seq2seq pretrain

$w_1\ w_2\ w_3\ w_4$

↑ ↑ ↑ ↑

[Model] → [Model]   reconstruct the input (self-supervised)

↑ ↑ ↑ ↑

$w_1\ w_2\ w_3\ w_4$ (corrupted)

怎么样 corrupt input? MASS (Masked seq to seq pre-training)
or BART (Bidirectional and autoregressive transformers)

MASS — 随机 mask token 还原 token

or delete — 还原 token

permutation token 还原 token

结果不行 < rotation token 还原 token

BART consistently good: Text infilling (加 token or 盖 token) 还原 token e.g. A ☐ B [SEP] ☐ E → A B [SEP] C D E (true)

UniLM ← encoder / decoder / seq2seq

BERT (bidirectional LM)

GPT (Left-to-right LM). 需要好好设计 attention. 每个 token 只能看到左边的 token

BART/MASS (seq2seq LM). segment 1 → encoder (互相看); segment 2 → decoder (只看左边的)

ELECTRA (efficiently learning an encoder that classifiers token replacements accurately)

No No Yes No No

☐ ☐ ☐ ☐ ☐

↑ ↑ ↑ ↑ ↑

Model

↑ ↑ ↑ ↑

the chef ate the meal

原来是 cooked

· predicting yes/no is easier than reconstruction.

every output position is used. (BERT 只 predict masked)

怎么样置换, 才能无法看出来... — 用 smaller BERT 来置换

$\frac{1}{4}$ size of XLNet 就实现 similar GLUE score

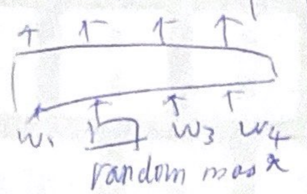
Predict next token

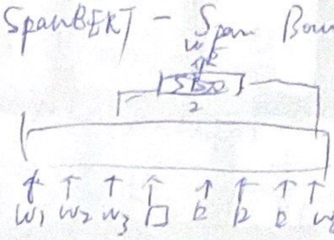
LSTM -> ELMO (双向：但是正向、逆向没有交汇，正向时没有看到后面的token)

self-attention -> GPT, Megatron, Turing NLG

Masking input { whole word
word
phrase-level / entity level — ERNIE
span (prob, span length) — SpanBERT

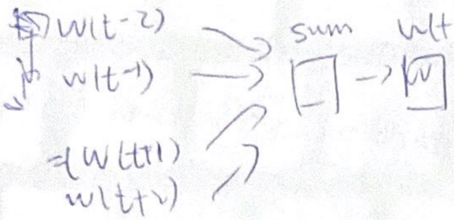
BERT





用于coreference

(CBOW (using context to predict the missing token)



window 20左右

XLNet

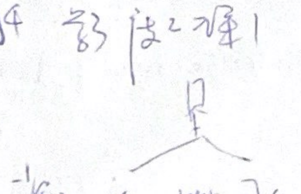
Transformer-XL

深1 度2 学3 习4

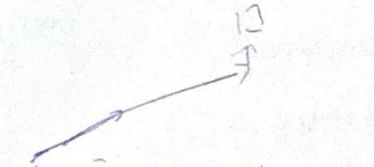
language model 向后

习4 学3 度2 深1

BERT 向后



random or



XLNet 只给positional info 不给content info. 作为特殊的处理 看paper

BERT cannot talk? (如果sequence本来就不是由左而右产生的, BERT也可以talk)

Given partial sequence, predict the next token (但BERT training不是这样 而是双向

∴BERT不善言辞。 ∴不适用于seq2seq的pre-trained model. (只能当作encoder/decoder预先做pretraining)

但MASS/BART 是pre-trained seq2seq model 用by self-supervised learning

BERT and its family

- Pretrain model

  word embedding.
  - word2vec
  - glove
  - fasttext. character-level, 可handle新词

  每个"狗"词义不一样

  contextualized word embedding.
  - ELMO
  - BERT

  看整句子之后再给 embedding.

Smaller BERT
- distill BERT
- Tiny BERT
- Mobile BERT
- Q8BERT
- ALBERT (最知名, 小且结果 improve)

Network Compression. (~~see~~ ~~refer~~ video for reference)

Network Architecture (希望读非常长的 sequence). BERT - 512 token

Transformer-XL: segment-level recurrence with state reuse

reformer
longformer → 当 sequence 很长, self-attention takes $O(n^2)$ 这两者用来减少 complexity of self-attention

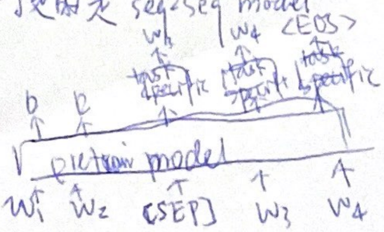
- How to fine-tune
  - ① fine tune task-specific
  - ② fine-tune 一起 pre-trained & task-specific
  - ③ adaptor: 只调 pretrain 中的一小部分 adaptor & task-specific

  BERT里ML讲过

  attention

但值得一提的是 seq2seq model v1 pretrain encoder → task-specific decoder. 问题是 decoder 没有 pretrain.

v2.



Sentence Level pretrain

(representation for sentence)

skip thought

$w_1\ w_2\ w_3 \rightarrow$ encoder $\rightarrow h \rightarrow$ decoder $\rightarrow w_4\ w_5\ w_6\ w_7$ (next sentence)

表示如果下一个 sentence 是相同的，

的 sentence embedding 相似

(但生成开面应要量比较大)

Quick thought

$w_1\ w_2\ w_3 \rightarrow$ encoder $\rightarrow h$

这段话 $w_4\ w_5\ w_6\ w_7 \rightarrow$ encoder $\rightarrow h$

如果相邻的 embedding 相近

BERT train [CLS] embedding 用一个 global 的任务：next sentence prediction

Robustly optimized BERT approach. (ROBERTa)

但结果不好 不能因为很简单

SOP: sentence order prediction (如果相接的话 yes.

更困难的 task ∴ 两个句子很相近 如果调换顺序的话 No.)

ALBERT

StructBERT (Alice)：结合 next sentence prediction 和 SOP.

T5 - ~~Comparison~~ 是一个 comparison

BERT + knowledge → ERNIE

(不是之前的 ERNIE)

another story

Audio BERT