ERNIE 2.0: A CONTINUAL PRE-TRAINING FRAMEWORK FOR LANGUAGE UNDERSTANDING

Authors: Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, Haifeng Wang

Presenter: Royal Sequeira
Facilitators: Raheleh Makki, Gordon Gibson
03 September, 2019
Contributions

- A framework for continuous incremental multi-task pre-training
- Outperforms BERT, XLNET on 16 tasks
Motivation

- Many existing models are based on co-occurrence of tokens and sequences
- ERNIE 2.0 incorporates lexical, syntactic, and semantic information
- A new task can be introduced any time during the training process
Multitask Learning

- Use large amounts of data across tasks and to learn a better representation of language
ERNIE 2.0: Training Process
ERNEIE 2.0: Training Process

Task 1

Task 1 architecture

Update weights

Shared Encoder

[CLS] Token 1 Token 2 Token 3 [SEP] Token 1 Token 2 Token 3
ERNIE 2.0: Training Process
ERNIE 2.0: Training Process
ERENIE 2.0: Training Process
ERNIE 2.0: Training Process
Framework

ERNIE 2.0: A Continual Pre-training framework for Language Understanding

Application
- Text Similarity
- Question Answering
- Sentiment Analysis
- Natural Language Inference

Fine-tuning

Continual Pre-Training
- Task N
- Task 2
- Task 1

Pre-training Tasks Construction
- Task 1
- Task 2
- Task 3
- Task N
- Big Data
- Prior Knowledge

Multi-Task Pre-training
- Pre-training Task 1
- Pre-training Task 2
- Pre-training Task 3
- Task N
ERANIE Model

- Word-aware Pre-training Task
  - Knowledge Masking
  - Token-Document Relation
  - Capital Prediction

- Structure-aware Pre-training Task
  - Sentences Reordering
  - Sentences Distance

- Semantic-aware Pre-training Task
  - Discourse Relation
  - IR Relevance

Transformer Encoder

- [CLS]
- token1
- token2
- token3
- [SEP]
- token1
- token2
- token3
- [SEP]
- token1
- token2
- token3
- [SEP]

Token embedding
Sentence embedding
Position embedding
Task Embedding
ERNIE Loss

Sequence-Level Loss

Token-Level Loss

Encoder

V₀

V₁

V₂

V₃

V₄

Sequence loss 1
Sequence loss 2
Sequence loss 3

Token loss 1
Token loss 2
Token loss 3

Token loss 1
Token loss 2
Token loss 3

CLS

Token1

Token2

Token3

Token4
Pre-training Tasks

- Word-aware pre-training tasks
- Structure-aware pre-training tasks
- Semantic-aware pre-training tasks
Word-aware Pre-training Tasks

- Knowledge Masking Task: phrase and entity masking
  - James was [MASK] by Jeremy
  - [MASK] [MASK] was written by George R. R. Martin
- Capitalization Prediction Task: capitalized or not?
  - james was kidnapped by jeremy
- Token-Document Relation Prediction Task: token appears in other segments?
  - A meme is an idea, behavior, or style that spreads from person to person within a culture
Structure-aware Pre-training Tasks

- Sentence Reordering Task: re-organize permuted sentences
- Sentence Distance Task:
  - 0: Two sentences are adjacent in the same document
  - 1: Two sentences are in the same document
  - 2: Two sentences are from two different documents
Semantic-aware Pre-training Tasks

- Discourse Relation Task
  - I took my umbrella this morning. [because] The forecast was rain in the afternoon

- IR Relevance Task
  - 0: Strong relevance
  - 1: Weak relevance
  - 2: Irrelevance
Experiments
Pre-training Data

- **English:**
  - Wikipedia
  - BookCorpus
  - Reddit
  - Discovery data (discourse relation data)

- **Chinese**
  - Data from Baidu Search Engine (news, IR, encyclopedia etc.)
Pre-training Settings

- **Base model**
  - 12 layers
  - 12 self-attention heads
  - 768-dimensional of hidden size

- **Large model**
  - 24 layers
  - 16 self-attention heads
  - 1024-dimensional of hidden size
Fine-tuning Tasks (English)

- GLUE (General Language Understanding Evaluation)
  - CoLA: syntax specification
  - SST-2: sentiment analysis
  - MNLI: multi-genre textual inference
  - RTE: recognizing textual entailment
  - WNLI: co-referencing information between two paragraphs
  - QQP: duplication of question pairs
  - MRPC: paraphrasing
  - STS-B: semantic text similarity
  - QNLI: natural language inference on question-answer pairs
  - AX: linguistic analysis of models
Pre-training Tasks (Chinese)

- Machine Reading Comprehension (MRC)
  - Chinese Machine Reading Comprehension 2018 (CMRC 2018)
  - Delta Reading Comprehension Dataset (DRCD)
  - DuReader
- Named Entity Recognition (NER)
- Natural Language Inference (NLI)
- Sentiment Analysis (SA)
- Semantic Similarity (SS)
- Question Answering (QA)
Results
# English Tasks

<table>
<thead>
<tr>
<th>Task (Metrics)</th>
<th>BASE model</th>
<th></th>
<th>LARGE model</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Test</td>
<td>Dev</td>
<td>Test</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BERT</td>
<td>ERNIE 2.0</td>
<td>BERT</td>
<td>ERNIE 2.0</td>
</tr>
<tr>
<td>CoLA (Matthew Corr.)</td>
<td>52.1</td>
<td>55.2</td>
<td>60.6</td>
<td>63.6</td>
</tr>
<tr>
<td>SST-2 (Accuracy)</td>
<td>93.5</td>
<td>95.0</td>
<td>93.2</td>
<td>95.6</td>
</tr>
<tr>
<td>MRPC (Accuracy/F1)</td>
<td>84.8/88.9</td>
<td>86.1/89.9</td>
<td>88.0/-</td>
<td>89.2/-</td>
</tr>
<tr>
<td>STS-B (Pearson Corr./Spearman Corr.)</td>
<td>87.1/85.8</td>
<td>87.6/86.5</td>
<td>90.0/-</td>
<td>91.8/-</td>
</tr>
<tr>
<td>QQP (Accuracy/F1)</td>
<td>89.2/71.2</td>
<td>89.8/73.2</td>
<td>91.3/-</td>
<td>91.8/-</td>
</tr>
<tr>
<td>MNLI-m/mm (Accuracy)</td>
<td>84.6/83.4</td>
<td>86.1/85.5</td>
<td>86.6/-</td>
<td>89.8/-</td>
</tr>
<tr>
<td>QNLI (Accuracy)</td>
<td>90.5</td>
<td>92.9</td>
<td>92.3</td>
<td>93.9</td>
</tr>
<tr>
<td>RTE (Accuracy)</td>
<td>66.4</td>
<td>74.8</td>
<td>70.4</td>
<td>83.8</td>
</tr>
<tr>
<td>WNLI (Accuracy)</td>
<td>65.1</td>
<td>65.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AX (Matthew Corr.)</td>
<td>34.2</td>
<td>37.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Score</td>
<td>78.3</td>
<td>80.6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Score: 80.5/83.6
## English Tasks

<table>
<thead>
<tr>
<th>Task (Metrics)</th>
<th>BASE model</th>
<th></th>
<th>LARGE model</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Test</td>
<td>Dev</td>
<td>Test</td>
<td></td>
</tr>
<tr>
<td>CoLA (Matthew Corr.)</td>
<td>52.1</td>
<td>60.6</td>
<td>65.4</td>
<td>60.5</td>
</tr>
<tr>
<td>SST-2 (Accuracy)</td>
<td>93.5</td>
<td>93.2</td>
<td>96.0</td>
<td>94.9</td>
</tr>
<tr>
<td>MRPC (Accurary/F1)</td>
<td>84.8/88.9</td>
<td>88.0/-</td>
<td>89.7/-</td>
<td>85.4/89.3</td>
</tr>
<tr>
<td>STS-B (Pearson Corr./Spearman Corr.)</td>
<td>87.1/85.8</td>
<td>90.0/-</td>
<td>92.3/-</td>
<td>87.6/86.5</td>
</tr>
<tr>
<td>QQP (Accuracy/F1)</td>
<td>89.2/71.2</td>
<td>91.3/-</td>
<td>92.5/-</td>
<td>89.3/72.1</td>
</tr>
<tr>
<td>MNLI-m/mm (Accuracy)</td>
<td>84.6/83.4</td>
<td>86.6/-</td>
<td>89.1/-</td>
<td>86.7/85.9</td>
</tr>
<tr>
<td>QNLI (Accuracy)</td>
<td>90.5</td>
<td>92.3</td>
<td>94.3</td>
<td>92.7</td>
</tr>
<tr>
<td>RTE (Accuracy)</td>
<td>66.4</td>
<td>70.4</td>
<td>85.2</td>
<td>70.1</td>
</tr>
<tr>
<td>WNLI (Accuracy)</td>
<td>65.1</td>
<td>-</td>
<td>-</td>
<td>65.1</td>
</tr>
<tr>
<td>AX (Matthew Corr.)</td>
<td>34.2</td>
<td>-</td>
<td>-</td>
<td>39.6</td>
</tr>
<tr>
<td>Score</td>
<td>78.3</td>
<td>-</td>
<td>-</td>
<td>80.5</td>
</tr>
</tbody>
</table>

Score: 80.5
## Chinese Tasks

<table>
<thead>
<tr>
<th>Task</th>
<th>Metrics</th>
<th>BERT&lt;sub&gt;BASE&lt;/sub&gt;</th>
<th>ERNIE 1.0&lt;sub&gt;BASE&lt;/sub&gt;</th>
<th>ERNIE 2.0&lt;sub&gt;BASE&lt;/sub&gt;</th>
<th>ERNIE 2.0&lt;sub&gt;LARGE&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Dev</td>
<td>Test</td>
<td>Dev</td>
<td>Test</td>
</tr>
<tr>
<td>CMRC 2018</td>
<td>EM/F1</td>
<td>66.3/85.9</td>
<td>-</td>
<td>65.1/85.1</td>
<td>-</td>
</tr>
<tr>
<td>DRCD</td>
<td>EM/F1</td>
<td>85.7/91.6</td>
<td>84.9/90.9</td>
<td>84.6/90.9</td>
<td>84.0/90.5</td>
</tr>
<tr>
<td>DuReader</td>
<td>EM/F1</td>
<td>59.5/73.1</td>
<td>-</td>
<td>57.9/72.1</td>
<td>-</td>
</tr>
<tr>
<td>MSRA-NER</td>
<td>F1</td>
<td>94.0</td>
<td>92.6</td>
<td>95.0</td>
<td>93.8</td>
</tr>
<tr>
<td>XNLI</td>
<td>Accuracy</td>
<td>78.1</td>
<td>77.2</td>
<td>79.9</td>
<td>78.4</td>
</tr>
<tr>
<td>ChnSentiCorp</td>
<td>Accuracy</td>
<td>94.6</td>
<td>94.3</td>
<td>95.2</td>
<td>95.4</td>
</tr>
<tr>
<td>LCQMC</td>
<td>Accuracy</td>
<td>88.8</td>
<td>87.0</td>
<td>89.7</td>
<td>87.4</td>
</tr>
<tr>
<td>BQ Corpus</td>
<td>Accuracy</td>
<td>85.9</td>
<td>84.8</td>
<td>86.1</td>
<td>84.8</td>
</tr>
<tr>
<td>NLPCC-DBQA</td>
<td>MRR/F1</td>
<td>94.7/80.7</td>
<td>94.6/80.8</td>
<td>95.0/82.3</td>
<td>95.1/82.7</td>
</tr>
</tbody>
</table>
## Chinese Tasks

<table>
<thead>
<tr>
<th>Task</th>
<th>Metrics</th>
<th>BERT\textsubscript{BASE}</th>
<th>ERNIE 1.0\textsubscript{BASE}</th>
<th>ERNIE 2.0\textsubscript{BASE}</th>
<th>ERNIE 2.0\textsubscript{LARGE}</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Dev</td>
<td>Test</td>
<td>Dev</td>
<td>Test</td>
</tr>
<tr>
<td>CMRC 2018</td>
<td>EM/F1</td>
<td>66.3/85.9</td>
<td>-</td>
<td>65.1/85.1</td>
<td>-</td>
</tr>
<tr>
<td>DRCD</td>
<td>EM/F1</td>
<td>85.7/91.6</td>
<td>84.9/90.9</td>
<td>84.6/90.9</td>
<td>84.0/90.5</td>
</tr>
<tr>
<td>DuReader</td>
<td>EM/F1</td>
<td>59.5/73.1</td>
<td>-</td>
<td>57.9/72.1</td>
<td>-</td>
</tr>
<tr>
<td>MSRA-NER</td>
<td>F1</td>
<td>94.0</td>
<td>92.6</td>
<td>95.0</td>
<td>93.8</td>
</tr>
<tr>
<td>XNLI</td>
<td>Accuracy</td>
<td>78.1</td>
<td>77.2</td>
<td>79.9</td>
<td>78.4</td>
</tr>
<tr>
<td>ChnSentiCorp</td>
<td>Accuracy</td>
<td>94.6</td>
<td>94.3</td>
<td>95.2</td>
<td>95.4</td>
</tr>
<tr>
<td>LCQMC</td>
<td>Accuracy</td>
<td>88.8</td>
<td>87.0</td>
<td>89.7</td>
<td>87.4</td>
</tr>
<tr>
<td>BQ Corpus</td>
<td>Accuracy</td>
<td>85.9</td>
<td>84.8</td>
<td>86.1</td>
<td>84.8</td>
</tr>
<tr>
<td>NLPCC-DBQA</td>
<td>MRR/F1</td>
<td>94.7/80.7</td>
<td>94.6/80.8</td>
<td>95.0/82.3</td>
<td>95.1/82.7</td>
</tr>
</tbody>
</table>
Key takeaways

- ERNIE 2.0: Multitask learning done sequentially
- Outperforms BERT, XLNET on 16 tasks
Discussion Points

- Is this a scalable approach?
- How much does the order of pre-training tasks affect results in the downstream tasks?
- What ablation studies would you like to see performed?
- How much improvement in the downstream tasks can be attributed to the novelty in the architecture vs size of training data?
- What are some other potential pre-training tasks that can be added?
References

- https://ademcan.net/blog/2013/04/10/how-to-convert-pdf-to-png-from-the-command-line-on-a-mac/
Loss Calculation

- Loss(instance) = Loss(sentence loss) + \text{avg}(\text{token losses task}_1) + \text{avg}(\text{token losses task}_2)+\ldots+\text{avg}(\text{token losses task}_n)
- Loss\_batch = \text{avg}(\text{Loss(instance)})

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Task</th>
<th>Token-Level Loss</th>
<th>Sentence-Level Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Knowledge Masking</td>
<td>Capital Prediction</td>
</tr>
<tr>
<td>Encyclopedia</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>BookCorpus</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>News</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Dialog</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>IR Relevance Data</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Discourse Relation Data</td>
<td>×</td>
<td>×</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 2: The Relationship between pre-training task and pre-training dataset. We use different pre-training dataset to construct different tasks. A type of pre-trained dataset can correspond to multiple pre-training tasks.