MASS: Masked Sequence to Sequence Pre-training for Language Generation

Tao Qin
Joint work with Kaitao Song, Xu Tan, Jianfeng Lu and Tie-Yan Liu
Microsoft Research Asia
Nanjing University of Science and Technology
Motivation

• BERT and GPT are very successful
  • BERT pre-trains an encoder for language understanding tasks
  • GPT pre-trains a decoder for language modeling.

• However, BERT and GPT are suboptimal on sequence to sequence based language generation tasks
  • BERT can only be used to pre-train encoder and decoder separately.
  • Encoder-to-decoder attention is very important, which BERT does not pre-train.

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without attention</td>
<td>26.71</td>
</tr>
<tr>
<td>With attention</td>
<td>36.15</td>
</tr>
</tbody>
</table>

MASS: Pre-train for Sequence to Sequence Generation

- MASS is carefully designed to jointly pre-train the encoder and decoder

- Mask k consecutive tokens (segment)
  - Force the decoder to attend on the source representations, i.e., encoder-decoder attention
  - Force the encoder to extract meaningful information from the sentence
  - Develop the decoder with the ability of language modeling
MASS vs. BERT/GPT

| $k = 1$ | $P(x^u|x_{\backslash u}; \theta)$ | masked LM in BERT |
| $k \in [1, m]$ | $P(x^{u:v}|x_{\backslash u:v}; \theta)$ | MASS |

| $k = m$ | $P(x^{1:m}|x_{\backslash 1:m}; \theta)$ | standard LM in GPT |
| $k \in [1, m]$ | $P(x^{u:v}|x_{\backslash u:v}; \theta)$ | MASS |
Unsupervised NMT

XLM: Cross-lingual language model pretraining, CoRR 2019
Low-resource NMT
Text summarization

Gigaword Corpus
Analysis of MASS: length of masked segment

(a), (b): PPL of the pre-trained model on En and Fr
(c): BLEU score of unsupervised En-Fr
(d): ROUGE of text summarization

- $K=50\%m$ is a good balance between encoder and decoder
- $K=1$ (BERT) and $K=m$ (GPT) cannot achieve good performance in language generation tasks.
Summary

• MASS jointly pre-trains the encoder-attention-decoder framework for sequence to sequence based language generation tasks
• MASS achieves significant improvements over the baselines without pre-training or with other pre-training methods on zero/low-resource NMT, text summarization and conversational response generation.
Thanks !
Backup
MASS pre-training

• Model configuration
  • Transformer, 6-6 layer, 1024 embedding.
  • Support cross-lingual tasks such as NMT, as well as monolingual tasks such as text summarization, conversational response generation.
  • English, German, French, Romanian, each language with a tag.

• Datasets
  • We use monolingual corpus from WMT News Crawl. Wikipedia data is also feasible.
  • 190M, 65M, 270M, 2.9M for English, French, German, Romanian.

• Pre-training details
  • K=50%m, 8 V100 GPUs, batch size 3000 tokens/gpu.
MASS \( (k=m) \Rightarrow \text{GPT} \)

<table>
<thead>
<tr>
<th>Length</th>
<th>Probability</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>( k = m )</td>
<td>( P(x^{1:m}</td>
<td>x^{1:m}; \theta) )</td>
</tr>
<tr>
<td>( k \in [1, m] )</td>
<td>( P(x^{u:v}</td>
<td>x^{u:v}; \theta) )</td>
</tr>
</tbody>
</table>
Analysis of MASS

- Ablation study of MASS

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU</th>
<th>Method</th>
<th>BLEU</th>
<th>Method</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discrete</td>
<td>26.76</td>
<td>Feed</td>
<td>25.56</td>
<td>MASS</td>
<td>27.41</td>
</tr>
</tbody>
</table>

- Discrete: instead of masking continuous segment, masking discrete tokens
- Feed: Feed the tokens to the decoder that appear in the encoder
Fine-tuning on conversation response generation

• We fine-tune the model on the Cornell movie dialog corpus, and simply use PPL to measure the performance of response generation.

<table>
<thead>
<tr>
<th>Method</th>
<th>Data = 10K</th>
<th>Data = 110K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>82.39</td>
<td>26.38</td>
</tr>
<tr>
<td>BERT+LM</td>
<td>80.11</td>
<td>24.84</td>
</tr>
<tr>
<td>MASS</td>
<td>74.32</td>
<td>23.52</td>
</tr>
</tbody>
</table>
Analysis of MASS: length of masked segment

(a), (b): PPL of the pre-trained model on En and Fr
(c): BLEU score of unsupervised En-Fr
(d), (e): ROUGE and PPL on text summarization and response generation

- K=50%m is a good balance between encoder and decoder
- K=1 (BERT) and K=m (GPT) cannot achieve good performance in language generation tasks.