Transformer Network for Abstractive Text Summarization

Aniruddha Bala 15655
Chittersu Raghu 15840
Ravi Raj Saxena 16326

Indian Institute of Science, Bangalore
Introduction

Given an input article the task is to summarize the article with novel sentences in the output.

Source Text

Munster have signed New Zealand international Francis Saili on a two-year deal. Utility back Saili, who made his All Blacks debut against Argentina in 2013, will move to the province later this year after the completion of his 2015 contractual commitments. The 24-year-old currently plays for Auckland-based Super Rugby side the Blues and was part of the New Zealand under-20 side that won the Junior World Championship in Italy in 2011. Saili's signature is something of a coup for Munster and head coach Anthony Foley believes he will be a great addition to their backline. Francis Saili has signed a two-year deal to join Munster and will link up with them later this year. ‘We are really pleased that Francis has committed his future to the province,’ Foley told Munster's official website. ‘He is a talented centre with an impressive skill-set and he possesses the physical attributes to excel in the northern hemisphere.’ I believe he will be a great addition to our backline and we look forward to welcoming him to Munster.’ Saili has been capped twice by New Zealand and was part of the under 20 side that won the Junior Championship in 2011. Saili, who joins all Black team-mates Dan Carter, Ma'a Nonu, Conrad Smith and Charles Piutau in agreeing to ply his trade in the northern hemisphere, is looking forward to a fresh challenge. He said: ‘I believe this is a fantastic opportunity for me and I am fortunate to move to a club held in such high regard, with values and traditions I can relate to from my time here in the Blues.’ This experience will stand to me as a player and I believe I can continue to improve and grow within the Munster set-up.’ As difficult as it is to leave the Blues I look forward to the exciting challenge ahead.’

Reference Summary

Utility back Francis Saili will join up with Munster later this year. The New Zealand international has signed a two-year contract. Saili made his debut for the All Blacks against Argentina in 2013.

http://www.abigailsee.com/2017/04/16/taming-rgns-for-better-summarization.html
Literature Survey

• The work by Nallapati et al provided 1st baselines for the task on CNN/DM data set. They used hierarchical RNN in their approach.

• See et al introduce a pointer generation mechanism to choose between generating a token & copying from input sequence, also keeping track of coverage to address repetition.

• **Transformers** replaces the recurrence and convolutions in neural models by self attention mechanism & are more parallelizable.

• **BERT**: learns bidirectional contextual representations.

Using Sequence-to-Sequence RNNs and Beyond (Nallapati et al., 2016)
See et al., 2017 Get to the Point: Summarization with pointer networks
Vaswani et al., 2017 Attention is all you need
Devlin et al., 2018 BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
Contribution

• Explored the effectiveness of Transformer models for abstractive summarization.

• Integrated the Pointer Generator mechanism with transformer model to handle out of vocabulary words.

• Performance improvements with transfer learning by using BERT encoder.
Datasets

- CNN Daily Mail dataset

<table>
<thead>
<tr>
<th>Task</th>
<th>Article Reference Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>287,726</td>
</tr>
<tr>
<td>Validation</td>
<td>13,368</td>
</tr>
<tr>
<td>Test</td>
<td>11,490</td>
</tr>
</tbody>
</table>

- ROUGE Metric
  ROUGE-1, ROUGE-2, ROUGE-L

Hermann et al., 2015 Teaching machines to read and comprehend
Nallapati et al. 2016 Abstractive text summarization using sequence to sequence rnns and beyond
Model-1

- **Key Features:**

- Input positional encodings
  - \( PE_{pos2i} = \sin(pos/10000^{2i/d_{model}}) \)
  - \( PE_{pos2i+1} = \cos(pos/10000^{2i/d_{model}}) \)

- Multi headed self attention
  - \( Q = W_Q X \)
  - \( K = W_K X \)
  - \( V = W_V X \)

- Layer norm and Residual connections

- Position wise feed forward network
  - \( FFN(X) = \max(0, XW_1 + b_1) W_2 + b_2 \)

- **Advantage:** No recurrence, one can leverage parallelism

---

Vaswani et al., 2017 Attention is all you need
Model-2

Generation probability

\[ P_{\text{gen}} = \text{sigmoid}(W_{\text{gen}}[X_d, X_c] + b_{\text{gen}}) \]

Vocab probability

\[ P_{\text{vocab}} = \text{softmax}(W_{\text{vocab}}[X_d] + b_{\text{vocab}}) \]

Final Extended vocabulary

\[ P_{\text{ext}} = P_{\text{gen}} \times P_{\text{vocab}} + (1 - P_{\text{gen}}) \times P_{\text{attn}} \]
Model-3

- In this we use the BERT base uncased pretrained PYTORCH model
- We replace our encoder with BERT encoder.
- We change our model hidden dimension to match BERT’s hidden dimension which is 768
- As we use BERT embedding we also use the BERT provided vocabulary which is of dimension 30,524
## Experimental Setup

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Batch size</td>
<td>8</td>
</tr>
<tr>
<td>Input sequence length</td>
<td>400</td>
</tr>
<tr>
<td>Output sequence length</td>
<td>100</td>
</tr>
<tr>
<td>Hidden dimension</td>
<td>512, 768 (for BERT)</td>
</tr>
<tr>
<td>FFN projection dimension</td>
<td>1024</td>
</tr>
<tr>
<td>Number of layers</td>
<td>6</td>
</tr>
<tr>
<td>Vocab size</td>
<td>50,000, 30,524 (for BERT)</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th>Model</th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
<th>ROUGE-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seq2seq + attn baseline (150k vocab)</td>
<td>30.49</td>
<td>11.17</td>
<td>28.08</td>
</tr>
<tr>
<td>Seq2seq + attn baseline (50k vocab)</td>
<td>31.33</td>
<td>11.89</td>
<td>28.83</td>
</tr>
<tr>
<td>Pointer Generator</td>
<td>36.44</td>
<td>15.66</td>
<td>33.42</td>
</tr>
<tr>
<td>Transformer</td>
<td>12.79</td>
<td>1.18</td>
<td>11.72</td>
</tr>
<tr>
<td>Transformer + Pointer Generator</td>
<td>28.52</td>
<td>7.02</td>
<td>26.14</td>
</tr>
<tr>
<td>BERT + Pointer Generator</td>
<td>30.98</td>
<td>12.25</td>
<td>28.93</td>
</tr>
</tbody>
</table>

See et al., 2017 Get to the Point: Summarization with pointer networks
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 gabbard 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 neighbor 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.
Analysis

• Transformer model performs poorly
  • learnt probability distributions over the vocabulary is not good enough

• Integration of Pointer Generator significantly improves performance
  • eliminates UNK’s from the output sequence with direct copy
  • increases the probability mass of the words in the vocab distribution which receive more attention in the input sequence

• With BERT encoder we are able to beat our baselines
  • better language representations learned from unsupervised pre-training helps
Conclusions

• Pointer Generator mechanism is a useful technique to augment and enhance the output probability distribution

• Transfer learning can ease the model’s workload of training end to end and help in achieving faster convergence with better performance.

• **Pros:**
  • Able to generate multi-sentence summaries
  • Faster training

• **Cons:**
  • Repetitive sentences in the output
Future Work

• Further work should be focused on integrating the coverage mechanism in Transformer models which is not trivial
• Recurrence formulation of the coverage formula restricts it’s use in Transformers
• Our proposed approach:
  • Use two layered attention
  • Use a lower triangular mask of ones and multiply with the attention probabilities obtained from the previous layer
• Future work can also be focused at investing ways to improve the Transformer model without Pointer Generator
THANK YOU