Self-Supervised Learning

Hung-yi Lee 李宏毅

https://www.sesameworkshop.org/what-we-do/sesame-streets-50th-anniversary
死臭酸宅本人
芝麻街
ELMo (Embeddings from Language Models)

ERNIE (Enhanced Representation through Knowledge Integration)

Big Bird: Transformers for Longer Sequences

BERT (Bidirectional Encoder Representations from Transformers)
BERT
340M parameters

Source of image:
https://leemeng.tw/attack_on_bert_transfer_learning_in_nlp.html
The models become larger and larger ...

ELMO (94M)
BERT (340M)
GPT-2 (1542M)

Source of image: https://huaban.com/pins/1714071707/
The models become larger and larger ...

GPT-3 is 10 times larger than Turing NLG.
BERT (340M)
GPT-3 (175B)
Switch
Transformer (1.6T)

https://arxiv.org/abs/2101.03961
Outline

BERT series

GPT series
Self-supervised Learning

Supervised

\[ y \xrightarrow{\text{label}} \hat{y} \]

\[ x \xleftarrow{} y \]

\[ \text{Model} \]

Self-supervised

\[ y \xrightarrow{} x'' \]

\[ x \xleftarrow{} x' \]

\[ \text{Model} \]

---

Yann LeCun
2019年4月30日

I now call it "self-supervised learning", because "unsupervised" is both a loaded and confusing term.

In self-supervised learning, the system learns to predict part of its input from other parts of it input. In other words a portion of the input is used as a supervisory signal to a predictor fed with the remaining portion of the input.
Masking Input

https://arxiv.org/abs/1810.04805

Randomly masking some tokens

 Transformer Encoder

 BERT

Softmax

Linear

| 學 | 0.1 |
| 灣 | 0.7 |
| 台 | 0.1 |
| 大 | 0.1 |

(all characters)
Masking Input

https://arxiv.org/abs/1810.04805

Randomly masking some tokens

Ground truth

minimize cross entropy

softmax

Transformer Encoder

BERT

Random

= Random (special token)

= MASK

一、天、大、小 ...
Next Sentence Prediction

- This approach is not helpful.
  Robustly optimized BERT approach (RoBERTa)  https://arxiv.org/abs/1907.11692

- **SOP**: Sentence order prediction
  Used in ALBERT
Masked token prediction
Next sentence prediction

BERT
• Masked token prediction
• Next sentence prediction

Self-supervised Learning
Pre-train

Fine-tune

Model for Task 1
Model for Task 2
Model for Task 3

Downstream Tasks
• The tasks we care
• We have a little bit labeled data.
General Language Understanding Evaluation (GLUE)

- Corpus of Linguistic Acceptability (CoLA)
- Stanford Sentiment Treebank (SST-2)
- Microsoft Research Paraphrase Corpus (MRPC)
- Quora Question Pairs (QQP)
- Semantic Textual Similarity Benchmark (STS-B)
- Multi-Genre Natural Language Inference (MNLI)
- Question-answering NLI (QNLI)
- Recognizing Textual Entailment (RTE)
- Winograd NLI (WNLI)

GLUE also has Chinese version (https://www.cluebenchmarks.com/)
BERT and its Family

• GLUE scores

How to use BERT – Case 1

Input: sequence
Output: class

Example:
Sentiment analysis
Random initialization
Init by pre-train
Better than random

This is the model to be learned.

This is good
Positive
Pre-train v.s. Random Initialization

(fine-tune)  (scratch)

How to use BERT – Case 2

Input: sequence
output: same as input

Example:
POS tagging

I saw a saw

N V DET N
How to use BERT – Case 3

Input: two sequences
Output: a class

Example:
Natural Language Inference (NLI)

premise: A person on a horse jumps over a broken down airplane
hypothesis: A person is at a diner.

contradiction
entailment
neutral
How to use BERT – Case 3

Input: two sequences
Output: a class
How to use BERT – Case 4

• Extraction-based Question Answering (QA)

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

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

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

**Answer**: \( A = \{d_s, \ldots, d_e\} \)

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 rain in a single location 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

$s = 2$

**Random Initialized**

**Softmax**

inner product

BERT

[CLS]  q₁  q₂  [SEP]  d₁  d₂  d₃

question  document
How to use BERT – Case 4

The answer is “d₂d₃”.

s = 2  e = 3
That's all!
Training BERT is challenging!

Training data has more than 3 billions of words.

3000 times of Harry Potter series

---

GLUE scores

8 days with TPU v3

---

Our ALBERT-base


---

Google’s ALBERT-base

Google’s BERT-base

---
BERT Embryology (胚胎學)


When does BERT know POS tagging, syntactic parsing, semantics?

The answer is counterintuitive!
Pre-training a seq2seq model

Reconstruct the input

Cross Attention

Encoder

Decoder

w_1  w_2  w_3  w_4

w_5  w_6  w_7  w_8

Corrupted
MASS / BART

https://arxiv.org/abs/1905.02450
T5 – Comparison

• Transfer Text-to-Text Transformer (T5)
• Colossal Clean Crawled Corpus (C4)
Why does BERT work?

Embedding: Represent the meaning of “大”

The tokens with similar meaning have similar embedding.

Context is considered.
Why does BERT work?

compute cosine similarity
Why does BERT work?

Contextualized word embedding

You shall know a word by the company it keeps

John Rupert Firth

[Diagram of BERT model with contextualized word embedding]

[Diagram of CBOW model]
Why does BERT work?

- Applying BERT to protein, DNA, music classification

This work is done by 高瑋聰

Why does BERT work?

This work is done by 高瑋聰

[Image of a network diagram with nodes labeled A, T, C, and G, and edges showing the flow of information from DNA sequence to BERT.]
Why does BERT work?

- Applying BERT to **protein, DNA, music classification**

<table>
<thead>
<tr>
<th></th>
<th>Protein</th>
<th></th>
<th>DNA</th>
<th></th>
<th>Music</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>location</td>
<td>stability</td>
<td>fluorescence</td>
<td>H3</td>
<td>H4</td>
</tr>
<tr>
<td>specific</td>
<td>69.0</td>
<td>76.0</td>
<td>63.0</td>
<td>87.3</td>
<td>87.3</td>
</tr>
<tr>
<td>BERT</td>
<td>64.8</td>
<td>74.5</td>
<td>63.7</td>
<td>83.0</td>
<td>86.2</td>
</tr>
<tr>
<td>re-emb</td>
<td>63.3</td>
<td>75.4</td>
<td>37.3</td>
<td>78.5</td>
<td>83.7</td>
</tr>
<tr>
<td>rand</td>
<td>58.6</td>
<td>65.8</td>
<td>27.5</td>
<td>75.6</td>
<td>66.5</td>
</tr>
</tbody>
</table>

This work is done by 高瑋聰

To Learn More ......

BERT (Part 1)  
https://youtu.be/1_gRK9EIQpc

BERT (Part 2)  
https://youtu.be/Bywo7m6ySlk
Multi-lingual BERT

Training a BERT model by many different languages.
Zero-shot Reading Comprehension

Training on the sentences of 104 languages

Train on English QA training examples

Test on Chinese QA test

Multi-BERT
Zero-shot Reading Comprehension

• English: SQuAD, Chinese: DRCD

<table>
<thead>
<tr>
<th>Model</th>
<th>Pre-train</th>
<th>Fine-tune</th>
<th>Test</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>QANet</td>
<td>none</td>
<td>Chinese</td>
<td></td>
<td>66.1</td>
<td>78.1</td>
</tr>
<tr>
<td></td>
<td>Chinese</td>
<td>Chinese</td>
<td></td>
<td>82.0</td>
<td>89.1</td>
</tr>
<tr>
<td>BERT</td>
<td>104 languages</td>
<td>Chinese</td>
<td>Chinese</td>
<td>81.2</td>
<td>88.7</td>
</tr>
<tr>
<td></td>
<td>English</td>
<td>Chinese</td>
<td></td>
<td>63.3</td>
<td>78.8</td>
</tr>
<tr>
<td></td>
<td>Chinese + English</td>
<td></td>
<td></td>
<td>82.6</td>
<td>90.1</td>
</tr>
</tbody>
</table>

F1 score of Human performance is 93.30%

This work is done by 劉記良、許宗嫄
Cross-lingual Alignment?

Multi-BERT

深 high
度 est
學 moun
習 tain

跳 jump
兔 rabbit
魚 fish

游 swim
Mean Reciprocal Rank (MRR):
Higher MRR, better alignment


Google’s Multi-BERT

Our Multi-BERT
200k sentences for each lang


How about 1000k?
The training is also challenging ...

Two days ......
(the whole training took one week)
Mean Reciprocal Rank (MRR):

Higher MRR, better alignment

The amount of training data is critical for alignment.

[Image of a bar chart showing MRR for different languages and models, with annotations for Google’s Multi-BERT (1000k sentences) and Our Multi-BERT (200k sentences for each language).]


Projection source: Shih Ting Wen's Master's Oral Examination Presentation
Weird???

If the embedding is language independent ...

How to correctly reconstruct?

There must be language information.

**Where is Language?**

There is a cat.
If this is true ...

This work is done by 劉記良、許宗嫄、莊永松

Unsupervised token-level translation 😊

<table>
<thead>
<tr>
<th>Input (en)</th>
<th>The girl that can help me is all the way across town. There is no one who can help me.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ground Truth (zh)</td>
<td>能帮助我的女孩在小镇的另一边。没有人能帮助我。</td>
</tr>
<tr>
<td>en→zh, α = 1</td>
<td>孩，can来我是 all the way across 市。。There 是无人 can help 我。</td>
</tr>
<tr>
<td>en→zh, α = 2</td>
<td>孩的的家我是这个人的市。。他是他人的到我。</td>
</tr>
<tr>
<td>en→zh, α = 3</td>
<td>。, 的的的他是的个的的，。：他是他，的。他。</td>
</tr>
</tbody>
</table>
Outline

BERT series

GPT series
Predict Next Token

Training data: “台灣大學”

Cross entropy

softmax

Linear Transform

from $w_t$ to $h_t$
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.
How to use GPT?

Description

A few example

例：

It's eight o'clock now. Sue _______ in her bedroom.
A. study
B. studies
C. studied
D. is studying

正確答案為 D，請在答案紙上塗黑作答。
"Few-shot" Learning

(no gradient descent)

<table>
<thead>
<tr>
<th></th>
<th>Translate English to French:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sea otter =&gt; loutre de mer</td>
</tr>
<tr>
<td>2</td>
<td>peppermint =&gt; menthe poivrée</td>
</tr>
<tr>
<td>3</td>
<td>plush giraffe =&gt; girafe peluche</td>
</tr>
<tr>
<td>4</td>
<td>cheese =&gt;</td>
</tr>
</tbody>
</table>

"In-context" Learning
Aggregate Performance Across Benchmarks

- Few Shot
- One Shot
- Zero Shot

Average of 42 tasks
To learn more ......

https://youtu.be/DOG1L9lvsDY
Beyond Text

Data Centric
- Position, 2015
- Jigsaw, 2017
- Rotation, 2018

Prediction
- RNNLM, 1997
- word2v, 2013
- Cutout, 2015
- BERT, 2018
- APC, 2019
- audio2v, 2019
- Mock, 2020
- TERA, 2020

Contrastive
- InfoNCE, 2017
- SimCLR, 2020
- MoCov2, 2020
- CPC, 2019
- MoCo, 2019
- BYOL, 2020
- SimSiam, 2020

NLP

Speech

CV

本投影片由劉廷緯同學提供
Image - SimCLR

https://github.com/google-research/simclr
Bootstrap your own latent:
A new approach to self-supervised Learning

Speech

BERT

Audio version

深度 學 習
Speech GLUE - SUPERB

- **Speech processing Universal PERformance Benchmark**
  - Will be available soon

- **Downstream**: Benchmark with 10+ tasks
  - The models need to know how to process content, speaker, emotion, and even semantics.

- **Toolkit**: A flexible and modularized framework for self-supervised speech models.
  - [https://github.com/s3prl/s3prl](https://github.com/s3prl/s3prl)
A Unified Interface

For All

Use upstreams with torch.hub even without cloning

Your own project

Masked Reconstruction
- Mockingjay
- TERA
- DeCoAR
- NPC

Autoregressive Reconstruction
- APC
- VQ-APC
- CPC

Contrastive
- wav2vec
- wav2vec 2.0

Multi-task
- PASE
- mfcc
- fbank

Baseline

Customize

Any Upstream

Any Downstream

Key Word
- Spoken Term
- Phone Libri
- Phone TIMIT
- ASR
- Intent Classification
- Slot Filling
- Speaker Classification
- Speaker Verification
- Diarization
- Emotion Classification
- Emotion Extraction
- Customization

Detection
- Recognition
- Semantic
- Speaker
- Emotion
- Emotion Representation

Appendix

(a joke)
Predict Next Token

They can do generation.

Keaton Patti @KeatonPatti 2019年8月13日
I forced a bot to watch over 1,000 hours of Batman movies and then asked it to write a Batman movie of its own. Here is the first page.

---

BATMAN
INT. TRADITIONAL BATCAVE

BATMAN stands next to his batmobile. He’s sometimes Bruce Wayne sometimes

BATMAN
This is a safe city. I punched a penguin into p:

ALFRED, Batman’s loyal batler, car:

ALFRED
Eat a dinner, Mattress W.

An explosion explodes. THE JOKER or Joker is a clown but insane. Two-Fs:

BATMAN
No! It is Two-Face and O.

They hate me for being a

Batman throws Alfred at Two-Face. I

a coin. Alfred lands heads up which

BATMAN (CONT'D
It is just you and I, the Bat versus clown. Moral

THE JOKER
I am such a freak. Society i

You drink water, I drink ana

BATMAN
I drink bats just like a bat

Batman looks around for his parents, b

This makes him have anger. He fires a

deflects it with his sick sense of hum

THE JOKER
I have never followed a rule

is my rule. Do you follow?

BATMAN
Alfred, give birth to Robin.

Alfred begins the process since it is

has a present in his hand. He juggles

THE JOKER
Happy batday, Birthman.

Batman opens the present since he’s a
coupon for new parents, but is expired
BATMAN

INT. TRADITIONAL BATCAVE

BATMAN stands next to his batmobile and uses his batcomputer. He’s sometimes Bruce Wayne sometimes Batman. Alltimes orphan.

    BATMAN
    This is now a safe city. I have
    punched a penguin into prison.

ALFRED, Batman’s loyal batler, carries a tray of goth ham.

    ALFRED
    Eat a dinner, Mattress Wayne.

An explosion explodes. THE JOKER and TWO-FACE enter the cave. Joker is a clown but insane. Two-Face is a man but attorney.

    BATMAN
    No! It is Two-Face and One-Face. They hate me for being a bat.

Batman throws Alfred at Two-Face. Two-Face flips Alfred like a coin. Alfred lands heads up which means Two-Face goes home.

    BATMAN (CONT’D)
    It is just you and I, the Joker.
    Bat versus clown. Moral enemies.
THE JOKER
I am such a freak. Society is bad.
You drink water, I drink anarchy.

BATMAN
I drink bats just like a bat would!

Batman looks around for his parents, but they are still dead.
This makes him have anger. He fires a batrocket. The Joker deflects it with his sick sense of humor. A clownly power.

THE JOKER
I have never followed a rule. That is my rule. Do you follow? I don’t.

BATMAN
Alfred, give birth to Robin.

Alfred begins the process since it is his job. The Joker now has a present in his hand. He juggles it over to Batman.

THE JOKER
Happy batday, Birthman.

Batman opens the present since he’s a good guy. It contains a coupon for new parents, but is expired. This is a Joker joke.
I forced a bot to watch over 1,000 hours of Olive Garden commercials and then asked it to write a Saw movie of its own. Here is the first page.

I forced a bot to watch over 1,000 hours of Jerry Springer and then asked it to write a Saw movie of its own. Here is the first page.