Implementing Deep Learning Models for Text Summarization

James Nguyen
UC Berkeley
james.n@
ischool.berkeley.edu

Tracy Huang
UC Berkeley
tracyhuang@
ischool.berkeley.edu

Michael Nielsen
UC Berkeley
michael.nielsen@
berkeley.edu

Abstract

In recent years, several key papers have introduced new techniques which continue to push the state of the art in abstractive text summarization to new limits. In this paper, we implement these algorithms, and then continue to apply our own techniques to reproduce and improve on existing results. We begin with a vanilla sequence-to-sequence model with attention; continue pointer mechanism to rare words directly from source data together with attention coverage mechanism to address the repetition issue. Finally, new mechanisms are implemented that further enhance quality and performance of summarization model, including our own mechanism to enhance beam search algorithm to drive out hypotheses that produce repetitive output, and implementing a recently developed intra-decoder attention mechanism. Our results are comparable or even better in Rouge-1 and Rouge-2 F-scores, compared to those from recently published papers in the domain.

Introduction

Summarizing long text into a coherent, informative compressed form is an active research topic in NLP and machine learning. Examples include advancements with deep reinforcement learning by Paulus, Xiong, and Socher (2017), enhancing seq2seq model for text summarization by Nallapati, Zhou, Santos, Gulcehre, and Xiang (2016), and innovations combining extractive techniques and abstractive seq2seq techniques by See, Liu, and Manning (2017). Although these papers each have their own special techniques to drive performance of text summarization, most share the following general approach:

- Start with a baseline sequence-to-sequence model, originating from the research by Sutskever, Vinyals, and Le (2014). The model laid the foundation for the successes in machine translation then extended to abstractive text summarization Rush et al. 2016 from Facebook .
- Address the three major issues from this seq2seq with attention based technique: (1) factual inaccuracy; (2) inability dealing with out-of-vocabulary (OOV) words; and, (3) word repetition.
- Introduce coverage enhancement techniques for the attention-based seq2seq model, forcing it to to learn from past decisions.
- Introduce a mechanism for copying words from the source text to minimize problems with out-of-vocabulary rare words.

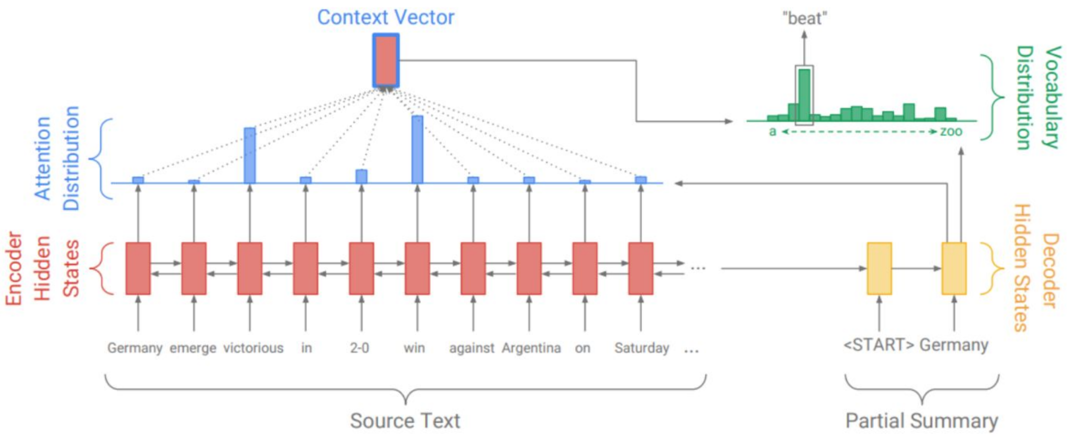
Implementation details differ among papers; in our implementation, we selectively applied algorithms from the aforementioned authors, then added our own techniques as follows:

1. Implement a sequence-to-sequence model with standard attention mechanism as a baseline, where the context vector is built from trainable weights assigned to all decoder state input time steps.
2. Next, we implemented history coverage mechanism similar to the idea of See, et al. (2017) , which combined past attention distribution on encoder's states in previous time steps with decoder inputs to feed to every decoder step. Furthermore, we enhance the beam search algorithm during decoding process to minimize the probability of hypotheses that lead to duplicate 3-grams in the output. We also use the copy mechanism to extend options for the decoder to produce output from source articles.
3. Lastly, we developed our final model with enhancement to the attention coverage mechanism by adding coverage for decoder states in past steps. This is motivated from Paulus, et al. (2017) usage of an intra decoder model. In this implementation, re-engineer the LSTM state computation at each step so that past decoder's states are included.

We found that the vanilla sequence to sequence model with attention, although popular for machine translation, performed quite poorly in text summarization, attaining a ROUGE-1 F-score of just 12%, ROUGE-2 F-score of 3% and ROUGE-L F-score of 11% on the test dataset. When we applied the source copy mechanism, performance jumped to 43.3, 20.44 and 29.7 respectively for the 2nd model.

Baseline implementation with attention based sequence to sequence model

This is the extension from standard sequence to sequence model, where the input to the decoder layer only comes from the final state of the decoder. In an attention-based seq2seq model, a context mechanism is introduced to incorporate states from past time steps. This helps the decoder have better visibility into the entire source sequence. We used this as a baseline without customization to understand how a vanilla attention-based seq2seq could perform against our test data in a text summarization scenario.



From "Get To The Point: Summarization with Pointer-Generator Networks" (See, et al., 2017)

This baseline model is implemented using standard tensorflow attention mechanism API. In particular

- We use a bidirectional LSTM cell as the encoder
- Article words are embedded as input into the encoder
- We used attention mechanism to using Bahdanau's formula to calculate attention weights which are used to calculate the context vector
- The context vector from decoder is fed together with abstract as the input to decoder in training time
- At inference time, we've explored both greedy search algorithm where the input to the next word in the decoder is generated by the only prediction from the previous word and beam search algorithm where k best probabilities from each previous word are used to search for the best algorithm (k=4)

With great success in translation, to our surprise, vanilla model perform quite poorly in text summarization. Despite 14 hours training, the output is very repetitive and poorly reflect the content of the source sentence; for an example, see Appendix A, Example 1.

Pointer model with attention coverage, enhanced beam search

Our next generation model is based on the idea of a combination between abstractive and extractive model by implementing a switch mechanism that allows the model to choose between the distribution from vocabulary or the distribution from encoder attention distribution.

These ideas are presented both by See, et al. (2017) and Paulus, et al. (2017). The switch, represented by *generation probability p*gen ∈ [0, 1] for timestep *t* is calculated from the context vector, the decoder state, and the decoder input:

$$p_{gen} = \sigma(context\_vector + decoder\_state + decoder\_input)$$

p_{gen} is used as a soft switch to choose between *generating* a word from the vocabulary by sampling from P_{vocab}, which is the distribution by softmax of the decoder output, or *copying* a word from the input sequence by sampling from the attention distribution, using the following:

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

Next, to address the repetition of generated output, we implemented the coverage mechanism which aims to inform the model about previous attention distribution decisions made in previous steps. Thus the model is aware of how positions in the decoder state were weighed, and avoid putting the same weight distribution to the input. The original decoder-encoder scoring function according to Badahnau's style used in vanilla attention model was:

$$e_i^t = v^T \tanh(W_h h_i + W_s s_t + b_{\text{attn}})$$

Where *h* is the encoder state and *s* is the decoder state. Following the direction of See, et al. (2017), we modified it to be included in the coverage vector *c* which contains all attention distribution from the previous step in the new scoring function:

$$e_i^t = v^T \tanh(W_h h_i + W_s s_t + w_c c_i^t + b_{\text{attn}})$$

where w_c is a learnable parameter of same length as *v*. This alone helps to reduce the repetition problem, yet we still occasionally found such occurrences such as:

> *Harry kirkham was run over by an officer on a routine patrol as he crossed the road outside erdington police station in birmingham . the teenager suffered a heart attack and was rushed to hospital but died the next day . the teenager suffered a heart attack and was rushed to hospital but died the next day .*

Beam search enhancement

We applied a new mechanism for the decoding process at testing time. During the beam search, we modified the standard probability calculation of hypothesis so that it's no longer simply the average of all step probabilities. Now, the log probability will be multiplied by 10000 to minimize the probability that such duplicates occur. (We do not set to zero as per Paulus, et al. (2017) suggest, because we suspect that we still need those duplicate hypothesis in extreme cases where good options are very limited). The result came out very good, rouge-1, rouge-2 and rouge-l f scores are 43.3, 20.44 and 29.7 respectively. See Appendix A, Example 2 for examples.

Intra-decoder attention mechanism

While encoder-based attention models ensure different parts of the encoded input sequence are used, our decoder can still generate repeated phrases based on its own hidden states, especially when generating long sequences. To prevent that, we can incorporate more information about the previously decoded sequence into the decoder. Looking back at previous decoding steps will allow our model to make more structured predictions and avoid repeating the same information, even if that information was generated many steps away. To achieve this, we introduce an intra-decoder attention mechanism. This mechanism is introduced in "A Deep Reinforced Model for Abstractive Summarization", by Paulus et al. (2017), and according to the author is not present in existing encoder-decoder models for abstractive summarization. We implemented this mechanism in using the same scoring function style by Bahdanau but in this case between the decode state at step *t* and previous decode states from *0→t-1*.

$$e_i^t = v^T \tanh(W_h h_i + W_s s_t + b_{\text{attn}})$$
$$a^t = \text{softmax}(e^t)$$

Once we have attention distribution a^t, we calculate the decoder state context vector $C_d = a^t \times H_d$. This decoder context vector is combined with LSTM internal cell state to create a new combined state that incorporates past time steps' states. For sample generated outputs of this last model, see Appendix A, Example 3.

Results

We evaluate our models with the standard ROUGE metric (Lin, 2004b), reporting the F_1 scores for ROUGE- 1, ROUGE-2 and ROUGE-L (which respectively measure the word-overlap, bigram-overlap, and longest common sequence between the reference summary and the summary to be evaluated).

1. Comparison of our different model configurations

Model: **History copy model with enhanced beam search...**	**ROUGE-1**			**ROUGE-2**			**ROUGE-L**		
	f-score	recall	precision	f-score	recall	precision	f-score	recall	precision
...without intra-decoder	43.34	44.65	45.02	20.44	21.05	21.29	29.71	30.71	30.80
...with intra-decoder	42.85			20.23			32.5		

2. Compared with other published papers' results in F-score

Model	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
Lead-3 (Nallapati et al., 2017) SummaRuNNer (Nallapati et al., 2017)	39.2 39.6	15.7 16.2	35.5 35.3
words-lvt2k-temp-att (Nallapati et al., 2016)	35.46	13.30	32.65
ML+RL, with intra-attention (Paulus, et al.)	39.87	15.82	36.90
Pointer to Pointer with Coverage (Abigail See et al. 2017)	39.53	17.28	36.38
Enhanced beam search without intra decoder	**43.34**	**20.44**	**29.71**
Enhanced beam search with intra decoder	**42.85**	**20.23**	**32.5**

Dataset & Training Infrastructure

We build and test our models on a modified version of the CNN/Daily Mail dataset (Hermann et al., (2015)), following the same pre-processing steps described in See et al., (2017). Our final dataset contains 146,113 training examples, 13,368 validation examples and 11,490 testing examples. We limit our vocabulary to 50,000 items and input length to 400 tokens, output length to 100 tokens. We used two cloud instances of GPU servers with M60 and K80. Each model iteration takes up to 4 days for training and evaluation.

Conclusion

We have applied different techniques from different papers including the new decoder attention mechanism, as well as our own beam search enhancement technique to achieve comparable or better results -- for at least the Rouge-1 and Rouge-2 F-scores -- using the CNN/Daily Mail dataset. We still believe that current techniques still fall short of practical applications in terms of requirements for specific domains which have stringent requirements for factual accuracy and completeness of key facts.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. arXiv:1409.0473

Sumit Chopra, Michael Auli, , and SEAS Harvard. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *HLT-NAACL*. http://www.aclweb.org/anthology/N16-1012

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In Advances in Neural Information Processing Systems.

Romain Paulus, Caiming Xiong, Richard Socher. 2017. A Deep Reinforced Model for Abstractive Summarization. arXiv:1705.04304

Sandhaus, Evan. 2008. The New York Times Annotated Corpus LDC2008T19. DVD. Philadelphia: Linguistic Data Consortium.

Abigail See, Peter J. Liu, Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. arxiv:1704.04368

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems (NIPS 2014)*. arXiv:1409.3215

Appendix A

Sample generated text summaries.

Example 1

Original text (truncated):
Foreign intervention in yemen ‘s chaos has dramatically raised the stakes in the arabian peninsula , threatening to expand what is already a civil war into a conflict pitting iran against saudi arabia and an arab coalition . the saudis launched operation `` decisive storm ‘’ last wednesday with dozens of airstrikes in an effort to blunt the advance of houthi militia and allied army units on the port of aden -- and to protect the last bastion of yemen ‘s internationally-recognized president abd-rabbu mansour hadi . there were also strikes in and around the capital , sanaa , which resumed early monday . many analysts were surprised at the speed and scale of the saudi air campaign , which the kingdom said would continue until the houthis -- a shia minority that has swept across the country in the last six months -- retreated and laid down their arms . essentially the saudis are trying to bomb the houthis to the negotiating table .

***Reference summary**: saudi arabia-led coalition launched airstrikes in yemen last wednesday . saudi intervention raises prospect of regional war between iran and arab states .*

***Model production**: new : egypt ’s foreign minister says it is a threat to syria . syria ’s foreign minister says it is a threat to syria . egypt ’s foreign minister says there is no threat.*
So obviously, the vanilla model cannot address challenges with repetitive content and truly learn the pattern of content summary which is much more abstractive than translation.

Example 2

Example 2a

***Reference Summary:** new research shows children from lower income families are more frequently consuming chips than they are salads . families with a higher income are more likely to eat salads and rice . latest figures from roy morgan also show more parents are trying to buy !!__additive-free__!! foods .*

***Model Generated Summary:** the latest figures from roy morgan show that socioeconomic status and family income have a significant impact on foods consumed . at the top end of the scale , children eat more salad and rice , but as income decreases more chips , wedges , chicken nuggets and noodles are consumed . new research shows children from lower income families are more frequently consuming chips than they are salads .*

Example 2b

Reference Summary: *woman filmed shouting at men after hearing them speak foreign language . she accused two men of only using english to ` get something ' in tirade . jay , 24 , filmed ` racist ' abuse on london underground at 11pm yesterday . onlookers accused woman of being a ` bigot ' and said she was ` disgusting ' british transport police has been notified about incident on district line .*

Model Generated Summary: *a woman was filmed hurling ` racist ' abuse at two men on the london underground and saying ` you can only speak in english while you 're on my train ' passengers said the woman took offence after hearing the men speaking in an eastern european language and accused them of only using english to ` get something ' in an expletive-laden tirade .*

Example 3

Example 3a

Original Article (truncated): *this is the moment that a controversial art project which was branded ` satanic ' by a church minister went up in flames . the intricately hand crafted wooden temple was the product of two years of hard work and planning by renowned american artist david best , who is best known for nevada 's annual burning man festival . the 70ft tower was made by catholics and protestants working together in the northern ireland city of londonderry and filled with messages written in memory of loved ones who had died . ahead of the bonfire presbyterian minister the reverend graeme orr had said he feared the burning could leave people open to the influence of satan . scroll down for video . this is the moment that thousands of people watched as a controversial art project went up in flames in londonderry , northern ireland . around 60,000 people had visited the intricately hand crafted wooden temple before it was set alight at dusk this evening . the wooden temple was the product of two years of hard work and planning by renowned american artist david best . the 70ft tower was made by catholics and protestants working together and filled with messages written in memory of loved ones who had died . but the artist creator david best said he believed it would help the bereaved end their pain . ` if this is __paganism__ then i am donald duck , ' he said . around 60,000 people visited the structure , made of carved plywood and featuring a giant wooden chandelier hanging down from a central spire , before it was torched this evening .*

Reference Summary*: messages written in memory of loved ones who have died went up in flames in londonderry , northern ireland . temple took two months to build and artist david best said he hoped it would ease the pain of the bereaved . around 60,000 people visited the 70ft hand crafted tower carved out of plywood before it was set alight . some of the messages inside the structure were about people who had died during the 30 year conflict . presbyterian minister reverend graeme orr said he feared the burning could leave people open to satan .*

Model Gen Summary: *the 70ft tower was made by catholics and protestants working together in the northern ireland city of londonderry and filled with messages written in memory of loved ones who had died . the wooden temple was the product of two years of hard work and planning by renowned american artist david best .*

Example 3b

Original Article: *a professional dancer has refused to have plastic surgery to remove her birthmark as she has chosen to embrace her individuality . cassandra __naud__ , 22 , says she loves her unique appearance and even credits her birthmark for making her memorable in her industry . the dancer , from alberta , canada , who was once told by a casting agent to digitally erase her birthmark from her __headshots__ , says that she believes it is a positive attribute . scroll down for video . cassandra refused to have her birthmark removed as she was told she would be left with severe scarring . ` my birthmark is a huge part of me , ' she says . ` it makes me unique and memorable , which is especially important for the career i 've chosen . ' when cassandra was born with a huge brown birthmark on her right cheek , her parents , richard , 60 , a power engineer , and france , 50 , a school caretaker , were given the option to have it removed , but concerned that it would leave her face heavily scarred , they decided against it . cassandra says she is happy with the decision that her parents made as the effects of surgery could have been extremely damaging . ` as my birthmark sunk through several layers of skin , plastic surgery was the only option for removal . doctors gave my parents a choice , warning them there could be scarring or i could be left with a lazy eye , ' she says . ` i 'm so glad my parents chose to leave my birthmark as it 's part of who i am . having a birthmark distinguishes me - and i do n't feel that it has ever held me back . ' cassandra __naud__ was born with a large birthmark under her right eye but rather than have it removed she has chosen to embrace her individuality . cassandra works as a professional dancer and believes that her birthmark makes her more memorable . cassandra , pictured with her mother and father , says she is thrilled that neither she or her parents went through with the removal of her birthmark . with her envious dancer body and gorgeous smile , cassandra is used to attracting attention from strangers , but she says people are often very curious when they notice her birthmark . `*

Reference Summary: *cassandra __naud__ was born with a large birthmark under her right eye . her parents decided not to have it removed as she would be left with scars . 22-year-old was bullied at school but now embraces her individuality . works as a dancer and says her birthmark helps her stand out .*

Model Gen Summary*: cassandra naud , 22 , says she loves her unique appearance and even credits her birthmark for making her memorable in her industry . the dancer , from alberta , canada , who was once told by a casting agent to digitally erase her birthmark from her headshots , says that she believes it is a positive attribute .*

Example 3c

Original Article: *charles darwin argued that animals can express emotions with their face and body , like humans do . now researchers have found that rats can recognise pain the faces of their fellow rodents . the study suggests that as well as expressing their emotional state , the pained faces of rats may have a ` __communicative__ function ' . they may even use expressions to warn other rats of danger or ask for help . researchers have found that rats can recognise pain the faces of their fellow rodents . researchers based at different institutions in tokyo , noted that rats flatten their ears , narrow their eyes and puff up their cheeks when they are in pain -lrb- shown in the images on the right -rrb- -- an observation in previous studies . researchers based at different institutions in tokyo , noted that rats flatten their ears , narrow their eyes and puff up their cheeks when they are in pain -- an observation in previous studies . the experts wanted to find out whether the rodents do this as a reflex reaction , or to communicate their discomfort to others . they found that __long-evans__ rats can read pain on other animal 's faces and use the information to makes decisions , popular science reported . the scientists first took photographs of rats who were in pain and others with neutral expressions . to test the rodents ' response to rats ' pained expressions , they showed them the photos in a specially made cage , to test whether they responded to the face , rather than the smell of another animal , for example . the rats chosen are described as ` naive ' in the study , published in the journal royal society open science , and were between eight and 12 months old . individuals were put into a cage composed of three __interconnecting__ rooms , one of which had images of a rat in pain , which had been given an electric shock . without any images on the walls and made their way to a compartment off the central room with pictures of pained rats on the walls and another offshoot room with photos of happier rats on the walls . each rat was allowed to wander freely around the cage for 10 minutes while the researchers recorded how long it spent in each room . they found that rats spent more time in the rooms where they did n't have to confront the image of a rat in pain . ' ... the rats stayed longer in the compartm*

.

***Reference Summary**: researchers in tokyo placed rats in a cage with three compartments . one room showed photos of rats in pain , and another with neutral faces . rats spent more time in the ` neutral ' room suggesting they recognise fear . experts believe facial expressions are used to communicate with others*

Model Gen Summary: *researchers based at different institutions in tokyo , noted that rats flatten their ears , narrow their eyes and puff up their cheeks when they are in pain -- an observation in previous studies . the scientists first took photographs of rats who were in pain and others with neutral expressions*