BERT
Hung-yi Lee
**1-of-N Encoding**

apple = [1 0 0 0 0]

bag = [0 1 0 0 0]

cat = [0 0 1 0 0]

dog = [0 0 0 1 0]

elephant = [0 0 0 0 1]

**Word Embedding**

dog, rabbit

cat, jump

tree, run, flower

**Word Class**

class 1

dog, cat, bird

Class 2

ranned, walked

Class 3

flower, tree, apple
A word can have multiple senses.

Have you paid that money to the bank yet?
It is safest to deposit your money in the bank.

The victim was found lying dead on the river bank.
They stood on the river bank to fish.

The hospital has its own blood bank.

The third sense or not?
More Examples

他是尼祿

這是加賀號護衛艦

她也是尼祿

這也是加賀號護衛艦
Contextualized Word Embedding

...money in the bank...

...the river bank...

...own blood bank...
Embeddings from Language Model (ELMO) [https://arxiv.org/abs/1802.05365]

- RNN-based language models (trained from lots of sentences)

  e.g. given “潮水 退了 就 知道 誰 沒穿 褲子”
ELMO

Each layer in deep LSTM can generate a latent representation.

Which one should we use???
我全都要
ELMO

Learned with the downstream tasks

\[
\alpha_1 + \alpha_2
\]
Bidirectional Encoder Representations from Transformers (BERT)

- BERT = Encoder of Transformer

Learned from a large amount of text without annotation
Training of BERT

- Approach 1: Masked LM
**Training of BERT**

**Approach 2: Next Sentence Prediction**

[CLS]: the position that outputs classification results

[SEP]: the boundary of two sentences

Approaches 1 and 2 are used at the same time.

Linear Binary Classifier

yes

BERT
Training of BERT

Approach 2: Next Sentence Prediction

[CLS]: the position that outputs classification results

[SEP]: the boundary of two sentences

Approaches 1 and 2 are used at the same time.
How to use BERT – Case 1

Input: single sentence, output: class
Example: Sentiment analysis (our HW), Document Classification

Linear Classifier

Trained from Scratch

BERT

Fine-tune

[CLS]

w_1 w_2 w_3

sentence

class
How to use BERT – Case 2

Input: single sentence, output: class of each word

Example: Slot filling

arrive Taipei on November 2nd
other dest other time time

(sentence)

[CLS] W₁ W₂ W₃

Linear Cls Linear Cls Linear Cls
How to use BERT – Case 3

Input: two sentences, output: class

Example: Natural Language Inference
Given a “premise”, determining whether a “hypothesis” is T/F/ unknown.
How to use BERT – Case 4

• Extraction-based Question Answering (QA) (E.g. SQuAD)

**Document:** \( D = \{d_1, d_2, \ldots, d_N\} \)

**Query:** \( Q = \{q_1, q_2, \ldots, q_N\} \)

\[ Q \xrightarrow{\text{QA Model}} s, e \]

output: two integers \((s, e)\)

**Answer:** \( A = \{q_s, \ldots, q_e\} \)

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rainfall that are scattered are called “showers”.

What causes precipitation to fall? gravity

\[ s = 17, e = 17 \]

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail? graupel

Where do water droplets collide with ice crystals to form precipitation? within a cloud

\[ s = 77, e = 79 \]
How to use BERT – Case 4

The answer is “d₂ d₃”.

s = 2, e = 3
How to use BERT – Case 4

s = 2, e = 3

The answer is “d₂ d₃”.

Learned from scratch

Softmax

dot product

s = 2, e = 3

The answer is “d₂ d₃”.

Learned from scratch

Softmax

dot product

s = 2, e = 3

The answer is “d₂ d₃”.

Learned from scratch

Softmax

dot product

s = 2, e = 3

The answer is “d₂ d₃”.

Learned from scratch

Softmax

dot product
<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BERT + DAE + AoA (ensemble)</td>
<td>87.147</td>
<td>89.474</td>
</tr>
<tr>
<td></td>
<td>Joint Laboratory of HIT and iFLYTEK Research</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>BERT + ConvLSTM + MTL + Verifier (ensemble)</td>
<td>86.730</td>
<td>89.286</td>
</tr>
<tr>
<td></td>
<td>Layer 6 AI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>BERT + N-Gram Masking + Synthetic Self-Training (ensemble)</td>
<td>86.673</td>
<td>89.147</td>
</tr>
<tr>
<td></td>
<td>Google AI Language</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><a href="https://github.com/google-research/bert">https://github.com/google-research/bert</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>XLNet (single model)</td>
<td>86.346</td>
<td>89.133</td>
</tr>
<tr>
<td></td>
<td>XLNet Team</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>SemBERT (ensemble)</td>
<td>86.166</td>
<td>88.886</td>
</tr>
<tr>
<td></td>
<td>Shanghai Jiao Tong University</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Enhanced Representation through Knowledge Integration (ERNIE)

• Designed for Chinese

Source of image:
https://zhuanlan.zhihu.com/p/59436589
What does BERT learn?

https://arxiv.org/abs/1905.05950
https://openreview.net/pdf?id=SJzSgnRcKX
Multilingual BERT

Trained on 104 languages

Task specific training data for English

- Class 1
- Class 2
- Class 3

Task specific testing data for Chinese

- ?
- ?
- ?
Generative Pre-Training (GPT)

Transformer Decoder

BERT (340M)

ELMO (94M)

GPT-2 (1542M)

Source of image: https://huaban.com/pins/1714071707/
Generative Pre-Training (GPT)

Many Layers...

退了

<i>BOS</i>
Generative Pre-Training (GPT)

Many Layers ...

\[ \alpha_{3,1}, \alpha_{3,2}, \alpha_{3,3} \]

\[ q^1, k^1, v^1, q^2, k^2, v^2, q^3, k^3, v^3 \]

\[ a^1, a^2, a^3, a^4 \]

\[ \langle \text{BOS} \rangle, \text{潮水}, \text{退了}, \text{就} \]
Zero-shot Learning?

- **Reading Comprehension**

  \[ d_1, d_2, \ldots, d_N, \]
  
  "Q:”, \( q_1, q_2, \ldots, q_N, \)
  
  “A:”

- **Summarization**

  \[ d_1, d_2, \ldots, d_N,” TL;DR:” \]

- **Translation**

  English sentence 1 = French sentence 1
  
  English sentence 2 = French sentence 2
  
  English sentence 3 =
Visualization

(The results below are from GPT-2)
In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

The scientist named the population, after their distinctive horn, Ovid’s Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.
Can BERT speak?

• Unified Language Model Pre-training for Natural Language Understanding and Generation
  • https://arxiv.org/abs/1905.03197
• BERT has a Mouth, and It Must Speak: BERT as a Markov Random Field Language Model
  • https://arxiv.org/abs/1902.04094
• Insertion Transformer: Flexible Sequence Generation via Insertion Operations
  • https://arxiv.org/abs/1902.03249
• Insertion-based Decoding with automatically Inferred Generation Order
  • https://arxiv.org/abs/1902.01370