KoreALBERT: Pretraining a Lite BERT Model for Korean Language Understanding

Paper ID: 1510

Hyunjae Lee, Jaewoong Yoon, Bonggyu Hwang, Seongho Joe, Seungjai Min, Youngjune Gwon

SAMSUNG SDS
Emerging of Pretrained Language Model (PLM)

- Many state-of-the-art results on natural language processing benchmarks have been accomplished by PLM, mainly BERT and BERT-derived models.
Motivation

Emerging of Pretrained Language Model (PLM)

- Many state-of-the-art results on natural language processing benchmarks have been accomplished by PLM, mainly BERT and BERT-derived models

Lack of Korean Language Model

- Most of research in PLM has been limited to the English language

Multi-lingual Pre-trained Model is not sufficient

- Google has released BERT multilingual model (M-BERT) pre-trained using 104 different languages including the Korean
- In most cases, Mono-lingual Model outperforms Multi-lingual Model
Background

Transformers

The Catcher in the Rye

Encoder

Self Attention

Feed Forward

Decoder

Context Vector

호밀 밭의 파수꾼 ❏</s>

Decoder

Feed Forward

Attention

Masked Self Attention

<s> 호밀 밭의 파수꾼</s>
Background

BERT (Bidirectional Encoder Representation from Transformers)
Background

ALBERT (A Lite BERT)

The Catcher in the Rye
Related Work

**Language-Specific BERT**

- FinBERT
- BETO
- CammemBERT
- ALBERTo
- BERTje
- Chinese-BERT
- Japanese-BERT

**BERT For Korean**

- **SKT KoBERT**
  - BERT Architecture, 92M Parameters
  - 320M words from Korean Wiki and News
  - SentencePiece Tokenizer, Vocab 8K

- **ETRI BERT**
  - BERT Architecture, 110M Parameters
  - 4.7B morphemes
  - Morpheme Based Tokenizer, Vocab 30K

- **No Available ALBERT practice**

※ Image Source: https://www.freepik.com/
Method

Korean + ALBERT = KoreALBERT

Corpus
- Korean Wiki, news, book review(43GB)
- General Domain, Written expressions
- SentencePiece Tokenizer, 32K of vocab size

Architecture
- Same architecture with ALBERT
- Size Comparison:
  BERT Base 110M, ALBERT Base 12M

Training Objectives
- We introduce an additional training objective, Word Order Prediction (WOP)

Masked Token Prediction
Word Order Prediction
Sentence Order Prediction
Method

KoreALBERT Pre-training architecture

Loss: 0/1 2 1 T₃ T’₁ 3’ 2’
WOP: — 2 1 — — — — — 3’ 2’ — — — —
SOP: 0/1 — — — — — — — — — — — —
MLM: — — — T₃ — — — — T’₁ — — — — — —

sum

Output: OCLS OT₂ OT₁ O_MASK … OTₚ OSEP O_MASK O’₁ O’₂ O’ₚ’ O’ₚ’’ … O’ₚ’’’ OSEP

Dense + Activation(Gelu) + Normalization

ALBERT

30% Re-ordering: CLS T₂ T₁ MASK … Tₚ SEP MASK T’₁ T’₂ T’ₚ’ T’ₚ’’ … T’ₚ’’’ SEP

15% Masking: CLS T₁ T₂ MASK … Tₚ SEP MASK T’₂ T’₃ … T’ₚ’’’ SEP
## Evaluation

### 6 Evaluation Tasks

<table>
<thead>
<tr>
<th>Task</th>
<th>Dataset</th>
<th>Input Example (Translated in English)</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentiment Analysis</td>
<td>NSMC</td>
<td>“I give it a MILLION STARS!!!”</td>
<td>1 (Positive)</td>
</tr>
<tr>
<td>Phraphrasing Detection</td>
<td>PD</td>
<td>“Katz was born in Sweden in 1947 and moved to New York City at the age of 1.”</td>
<td>1 (Positive)</td>
</tr>
</tbody>
</table>
| Sentence Similarity        | KorSTS  | “A man is eating food.”  
“A man is eating something”                                                                                                                                                                                                       | 0.84 (Spearman) |
| Natural Language Inference | KorNLI  | “You don’t have to stay there”  
“You can leave.”                                                                                                                                                                                                                   | Entailment   |
| Reading Comprehension      | KorQuAD1.0 | Pair (Paraphrase, Question)                                                                                                                                                                                                       | Start/End Index of Answer |
| Named Entity Recognition   | NER     | “I went to the Changwon Univ. to watch the Daedongjae”                                                                                                                                                                                | Changwon Univ/LOC_B, Daedongjae/EVT_B |
Evaluation

✓ Compatibility of Word Order Prediction (WOP)

• WOP hardly hurts the performance of MLM or SOP
• WOP has been proved to improve performance of downstream tasks when it is added to other two objectives

<table>
<thead>
<tr>
<th>Objectives</th>
<th>MLM acc</th>
<th>SOP acc</th>
<th>WOP acc</th>
<th>KorNLI acc</th>
<th>KorSTS spearman</th>
<th>NSMC acc</th>
<th>PD acc</th>
<th>NER acc</th>
<th>KorQuAD1.0 f1</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLM + SOP</td>
<td>35.3</td>
<td>79.8</td>
<td>-</td>
<td>76.4</td>
<td>75.6</td>
<td><strong>88.6</strong></td>
<td>92.9</td>
<td>80.7</td>
<td>89.5</td>
</tr>
<tr>
<td>MLM + SOP + WOP</td>
<td>35.1</td>
<td>79.1</td>
<td>80.7</td>
<td><strong>76.9</strong></td>
<td><strong>76.6</strong></td>
<td>88.4</td>
<td><strong>93.2</strong></td>
<td><strong>91.2</strong></td>
<td><strong>89.8</strong></td>
</tr>
<tr>
<td>MLM + WOP</td>
<td>35.6</td>
<td>-</td>
<td>84.0</td>
<td>76.8</td>
<td>73.3</td>
<td>88.5</td>
<td>92.3</td>
<td>81.0</td>
<td>89.3</td>
</tr>
</tbody>
</table>

* avg. score of 10 different seeds(50K steps)
Evaluation

✔ Final Results on 6 Downstream tasks
  • KoreALBERT consistently outperforms multi and monolingual baselines on 6 downstream NLP tasks while having much fewer parameters

<table>
<thead>
<tr>
<th>Model</th>
<th>Params (M)</th>
<th>Speedup</th>
<th>KorNLI acc</th>
<th>KorSTS spearman</th>
<th>NSMC acc</th>
<th>PD acc</th>
<th>NER acc</th>
<th>KorQuAD1.0 f1</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multilingual BERT</td>
<td>172</td>
<td>1.0</td>
<td>76.8</td>
<td>77.8</td>
<td>87.5</td>
<td>91.1</td>
<td>80.3</td>
<td>86.5</td>
<td>83.3</td>
</tr>
<tr>
<td>XLM-R Base</td>
<td>270</td>
<td>0.5x</td>
<td>80.0</td>
<td>79.4</td>
<td>90.1</td>
<td>92.6</td>
<td>83.9</td>
<td>92.3</td>
<td>86.4</td>
</tr>
<tr>
<td>KoBERT</td>
<td>92</td>
<td>1.2x</td>
<td>78.3</td>
<td>79.2</td>
<td>90.1</td>
<td>91.1</td>
<td>82.1</td>
<td>90.3</td>
<td>85.2</td>
</tr>
<tr>
<td>ETRI BERT</td>
<td>110</td>
<td>-</td>
<td>79.5</td>
<td>80.5</td>
<td>88.8</td>
<td>93.9</td>
<td>82.5</td>
<td>94.1</td>
<td>86.6</td>
</tr>
<tr>
<td>KoreALBERT Base</td>
<td>12</td>
<td>5.7x</td>
<td>79.7</td>
<td>81.2</td>
<td>89.6</td>
<td>93.8</td>
<td>82.3</td>
<td>92.6</td>
<td>86.5</td>
</tr>
<tr>
<td>KoreALBERT Large</td>
<td>18</td>
<td>1.3x</td>
<td><strong>81.1</strong></td>
<td><strong>82.1</strong></td>
<td><strong>94.1</strong></td>
<td>83.7</td>
<td><strong>94.5</strong></td>
<td></td>
<td><strong>87.5</strong></td>
</tr>
</tbody>
</table>
Evaluation

Final Results on 6 Downstream tasks

- KoreALBERT consistently outperforms multi and monolingual baselines on 6 downstream NLP tasks while having much fewer parameters

<table>
<thead>
<tr>
<th>Model</th>
<th>Params (M)</th>
<th>Speedup</th>
<th>KorNLI acc</th>
<th>KorSTS spearman</th>
<th>NSMC acc</th>
<th>PD acc</th>
<th>NER acc</th>
<th>KorQuAD1.0 f1</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multilingual BERT</td>
<td>172</td>
<td>1.0</td>
<td>76.8</td>
<td>77.8</td>
<td>87.5</td>
<td>91.1</td>
<td>80.3</td>
<td>86.5</td>
<td>83.3</td>
</tr>
<tr>
<td>XLM-R Base</td>
<td>270</td>
<td>0.5x</td>
<td>80.0</td>
<td>79.4</td>
<td>90.1</td>
<td>92.6</td>
<td>83.9</td>
<td>92.3</td>
<td>86.4</td>
</tr>
<tr>
<td>KoBERT</td>
<td>92</td>
<td>1.2x</td>
<td>78.3</td>
<td>79.2</td>
<td>90.1</td>
<td>91.1</td>
<td>82.1</td>
<td>90.3</td>
<td>85.2</td>
</tr>
<tr>
<td>ETRI BERT</td>
<td>110</td>
<td>-</td>
<td>79.5</td>
<td>80.5</td>
<td>88.8</td>
<td>93.9</td>
<td>82.5</td>
<td>94.1</td>
<td>86.6</td>
</tr>
<tr>
<td>KoreALBERT Base</td>
<td>12</td>
<td>5.7x</td>
<td>79.7</td>
<td>81.2</td>
<td>89.6</td>
<td>93.8</td>
<td>82.3</td>
<td>92.6</td>
<td>86.5</td>
</tr>
<tr>
<td>KoreALBERT Large</td>
<td>18</td>
<td>1.3x</td>
<td><strong>81.1</strong> +4.3</td>
<td><strong>82.1</strong> +4.3</td>
<td>89.7 +2.2</td>
<td><strong>94.1</strong> +3</td>
<td>83.7 +3.4</td>
<td><strong>94.5</strong> +8</td>
<td><strong>87.5</strong></td>
</tr>
</tbody>
</table>
Conclusion

We have introduced KoreALBERT, a pre-trained monolingual ALBERT model for Korean language understanding.

We have proposed a word order prediction loss, a new training objective, which is compatible with the original MLM and SOP objectives of ALBERT.

Further questions and ideas, Contact us: h8.lee@samsung.com