Just News It: Abstractive Text Summarization with a Pointer-Generator Transformer

Alexandre Bucquet (bucqueta@stanford.edu)

Vrinda Vasavada (vrindav@stanford.edu)

Motivation

- **Abstractive text summarization** task
  - Create a summary which is shorter than the source document but contains its main ideas
  - More powerful than extractive text summarization which constrains summary to only contain words from input
  - Produces more complex, human-like summaries
- Evaluate See et al.’s **pointer-generator** model on complex pieces (New York Times op-eds) to understand areas of improvement for current models
- Experiment with combining this model with a **strictly experiment with combining this model with a**
  - **Abstractive text summarization** task
  - Attractive because can be trained in parallel and proved in other tasks to **better model long-term dependencies**
  - Strictly avoid source article shortening unlike most other recent work

Data

- CNN/Daily Mail dataset: multi-line summaries (56 tokens on average) for news articles (781 tokens on average)
- Manually collected 20 **op-eds** from the New York Times

Baseline Model

- Used See et al.’s **pointer-generator** network for abstractive text summarization
- **LSTM encoder/decoder** model with **additive attention**
  - Attention computed from encoder hidden state h, decoder state s, and coverage vector c
  - Uses **generation probability** to decide whether to copy words from the source or generate a new word from vocabulary, enabling preservation of factual details
  - **Coverage loss** to disincentivize repetition of phrases
  - Uses **coverage** from (Neural Machine Translation) to represent how much each word has contributed to the output summary thus far, disincentivizing repetition of phrases

Model Architecture

- **Transformer** encoder/decoder with **additive attention**
- Uses **attention** computed from encoder hidden state h, decoder state s, and coverage vector c
- **Coverage loss** to disincentivize repetition of phrases

Quantitative Analysis

- Transformer Model Performance
  - **Failed to match pointer-generator performance**
  - Learns average word distribution of the whole dataset, but never commits to a prediction
  - All experiments struggle to learn past this local optimum
  - **Underfitting** might be a result of the long input sequences
  - All past applications of transformers for this task have **artificially shortened input articles**
  - No existing solution allows transformers to take in full input sequences

Transformed Model Performance

- Learns average word distribution of the whole dataset, but never commits to a prediction
- All experiments struggle to learn past this local optimum
- **Underfitting** might be a result of the long input sequences
- All past applications of transformers for this task have **artificially shortened input articles**
- No existing solution allows transformers to take in full input sequences

Qualitative Analysis

- **Pointer-Generator Performance on Test Set**
  - Model Output: Paul walker’s death in November 2013 at the age of 40 after a car crash, the actor was on a run at the age of 40 after a car crash in November 2013, the actor was on a run from broken film “**Furious 7**” on Friday.
  - Reference: Hardy was convicted of domestic abuse against ex-girlfriend Nicki holder and was suspended from the Dallas Cowboys for 10 days by the NFL charges were eventually dropped after holder could not be located when Hardy’s lawyers appealed the decision and asked for a jury trial, this week he got stuck in his Bentley in deep flood waters in Dallas. Hardy was forced to abandon his car and it was towed away hours later.
  - Model Output: Greg Hardy was suspended from 10 NFL games. He was fined $[UNK] a week. He was dropped by his previous team, the sheriff says.
  - **Summaries often shorter than “gold” summaries**
  - **Struggles to attribute entities and identify sources**
  - Often produces a somewhat **coherent but irrelevant summary**

- **Pointer-Generator Performance on Op-Eds**
  - **First Paragraph Snippet:** I was 15 when I started smoking, and so were most of my friends. Because cigarettes were both forbidden and easy to get: ten quarters in a cigarette vending machine, which you could still find in most pizza joints and doughnut shops in suburban New Jersey in the early 1990s...
  - Transformer: Wabtec is willing to keep the existing workers at the current average of $35 per hour...
  - **Pointer-Generator Performance on Test Set**
  - **Quantitative Analysis**
    - **Failure to capture overall context:** lower scores for both gains and losses.
    - **Use of hyperparameter exploration** to reduce the number of neighbor words that are considered in attention computations

Future Work

- **Further hyperparameter exploration**
  - Apply transformers to the **easier task of extractive summarization**, in order to understand how transformers behave on long inputs
  - Experiment with **adaptive attention windows** to reduce the number of neighbor words that are considered in attention computations

References

- A. Vaswani et al., attention is all you need. https://arxiv.org/abs/1706.03762
- A. Kumar, Pointer Summarizer (GitHub repository). https://github.com/atulkum/pointer_summarizer