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.
ERNEIE 2.0: Training Process
ERNIE 2.0: Training Process
ERNIE 2.0: Training Process
ERNIE 2.0: Training Process
ERNEIE 2.0: Training Process
ERNIE 2.0: Training Process

Diagram showing the training process of ERNIE 2.0, including tasks, architectures, data, and a shared encoder. Each task is connected to a GPU, and there is an average loss calculated from the task losses. The diagram also illustrates the flow of data and weights update.
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
- Sequentially:
  - 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
- ...
- Pre-training Task N
# ERNIE Model

| Word-aware Pre-training Task | Structure-aware Pre-training Task | Semantic-aware Pre-training Task |
|-----------------------------|----------------------------------|----------------------------------|
| Knowledge Masking           | Sentences Reordering             | Discourse Relation               |
| Token-Document Relation     | Sentences Distance               | IR Relevance                      |
| Capital Prediction          |                                  |                                  |

## Transformer Encoder

|   | [CLS] | token1 | token2 | token3 | [SEP] | token1 | token2 | token3 | [SEP] | token1 | token2 | token3 | [SEP] |
|---|-------|--------|--------|--------|-------|--------|--------|--------|-------|--------|--------|--------|-------|
|   | +     | +      | +      | +      | 0     | +      | +      | +      | +     | +      | +      | +      | +     |
|   | A     | A      | A      | A      | A     | A      | B      | B      | B     | C      | C      | C      | C     |
|   | +     | +      | +      | +      | +     | +      | +      | +      | +     | +      | +      | +      | +     |
|   | 0     | 1      | 2      | 3      | 4     | 5      | 6      | 7      | 8     | 9      | 10     | 11     | 12    |
|   | +     | +      | +      | +      | +     | +      | +      | +      | +     | +      | +      | +      | +     |
|   | 3     | 3      | 3      | 3      | 3     | 3      | 3      | 3      | 3     | 3      | 3      | 3      | 3     |

- **Token embedding**
- **Sentence embedding**
- **Position embedding**
- **Task Embedding**
ERNIE Loss

Sequence-Level Loss

Token-Level Loss

Encoder
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

| Task(Metrics)                  | BASE model                     | LARGE model                     |       |
|-------------------------------|--------------------------------|---------------------------------|-------|
|                               | Test                           | Dev                             | Test  |
|                               | BERT  | ERNIE 2.0 | BERT  | XLNet | ERNIE 2.0 | BERT  | ERNIE 2.0 |
| CoLA (Matthew Corr.)          | 52.1  | 55.2      | 60.6  | 63.6  | 65.4      | 60.5  | 63.5       |
| SST-2 (Accuracy)              | 93.5  | 95.0      | 93.2  | 95.6  | 96.0      | 94.9  | 95.6       |
| MRPC (Accuracy/F1)            | 84.8/88.9 | 86.1/89.9 | 88.0/-| 89.2/-| 89.7/-    | 85.4/89.3 | 87.4/90.2 |
| STS-B (Pearson Corr./Spearman Corr.) | 87.1/85.8 | 87.6/86.5 | 90.0/-| 91.8/-| 92.3/-    | 87.6/86.5 | 91.2/90.6 |
| QQP (Accuracy/F1)             | 89.2/71.2 | 89.8/73.2 | 91.3/-| 91.8/-| 92.5/-    | 89.3/72.1 | 90.1/73.8 |
| MNLI-m/mm (Accuracy)          | 84.6/83.4 | 86.1/85.5 | 86.6/-| 89.8/-| 89.1/-    | 86.7/85.9 | 88.7/88.8 |
| QNLI (Accuracy)               | 90.5  | 92.9      | 92.3  | 93.9  | 94.3      | 92.7  | 94.6       |
| RTE (Accuracy)                | 66.4  | 74.8      | 70.4  | 83.8  | 85.2      | 70.1  | 80.2       |
| WNLI (Accuracy)               | **65.1** | 65.1      | -     | -     | -         | 65.1  | 67.8       |
| AX(Matthew Corr.)             | 34.2  | 37.4      | -     | -     | -         | 39.6  | 48.0       |
| Score                         | 78.3  | 80.6      | -     | -     | -         | 80.5  | 83.6       |
# English Tasks

| Task (Metrics) | BASE model |  | LARGE model |  |
|---------------|------------|------------|-------------|------------|
|               | Test       | Dev        | Test        |  |
|               | BERT       | ERNIE 2.0  | BERT        | XLNet      | ERNIE 2.0  | BERT       | ERNIE 2.0  |
| CoLA (Matthew Corr.) | 52.1/55.2  | 60.6/63.6  | 65.4/69.0   | 60.5/63.5  |
| SST-2 (Accuracy)    | 93.5/95.0  | 93.2/95.6  | 96.0/98.0   | 94.9/95.6  |
| MRPC (Accuracy/F1)  | 84.8/88.9  | 88.0/89.2  | 89.7/91.8   | 85.4/89.3  | 87.4/90.2  |
| STS-B (Pearson Corr./Spearman Corr.) | 87.1/85.8  | 90.0/91.8  | 92.3/94.3   | 87.6/86.5  | 91.2/90.6  |
| QQP (Accuracy/F1)   | 89.2/71.2  | 91.3/91.8  | 92.5/94.3   | 89.3/72.1  | 90.1/73.8  |
| MNLI-m/mm (Accuracy)| 84.6/83.4  | 86.6/89.8  | 89.1/91.6   | 86.7/85.9  | 88.7/88.8  |
| QNLI (Accuracy)     | 90.5/92.9  | 92.3/93.9  | 94.3/96.3   | 92.7/94.6  | 94.6/96.2  |
| RTE (Accuracy)      | 66.4/74.8  | 70.4/83.8  | 85.2/88.2   | 70.1/80.2  | 80.2/82.8  |
| WNLI (Accuracy)     | 65.1/65.1  | -          | -           | 65.1/67.8  |
| AX (Matthew Corr.)  | 34.2/37.4  | -          | -           | 39.6/48.0  |
| Score               | 78.3/80.6  | -          | -           | 80.5/83.6  |
# Chinese Tasks

| Task          | Metrics     | BERT\textsubscript{BASE} | ERNIE 1.0\textsubscript{BASE} | ERNIE 2.0\textsubscript{BASE} | ERNIE 2.0\textsubscript{LARGE} |
|---------------|-------------|---------------------------|-------------------------------|-------------------------------|---------------------------------|
|               |             | Dev | Test | Dev | Test | Dev | Test | Dev | Test | Dev | Test | Dev | Test |
| CMRC 2018     | EM/F1       | 66.3/85.9 | -    | 65.1/85.1 | -   | 69.1/88.6 | -   | 71.5/89.9 | -   | 89.7/94.7 | 89.0/94.2 |
| DRCD          | EM/F1       | 85.7/91.6 | 84.9/90.9 | 84.6/90.9 | 84.0/90.5 | 88.5/93.8 | 88.0/93.4 | 61.3/74.9 | -   | 64.2/77.3 | -   |
| DuReader      | EM/F1       | 59.5/73.1 | -    | 57.9/72.1 | -   | 95.2 | 93.8 | 95.2 | 93.8 | 96.3 | 95.0 | 82.6 | 81.0 |
| MSRA-NER      | F1          | 94.0 | 92.6 | 95.0 | 93.8 | 95.7 | 95.5 | 96.1 | 95.8 | 96.1 | 95.8 |         |      |
| XNLI          | Accuracy    | 78.1 | 77.2 | 79.9 | 78.4 | 81.2 | 79.7 | 82.6 | 81.0 |         |      |         |      |
| ChnSentiCorp  | Accuracy    | 94.6 | 94.3 | 95.2 | 95.4 | 95.7 | 95.5 | 96.1 | 95.8 |         |      |         |      |
| LCQMC         | Accuracy    | 88.8 | 87.0 | 89.7 | 87.4 | 90.9 | 87.9 | 90.9 | 87.9 |         |      |         |      |
| BQ Corpus     | Accuracy    | 85.9 | 84.8 | 86.1 | 84.8 | 86.4 | 85.0 | 86.5 | 85.2 |         |      |         |      |
| NLPCC-DBQA    | MRR/F1      | 94.7/80.7 | 94.6/80.8 | 95.0/82.3 | 95.1/82.7 | 95.7/84.7 | 95.7/85.3 | 95.9/85.3 | 95.8/85.8 |         |      |         |      |
| Task          | Metrics | **BERT_{BASE}** | **ERNIE 1.0_{BASE}** | **ERNIE 2.0_{BASE}** | **ERNIE 2.0_{LARGE}** |
|---------------|---------|----------------|----------------------|----------------------|-----------------------|
|               |         | Dev/ Test      | Dev/ Test            | Dev/ Test            | Dev/ Test             |
| CMRC 2018     | EM/F1   | 66.3/85.9      | 65.1/85.1            | 69.1/88.6            | 71.5/89.9             |
| DRCD          | EM/F1   | 85.7/91.6      | 84.6/90.9            | 88.5/93.8            | 89.7/94.7             |
| DuReader      | EM/F1   | 59.5/73.1      | 57.9/72.1            | 61.3/74.9            | 64.2/77.3             |
| MSRA-NER      | F1      | 94.0/92.6      | 95.0/93.8            | 95.2/93.8            | 96.3/95.0             |
| XNLI          | Accuracy| 78.1/77.2      | 79.9/78.4            | 81.2/79.7            | 82.6/81.0             |
| ChnSentiCorp  | Accuracy| 94.6/94.3      | 95.2/95.4            | 95.7/95.5            | 96.1/95.8             |
| LCQMC         | Accuracy| 88.8/87.0      | 89.7/87.4            | 90.9/87.9            | 90.9/87.9             |
| BQ Corpus     | Accuracy| 85.9/84.8      | 86.1/84.8            | 86.4/85.0            | 86.5/85.2             |
| NLPCC-DBQA    | MRR/F1  | 94.7/80.7      | 95.0/82.3            | 95.7/84.7            | 95.9/85.3             |
|               |         | 94.6/80.8      | 95.1/82.7            | 95.7/85.3            | 95.8/85.8             |
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://arxiv.org/pdf/1907.12412.pdf
- 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 + \( \text{avg}(\text{token losses task}_n) \)
- \( \text{Loss\_batch} = \text{avg}(\text{Loss(instance)}) \)

| Task                  | Token-Level Loss | Sentence-Level Loss |
|-----------------------|------------------|---------------------|
|                       | Knowledge Masking| Sentence Reordering |
| Encyclopedia          | ✓                | ✓                   |
| BookCorpus            | ✓                | ✓                   |
| News                  | ✓                | ✓                   |
| Dialog                | ✓                | ✓                   |
| IR Relevance Data     | ✗                | ✓                   |
| Discourse Relation Data | ✗            | ✓                   |

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.