M4R Oral Presentation

Pre-trained Language Models in NLP (M4R)

Jiacheng Xu
Supervised by: Dr. Prasun Ray

Imperial College London

Outline

What is NLP?

Deifinition

Natural language processing (NLP) is an interdisciplinary subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data.

What is NLP?

Deifinition

Natural language processing (NLP) is an interdisciplinary subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data.

Some common NLP tasks:

What is NLP?

Deifinition

Natural language processing (NLP) is an interdisciplinary subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data.

Some common NLP tasks:

- Parsing

What is NLP?

Deifinition

Natural language processing (NLP) is an interdisciplinary subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data.

Some common NLP tasks:

- Parsing
- Named entity recognition

What is NLP?

Deifinition

Natural language processing (NLP) is an interdisciplinary subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data.

Some common NLP tasks:

- Parsing
- Named entity recognition
- Question answering

What is NLP?

Deifinition

Natural language processing (NLP) is an interdisciplinary subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data.

Some common NLP tasks:

- Parsing
- Named entity recognition
- Question answering
- Text classification

Pre-training - Fine-tuning

Pre-training - Fine-tuning

Definition

'Pre-training - Fine-tuning' is an approach of model training and application in the field of artificial intelligence, which is composed of 'pre-training' and 'fine-tuning' stages.

Pre-training - Fine-tuning

Definition

'Pre-training - Fine-tuning' is an approach of model training and application in the field of artificial intelligence, which is composed of 'pre-training' and 'fine-tuning' stages.

Pre-training:

Pre-training - Fine-tuning

Definition

'Pre-training - Fine-tuning' is an approach of model training and application in the field of artificial intelligence, which is composed of 'pre-training' and 'fine-tuning' stages.

Pre-training:
self-supervised training

Pre-training - Fine-tuning

Definition

'Pre-training - Fine-tuning' is an approach of model training and application in the field of artificial intelligence, which is composed of 'pre-training' and 'fine-tuning' stages.

Pre-training:
self-supervised training

Fine-tuning:

Pre-training - Fine-tuning

Definition

'Pre-training - Fine-tuning' is an approach of model training and application in the field of artificial intelligence, which is composed of 'pre-training' and 'fine-tuning' stages.

Pre-training:
self-supervised training

Fine-tuning:
supervised training, downstream tasks

ELMo, GPT and BERT

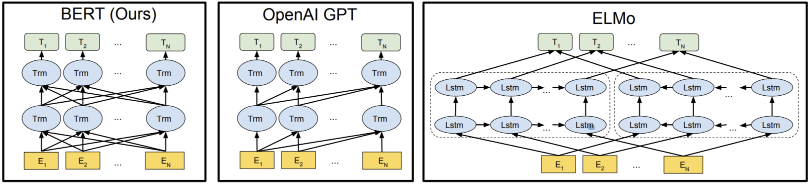

Figure: Display of the three pre-trained model architectures

ELMo, GPT and BERT

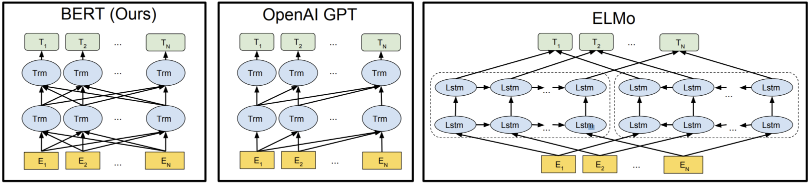

Figure: Display of the three pre-trained model architectures

Long short-term memory (LSTM)

ELMo, GPT and BERT

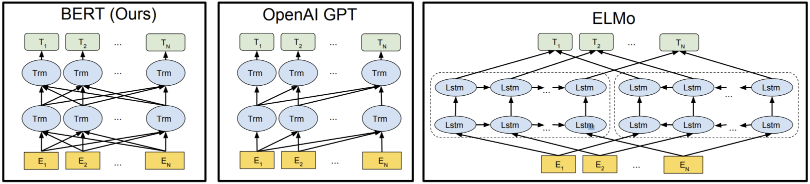

Figure: Display of the three pre-trained model architectures

Long short-term memory (LSTM)
Transformer:

ELMo, GPT and BERT

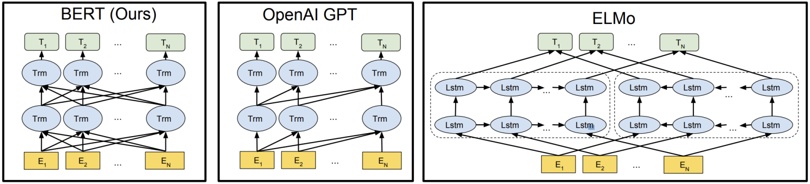

Figure: Display of the three pre-trained model architectures

Long short-term memory (LSTM)
Transformer: Self-attention

Input representations

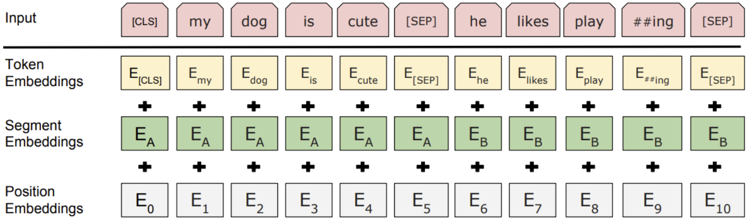

Figure: Input representations of BERT

Masked LM

Input sentence: `My dog is hairy.`

Masked LM

Input sentence: `My dog is hairy.`
Masking: `My dog is [MASK].`

Masked LM

Input sentence: `My dog is hairy.`
Masking: `My dog is [MASK].`

- 80% of the time: Replace the word with the `[MASK]` token, `my dog is [MASK].`

Masked LM

Input sentence: `My dog is hairy.`
Masking: `My dog is [MASK].`

- 80% of the time: Replace the word with the `[MASK]` token, `my dog is [MASK].`
- 10% of the time: Replace the word with a random word, `my dog is apple.`

Masked LM

Input sentence: `My dog is hairy.`
Masking: `My dog is [MASK].`

- 80% of the time: Replace the word with the `[MASK]` token, `my dog is [MASK].`
- 10% of the time: Replace the word with a random word, `my dog is apple.`
- 10% of the time: Keep the word unchanged, `my dog is hairy.` The purpose of this is to bias the representation towards the actual observed word.

Next Sentence Prediction

Input = `[CLS] the man went to [MASK] store [SEP]he bought a gallon [MASK] milk [SEP]`
Label = `IsNext`

Input = `[CLS] the man [MASK] to the store [SEP]penguin [MASK] are flight less birds[SEP]`
Label = `NotNext`

Initiative

Table 2: Knowledge-infused methods based on training objective design

Model name	Knowledge resource	Training objectives for knowledge infusion
ERNIE[26]	Chinese word segmentation, entity words	Improved BERT MLM masking strategy: Full word masking, Entity word masking
MacBERT[27]	Synonym list	Improved BERT MLM masking strategy: Synonym masking
ERNIE2.0[28]	Grammar rules, search logs	Improved BERT NSP training objectives: Rhetoric relationship prediction, Retrieval relationship prediction
SentiLR[29]	SentiWordNet	Improved MLM training objectives: Sentiment word masking prediction
KEPLER[30]	Wikidata;Wikipedia	Knowledge representation learning
WKLM[31]	Wikidata;Wikipedia	Entity replacement prediction, Entity boundary prediction
KnowBert[32]	WordNet;Wikipedia	Entity alignment
EMBERT[33]	Entity Triples	Entity replacement prediction, Entity segmentation prediction
K-Adapter[34]	Wikipedia;Wikidata; Standford Parser	Entity relationship classification Dependency relationship classification
LIBERT[35]	WordNet; Roget's Thesaurus	Grammatical relationship classification
SenseBERT[36]	WordNet	Word sense prediction

Deficiencies of Original BERT

Mask LM

Deficiencies of Original BERT

Mask LM

- Masking individual words is not conducive to the learning of complete word semantic features.

Deficiencies of Original BERT

Mask LM

- Masking individual words is not conducive to the learning of complete word semantic features.
- Domain-specific terms are more densely distributed in scientific field corpora.

Deficiencies of Original BERT

Mask LM

- Masking individual words is not conducive to the learning of complete word semantic features.
- Domain-specific terms are more densely distributed in scientific field corpora.

NSP

Deficiencies of Original BERT

Mask LM

- Masking individual words is not conducive to the learning of complete word semantic features.
- Domain-specific terms are more densely distributed in scientific field corpora.

NSP

- Abstracts of scientific papers have a clear structure of rhetorical steps.

Deficiencies of Original BERT

Mask LM

- Masking individual words is not conducive to the learning of complete word semantic features.
- Domain-specific terms are more densely distributed in scientific field corpora.

NSP

- Abstracts of scientific papers have a clear structure of rhetorical steps.
- The order of rhetorical steps conveys a certain logical relationship.

Term-Masked LM (T-MLM)

Input sentence: `Exploring the causes of pneumonia`

Term-Masked LM (T-MLM)

Input sentence: `Exploring the causes of pneumonia`
Masking: `Exploring the causes of [MASK]`

Term-Masked LM (T-MLM)

Input sentence: `Exploring the causes of pneumonia`
Masking: `Exploring the causes of [MASK]` In tokens:
`[Exploring] [the] [causes] [of] [pneu] [monia]`

Term-Masked LM (T-MLM)

Input sentence: `Exploring the causes of pneumonia`
Masking: `Exploring the causes of [MASK]` In tokens:
`[Exploring] [the] [causes] [of] [pneu] [monia]`

- In 80% of cases, the token is replaced with `[MASK]`. If this token is part of a scientific term, other tokens belonging to this term are also masked: `Exploring the causes of [MASK]`

Term-Masked LM (T-MLM)

Input sentence: `Exploring the causes of pneumonia`
Masking: `Exploring the causes of [MASK]` In tokens:
`[Exploring] [the] [causes] [of] [pneu] [monia]`

- In 80% of cases, the token is replaced with `[MASK]`. If this token is part of a scientific term, other tokens belonging to this term are also masked: `Exploring the causes of [MASK]`
- 10% of cases, the token is replaced with a random word: `Exploring the causes of apple##monia`

Term-Masked LM (T-MLM)

Input sentence: `Exploring the causes of pneumonia`
Masking: `Exploring the causes of [MASK]` In tokens:
`[Exploring] [the] [causes] [of] [pneu] [monia]`

- In 80% of cases, the token is replaced with `[MASK]`. If this token is part of a scientific term, other tokens belonging to this term are also masked: `Exploring the causes of [MASK]`
- 10% of cases, the token is replaced with a random word: `Exploring the causes of apple##monia`
- 10% of cases, keep the token unchanged: `Exploring the causes of pneumonia`

Move-Sentence Order Prediction (M-SOP)

Directly inputs two adjacent sentences, though their order may be shuffled. For the input sample, the model needs to make judgments in the following four situations for classification:

Move-Sentence Order Prediction (M-SOP)

Directly inputs two adjacent sentences, though their order may be shuffled. For the input sample, the model needs to make judgments in the following four situations for classification:

- The sentence pair is in correct order and in the same move.
- The sentence pair is in correct order but in different moves.
- The sentence pair is in incorrect order but in the same move.
- The sentence pair is in incorrect order and in different moves.

Results

		Classification Task MOVE	CLC	Sequence Task AKE	Tagging NER	Non-scientific Paper Dataset chip-ctc	ccks
Baseline Model	BERT-base	93.9	82.55	83.89	69.77	85.46	78.76
Model in This Study	CsciMedBERT	95.3	85.3	88.87	74.87	85.82	79.03
Other Models	MacBERT	94.16	83.66	84.43	69.95	85.83	78.9
	Medbert	93.84	84.46	84.62	71.38	86.11	78.68
	MC-BERT	93.99	83.44	84.43	70.53	85.98	79.36