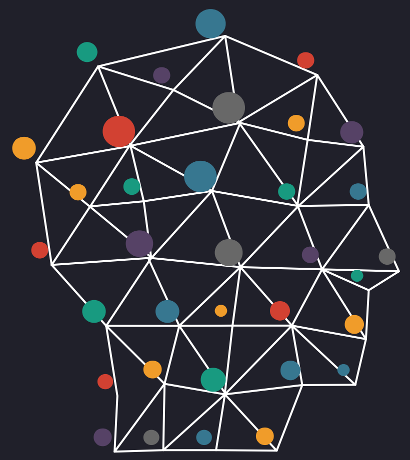
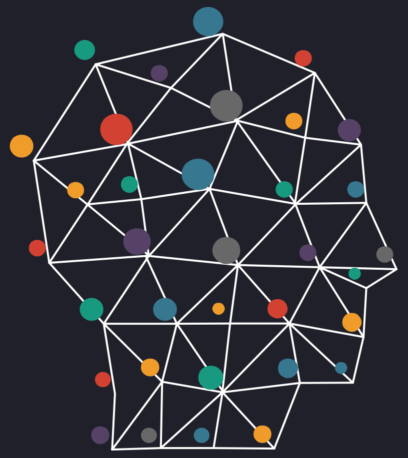
Pre-trained Models for Natural Language Processing: A Survey
自然语言处理预训练模型综述

周梅、周雨慧、陈悦

2021年10月28日

CONTENTS





1. Introduction

1.1 为什么需要预训练模型

- 从CV领域中来
- 标注数据昂贵
- 性能极大提升

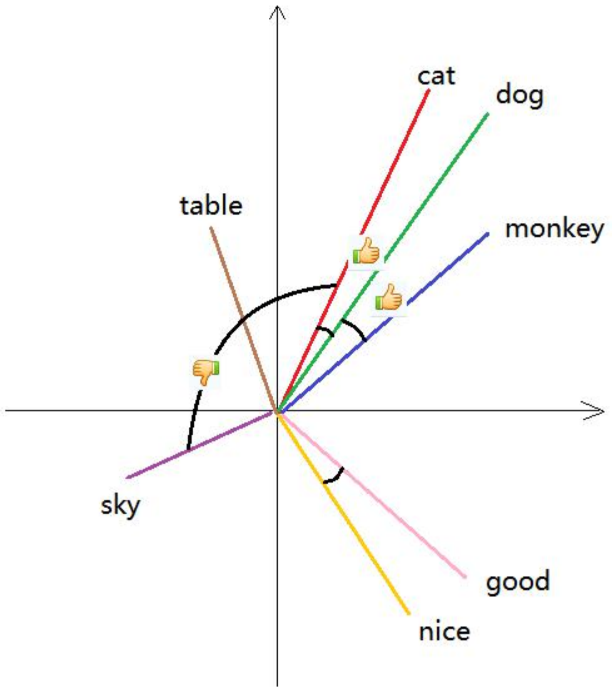
Edges (layer conv2d0) **Textures** (layer mixed3a) **Patterns** (layer mixed4a) **Parts** (layers mixed4b & mixed4c)

- 在庞大的无标注数据上进行预训练可以获取更通用的语言表示，并有利于下游任务
- 为模型提供了一个更好的初始化参数，在目标任务上具备更好的泛化性能、并加速收敛
- 是一种有效的正则化手段，避免在小数据集上过拟合（一个随机初始化的深层模型容易对小数据集过拟合）

1.2 词嵌入（word embedding）

	o_ENE	o_ESE	o_East	o_NE	o_NNE	o_NNW	o_NW	o_North	o_SE	o_SSE	o_SSW	o_SW	o_South	o_Variable	o_WSW
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
4	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
10	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0

One-Hot



Projection of the embedding vectors to 2-D

- 机器学习时代：矩阵分解（SVD）、LSA、LDA
- 深度学习时代：Word2Vec、Bert、GPT

1.3 预训练模型两大范式

1. 浅层词嵌入（ Non-Contextual Embeddings）

词嵌入	训练目标	全局/局部语料
NNLM	语言模型	局部语料
word2vec	非语言模型（窗口上下文）	局部语料
Glove	非语言模型（词共现矩阵）	全局语料

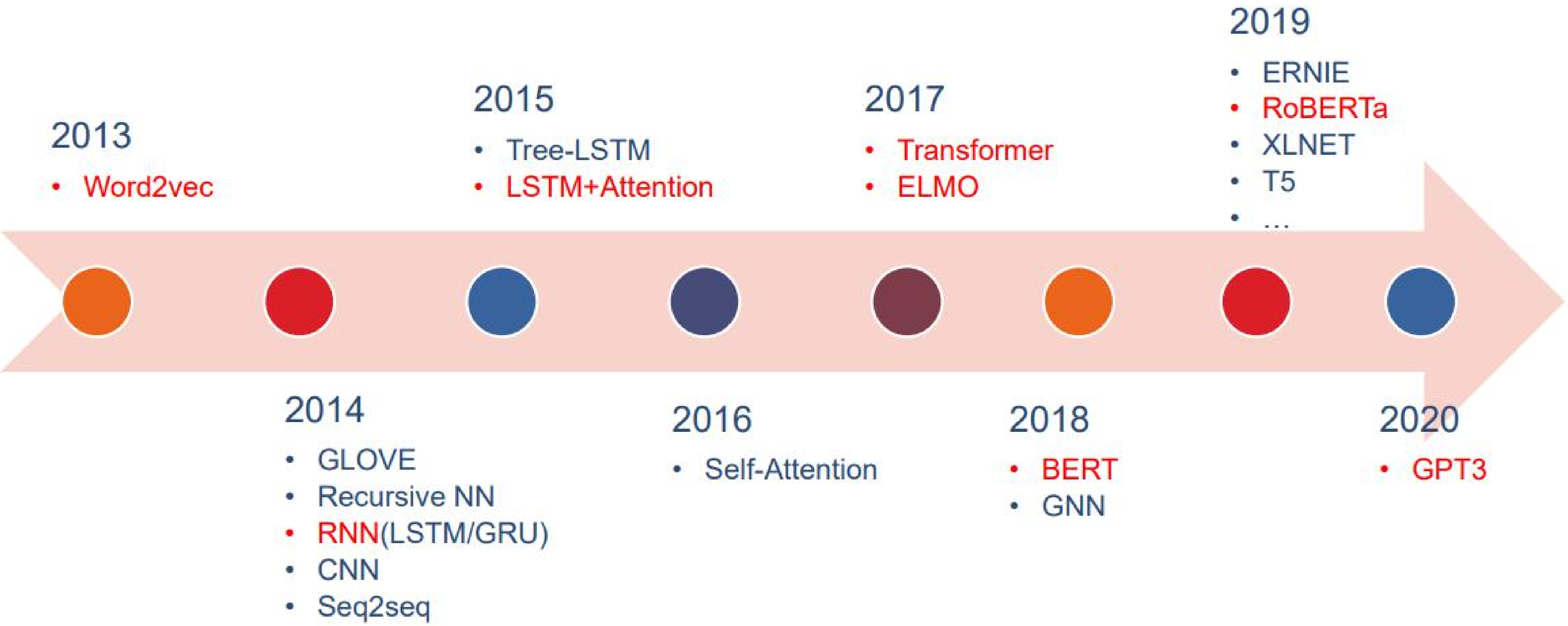
浅层词嵌入的主要缺陷为：词嵌入与上下文无关，每个单词的嵌入向量始终是相同，因此不能解决一词多义的问题。

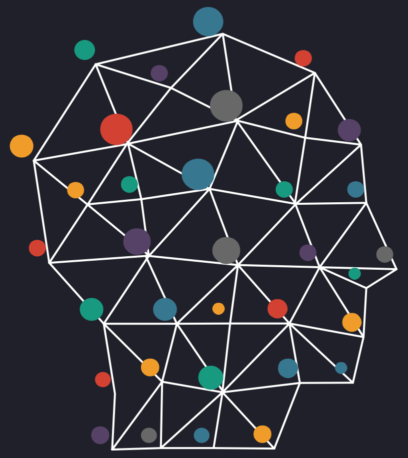
2.预训练编码器（Contextual Embeddings）

编码器	PTMs代表	计算方式
MLP	NNLM/word2vec	前馈+并行
CNNs		前馈+并行
RNNs	ELMO	循环+串行
Transformer	GPT（Decoder） BERT（Encoder）	前馈+并行
Transformer-XL	XLNet	循环+串行
长距离依赖建模能力		

预训练编码器通常采用LSTM和Transformer两种特征提取器

1.4 PTMs的发展历程





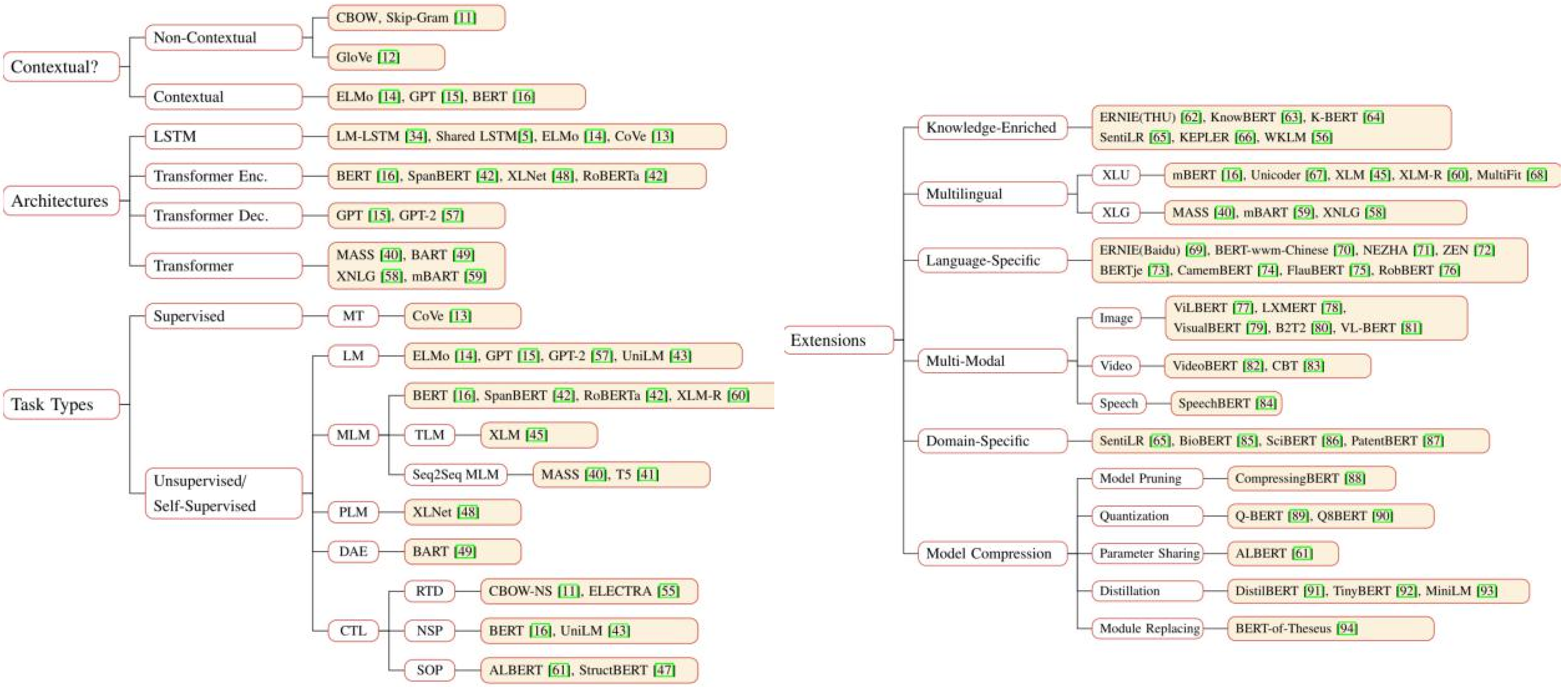
2.Classification

2.1 预训练任务类型

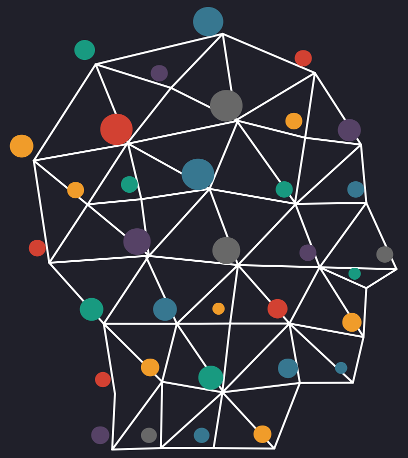
Task	Loss Function	Description
LM	$\mathcal{L}_{\mathrm{LM}} = -\sum_{t=1}^{T} \log p(x_t \mid \mathbf{x}_{<t})$	$\mathbf{x}_{<t} = x_1, x_2, \cdots, x_{t-1}$.
MLM	$\mathcal{L}_{\mathrm{MLM}} = -\sum_{\hat{x} \in m(\mathbf{x})} \log p\left(\hat{x} \mid \mathbf{x}_{\backslash m(\mathbf{x})}\right)$	$m(\mathbf{x})$ and $\mathbf{x}_{\backslash m(\mathbf{x})}$ denote the masked words from $\mathbf{x}$ and the rest words respectively.
Seq2Seq MLM	$\mathcal{L}_{\mathrm{S2SMLM}} = -\sum_{t=i}^{j} \log p\left(x_t \mid \mathbf{x}_{\backslash \mathbf{x}_{i:j}}, \mathbf{x}_{i:t-1}\right)$	$\mathbf{x}_{i:j}$ denotes an masked n-gram span from i to j in $\mathbf{x}$.
PLM	$\mathcal{L}_{\mathrm{PLM}} = -\sum_{t=1}^{T} \log p(z_t \mid \mathbf{z}_{<t})$	$\mathbf{z} = perm(\mathbf{x})$ is a permutation of $\mathbf{x}$ with random order.
DAE	$\mathcal{L}_{\mathrm{DAE}} = -\sum_{t=1}^{T} \log p(x_t \mid \hat{\mathbf{x}}, \mathbf{x}_{<t})$	$\hat{\mathbf{x}}$ is randomly perturbed text from $\mathbf{x}$.
DIM	$\mathcal{L}_{\mathrm{DIM}} = s(\hat{\mathbf{x}}_{i:j}, \mathbf{x}_{i:j}) - \log \sum_{\tilde{\mathbf{x}}_{i:j} \in \mathcal{N}} s(\hat{\mathbf{x}}_{i:j}, \tilde{\mathbf{x}}_{i:j})$	$\mathbf{x}_{i:j}$ denotes an n-gram span from i to j in $\mathbf{x}$, $\hat{\mathbf{x}}_{i:j}$ denotes a sentence masked at position i to j, and $\tilde{\mathbf{x}}_{i:j}$ denotes a randomly-sampled negative n-gram from corpus.
NSP/SOP	$\mathcal{L}_{\mathrm{NSP/SOP}} = -\log p(t \mid \mathbf{x}, \mathbf{y})$	$t = 1$ if $\mathbf{x}$ and $\mathbf{y}$ are continuous segments from corpus.
RTD	$\mathcal{L}_{\mathrm{RTD}} = -\sum_{t=1}^{T} \log p(y_t \mid \hat{\mathbf{x}})$	$y_t = \mathbf{1}(\hat{x}_t = x_t)$, $\hat{\mathbf{x}}$ is corrupted from $\mathbf{x}$.

[1] $\mathbf{x} = [x_1, x_2, \cdots, x_T]$ denotes a sequence.

2.2 预训练模型分类



Contextual?
Non-Contextual
CBOW, Skip-Gram [11]
GloVe [12]
Contextual
ELMo [14], GPT [15], BERT [16]
Architectures
LSTM
LM-LSTM [34], Shared LSTM[5], ELMo [14], CoVe [13]
Transformer Enc.
BERT [16], SpanBERT [42], XLNet [48], RoBERTa [42]
Transformer Dec.
GPT [15], GPT-2 [57]
Transformer
MASS [40], BART [49]
XNLG [58], mBART [59]
Task Types
Supervised
MT
CoVe [13]
Unsupervised/
Self-Supervised
LM
ELMo [14], GPT [15], GPT-2 [57], UniLM [43]
MLM
BERT [16], SpanBERT [42], RoBERTa [42], XLM-R [60]
TLM
XLM [45]
Seq2Seq MLM
MASS [40], T5 [41]
PLM
XLNet [48]
DAE
BART [49]
CTL
RTD
CBOW-NS [11], ELECTRA [55]
NSP
BERT [16], UniLM [43]
SOP
ALBERT [61], StructBERT [47]
Extensions
Knowledge-Enriched
ERNIE(THU) [62], KnowBERT [63], K-BERT [64]
SentiLR [65], KEPLER [66], WKLM [56]
Multilingual
XLU
mBERT [16], Unicoder [67], XLM [45], XLM-R [60], MultiFit [68]
XLG
MASS [40], mBART [59], XNLG [58]
Language-Specific
ERNIE(Baidu) [69], BERT-wwm-Chinese [70], NEZHA [71], ZEN [72]
BERTje [73], CamemBERT [74], FlauBERT [75], RobBERT [76]
Multi-Modal
Image
ViLBERT [77], LXMERT [78],
VisualBERT [79], B2T2 [80], VL-BERT [81]
Video
VideoBERT [82], CBT [83]
Speech
SpeechBERT [84]
Domain-Specific
SentiLR [65], BioBERT [85], SciBERT [86], PatentBERT [87]
Model Compression
Model Pruning
CompressingBERT [88]
Quantization
Q-BERT [89], Q8BERT [90]
Parameter Sharing
ALBERT [61]
Distillation
DistilBERT [91], TinyBERT [92], MiniLM [93]
Module Replacing
BERT-of-Theseus [94]

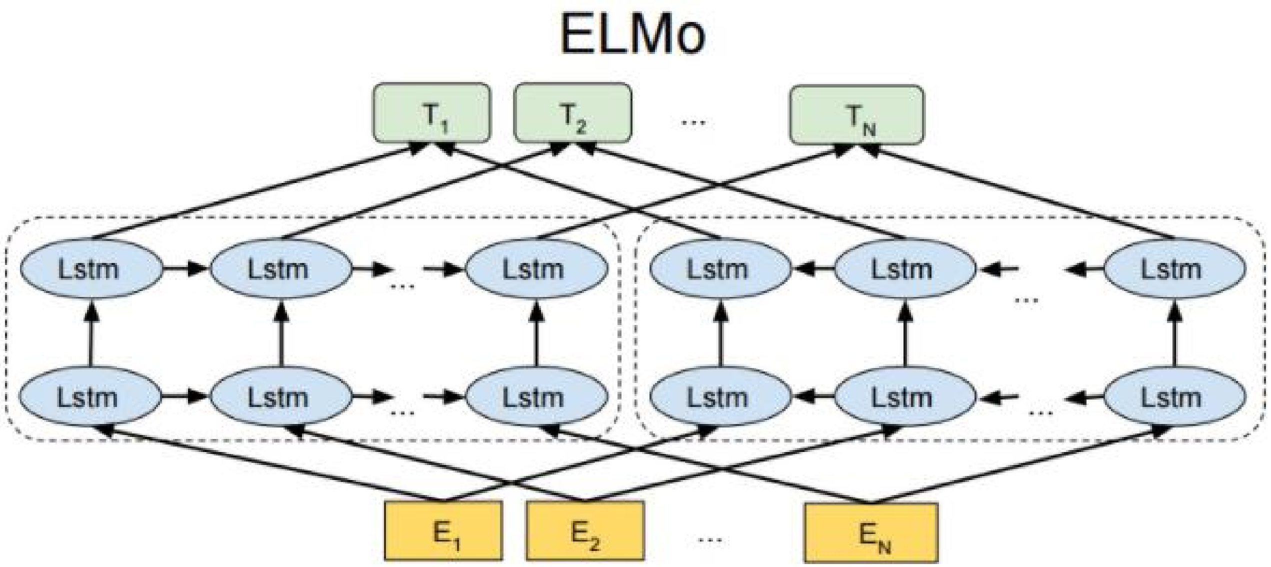


3. Main models

3.1 AllenNLP ELMo: Embeddings from Language Models

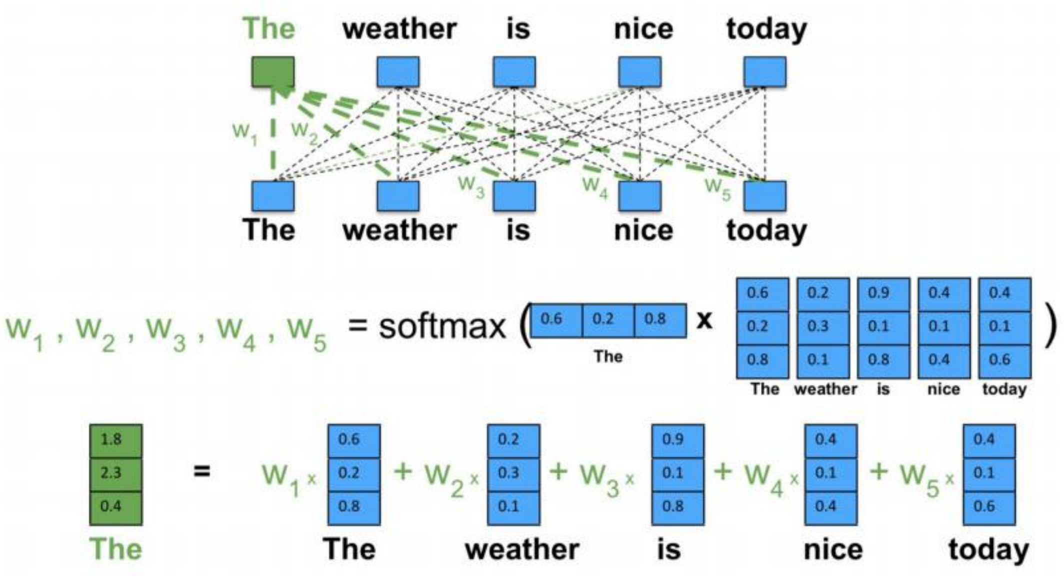
- ✓ Embedding size: 512
  - ✓ 2048 character n-gram convolutional filters
- ✓ BiLSTM layers: 2
- ✓ BiLSTM hidden states : 4096
- ✓ Residual Connection

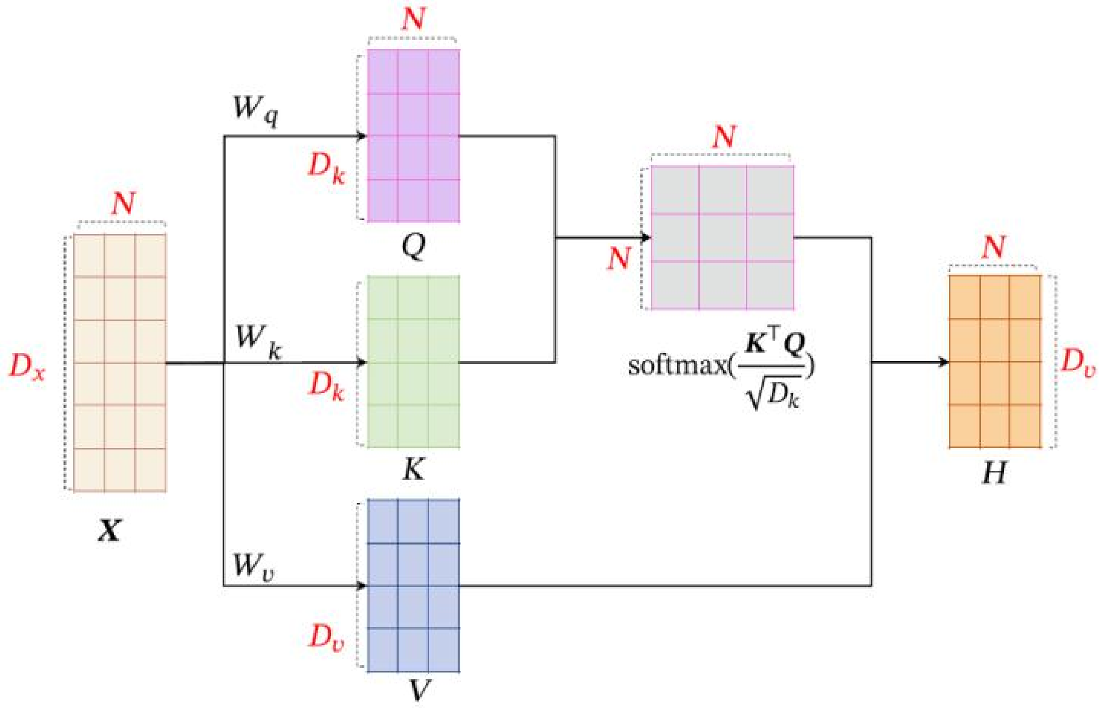
分别训练前向语言模型和反向语言模型



Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations. arXiv preprint arXiv:1802.05365, 2018.

3.2 自注意力（Self- Attention ）

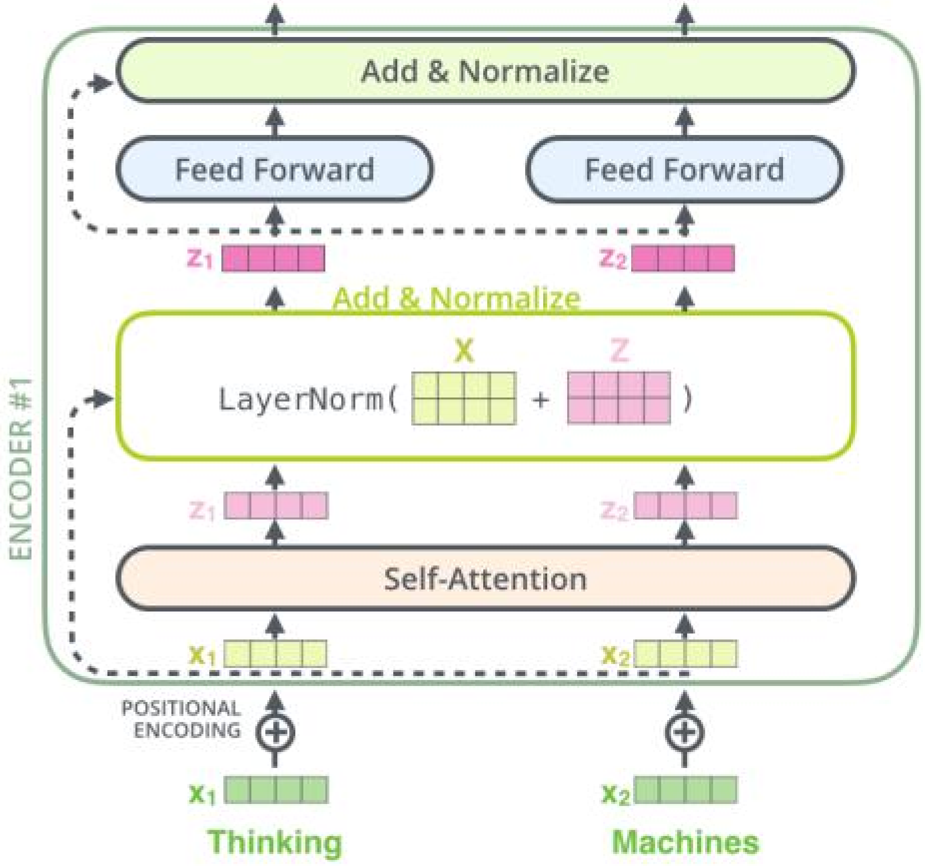




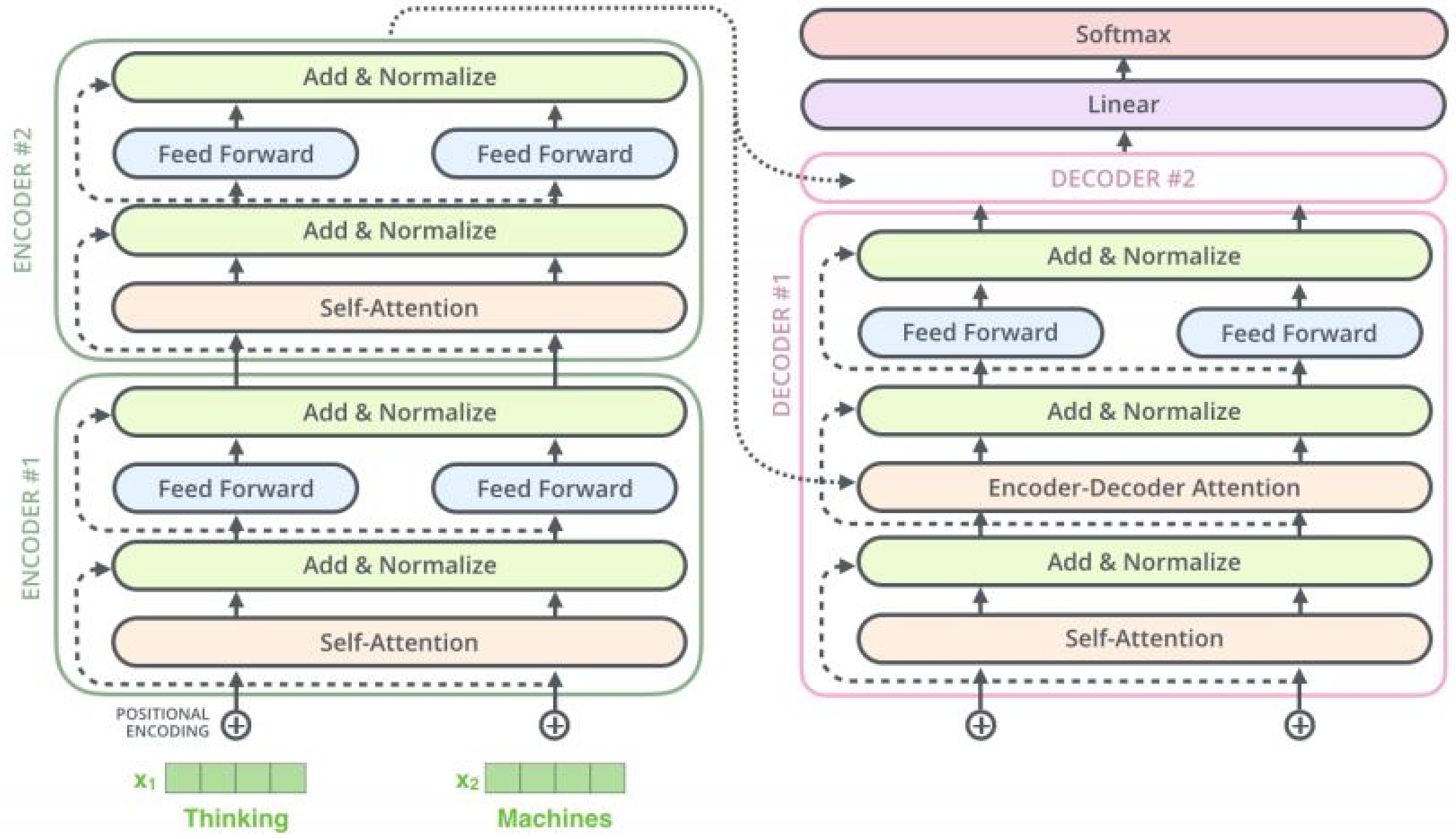
3.3 Transformer：可能是目前为止最适合NLP的模型

◆ 广义的**Transformer**指一种基于自注意力的全连接神经网络

- 核心组件
  - 自注意力（Self-Attention）
- 仅仅自注意力还不够，包括其它操作
  - 位置编码
  - 层归一化
  - 直连边
  - 逐位的FNN

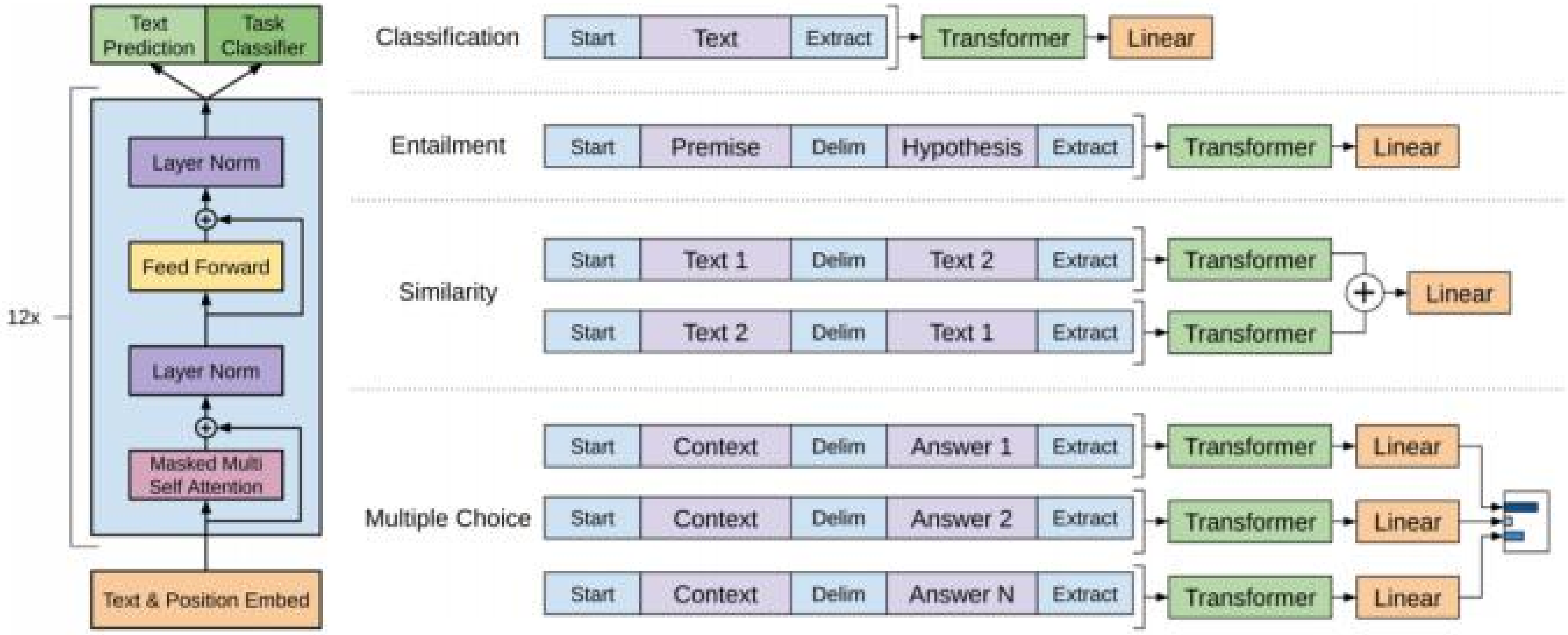


3.3 Transformer完整结构



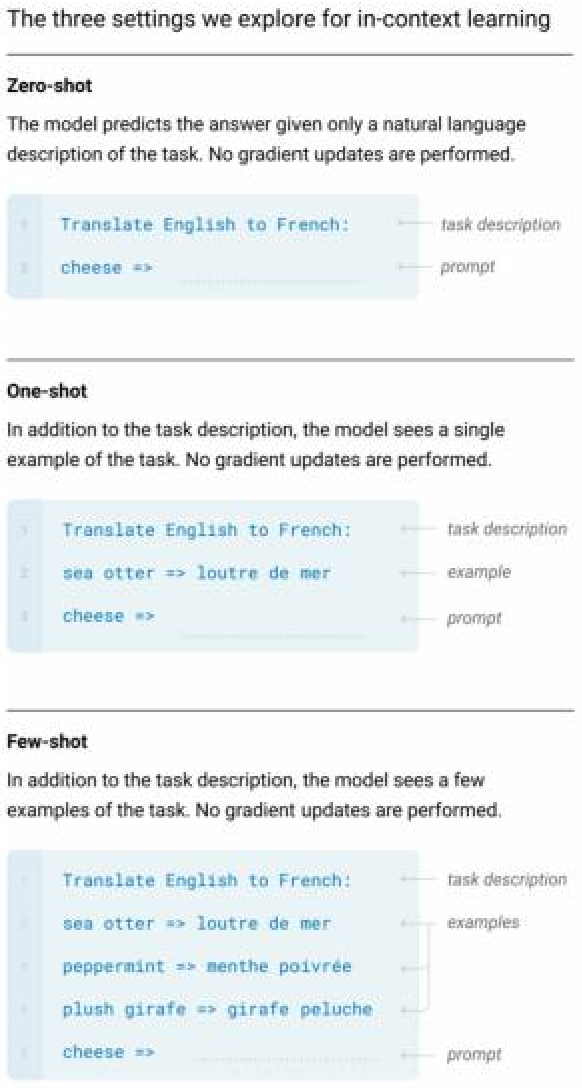
3.4 OpenAI GPT: Generative Pre-Training

- ✓ BPE tokens: 7,000
- ✓ Embedding size: 512
- ✓ Transformer layers: 12
- ✓ Attention heads: 12
- ✓ Attention hidden states: 768
- ✓ FFN hidden states : 3072

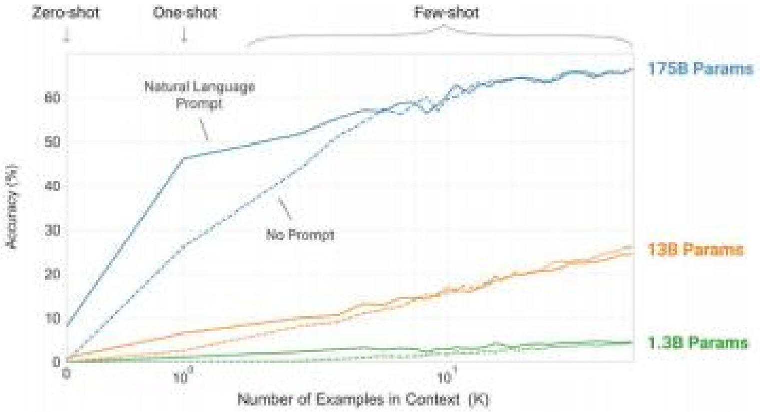


Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training. 2018.

3.4 GPT-3: Language Models are Few-Shot Learners

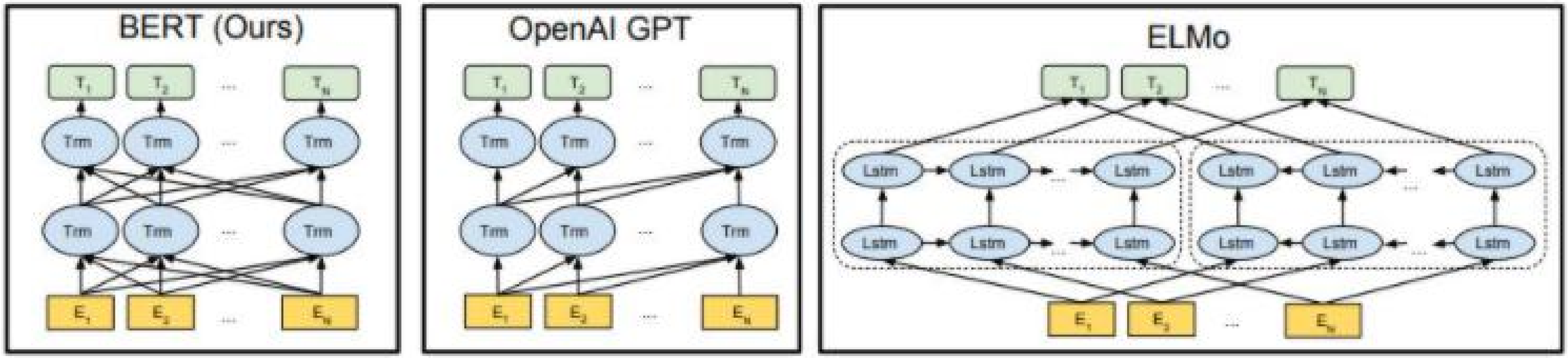


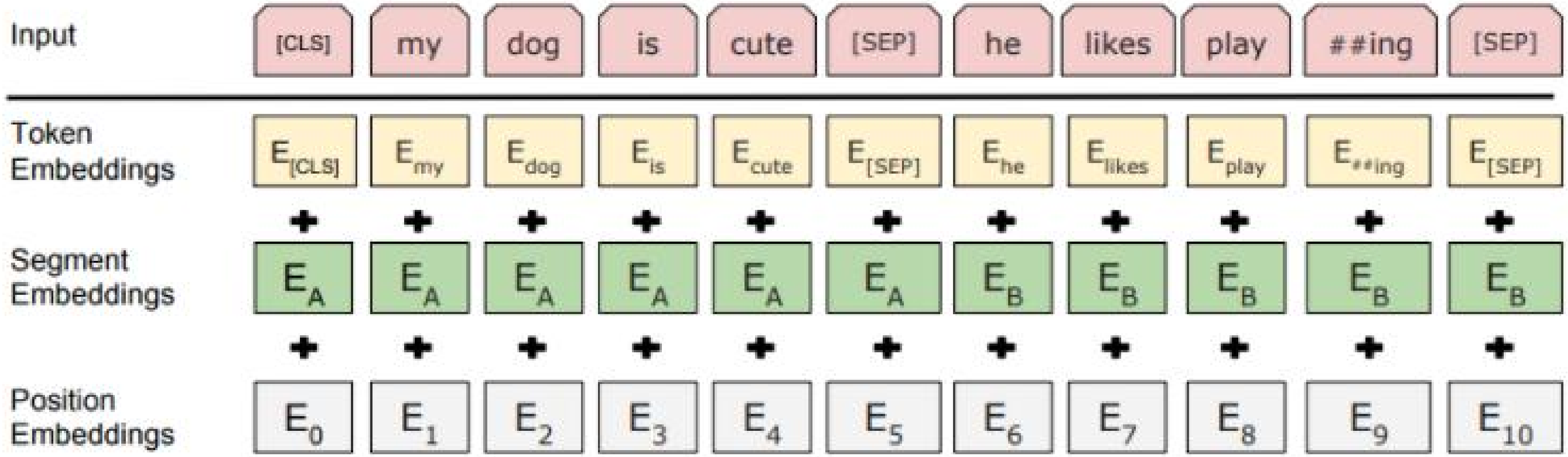
Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4



3. Main models

3.5 Bert





3.5 Bert

Task1. Masked Language Model

- 80%: `my dog is hairy -> my dog is [mask]`
- 10%: `my dog is hairy -> my dog is apple`
- 10%: `my dog is hairy -> my dog is hairy`

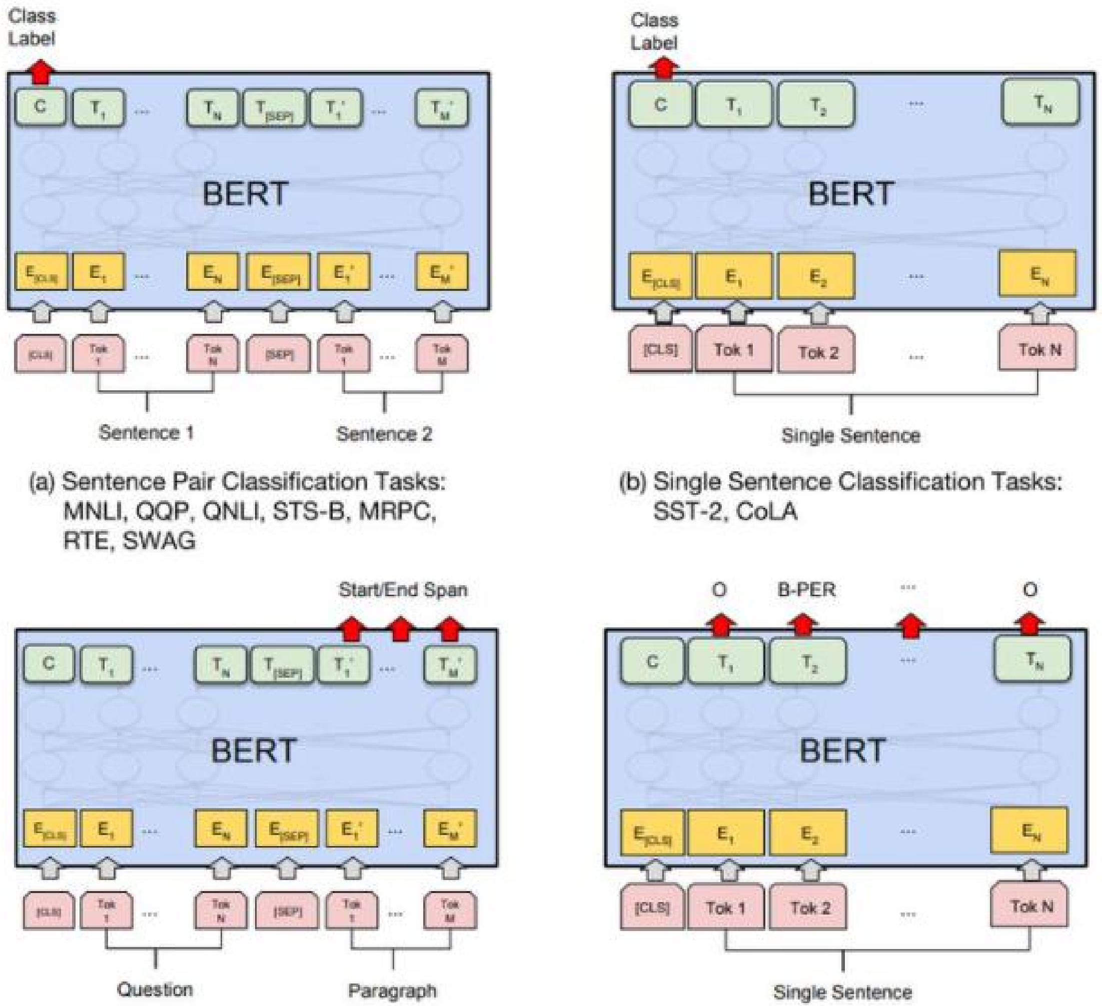
Task2. Next Sentence Prediction

目的是让模型理解两个句子之间的联系。训练的输入是句子A和B，B有一半的几率是A的下一句，输入这两个句子，模型预测B是不是A的下一句。

3. Main models

3.5 Bert

- ◆ 单个序列文本分类任务(SST-2, CoLA)
- ◆ 两个序列文本分类任务(MNLI, QQP, QNLI, STS-B, MRPC, RTE)
- ◆ 阅读理解任务(SQuAD)
- ◆ 序列标注任务(CoNLL-2003 NER)



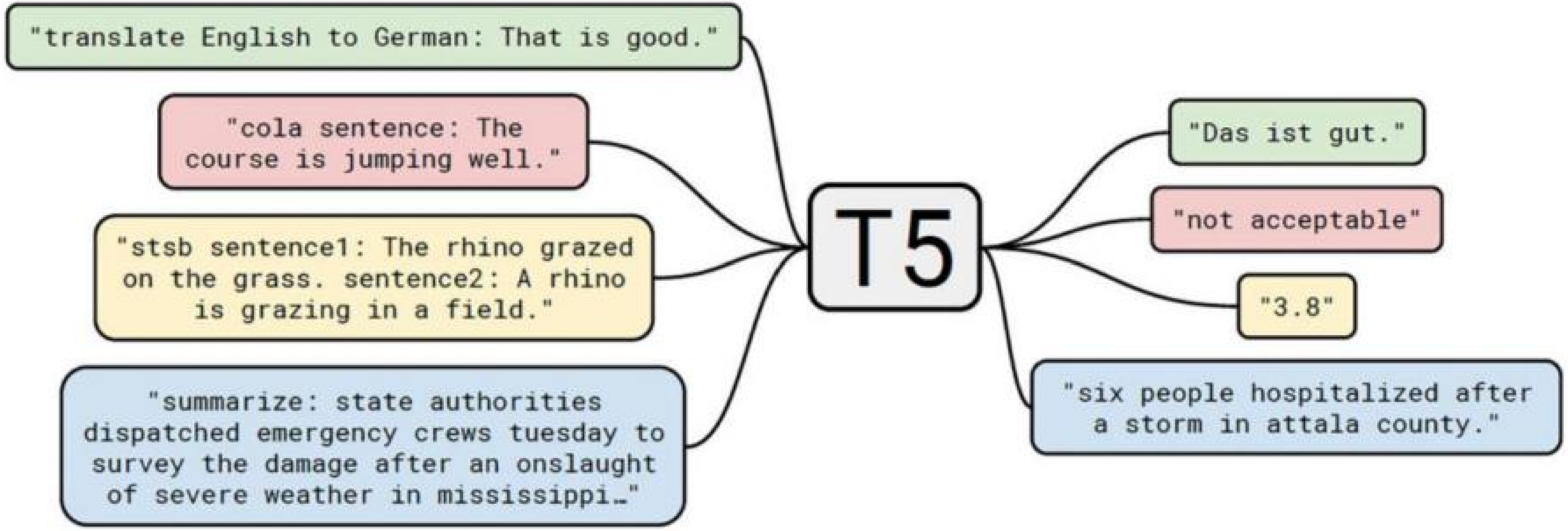
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

(b) Single Sentence Classification Tasks: SST-2, CoLA

(c) Question Answering Tasks: SQuAD v1.1

(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

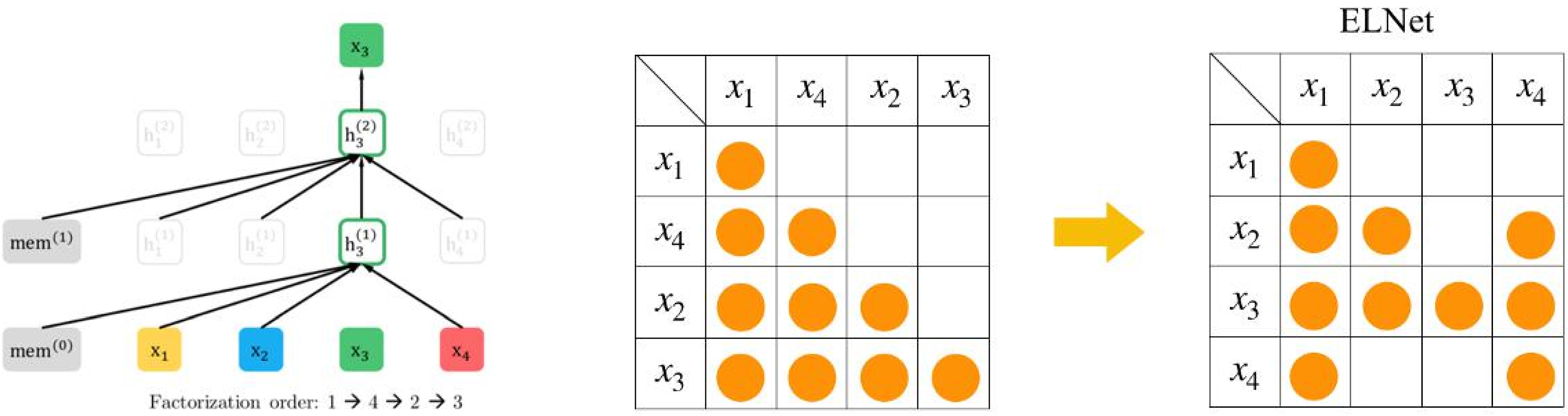
3.6 T5——Seq2Seq Masked Language Modeling



Text-to-Text Transfer Transformer (T5)

Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer, 2019. https://arxiv.org/abs/1910.10683

3.7 XLNet——Permutation Language Modeling



Yang Z, Dai Z, Yang Y, et al. Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems. 2019.

3.7 XLNet

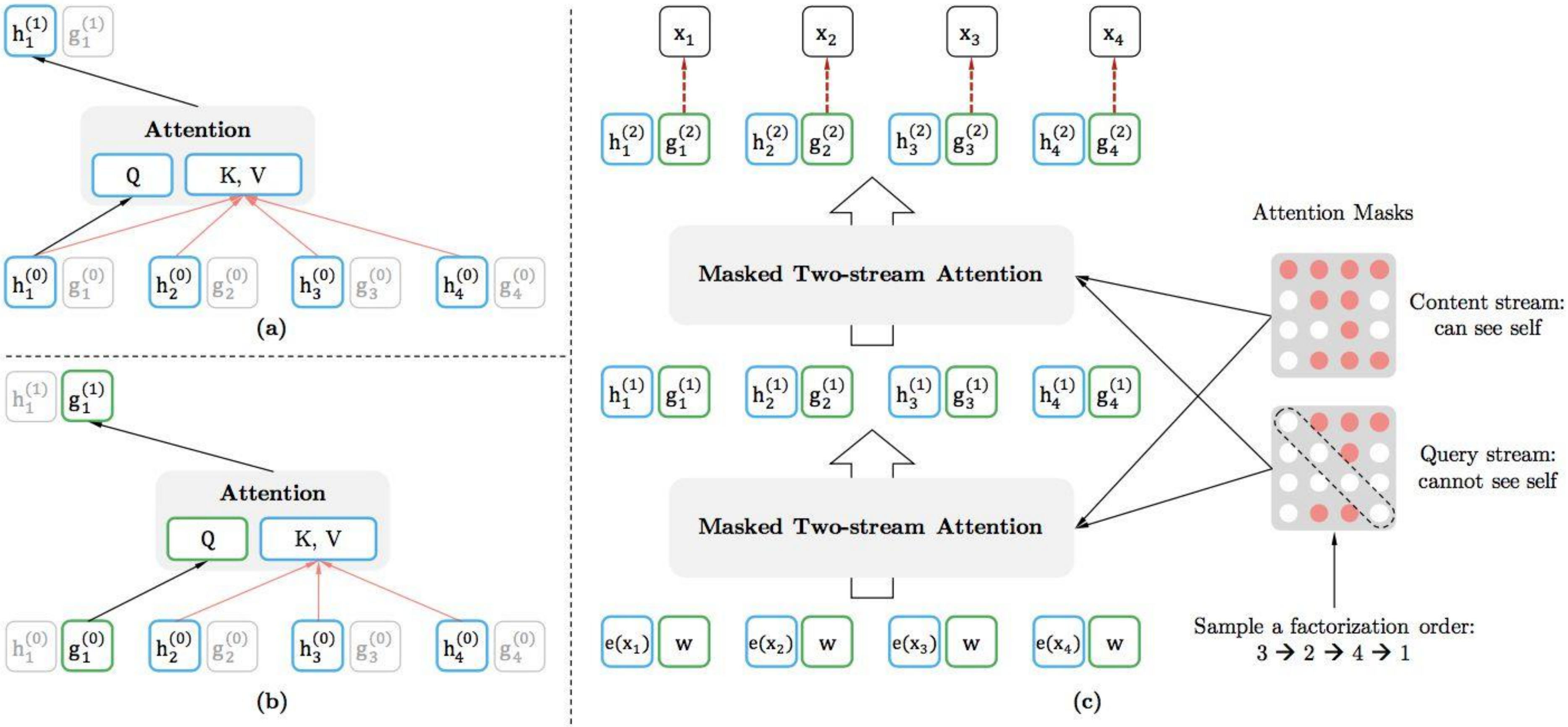


Figure 2: (a): Content stream attention, which is the same as the standard self-attention. (b): Query stream attention, which does not have access information about the content x_{z_t}. (c): Overview of the permutation language modeling training with two-stream attention.

Yang Z, Dai Z, Yang Y, et al. Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems. 2019.

3. Main models

3.7 XLNet

阅读理解（RACE）：性能比较

Bert效果

RACE	Accuracy	Middle	High
GPT [25]	59.0	62.9	57.4
BERT [22]	72.0	76.6	70.1
BERT+OCN* [28]	73.5	78.4	71.5
BERT+DCMN* [39]	74.1	79.5	71.8
XLNet	**81.75**	**85.45**	**80.21**

XLNet效果：效果提升明显

阅读理解（SQuAD）：性能比较

较长文档：效果提升明显

SQuAD1.1	EM	F1	SQuAD2.0	EM	F1
Dev set results without data augmentation					
BERT [10]	84.1	90.9	BERT† [10]	78.98	81.77
XLNet	**88.95**	**94.52**	XLNet	**86.12**	**88.79**
Test set results on leaderboard, with data augmentation (as of June 19, 2019)					
Human [27]	82.30	91.22	BERT+N-Gram+Self-Training [10]	85.15	87.72
ATB	86.94	92.64	SG-Net	85.23	87.93
BERT* [10]	87.43	93.16	BERT+DAE+AoA	85.88	88.62
XLNet	**89.90**	**95.08**	XLNet	**86.35**	**89.13**

综合NLP任务（GLUE）：性能比较

重点看这组数据：效果也有提升

Model	MNLI	QNLI	QQP	RTE	SST-2	MRPC	CoLA	STS-B	WNLI
Single-task single models on dev									
BERT [2]	86.6/-	92.3	91.3	70.4	93.2	88.0	60.6	90.0	-
XLNet	**89.8/-**	**93.9**	**91.8**	**83.8**	**95.6**	**89.2**	**63.6**	**91.8**	-
Single-task single models on test									
BERT [10]	86.7/85.9	91.1	89.3	70.1	94.9	89.3	60.5	87.6	65.1
Multi-task ensembles on test (from leaderboard as of June 19, 2019)									
Snorkel* [29]	87.6/87.2	93.9	89.9	80.9	96.2	91.5	63.8	90.1	65.1
ALICE*	88.2/87.9	95.7	**90.7**	83.5	95.2	92.6	**68.6**	91.1	80.8
MT-DNN* [18]	87.9/87.4	96.0	89.9	**86.3**	96.5	92.7	68.4	91.1	89.0
XLNet*	**90.2/89.7**†	**98.6**†	90.3†	**86.3**	**96.8**†	**93.0**	67.8	**91.6**	**90.4**

3.7 XLNet

文本分类任务：性能比较

Model	IMDB	Yelp-2	Yelp-5	DBpedia	AG	Amazon-2	Amazon-5
CNN [14]	-	2.90	32.39	0.84	6.57	3.79	36.24
DPCNN [14]	-	2.64	30.58	0.88	6.87	3.32	34.81
Mixed VAT [30, 20]	4.32	-	-	0.70	4.95	-	-
ULMFiT [13]	4.6	2.16	29.98	0.80	5.01	-	-
BERT [35]	4.51	1.89	29.32	0.64	-	2.63	34.17
XLNet	**3.79**	**1.55**	**27.80**	**0.62**	**4.49**	**2.40**	**32.26**

重点看这组数据：效果有提升，幅度不算大

信息检索任务：性能比较

Model	NDCG@20	ERR@20
DRMM [12]	24.3	13.8
KNRM [8]	26.9	14.9
Conv [8]	28.7	18.1
BERT†	30.53	18.67
XLNet	**31.10**	**20.28**

重点看这组数据：效果有提升，幅度不算大

3.7 XLNet

1. 与Bert采取De-noising Autoencoder方式不同的新的预训练目标：Permutation Language Model(简称PLM)；打开了NLP中两阶段模式潮流的一个新思路。

2. 引入了Transformer-XL的主要思路：相对位置编码以及分段RNN机制。实践已经证明这两点对于长文档任务是很有帮助的；

3. 加大增加了预训练阶段使用的数据规模；Bert使用的预训练数据是BooksCorpus和英文Wiki数据，大小13G。XLNet除了使用这些数据外，另外引入了Giga5，ClueWeb以及Common Crawl数据，并排掉了其中的一些低质量数据，大小分别是16G,19G和78G。

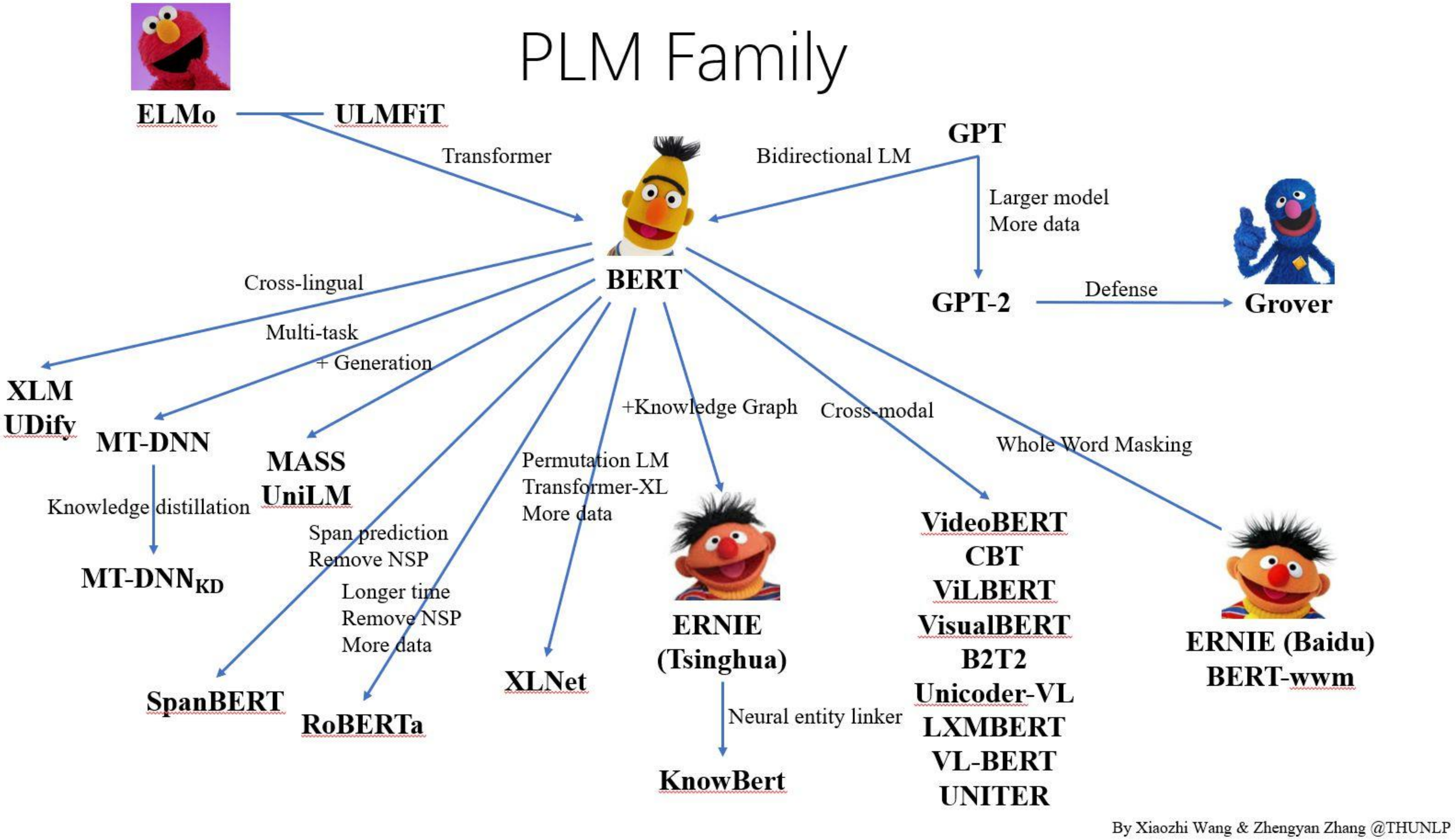
XLNet与Bert纯粹模型比较：性能比较

#	Model	RACE	SQuAD2.0 F1	SQuAD2.0 EM	MNLI m/mm	SST-2
1	BERT-Base	64.3	76.30	73.66	84.34/84.65	92.78
2	DAE + Transformer-XL	65.03	79.56	76.80	84.88/84.45	92.60
3	XLNet-Base ($K = 7$)	66.05	**81.33**	**78.46**	**85.84/85.43**	92.66
4	XLNet-Base ($K = 6$)	66.66	80.98	78.18	85.63/85.12	**93.35**
5	- memory	65.55	80.15	77.27	85.32/85.05	92.78
6	- span-based pred	65.95	80.61	77.91	85.49/85.02	93.12
7	- bidirectional data	66.34	80.65	77.87	85.31/84.99	92.66
8	+ next-sent pred	**66.76**	79.83	76.94	85.32/85.09	92.89

与Bert相比：长文档阅读理解提升幅度大，其它任务还好

哈工大讯飞联合实验室（HFL）——中文XLNet：https://github.com/ymcui/Chinese-XLNet

3.8 小结


PLM Family
ELMo
ULMFiT
Transformer
BERT
Bidirectional LM
GPT
Larger model
More data
GPT-2
Defense
Grover
Cross-lingual
XLM
UDify
Multi-task
MT-DNN
Knowledge distillation
MT-DNN$_{KD}$
+ Generation
MASS
UniLM
Span prediction
Remove NSP
SpanBERT
Longer time
Remove NSP
More data
RoBERTa
Permutation LM
Transformer-XL
More data
XLNet
+Knowledge Graph
ERNIE
(Tsinghua)
Neural entity linker
KnowBert
Cross-modal
VideoBERT
CBT
ViLBERT
VisualBERT
B2T2
Unicoder-VL
LXMBERT
VL-BERT
UNITER
Whole Word Masking
ERNIE (Baidu)
BERT-wwm
By Xiaozhi Wang & Zhengyan Zhang @THUNLP

3. Main models

3.8 小结

PTMs	Architecture†	Input	Pre-Training Task	Corpus	Params	GLUE‡	FT?♯
ELMo [14]	LSTM	Text	BiLM	WikiText-103			No
GPT [15]	Transformer Dec.	Text	LM	BookCorpus	117M	72.8	Yes
GPT-2 [58]	Transformer Dec.	Text	LM	WebText	117M ~ 1542M		No
BERT [16]	Transformer Enc.	Text	MLM & NSP	WikiEn+BookCorpus	110M ~ 340M	81.9*	Yes
InfoWord [55]	Transformer Enc.	Text	DIM+MLM	WikiEn+BookCorpus	=BERT	81.1*	Yes
RoBERTa [43]	Transformer Enc.	Text	MLM	BookCorpus+CC-News+OpenWebText+ STORIES	355M	88.5	Yes
XLNet [49]	Two-Stream Transformer Enc.	Text	PLM	WikiEn+ BookCorpus+Giga5 +ClueWeb+Common Crawl	≈BERT	90.5§	Yes
ELECTRA [56]	Transformer Enc.	Text	RTD+MLM	same to XLNet	335M	88.6	Yes
UniLM [44]	Transformer Enc.	Text	MLM+ NSP	WikiEn+BookCorpus	340M	80.8	Yes
MASS [41]	Transformer	Text	Seq2Seq MLM	*Task-dependent			Yes
BART [50]	Transformer	Text	DAE	same to RoBERTa	110% of BERT	88.4*	Yes
T5 [42]	Transformer	Text	Seq2Seq MLM	Colossal Clean Crawled Corpus (C4)	220M ~ 11B	89.7*	Yes
ERNIE(THU) [76]	Transformer Enc.	Text+Entities	MLM+NSP+dEA	WikiEn + Wikidata	114M	79.6	Yes
KnowBERT [77]	Transformer Enc.	Text	MLM+NSP+EL	WikiEn + WordNet/Wiki	253M ~ 523M		Yes
K-BERT [78]	Transformer Enc.	Text+Triples	MLM+NSP	WikiZh + WebtextZh + CN-DBpedia + HowNet + MedicalKG	=BERT		Yes
KEPLER [80]	Transformer Enc.	Text	MLM+KE	WikiEn + Wikidata/WordNet			Yes
WKLM [57]	Transformer Enc.	Text	MLM+ERD	WikiEn + Wikidata	=BERT		Yes
CoLAKE [81]	Transformer Enc.	Text+Triples	MLM	WikiEn + Wikidata	=RoBERTa	86.3	Yes

4. Prospect

4.1如何对预训练模型进行迁移学习?

1）选择合适的预训练任务:

语言模型PTM是最为流行的预训练任务；预训练任务有其自身的偏置，并且对不同的任务会产生不同的效果。例如，NSP任务可以使诸如问答（QA）和自然语言推论（NLI）之类的下游任务受益。

2）选择合适的模型架构:

例如BERT采用的MLM策略和Transformer-Encoder结构，导致其不适合直接处理生成任务。

3）选择合适的数据:

下游任务的数据应该近似于PTMs的预训练任务，现在已有有很多现成的PTMs可以方便地用于各种特定领域或特定语言的下游任务。

4）选择合适的layers进行transfer:

主要包括Embedding迁移、top layer迁移和all layer迁移。如word2vec和Glove可采用Embedding迁移，BERT可采用top layer迁移，Elmo可采用all layer迁移。

5）特征集成还是fine-tune?

对于特征集成预训练参数是freeze的，而fine-tune是unfreeze的。特征集成方式却需要特定任务的体系结构，fine-tune方法通常比特征提取方法更为通用和方便。

4.2 预训练模型还有哪些问题需要解决?

1、PTMs的上限

大多数的PTMs可通过使用更长训练步长和更大数据集来提升其性能。

2、面向任务的预训练和模型压缩

在实践中，不同的目标任务需要 PTMs拥有不同功能。而 PTMs与下游目标任务间的差异通常在于两方面：模型架构与数据分布。

3、PTMs的架构设计

对于PTMs，Transformer 已经被证实是一个高效的架构。然而 Transformer 最大的局限在于其计算复杂度（输入序列长度的平方倍）。

4、finetune中的知识迁移

finetune是目前将 PTM 的知识转移至下游任务的主要方法，但效率却很低，每个下游任务都需要有特定的finetune参数。

5、PTMs 的解释性与可靠性

PTMs 的可解释性与可靠性仍然需要从各个方面去探索，它能够帮助我们理解 PTM 的工作机制，为更好的使用及性能改进提供指引。

Pre-trained Models for Natural Language Processing: A Survey

自然语言处理预训练模型综述

Thanks!