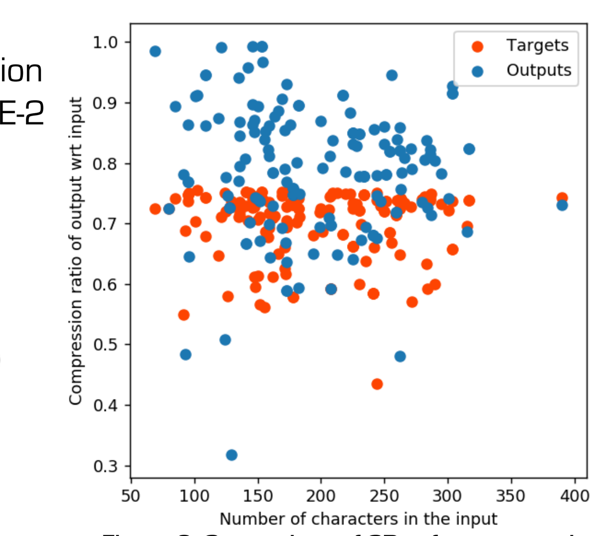


Figure 2: Comparison of CRs of outputs and Targets

Conclusions

- **Average CR: 0.77 and average ROUGE-2 F1: 0.54**
- Is the Transformer is suitable for "Abstractive" Sentence Compression? YES!
- **Fine tuning: simple transformer works better**
  - Reduced number of layers
  - Reduces hidden size
  - Increased Learning rate
  - Increased Attention Dropout
  - Increased layer pre-post process dropout
- **Data quality and input length are important considerations**

Dataset

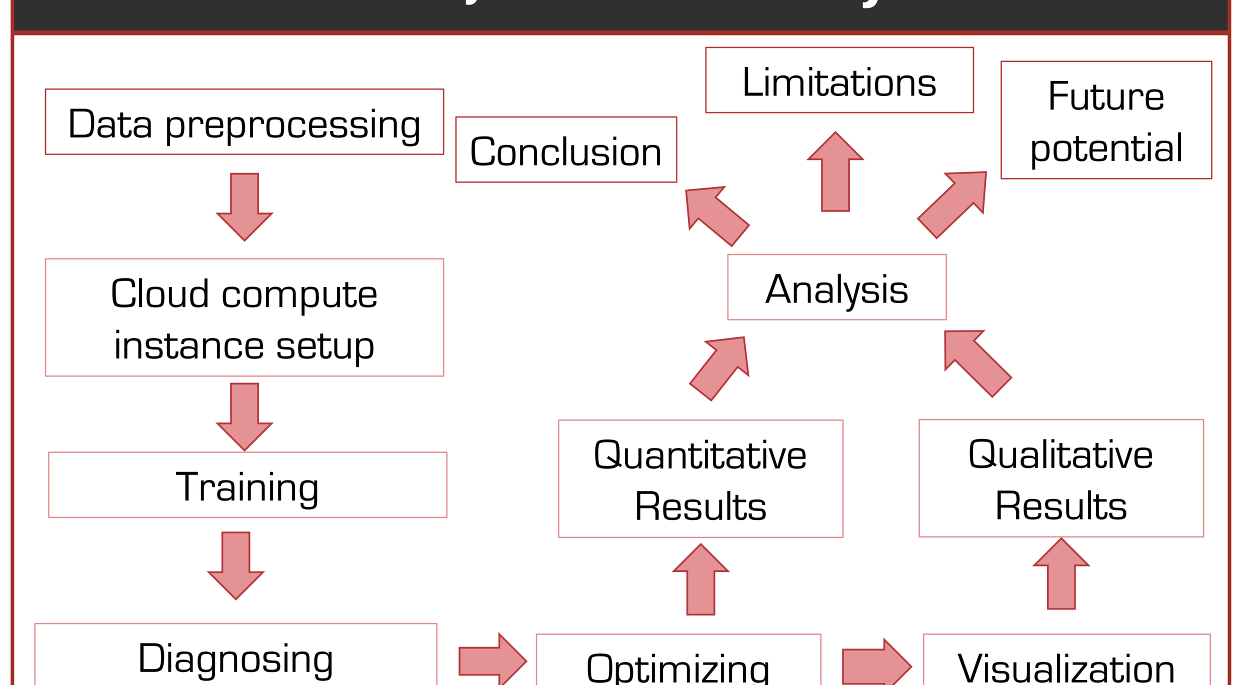
Used the Microsoft Research Abstractive Text Compression dataset:

- Used in Yu et 2018
- 26,000 pairs, 6,000 unique sentences
- Diverse: From news journals, business letters and technical documents
- I utilized the "best" rated compressions from this dataset
- That is a subset containing **8,290 pairs for training** and **921 pairs for evaluation**
- Open source and available online at: https://www.microsoft.com/en-us/research/project/intelligent-editing/

Acknowledgements

- **Thank you Karthik Narasimhan for guidance**
- **Thank you Princeton COS dept. for funding and support**

Project summary



Metrics

$$\text{ROUGE-2}_{\text{Recall}} = \frac{\text{Number of overlapping bigrams in the model outputs and targets}}{\text{Number of bigrams in the targets}}$$

$$\text{ROUGE-2}_{\text{Precision}} = \frac{\text{Number of overlapping bigrams in the model outputs and targets}}{\text{Number of bigrams in the outputs}}$$

$$\text{ROUGE-2}_{\text{F1}} = \left(\frac{\text{ROUGE-2}_{\text{Recall}}{}^{-1} + \text{ROUGE-2}_{\text{Precision}}{}^{-1}}{2}\right)^{-1}$$

$$\text{Compression Ratio (CR)} = \frac{\text{Number of characters in the output}}{\text{Number of characters in the input}}$$

Additionally reported:

- **BLEU score:** standard in Machine translation
- **ROUGE-L:** same as ROUGE-2 except uses length of longest common sequence
- **ROUGE-1 scores:** same as ROUGE-2 except uses unigrams

Project Future Work

- Sophisticated model to handle input sequences of different lengths differently
- Model that accepts "desired" level of compression from user that will help find the right balance between compression amount and compression quality
- **Hybrid rule-based and learning model**
  - Introduction ILP constraints in the decoder
  - Thresholding for the encoder, i.e. Leaving outlier sentences unchanged
- Making the model usable, developing scripts for correcting grammar and spelling for predictions

References

Image sources in order:

1. https://www.youtube.com/watch?v=cS2UZKHq4i4&t=1905s
2. http://colah.github.io/posts/2015-08-Understanding-LSTMs/
3. https://medium.com/syncedreview/a-brief-overview-of-attention-mechanism-13c578ba9129
4. https://arxiv.org/abs/1706.03762
5. https://arxiv.org/pdf/1706.03762.pdf

References:

1. D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014. Available: http://arxiv.org/abs/1409.0473
2. A. Vaswani et al., "Attention is all you need," CoRR, vol. abs/1706.03762, 2017. Available: http://arxiv.org/abs/1706.03762
3. N. Yu et al., "An operation network for abstractive sentence compression." Available: https://www.aclweb.org/anthology/C18-1091