NLP tasks
- Question Answering (问答)
- Machine Translation (机器翻译)
- Reading Comprehension (阅读理解)
- Summarization (总结)
- Commonsence Reasoning (常识推理)
- Sentiment Analysis (情感分析)

Zero-shot: 不进行任何的系统架构/参数的修改，直接将model应用于任务之上

WSC (Winograd Schema Challenge): Winograd 教授提出的机器智能测试，通过向machine询问一系列选择题来测试其智能.

Autoregressive LM: LM that ① [predicts the next word based on the before-context] or ② [predicts the previous word based on the after-context.] ⇒ 即自动匹配任务来预测 next word.

AutoEncoder: A model composed by an encoder and a decoder sub-models.
- encoder: compress input
- decoder: recreate input from compressed representation

After training, the encoder is saved while the decoder is discarded.

Pre-training Technique

I. Word2vec ⇒ Skip. more details see another note.

II. ELMo

What:

ELMo is a dynamic word embedding, compared to Word2vec, which is static

an auto-regressive model.

Why propose ELMo:

Word2vec is static, it can't handle polysemous words.

多义词，需结合context才能判断其语义.

How:

ELMo's network architecture adds two Bi-LSTM layers on top of the word embedding layer

⇓

learned by Word2vec.

Then all three embeddings are weighted and summed up

⇓

1 word embedding layer + 2 Bi-LSTM layer

to get new ELMo embedding.

△ Two Bi-LSTM layers are trained through specific context.

III. GPT

Transformer: Encoder / Decoder

language model

What:

Transformer-decoder-based autoregressive LM.

自回归

How: Put a certain number of Transformer-decoder layers on top of embedding layer.

learned by Word2vec.

Why: Add context information to a word's embedding throught Transformer-decoder layer.

IV. BERT ⟹ Skip, more details see another note

V. XLNet

Denoise AutoEncoder

What:

XLNet is a combination of Autoregressive model and DAE

How:

Introduce the bidirectional LM into Autoregressive LM.

DAE

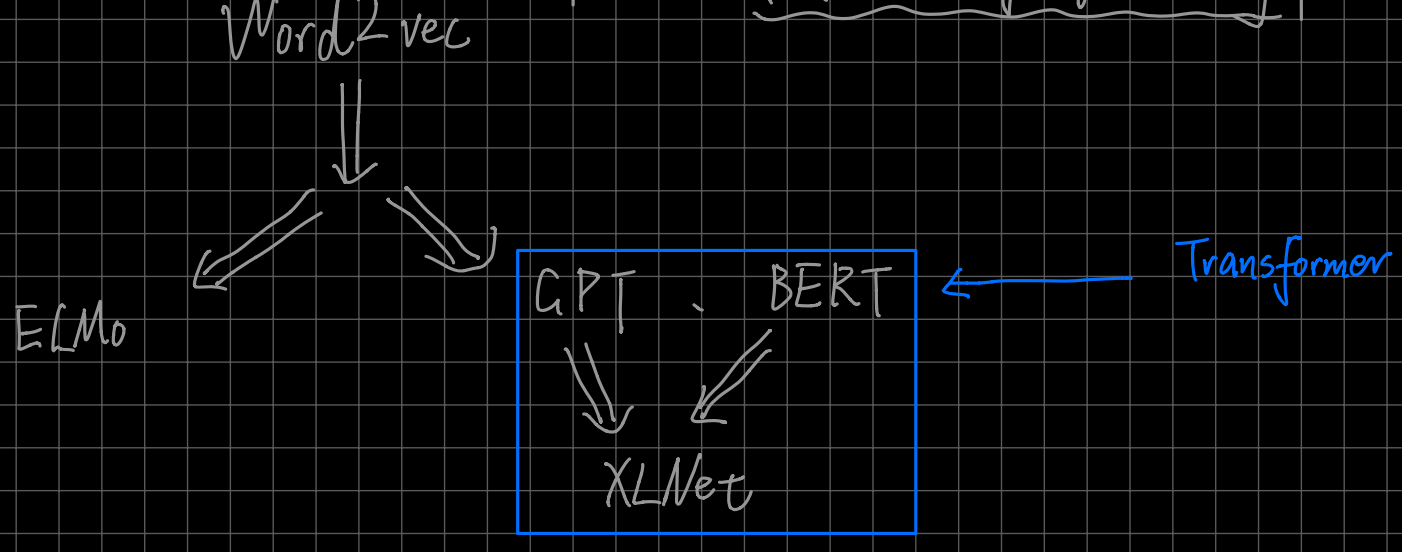
⟹ For an input sentence, do permutation to get a list of possible sentences. Randomly choose some sentences as input for model-pretraining (Note that the current word position in each sentence is fixed, in order to remain the Autoregressive LM architecture unchanged.

全排列

Comparison :

	Word2vec	ELMo	GPT	BERT	XLNet
Method:		feature-based	Transformer-decoder based + fine-tuning based	Transformer-encoder based + fine-tuning based	ALM + DAE
Network:		Word2vec + 2 Bi-LSTM	Word2vec + several Transformer-decoder layers	Word2vec + several Transformer-encoder layers	ALM archit + DAE permutation layer

区别：① GPT use transformer's decoder
BERT use transformer's encoder
② GPT is unidirectional LM
BERT is bi-directional LM
③ GPT embedding ⟺ [Word embedding + attention information]
BERT embedding ⟺ Token embedding + Position embedding + Segment embedding



Common datasets for LMs

Note: NLP model 的 performance 常用一些公认的 dataset 来测估 model 于各方面的能力，可以帮助：① 发现 model 的不足之处 ② 评估 model 与 SOTA 的差距/领先 SOTA 的量.

1. Children's Book Test: model 在 common nouns, name entities 等 (verbs, adjs...) 不同类别的词上的表现 ⇒ omitted word selection accuracy
2. LAMBADA: ability to model long-range dependencies in text ⇒ predict the final word of sentences

词准确性

3. CoQA / SQUAD: 阅读理解能力 ⇒ F1 values of prediction
4. CNN and Daily Mail: 总结能力 ⇒ first 3 sentences in 100 generated tokens (by Top-k random sampling) as summary, use ROUGE 1,2,L metrics to measure generative sumaries.
5. WMT-14 French-English / English-French: 翻译能力 ⇒ BLEU metrics for translated sentence and target sentence.
6. Natural Questions: 问答能力 ⇒ exact match metric of the real answers and expected answers
7. WSC: 常识推理能力 ⇒ by measuring its ability to resolve ambiguities in the text

△ 在 test evaluation model performance 时要注意 generalization ability 是否真正达到，是否有 Memorization (e.g overlap between train-test set) 导致的 over-reporting.