EXTRAS: Exploring Transformer models for Abstractive Text Summarization

Aniruddha Bala
Indian Institute of Science
Bangalore, India
aniruddhab@iisc.ac.in

Raghu Chittersu
Indian Institute of Science
Bangalore, India
chittersuv@iisc.ac.in

Abstract

Sequence to Sequence models have been widely used for text summarization tasks. Transformer models in machine translation have proved that one doesn't need recurrent structures to capture sequential information. Also, Transformers make use of extensive parallelism and have been shown to outperform sequence to sequence models. So in this paper, we explore the transformer models for the task of Abstractive text summarization and propose a way to integrate the pointer generator mechanism into transformer models. Using the Transformer model with BERT encoder and pointer mechanism we are able to beat the sequence to sequence baselines.

1 Introduction and Problem Statement

Text summarization can be defined as the concise and fluent presentation of information present in a document. There are essentially two approaches to this extractive summarization and abstractive summarization. Extractive summarization techniques copy direct sentences from the paragraph based on their importance whereas abstractive techniques include novel phrases and make the summary look more natural. Some of the recent approaches of abstractive text summarization attempt to solve problems related to the repetitive occurrence of phrases and out of vocabulary words. However, there are still a few unsolved problems such as incorrect reproduction of facts due to loss of positional information, generation of non-sensical text.

In our work, we verify the performance of recent models such as Transformers and BERT which have been shown to be effective in other tasks but remain unexplored in the domain of text summarization.

2 Related Work

Almost all the recent literature on Abstractive Text Summarization is based on the sequence to sequence models. (Nallapati et al., 2016) provided the first baselines for the task on CNN/Daily Mail dataset with Sequence to sequence models. This is the first time Abstractive text Summarization had been applied on large sentences. The same authors used hierarchical RNNs in their new approach and outperformed their previous result.

See et al. (2017) uses Pointer Generator mechanism to choose between generating a token and copying from input sequences, this helps particularly in case of generating rare words. They also used Coverage Mechanism along with coverage loss to address repetition.

So far we have seen sequence to sequence models used as transducers. Vaswani et al. (2017) proposed a novel architecture which replaces the recurrence and convolutions in neural models by the self-attention mechanism. The work shows significant improvements in BLEU score for the Machine Translation task. It is also shown that these models are more parallelizable and train faster.

BERT(Bidirectional Encoder Representation from Transformer) the work by Devlin et al. (2018) uses Transformer model to learn bidirectional contextual representations. The learned transformer encoder has been used in various downstream tasks to obtain state of the art results.

3 Contributions

In this work, we explore the effectiveness of Transformer models for Abstractive Summarization. To the best of our knowledge transformer models haven't been applied to abstractive text summarization task yet. The novel contribution of this work is that we integrate the Pointer Generator mechanism (See et al., 2017) with the Trans-

former model to handle out of vocabulary words. We further show performance improvements with transfer learning by using BERT encoder.

4 Methodology

We have implemented the following three models which we describe in the subsequent sections. The code for our models are available here.[1]

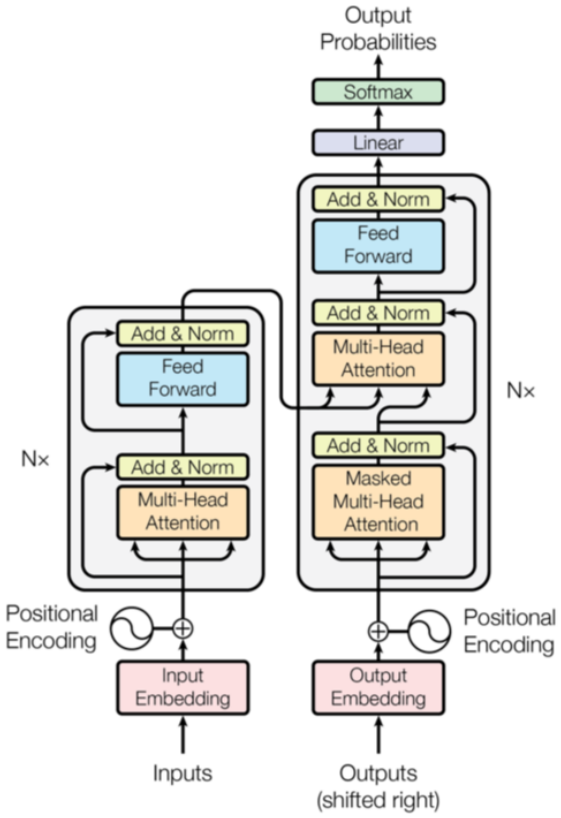

Figure 1: Transformer model architecture (Vaswani et al., 2017)

4.1 Transformer

In this architecture like any other transduction model, we have an encoder to capture source representation and decoder to generate the target sequence.

Input: The Transformer has access to all the input sequence at once, hence to capture the relative positional information between words positional encodings are used calculated using the following formula

$$PE_{pos,2i} = sin(pos/10000^{2i/d_{model}})$$

$$PE_{pos,2i+1} = cos(pos/10000^{2i/d_{model}})$$

[1] https://github.com/ani555/EXTRAS

where pos is the position and i is the dimension and d_{model} is the hidden dimension. This positional encodings are added to the input embeddings of the words.

Multi Headed Attention Layer: For each attention head the queries, keys and values are generated by projecting the representations at any layer for encoder/decoder using weight matrices W_Q, W_K and W_V respectively as follows.

$$Q = W_Q X$$

$$K = W_K X$$

$$V = W_V X$$

Given this matrices the attention is calculated using scaled dot product attention as follows

$$X_c = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

where X_c is matrix of context vectors.

Position Wise Feed Forward Network: The position wise feed forward layer is implemented as per the formula

$$FFN(X) = max(0, XW_1 + b_1)W_2 + b_2$$

. Here we project the input X to d_{ff} dimension vector using W_1 and back to d_{model} using W_2

Encoder Layer: Each encoder layer comprises of a multi-headed self-attention layer followed by layer normalization and position wise feedforward network. There are also intra-layer residual connections. In the multi-headed self-attention layer the queries, keys, and values are generated from the input encoder representation for that layer.

Decoder Layer: Each decoder layer comprises of a multi-headed decoder-decoder self-attention followed by multi-headed encoder-decoder attention and position wise feedforward network. Here also intra-layer residual connections are used. To calculate the decoder-decoder self-attention the queries, keys, and values, are generated from the decoder input representation, and for encoder-decoder attention, the queries are the decoder states and keys and values are the final encoder states. The final model comprises of stacks of encoder layer connected with stacks of decoder layers.

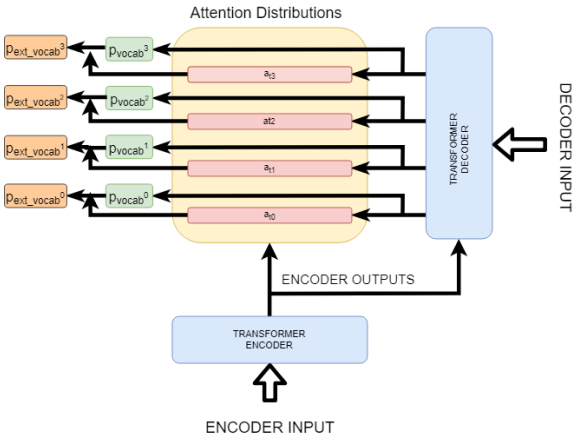

Figure 2: Figure showing proposed way of integration of Pointer Generator with Transformer

4.2 Transformer with Pointer Generator

Pointer Generator Here we describe our method of integrating the pointer generator mechanism (See et al., 2017) with the above model (refer figure 1). We first apply as single-headed attention using the final output of the encoder and query as the final output of the decoder. These gives the attention probabilities on the encoder sequence for each decoder time step. The generation probabilities are calculated in the following way

$$p_{gen} = sigmoid(W_{gen}[X_d, X_c] + b_{gen})$$

where X_d is the final decoder output, X_c is context obtained after attention layer and W_{gen} and b_{gen} are parameters of linear layers. The output distribution over the vocabulary is calculated as

$$P_{vocab} = softmax(W_{vocab}X_d + b_{vocab})$$

where W_{vocab} and b_{vocab} are parameters of linear layer.

The probability distribution over the extended vocab is calculated as

$$P_w = p_{gen} * P_{vocab} + (1 - p_{gen}) * \sum_{i:w_i=w} a_i^t$$

where a is the attention vector t is the decoder time step and i is the encoder sequence index.

4.3 Transformer Decoder with Pointer Generator and Bert Encoder

In this setup, we replace our transformer encoder with pre-trained BERT encoder. Here we use the BERT base uncased model. We change the model hidden dimension d_{model} to match with the BERT d_{model} which is 768 and also the vocabulary size to 30524 to match the BERT vocab size. We then freeze the BERT encoder and train our decoder.

In the above three models, we use loss function as label smoothing with KL divergence between the true and the predicted distribution.

5 Datasets and Metrics

To run our experiments we use the CNN/Daily Mail dataset (Hermann et al., 2015; Nallapati et al., 2016) which is used extensively for abstractive summarization task. The dataset contains a multi-sentence abstractive summary for news articles on CNN /Daily Mail websites. We have used the scripts supplied (See et al., 2017) to obtain the same version of data used by previous others which has 287,226 training pairs, 13,368 validation pairs, and 11,490 test pairs.

For evaluation, we use ROUGE scores which is a standard metric used for evaluating summarization tasks. We have reported the scores for ROUGE-1, ROUGE-2, and ROUGE-L (which respectively measure the unigram-overlap, bigram-overlap, and longest common sequence between the reference summary and the summary to be evaluated).

6 Baselines

We refer to the work by See et al. (2017) and consider 3 different models as our baselines.

6.1 Sequence-to-sequence attention model

The sequence to sequence attention model uses single layer bi-directional LSTM as encoder and a single uni-directional LSTM as decoder. Bahdanau attention is used to generate context vector which is concatenated with decoder output to make the prediction.

The above model is used as baseline with two different vocabulary sizes 150K and 50K. Both of them were trained for 600K iterations with batch size 16.

6.2 Pointer-generator network

The Pointer Generator network is augmentation of Pointer Mechanism to the above model. The final distribution is based on both the generation probability and pointer probability.

The above model has vocabulary size of 50K and has been trained for 230K iterations with

Model	ROUGE-1	ROUGE-2	ROUGE-L
Seq2seq + attn baseline (150k vocab) (See et al., 2017)	30.49	11.17	28.08
Seq2seq + attn baseline (50k vocab) (See et al., 2017)	31.33	11.89	28.83
Pointer Generator (See et al., 2017)	36.44	15.66	33.42
Transformer (**Our Model**)	12.79	1.18	11.72
Transformer + Pointer Generator (**Our Model**)	28.52	7.02	26.14
BERT + Pointer Generator (**Our Model**)	30.98	12.25	28.93

Table 1: Result shows a comparison of our model performance with three baselines

batch size 16.

In both models the hidden dimension is of size 256 and embedding dimension of size 128 were used.

7 Experimental Setup

For our experiments, we follow similar setup as (See et al., 2017) i.e. we truncate our input articles to length 400 and the target summaries to length 100. Hence the encoder sequence length and the decoder sequence length are 400 and 100 respectively. We use vocab size of 50,000 for models 1 and 2 and use BERT vocab of size 30524 for model 3.

For models 1 and 2 we use $d_{model} = 512$ and $d_{ff} = 1024$ and for model 3 we use $d_{model} = 768$ matching the BERT encoder hidden dimension. For both encoder and decoder we use $N = 6$ layers. We use dropout probability of 0.1 during training. We use Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-9}$ and vary the learning rate according to the formula

$$lr = d_{model}^{-0.5} * min(step^{-0.5}, step * warmup^{-1.5})$$

same as (Vaswani et al., 2017). The variation of the learning rate is shown in figure -.

During training we use $batch\_size = 8$ and $max\_train\_steps = 200000$

8 Results

The table [1] shows the comparison between seq2seq baselines and our models.

9 Discussions

We see that for the Transformer model without Pointer Generator the ROUGE scores are very low. We suspect that the learnt probability distribution over the vocabulary is not good enough. We are able to say this because on addition of Pointer

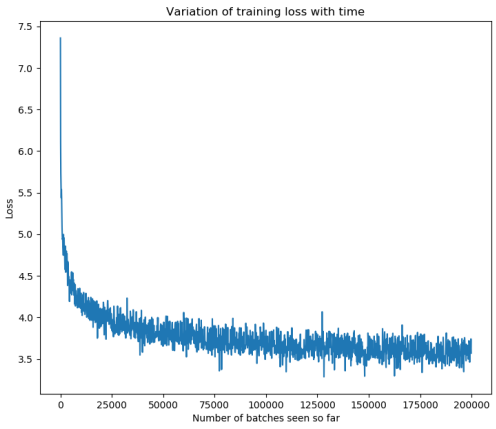

Figure 3: Variation of train loss with time

mechanism in which we superimpose the attention distribution with the vocabulary distribution, the results significantly improve. We provide here two explanations for the performance gains. First, the attention distribution on the input sequence helps by eliminating unknowns in the output sequence by directly copying the unknown from the input sequence. Second, it increases the probability mass of the words in the output distribution which receive more attention in the input sequence.

By replacing our Transformer encoder with the pre-trained BERT encoder we see further improvements in ROUGE scores. This is because the BERT model which has been trained on large corpus of text learns better language representations. Use of such pre-trained input representations also reduces the model training time, as we only need to fine-tune our decoder on top of the encoder.

10 Conclusions and Future Work

From the above results, we observe that Transformers for Abstractive Summarization show promising results. From our observations we can

draw the following two conclusions. First, the Pointer Generator mechanism is a useful technique to augment and enhance the output probability distributions. Second, transfer learning can ease the model's workload of training end to end and help in achieving faster convergence with better performance.

Our models were trained for almost half the number of epochs as compared to that by See et al. (2017) yet the models 2 and 3 are able to generate meaningful multi-sentence summaries with comparable ROUGE scores. Also for model 1, due to constraint of time we could not train our Transformer model for larger epochs, as done by See et al. (2017) for their baseline models. Future work should focus on improving the performance for the Transformer model alone by exploring different settings of hyper-parameters and exploring more complex models and training for larger epochs.

There are also a few shortcomings of our model. Our model may sometimes represent facts incorrectly in the output summary. This problem is inherent in present day state of the art abstractive summarizers. Our hypothesis was that using transformers which use self attention mechanism to model the dependencies between input sequence tokens we would be able to solve the problem. However that did not seem to work and the problem remains unsolved. On some occassions it also produces repetitive summaries. A solution to this as proposed by See et al. (2017) is to maintain a coverage vector and include an additional term in the loss to penalize the model for generating repetitive phrases. Such a mechanism can not be directly adapted for Transformers as the coverage vector is depended on the attention outputs from previous time steps. This introduces recurrence which violates the key principles on which Transformers are built.

To exploit coverage mechanism for Transformers we propose to use two layers of attention at the end of the final decoder layer. The attention output probabilities from the penultimate attention layer will be used by the final attention layer to calculate the coverage vector. To calculate the coverage vector we use a upper triangular mask of ones which we multiply with the attention distributions obtained from the previous layer instead of summing across the previous time steps. This gives the coverage vector for all the decoder time steps at once. Although, we are not able to include the results of this experiment, but this opens up further scope for future work.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1693–1701. Curran Associates, Inc.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

A Appendices

The appendix section shows the comparison of our model's predicted summaries with reference summaries for different articles. We show three different examples in table [2], the analysis for the same is presented below.

Example 1: We observe that the predicted summary is actually better than the reference summary. The reference summary is not appropriate for the given article whereas our model gives a better gist of the article. In such cases the ROUGE score obtained is low and affects the model overall performance.

Example 2: In this example the reference summary turns out to be better but the predicted summary is also not bad. One thing to note here is that our model confuses "gerardo's" age as 41 instead of 44 in the second line predicted. (An example of incorrect production of facts.)

Example 3: In this example the model repeat's the phrase "will be visible on saturday".

Article: a kentucky man has been arrested after police say he was found under the influence while riding a horse on us 23 . michael kimmel , 40 , was taken into custody by kentucky state police on monday evening after they received a 911 call about an intoxicated horse rider . trooper j. gabbard 's report says that kimmel was ordered to stop , but instead dismounted and ran away wearing only a brown hat , jeans and boots . trooper gababrd launched a manhunt for kimmel and later found him on horseback again and according to the floydcountytimes , he resisted arrest , saying , ' i did n't do s *** , i was just riding my horse . ' according to the arrest report , kimmel would not take a sobriety test and refused a breath and blood alcohol test . however , officers said he had slurred speech , smelt of alcohol and was unsteady on his feet . " subject made threats to ' -lsb- expletive -rsb- -lsb- expletive -rsb- up driving drunk in a car next time and he would give me something to worry about , "' the arrest citation quotes kimmel as saying . kimmel , who has used the alias " mike bicycle , " is currently on probation for a prior conviction for burglary . he is currently + in the floyd county jail on $ 5000 cash bond . kimmel faces dui , fleeing or evading police , and other possible charges . **Reference Summary:** michael kimmel was taken into custody wearing only boots , jeans and a cowboy hat . **Predicted Summary:** michael kimmel , 40 , was taken into custody on monday . police received 911 call police call about an intoxicated horse rider . he was arrested after they received a 911 call about an intoxicated horse rider .
Article: a mother-of-three was shot and killed by her husband who then turned the gun on himself , according to police in tulare , california . the couple have been named as georgina rojas-medina , 41 , and her common-law husband , gerardo tovar , 44 . neighbors say they were alerted to the bodies by the couple 's 4-year-old daughter just after midnight on saturday . mother-of-three georgina rojas-medina , 41 , was shot and killed by her common-law husband , gerardo tovar , 44 , who then turned the gun on himself on saturday night , according to police in tulare , california . bloody footprints show the path the 4-year-old girl had to make after finding her mother shot to death and a father who 'd killed himself . bloody footprints show the path the 4-year-old girl had to make after finding her mother shot to death and a father who 'd killed himself , reports abc30 . ' the 4-year-old who was present on scene at the time of the incident was not injured , ' said sgt. andrew garcia of the tulare police department . **Reference Summary:** mother-of-three georgina rojas-medina , 41 , was shot and killed by her common-law husband on saturday night . gerardo tovar , 44 , then turned the gun on himself in tulare , california . the couple 's bodies were discovered by their 4-year-old daughter who walked out of the house and asked a neightbor to call the police . the 4-year-old girl and two other siblings , ages 10 and 12 , were not hurt in the shooting . **Predicted Summary:** mother-of-three georgina rojas-medina , 41 , was shot and killed by her husband , gerardo , 44 , 44 , on saturday night . her husband , gerardo , 41 , then turned the gun on himself .
Article: the moon is set to turn a sinister-looking blood red this easter weekend . the total lunar eclipse will transform the moon on saturday night and will be visible in the skies of north america , asia and australia . according to one us pastor , the event was predicted in the bible and hints at an imminent world-changing event , but nasa is quick to point out that the change in hue is entirely harmless . scroll down for video . saturday 's total lunar eclipse , which will turn the moon a burnt reddish orange , will be visible in in the skies of north america , asia and australia . a previous total lunar eclipse as seen from mexico city on december 21 , 2010 is pictured . the eclipse is the third in a series of four blood moons , with the final one expected on september 28 . pastor john hagee told the mirror that the sight suggests a world-changing event will take place , as predicted in the bible . according to the king james bible : ' the sun shall be turned into darkness , and the moon into blood , before the great and the terrible day of the lord comes . ' mr hagee , who has written a book on the tetrad called ' four blood moons ' , said that because the blood moon falls on easter weekend , it 's a sign that ' something dramatic will happen which will change the whole world ' . the event is called a tetrad since it involves four successive total blood red lunar eclipses each followed by six full moons . **Reference Summary:** blood moon will be visible in the skies of north america , asia and australia . it 's the third blood moon in the ' tetrad ' that will end in september . red moons are predicted in the bible and signify important events . us pastor says strange event could predict the second coming . **Predicted Summary:** the total lunar eclipse will be visible on saturday night of north america , asia and australia and australia will be visible on saturday 's . the moon is the third in a series of blood moon .

Table 2: Table shows the articles and the reference summaries with our predicted summaries