Self-supervised Learning

Unsupervised Learning: there is no labels y

- Self-supervised Learning: without labels y given, the model learn to

BERT

Bidirectional encoder representations from transformers
Basically, BERT is a transformer encoder, so the length of input sequence is equal to the length of output sequence.

Pre-train (self-supervised) + Fine-tune(supervised) --> BERT (semi-supervised)

Pre-train

BERT was pre-trained in 2 tasks:

- Masking Language Model: predict the masked token given its context
  the way of masking is also ramdomly chosen:
  - replace the token with a special token [Mask]
  - replace the token with another random token
- Next Sentence Prediction: predict if the two sentence appeared sequentially in the training corpus
  2 special tokens in NSP
  - [CLS] (for classify) to mark the beginning of the sentence (in pre-train) and represent the sentiment information (in fine-tuning)
  - [SEP] (for separate) to separate sentences

Fine-tune

We can find-tune the pre-trained BERT to gain the capability in **downstream tasks**

Task sets: a collection of well-defined tasks used to evaluate the performance of pre-trained models
e.g. General Language Understanding Evaluation (GLUE)

Finetuned tasks for BERT

- Sentiment classification
- Sentence classification

- Answering multiple-choice questions
- Part-of-speech tagging

Embedding:
How to fine-tune

1. we usually input two sentence , [CLS] and [SEP]
2. BERT encodes tokens into embeddings. e.g. the token embedding of [CLS] represents sentiment information of the whole seqeunce
3. we select tokens according to our understanding of the downstream task.
   e.g. For classification, [CLS] embedding should be selected because the sentiment information is required
4. input the token embeddings into a Softmax layer for multi-classification

Why BERT works

The conventional explanation: A word's meaning is determined by the company it keeps (the context). BERT learn the word's meaning in Masking Language Model task, so it can understand each word well.
Continous Bag of Words (CBOW): a ML model that predicts the masked token by the context.
BERT can be seen as the DL version of CBOW.