Deep Learning Models for Text Summarization

Implementation & Evaluation of Abstractive Text Summarization Models

Project Overview

- Implement existing state-of-the-art algorithms in abstractive text summarization
- Add to this foundation additional techniques to reproduce or improve results
  - Enhanced beam search algorithm
  - Decoder attention mechanism

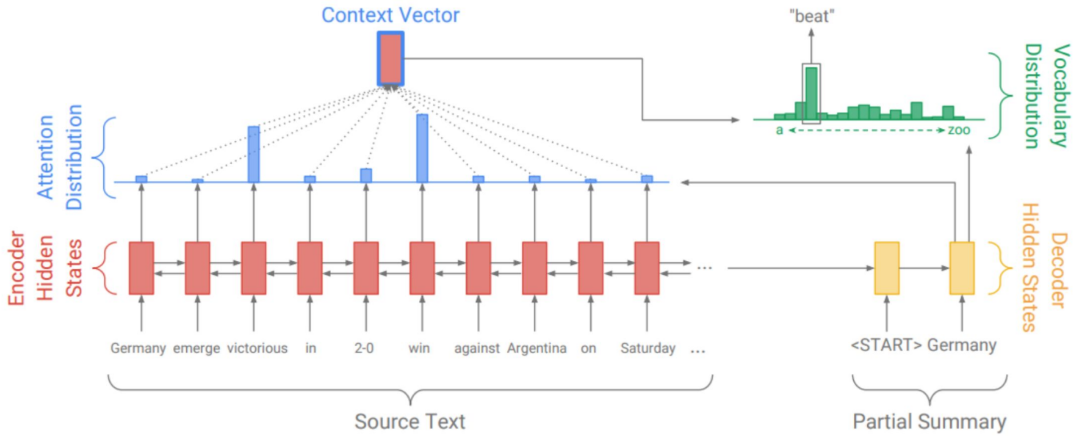
Common approach from recent researches in text summarization

- Start with a baseline sequence-to-sequence model, originating from Sutskever, Vinyals, and Le (2014) which is the foundation for the successes in machine translation then extended to abstractive text summarization Rush et al. 2016 from Facebook .
- Address 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.

Our approach to text summarization

- Implement a sequence-to-sequence model with standard attention mechanism as a baseline.
- Next, we implemented attention 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. 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.
- 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.

Baseline Implementation



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

- We use a bidirectional LSTM cell as 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 and beam search algorithm with width 6

Baseline implementation result

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 6% and ROUGE-L F-score of 8% on the test dataset

- ***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 .*
- ***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.*

Lots of repetition

Next model v1.0: Pointer, attention coverage and enhanced beam search

- Adding source copy mechanism by introducing switch mechanism to choose between vocab distribution generated by seq2seq' decoder and source article words distribution (See, et al. (2017) and Paulus, et al. (2017))
  - p_{gen} = σ(*context_vector* + *decoder_state* + *decoder_input*) ⇨ $P(w) = p_{\text{gen}} P_{\text{vocab}}(w) + (1 - p_{\text{gen}}) \sum_{i:w_i=w} a_i^t$
- Modifying the attention scoring function to include the history of attention decisions to decoder input (See, et al. (2017)). The context is also merged to decoder input

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

Much better result but repetition still occur occasionally

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 .

Next model v1.0: Pointer, attention coverage and enhanced beam search

- **Beam search enhancement**
  - We applied a new mechanism for the decoding process at infering 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)

Minimize repetition

harry kirkham *was run over by an officer on a routine patrol as he crossed the road outside erdington police station in birmingham .the officer , from sutton coldfield police station , gave the youngster first aid along with a passing paramedic who witnessed the crash .'*

Final Model v2.0: Intra-Decoder Attention

- Encoder-based attention models while ensure different parts of the encoded input sequence 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 (introduced in "A Deep Reinforced Model for Abstractive Summarization", by Paulus).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 $C_t = a^t \times H_t$. This decoder context vector is combined with LSTM internal cell state to create a new combined state that incorporates past time steps' states

Minimize repetition while ensure

harry kirkham *was run over by an officer on a routine patrol as he crossed the road outside erdington police station in birmingham .\n'b'a 15-year-old schoolboy was knocked down and killed by a police car while walking home from church with his younger brother .\n'b'the teenager suffered a heart attack and was rushed to hospital but died the next day .'*

Additional example by Model 2.-

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*

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 .*

ROUGE Scores

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	44.83	43.70	20.23	20.74	20.36	31.1	31.88	30.43

Analysis: Compared to other results

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**

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.

Reference

- 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

Reference

- Full paper is at: https://docs.google.com/document/d/143C5YOmgCUsvs0WaHcwDQIh0zW5tTCBLnOjryWG24e0/edit#
- Code is available at: https://github.com/michaeln-cal/w266-text-summarize