2020年度暑期强化课程

预训练模型：过去、现在与未来

授课人：曹亚男

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

6. 预训练模型：过去、现在与未来

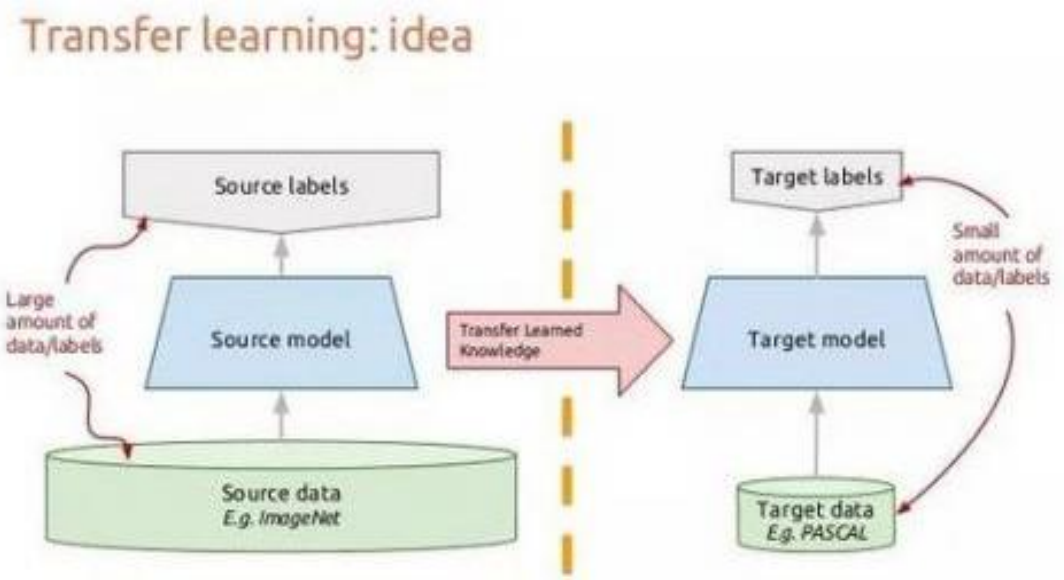
DL4NLP核心问题

如何在有限的人工标注数据集下训练有效的深度神经网络模型？

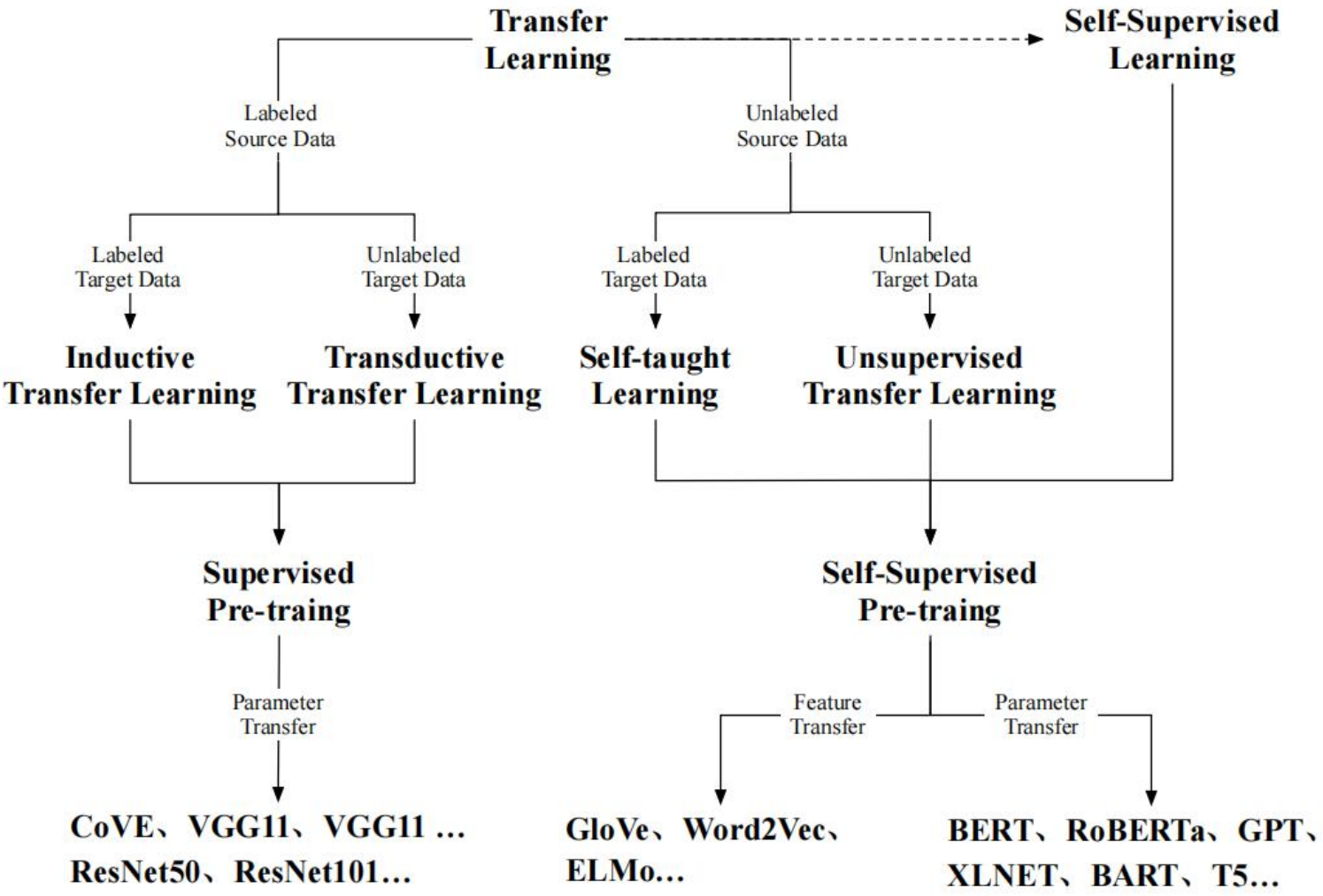
迁移学习

迁移学习

- 源任务和目标任务具有完全不同的数据领域和任务设置，但是处理任务所需的知识是一致的
- 迁移学习有两种常用的预训练方法，包括特征迁移和参数迁移
  - 特征迁移是在预训练阶段训练有效的特征表示（跨领域和任务地预编码知识），然后在目标任务使用（例如word2vec\elmo\glove等）
  - 参数迁移假设源任务和目标任务可以共享模型参数或超参数的先验分布。在预训练阶段将知识编码进共享的模型参数中，在目标领域微调参数（Bert、 RoBERTa、GPT、XLNET、T5）

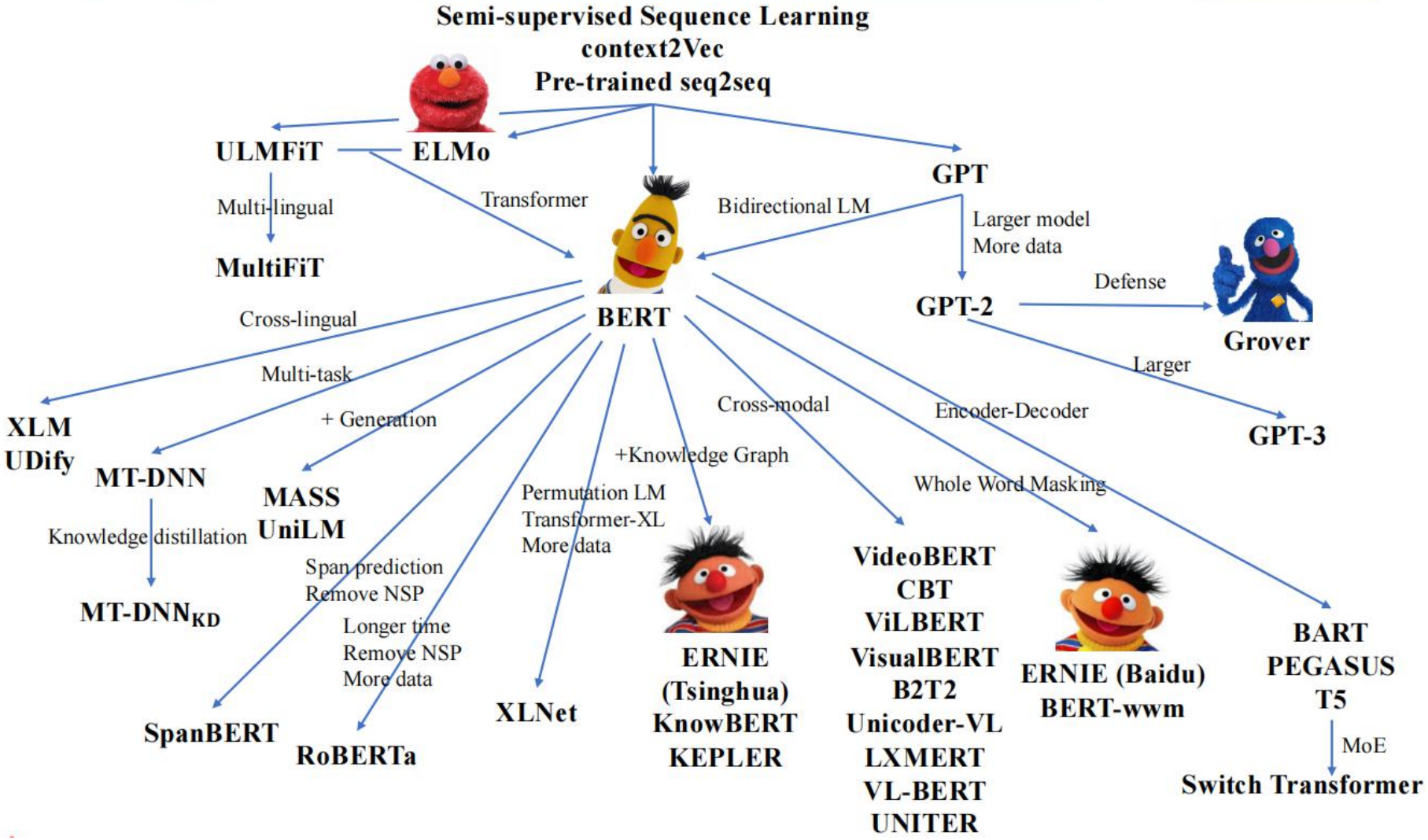


预训练模型 & 迁移学习

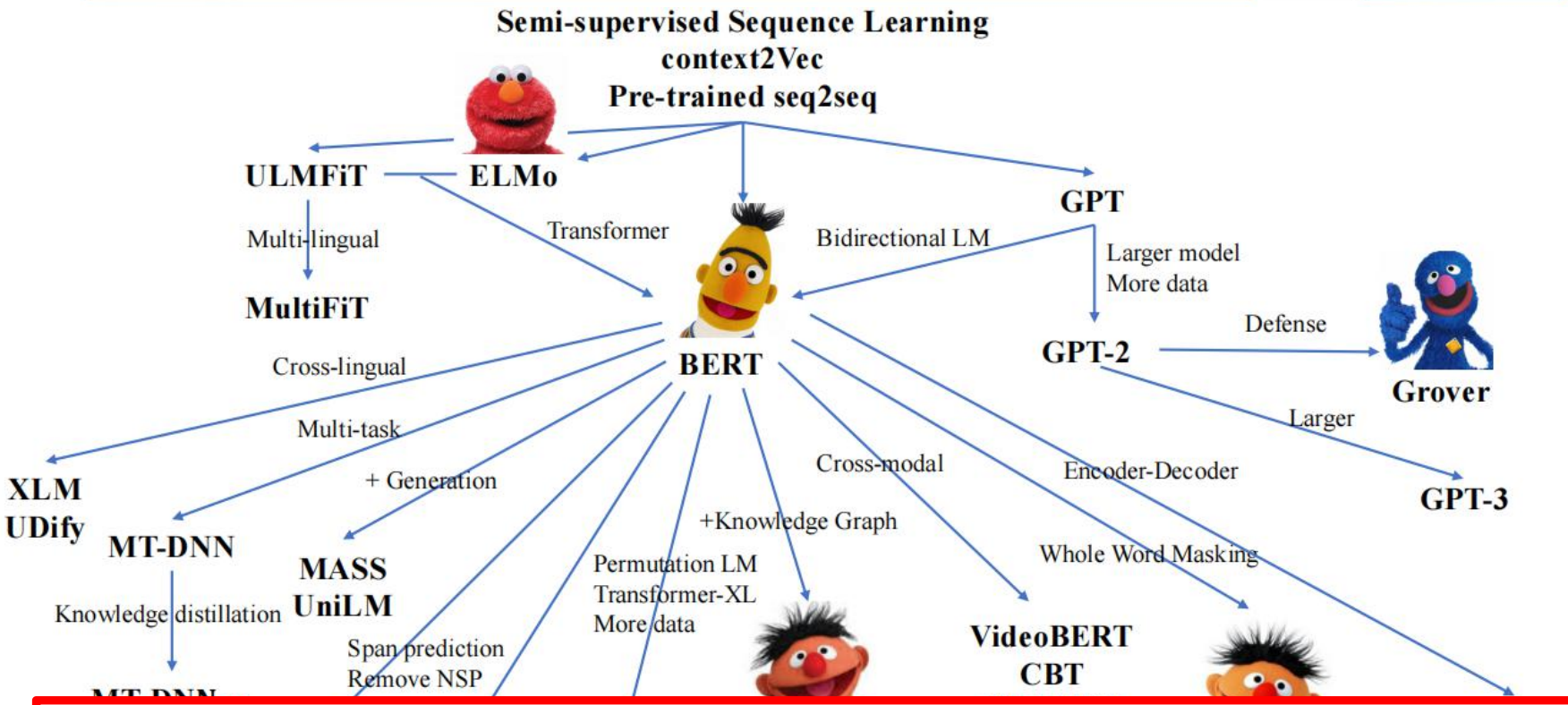

Transfer Learning
Self-Supervised Learning
Labeled Source Data
Unlabeled Source Data
Labeled Target Data
Unlabeled Target Data
Labeled Target Data
Unlabeled Target Data
Inductive Transfer Learning
Transductive Transfer Learning
Self-taught Learning
Unsupervised Transfer Learning
Supervised Pre-traing
Self-Supervised Pre-traing
Parameter Transfer
Feature Transfer
Parameter Transfer
CoVE、VGG11、VGG11 ...
ResNet50、ResNet101...
GloVe、Word2Vec、
ELMo...
BERT、RoBERTa、GPT、
XLNET、BART、T5...

6. 预训练模型：过去、现在与未来

预训练模型家族


Semi-supervised Sequence Learning
context2Vec
Pre-trained seq2seq
ULMFiT
ELMo
GPT
Multi-lingual
MultiFiT
Transformer
Bidirectional LM
Larger model
More data
Defense
GPT-2
Grover
Cross-lingual
BERT
Multi-task
+ Generation
Larger
GPT-3
XLM
UDify
MT-DNN
MASS
UniLM
Cross-modal
Encoder-Decoder
+Knowledge Graph
Whole Word Masking
Permutation LM
Transformer-XL
More data
Knowledge distillation
MT-DNN$_{KD}$
Span prediction
Remove NSP
Longer time
Remove NSP
More data
SpanBERT
RoBERTa
XLNet
ERNIE
(Tsinghua)
KnowBERT
KEPLER
VideoBERT
CBT
ViLBERT
VisualBERT
B2T2
Unicoder-VL
LXMERT
VL-BERT
UNITER
ERNIE (Baidu)
BERT-wwm
BART
PEGASUS
T5
MoE
Switch Transformer

预训练模型家族



- 大规模预训练模型的效果源于其精细化的预训练目标和大量的模型参数
  - 在预训练阶段，将学习到的大量知识均存储在参数中
  - 在fine-tuning阶段，将这些知识（学习到的模型）应用到下游任务中，并发挥作用

预训练模型的代表性团队

机构名称	代表人物	代表作品	目前方向
清华大学	唐杰、刘知远	CPM、ERNIE	中文大规模预训练模型，包括文本、融入知识和多模态
哈工大	崔一鸣、刘挺	中文ELECTRA等	中文特色预训练模型
百度	孙宇	ERNIE	跨语言和多模态
微软	韦福如、董力	UniLM、MiniLM、LayoutLM	模型蒸馏、多模态预训练模型
谷歌	Jacob Devlin	BERT、GPT-3、T5	超大规模预训练模型、结构创新
脸书	Yann LeCun	Roberta、SpanBERT、XLM	跨语言、任务优化
阿里巴巴	杨红霞	M6、FashionBERT	超大规模模型、多模态预训练
华为	尚利峰	Pangu、TinyBERT	超大规模预训练模型、模型蒸馏裁剪
字节跳动	李航	AMBERT	预训练模型加速
英伟达	先超	Megatron	模型加速、超大规模模型训练并行方案

改进模型结构和预训练任务

统一序列建模

认知驱动建模

新的预训练任务

- 融合自回归和自编码模型
- 使用通用编解码器

三类NLP任务

- 自然语言理解（包括句法分析、语法分析、分类、问答、常识推理等）
- 开放式文本生成（对话生成、故事生成、数据到文本生成等）
- 非开放式文本生成（机器翻译、摘要生成、完形填空）

自回归 vs. 自编码

- 自回归语言模型：

$$\max_{\theta} \quad \log p_\theta(\mathbf{x}) = \sum_{t=1}^{T} \log p_\theta(x_t \mid \mathbf{x}_{<t}) = \sum_{t=1}^{T} \log \frac{\exp\left(h_\theta(\mathbf{x}_{1:t-1})^\top e(x_t)\right)}{\sum_{x'} \exp\left(h_\theta(\mathbf{x}_{1:t-1})^\top e(x')\right)}$$

GPT、ELMo：难以有效捕获单词的上下文信息

- 自编码语言模型：

$$\max_{\theta} \quad \log p_\theta(\bar{\mathbf{x}} \mid \hat{\mathbf{x}}) \approx \sum_{t=1}^{T} m_t \log p_\theta(x_t \mid \hat{\mathbf{x}}) = \sum_{t=1}^{T} m_t \log \frac{\exp\left(H_\theta(\hat{\mathbf{x}})_t^\top e(x_t)\right)}{\sum_{x'} \exp\left(H_\theta(\hat{\mathbf{x}})_t^\top e(x')\right)}$$

BERT：训练和测试阶段不一致

自回归 vs. 自编码

- 自回归语言模型：

$$\max_{\theta} \quad \log p_\theta(\mathbf{x}) = \sum_{t=1}^{T} \log p_\theta(x_t \mid \mathbf{x}_{<t}) = \sum_{t=1}^{T} \log \frac{\exp\left(h_\theta(\mathbf{x}_{1:t-1})^\top e(x_t)\right)}{\sum_{x'} \exp\left(h_\theta(\mathbf{x}_{1:t-1})^\top e(x')\right)}$$

GPT、ELMo：难以有效捕获单词的上下文信息

- 自编码语言模型：

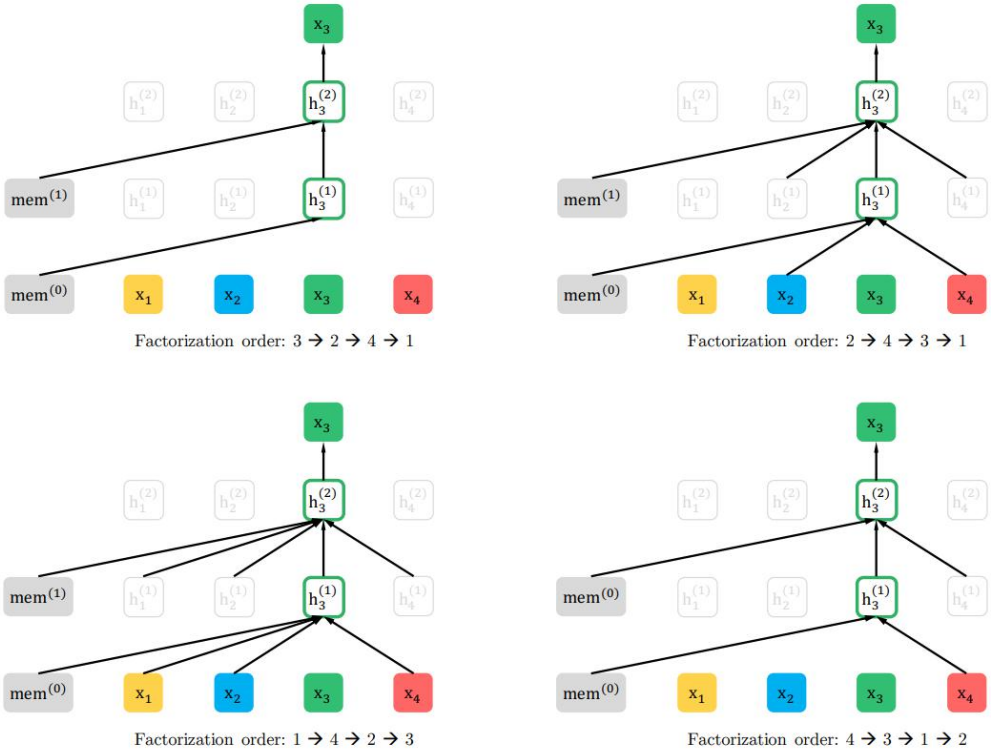
$$\max_{\theta} \quad \log p_\theta(\bar{\mathbf{x}} \mid \hat{\mathbf{x}}) \approx \sum_{t=1}^{T} m_t \log p_\theta(x_t \mid \hat{\mathbf{x}}) = \sum_{t=1}^{T} m_t \log \frac{\exp\left(H_\theta(\hat{\mathbf{x}})_t^\top e(x_t)\right)}{\sum_{x'} \exp\left(H_\theta(\hat{\mathbf{x}})_t^\top e(x')\right)}$$

自回归+双向信息？

XLNet

- 排列语言模型（Permutation LM）：考虑T!个语言模型

$$max_{\theta} E_{z \sim Z_T} [\sum_{t=1}^{T} log p_{\theta}(x_{z_t} | x_{z<t})]$$



XLNet

- 模型结构：双流自注意力机制

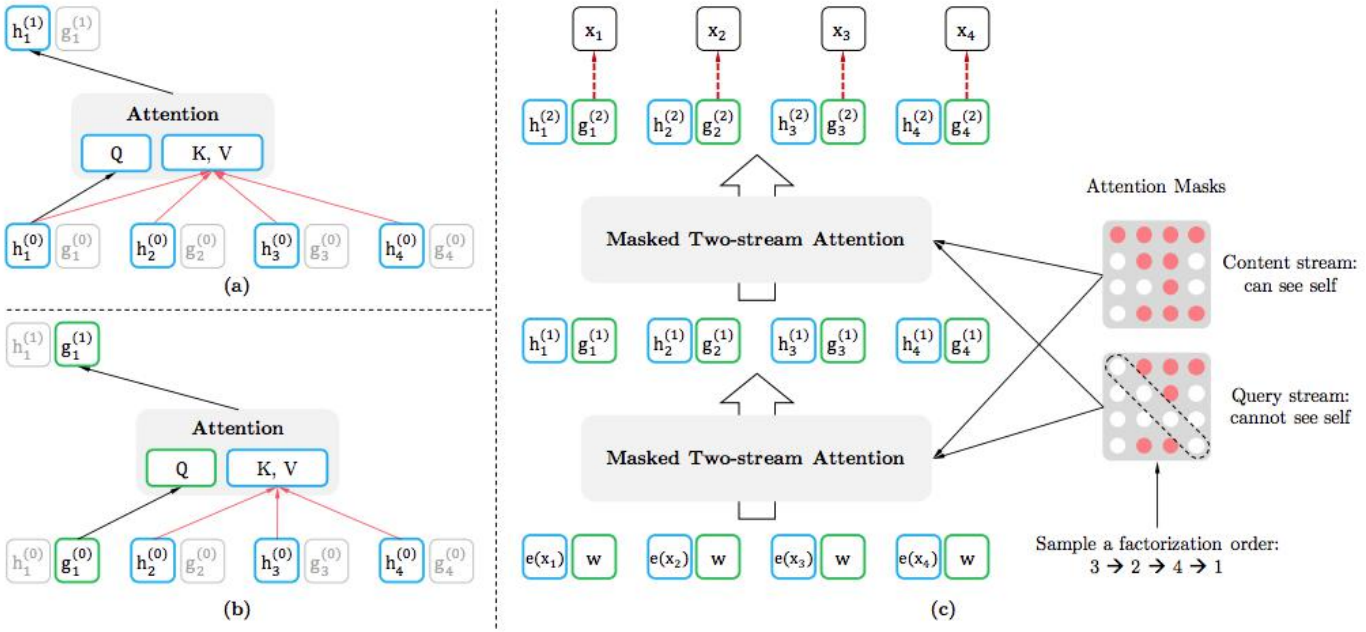
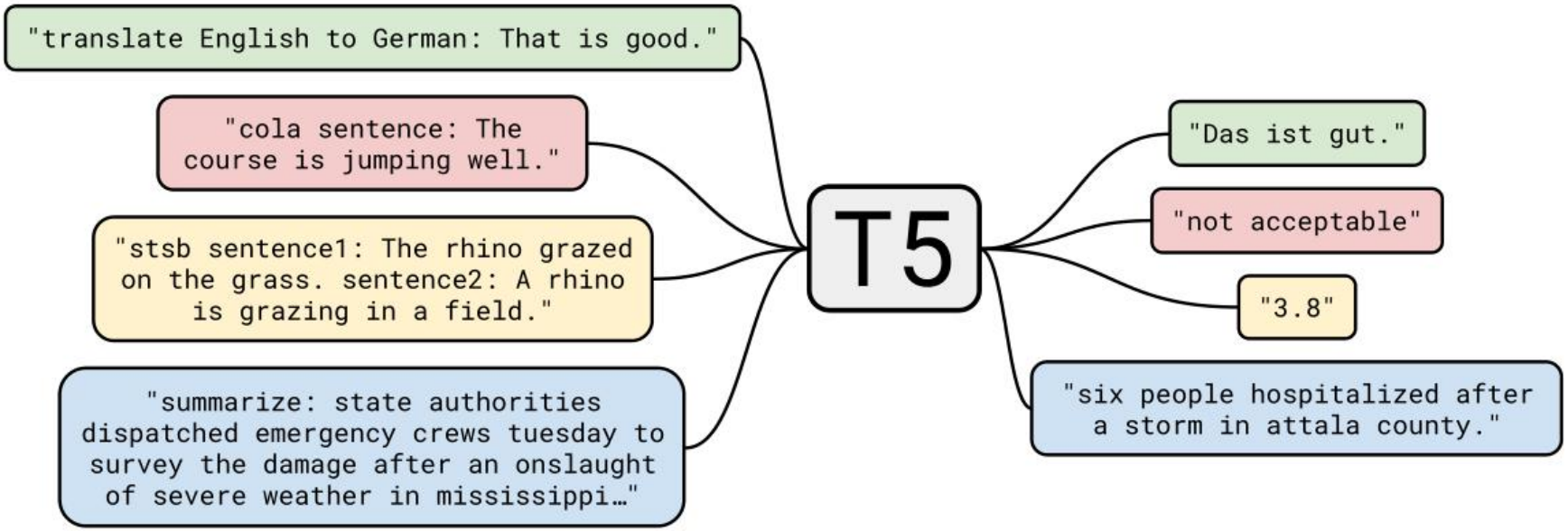


Figure 1: (a): Content stream attention, which is the same as the standard self-attention. (b): Query stream attention, which does not have access information about the content x_{z_t}. (c): Overview of the permutation language modeling training with two-stream attention.

T5

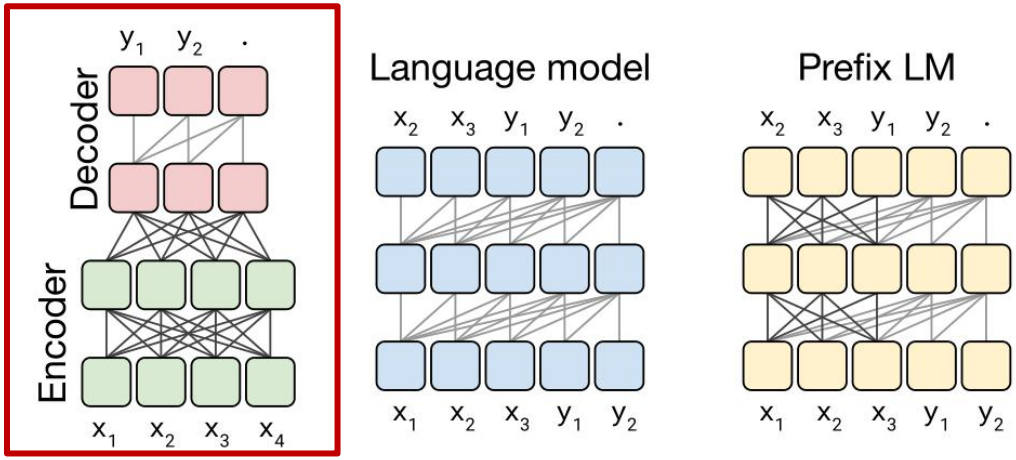
- 将所有 NLP 任务都转化成 Text-to-Text（文本到文本）任务：机器翻译、问答、摘要生成和文本分类



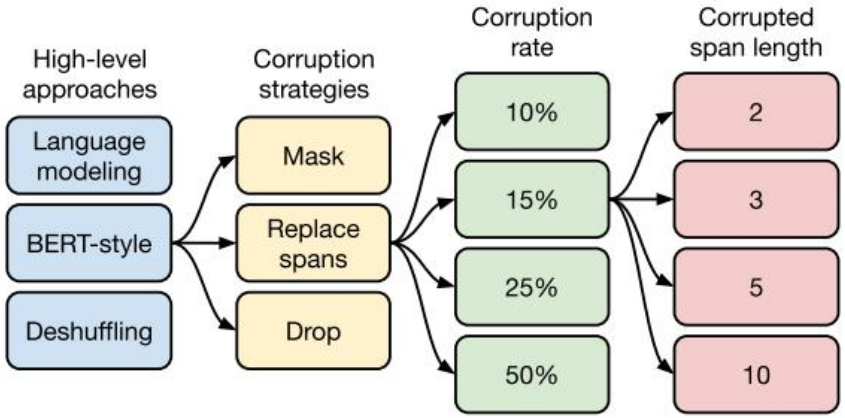
Raffel, Colin et al. “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.” ArXiv abs/1910.10683 (2020): n. pag.

T5

- 预训练模型框架选择：



- 文本恢复策略：



ERNIE（Baidu）

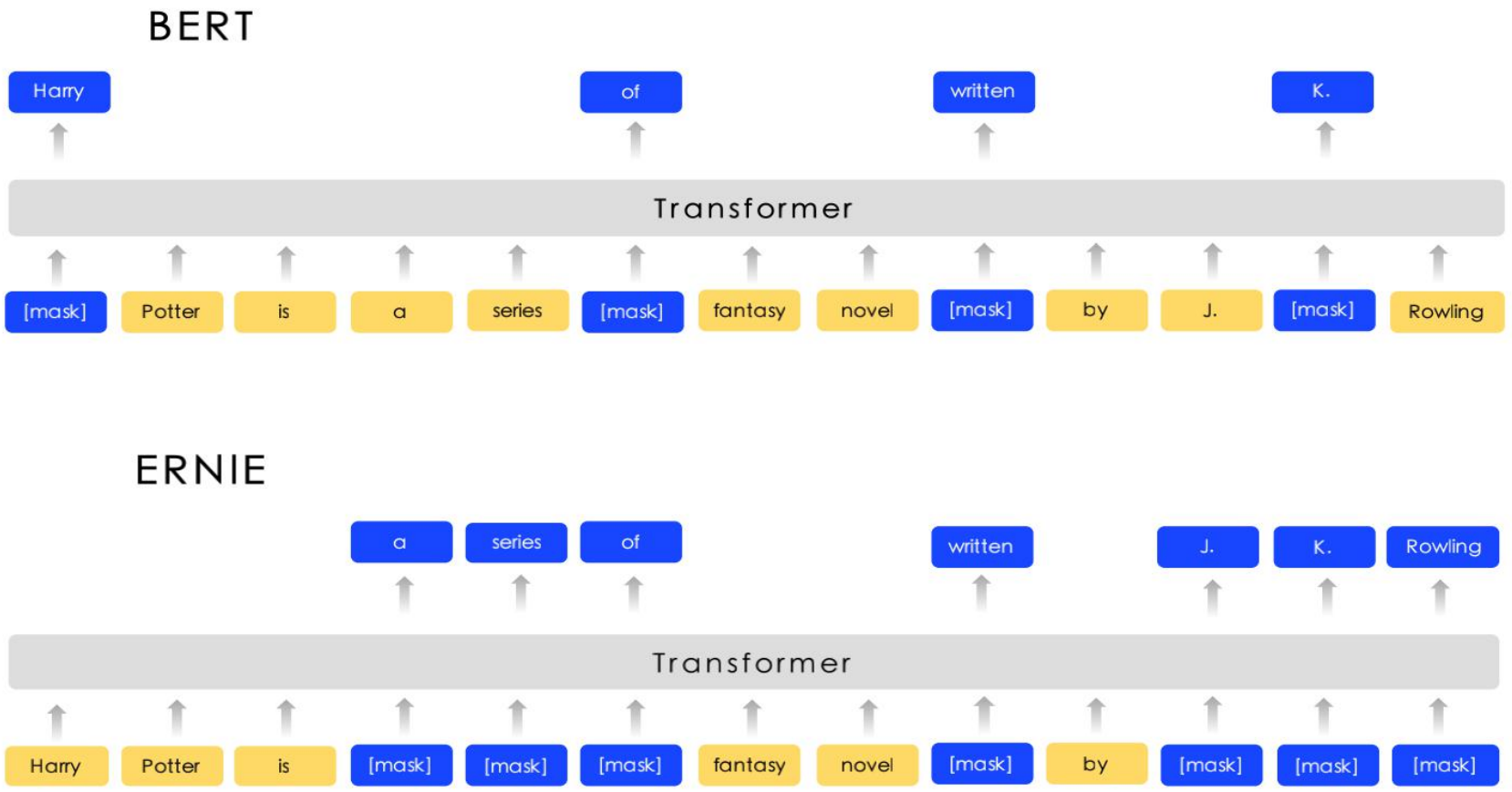


Figure 1: The different masking strategy between BERT and ERNIE

Yu Sun. et.al., ERNIE: Enhanced Representation through Knowledge Integration

ERNIE 1.0

- 两种新的 masking 策略
  - phrase-level masking：短语类如a series of, written等
  - entity-level masking（人名，位置, 组织，产品等名词，如Apple, J.K. Rowling等）

Sentence	Harry	Potter	is	a	series	of	fantasy	novels	written	by	British	author	J.	K.	Rowling
Basic-level Masking	[mask]	Potter	is	a	series	[mask]	fantasy	novels	[mask]	by	British	author	J.	[mask]	Rowling
Entity-level Masking	Harry	Potter	is	a	series	[mask]	fantasy	novels	[mask]	by	British	author	[mask]	[mask]	[mask]
Phrase-level Masking	Harry	Potter	is	[mask]	[mask]	[mask]	fantasy	novels	[mask]	by	British	author	[mask]	[mask]	[mask]

Figure 2: Different masking level of a sentence

Yu Sun. et.al., ERNIE: Enhanced Representation through Knowledge Integration

ERNIE 2.0

- 预训练连续学习(Continual Learning)
  - 用大量的数据与先验知识构建不同的预训练任务
  - 用多个预训练任务顺序更新ERNIE 模型

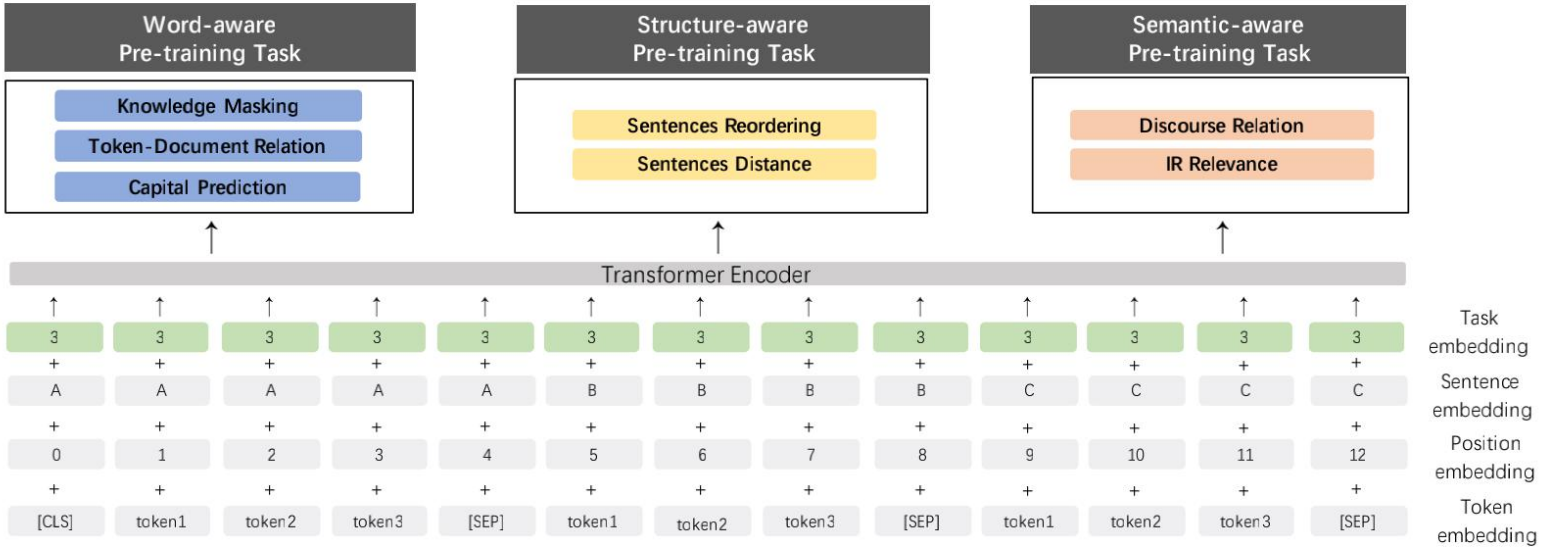


Figure 3: The structure of the ERNIE 2.0 model. The input embedding contains the token embedding, the sentence embedding, the position embedding and the task embedding. Seven pre-training tasks belonging to different kinds are constructed in the ERNIE 2.0 model.

RoBERTa

- 去掉了NSP任务
- 更多训练步骤、更大的batch size、更大规模数据；
- 更长的训练序列
- 动态修改了 [MASK]模式

Model	data	bsz	steps	SQuAD (v1.1/2.0)	MNLI-m	SST-2
RoBERTa						
with BOOKS + WIKI	16GB	8K	100K	93.6/87.3	89.0	95.3
+ additional data (§3.2)	160GB	8K	100K	94.0/87.7	89.3	95.6
+ pretrain longer	160GB	8K	300K	94.4/88.7	90.0	96.1
+ pretrain even longer	160GB	8K	500K	**94.6/89.4**	**90.2**	**96.4**
$\text{BERT}_{\text{LARGE}}$						
with BOOKS + WIKI	13GB	256	1M	90.9/81.8	86.6	93.7
$\text{XLNet}_{\text{LARGE}}$						
with BOOKS + WIKI	13GB	256	1M	94.0/87.8	88.4	94.4
+ additional data	126GB	2K	500K	94.5/88.8	89.8	95.6

Table 4: Development set results for RoBERTa as we pretrain over more data (16GB → 160GB of text) and pretrain for longer (100K → 300K → 500K steps). Each row accumulates improvements from the rows above. RoBERTa matches the architecture and training objective of $\text{BERT}_{\text{LARGE}}$. Results for $\text{BERT}_{\text{LARGE}}$ and $\text{XLNet}_{\text{LARGE}}$ are from

利用更丰富的数据资源

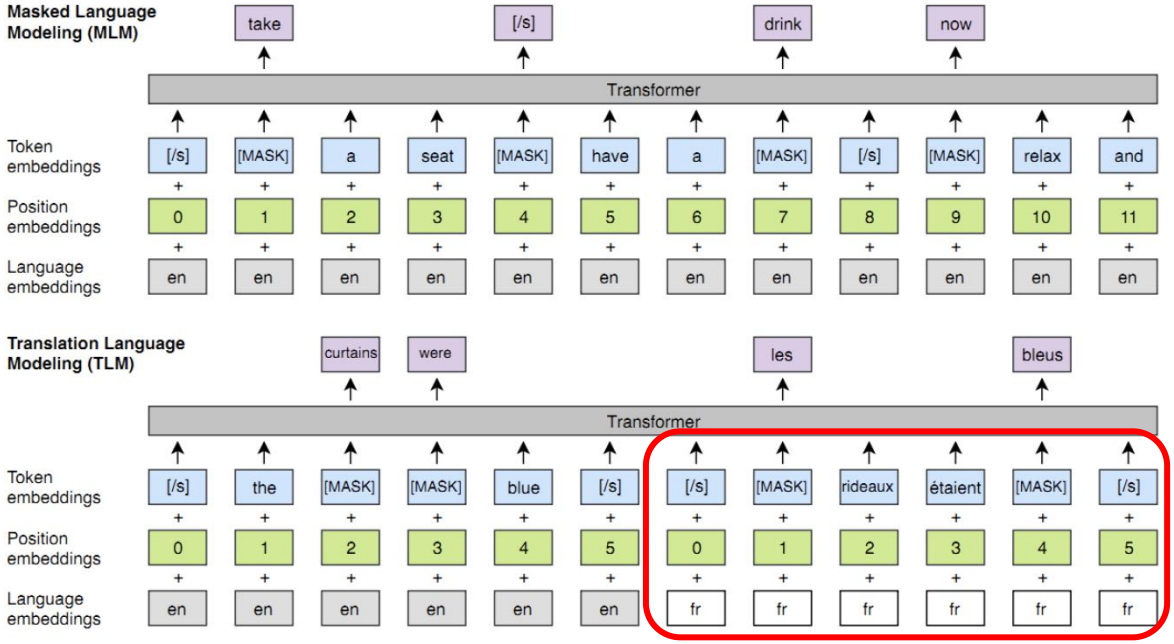
多语言语料

知识图谱

多模态数据

XLM

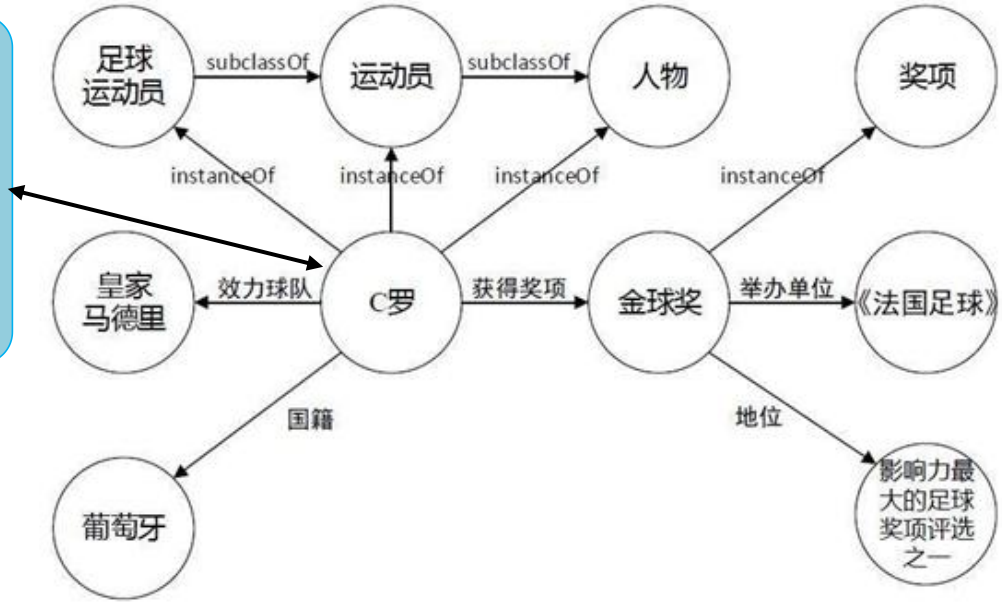
- **因果语言模型(CLM)：** 使用Transformer在给定前序词语的情况下预测下一个词的概率
- **掩蔽语言模型(MLM)：** 使用由任意数量的句子(每个句子截断为256个token)组成的文本流代替成对的句子，去除NSP任务
- **翻译语言模型(TLM)：** 直接连接平行的句子而不是只考虑单语言的文本流



什么是外部知识

- **知识图谱（Knowledge Graph，KG）**
  - 知识图谱是结构化的语义知识库，用于以符号形式描述现实世界中的概念及其相互关系。知识图谱的基本组成单位是“**实体-关系-实体**”三元组，实体间通过关系相互联结，构成有向知识图结构

C罗，效力球队，皇家马德里
- 头实体：C罗
- 关系：效力球队
- 尾实体：皇家马德里
- 起始时间：2009年6月
- 终止时间：2018年7月

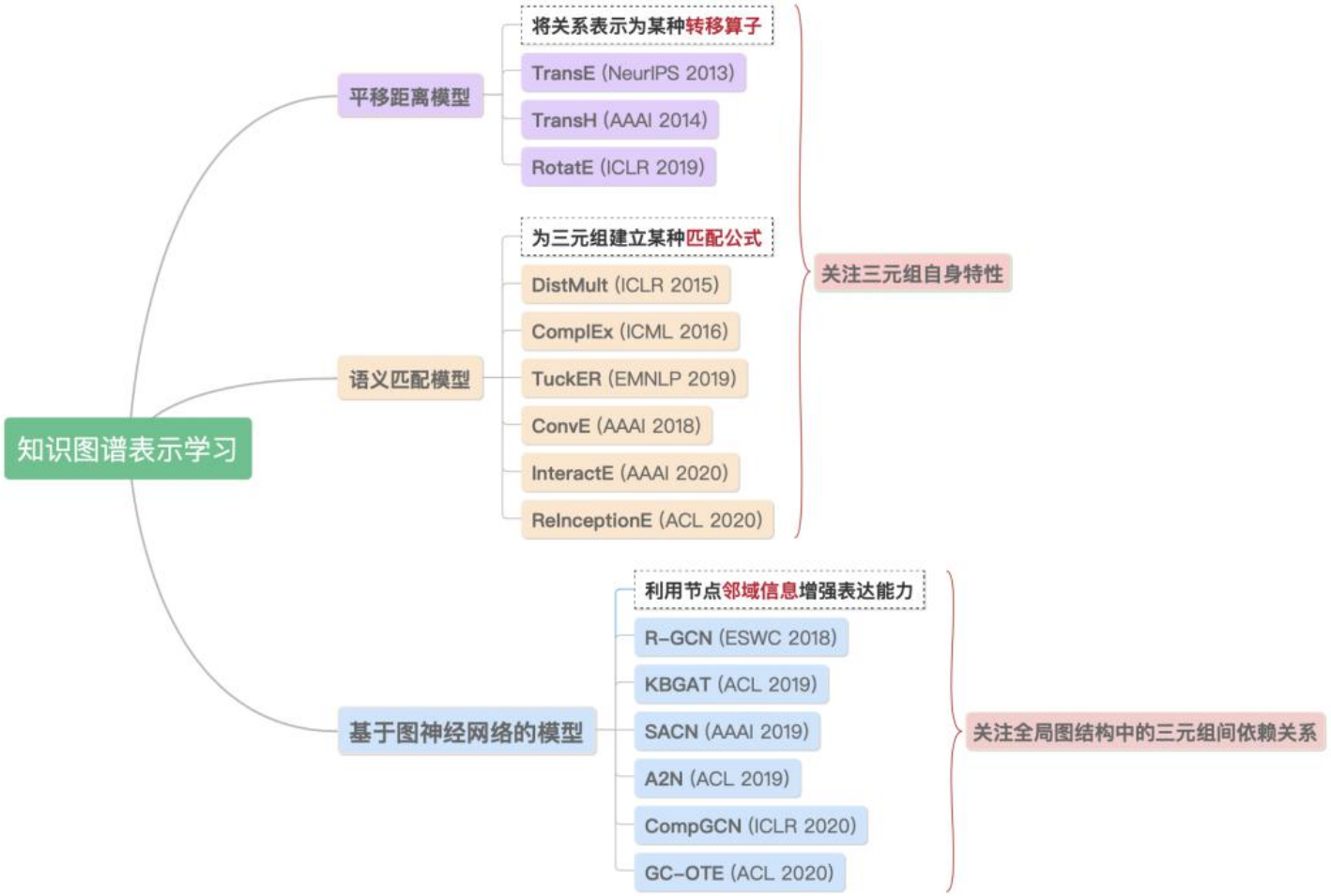


知识图谱

名称	语言	时间	机构	实体	关系	属性	领域	其他信息
CN-DBpedia	中	2017	复旦	900w	6700w	-	Open-domain	mention2entity信息110万+，摘要信息400万+，标签信息1980万+，infobox信息4100万
HowNet	中	2006	知网	6w	-	-	词典	
MedicalKG	中	-	-	-	13864	-	Medical	
OpenBase	中	-	-	1151w	10726w	9224w	Open-domain	
BigCilin	中	2016	哈工大	-	-	-	Open-domain	
Wordnet	英	2005	Princeton	15w	20w	-	词典	最常用的词典知识库，主要定义了名词、动词、形容词和副词之间的语义关系
Wikidata	英	2012	Wikipedia	9245w	-	-	Open-domain	
ConceptNet	英	1999	MIT	-	2800w	-	Common Sense	常识图谱
freeBase	英	-	MetaWeb	-	-		Open-domain	关闭，整合入Wikidata
DBPedia	英	-		-	30亿	-	Open-domain	来源于wikipedia
Microsoft ConceptGraph	英	-	MS	1255w	8760w	-	Common Sense	概念化常识知识图谱

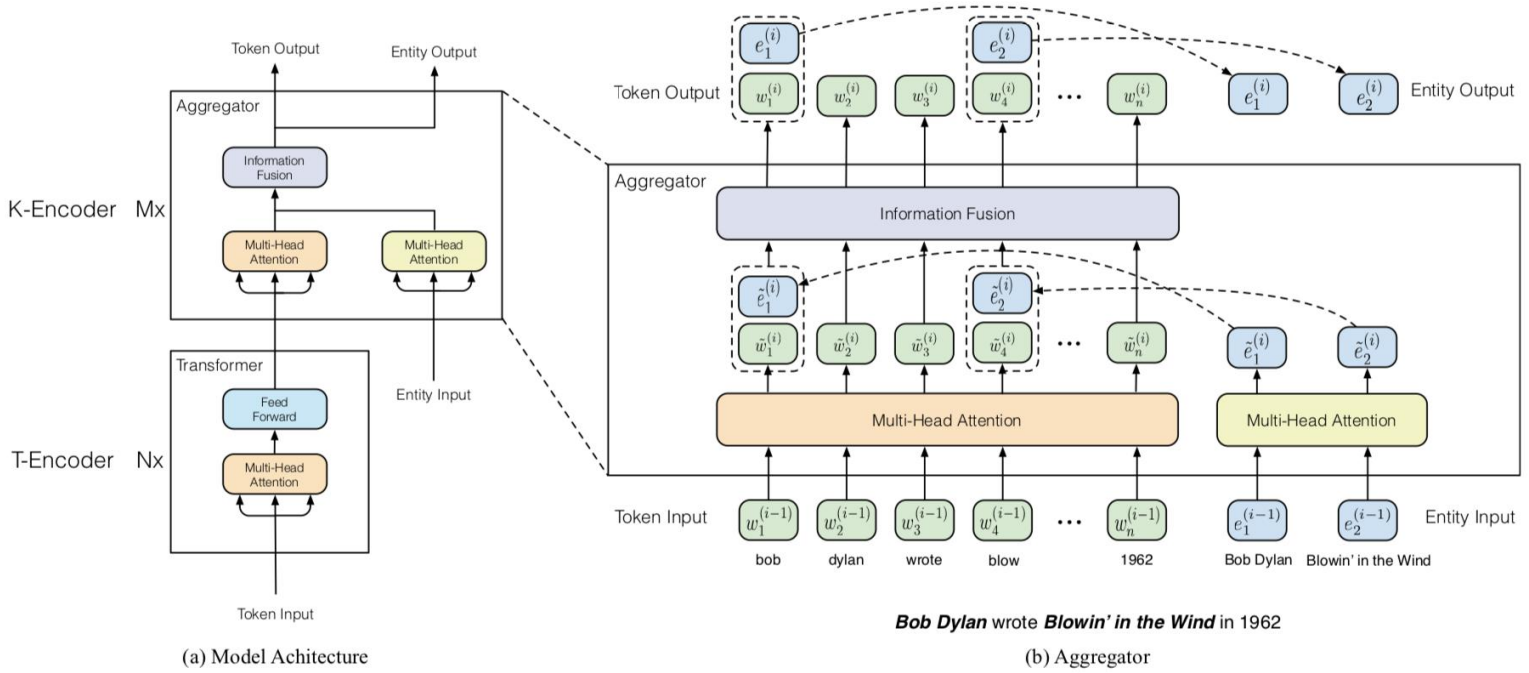
知识图谱表示学习方法

- **知识图谱表示学习**：为知识图谱中的实体、关系学习低维向量表征，以将符号化的知识融入到数值计算模型中，来支撑众多下游任务



ERNIE （Tsinghua）

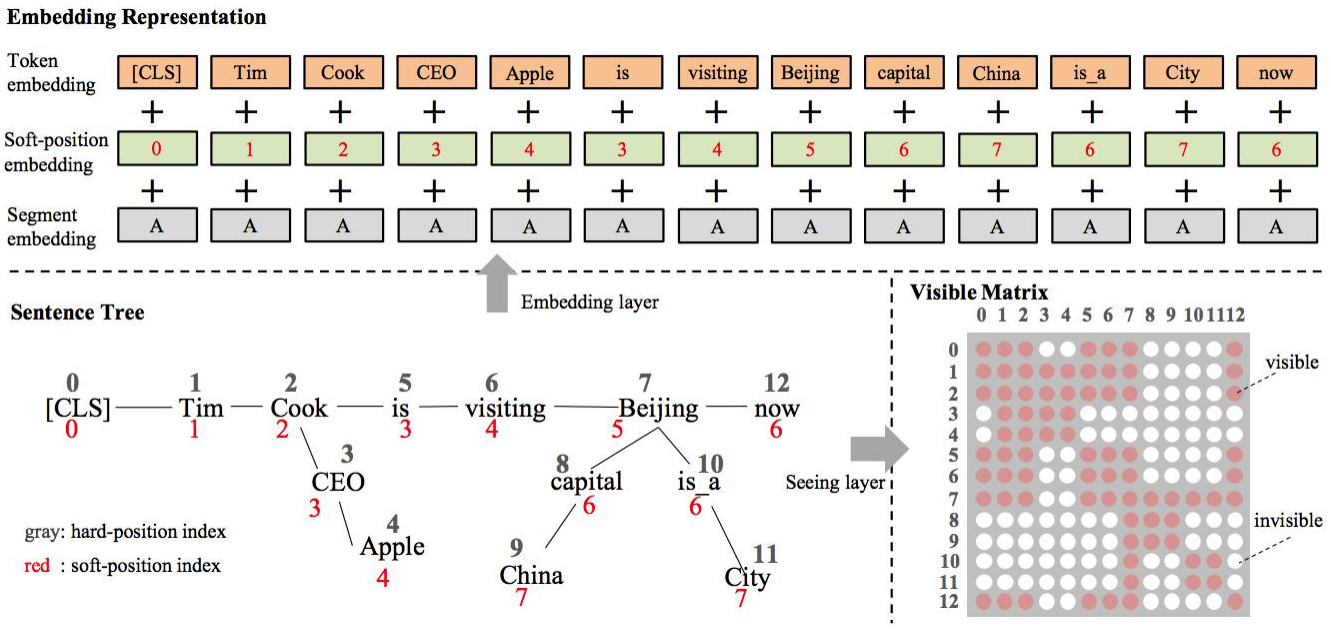
- 在BERT的语言模型中引入了知识图谱中的命名实体的先验知识，通过网络结构的改造进行KG实体的语义对齐
  - **T-encoder**：对文本的**Token**输入进行编码
  - **K-Encoder**：对知识图谱的实体嵌入特征进行编码
  - **Aggregator**：对文本特征和实体特征的融合



Zhang, Zhengyan et al. “ERNIE: Enhanced Language Representation with Informative Entities.” ACL (2019).

K-BERT

- 修改TRANSFORMER中的ATTENTION机制，通过特殊的MASK方法将知识图谱中的相关边考虑到编码过程中，进而增强预训练模型的效果
- 对每一个句子中包含的实体抽取其相关的三元组，这里的三元组被看作是一个短句（首实体，关系，尾实体），与原始的句子合并一起输入给TRANSFORMER模型
- 引入MASK ATTENTION机制，解决知识库引入噪声的问题



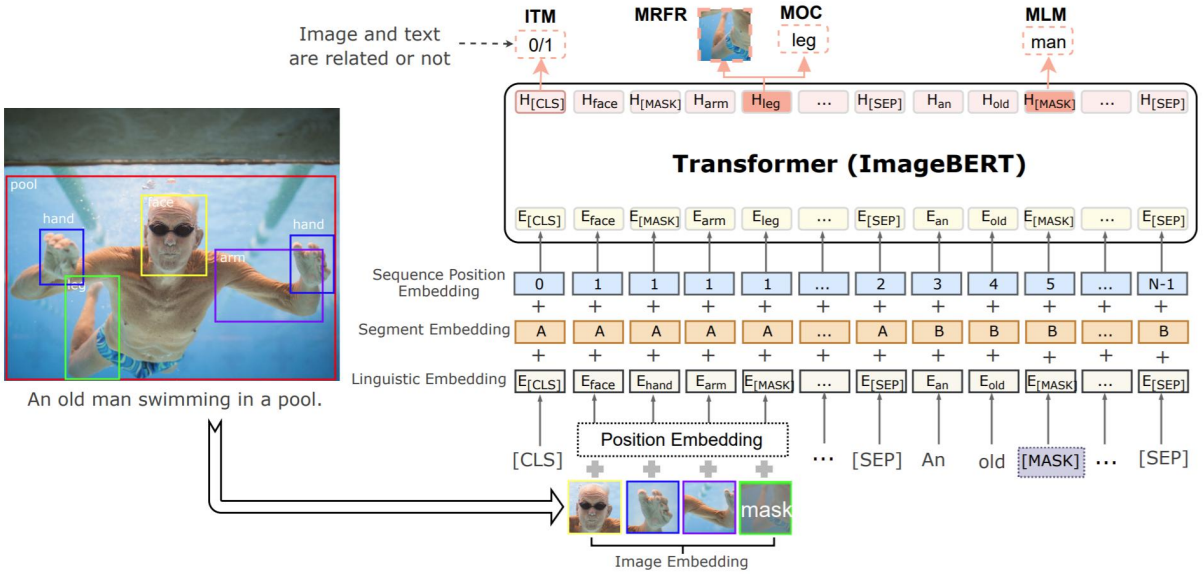
K-BERT: Enabling Language Representation with Knowledge Graph. AAAI 2019.

多模态数据

- 图像、文本、视频、音频等构成相互融合的多媒体形态：形式上多源异构，语义上相互关联

IMAGEBERT

- 预训练任务
  - 掩蔽文本预测（MLM）
  - 掩蔽图像类别预测（MOC）：MLM任务的扩展，通过对视觉对象进行MASK，期望模型预测出被MASK的图像token的类别
  - 掩蔽图像特征回归（MFR）：预测被掩MASK的视觉对象的嵌入特征，在相应位置的输出特征向量后添加一个全连接层，投影到与原始对象特征相同的维度上
  - 图片-文本匹配（ITM）：对于每个训练样本，对每个图像随机抽取负例句子，对每个句子随机抽取负例图像，生成负例训练数据，判断给定的图像文本对是否对应



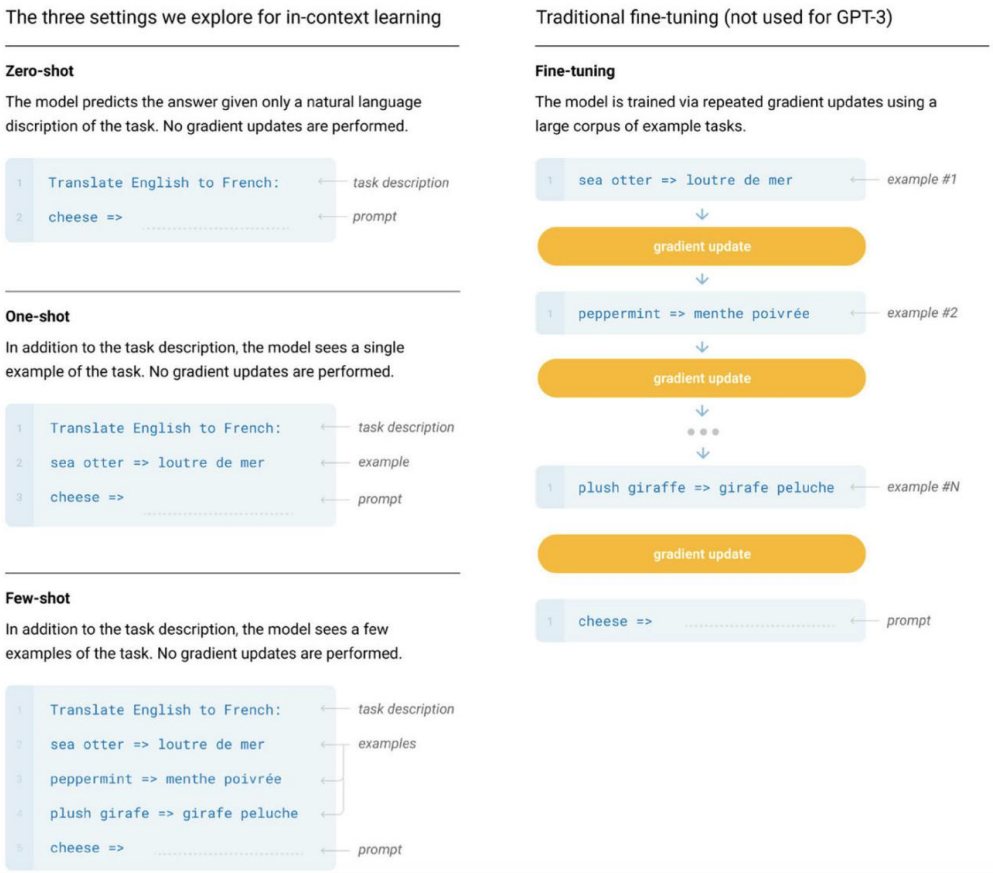
设计更大的模型

超大规模预训练模型

模型名称	参数规模	训练数据	机构名称	模型结构	是否支持多模态	是否支持多语言	是否需要精调	是否适用生成任务
GPT-3	1750亿	45TB	Google	Unidirectional	否	否	是	是
T5	110亿	750GB	Google	Encoder-Decoder	否	是	是	是
M6	1000亿	2TB	Alibaba	半自回归	是	否	是	是
Switch Transformer	1.6万亿	同T5	Google	Encoder-Decoder	否	是	是	否
Megatron	39亿	174GB	英伟达	Bidirectional	否	否	是	否
CPM	26亿	100GB	智源	同GPT-3	否	否	是	是
Pangu	2000亿	1.1TB	Huawei	Unidirectional	否	否	是	是

GPT-3

- GPT模型 尺寸增大到了1750亿，使用45TB数据进行训练，可直接用于零样本和少样本任务



Brown, Tom B. et al. “Language Models are Few-Shot Learners.” ArXiv abs/2005.14165 (2020)

Pangu

- 截至目前最大规模的中文预训练模型，具有2000亿可训练参数
- 采用数据并行、模型并行、优化器模型并行等并行化方案
- 在CLUE榜单上达到新的SOTA

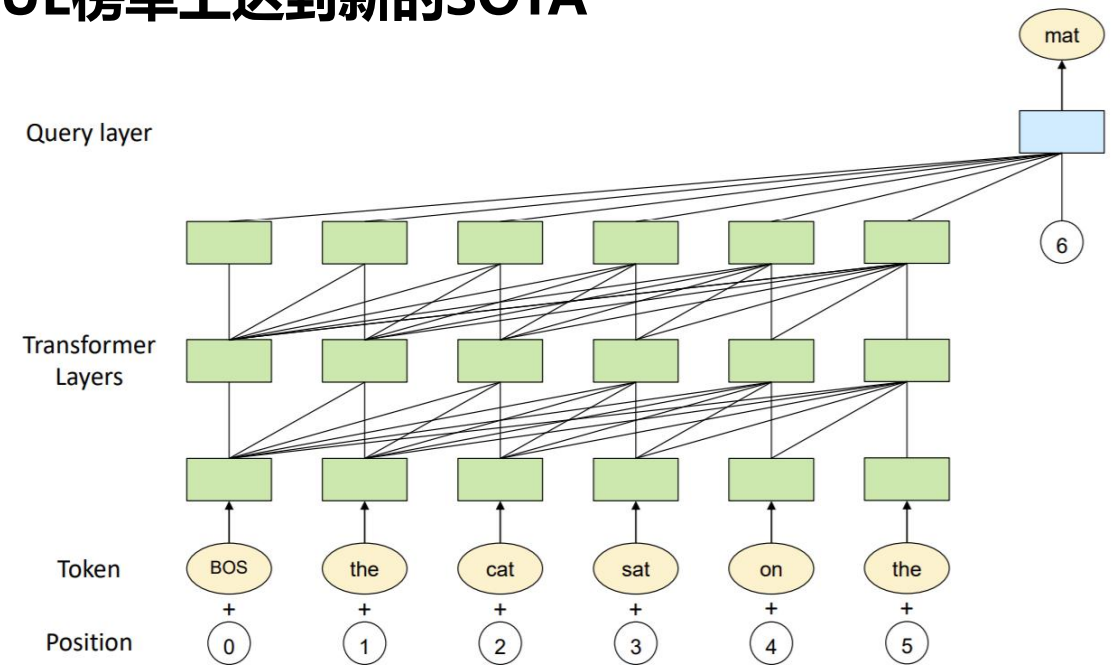


Figure 1: The architecture of PanGu-α. The model is based on a uni-directional Transformer decoder. A query layer is stacked on top of Transformer layers with the position embedding as the query in the attention mechanism to generate the token at the next position.

PANGU-α: LARGE-SCALE AUTOREGRESSIVE PRETRAINED CHINESE LANGUAGE MODELS WITH AUTO-PARALLEL COMPUTATION. 2021.04.26

优化计算效率和性能

系统级优化

- 数据并行
- 模型并行

学习算法优化

- 优化预训练策略
- 优化模型结构

模型压缩

- 参数共享
- 模型剪枝
- 知识蒸馏
- 模型量化

ALBert

- 瘦身版Bert：
  - 所有Transformer层之间实现参数共享，减少训练参数，提升模型训练效率
  - 将输入的词嵌入矩阵分解成两个更小的矩阵
  - 使用句子顺序预测任务（SOP）去代替Bert中的NSP任务

Model		Parameters	SQuAD1.1	SQuAD2.0	MNLI	SST-2	RACE	Avg	Speedup
BERT	base	108M	90.4/83.2	80.4/77.6	84.5	92.8	68.2	82.3	4.7x
	large	334M	92.2/85.5	85.0/82.2	86.6	93.0	73.9	85.2	1.0
ALBERT	base	12M	89.3/82.3	80.0/77.1	81.6	90.3	64.0	80.1	5.6x
	large	18M	90.6/83.9	82.3/79.4	83.5	91.7	68.5	82.4	1.7x
	xlarge	60M	92.5/86.1	86.1/83.1	86.4	92.4	74.8	85.5	0.6x
	xxlarge	235M	**94.1/88.3**	**88.1/85.1**	**88.0**	**95.2**	**82.3**	**88.7**	0.3x

加速比和部分实验效果

Lan, Zhenzhong et al. “ALBERT: A Lite BERT for Self-supervised Learning of Language Representations.” ArXiv abs/1909.11942 (2020): n. pag.

6. 预训练模型：过去、现在与未来

预训练模型的未来

- 模型结构和预训练方法
- 多模态和多语言预训练
- 计算效率
- 理论基础
- 认知学习
- 模型应用

参考文献

- Pre-Trained Models: Past, Present and Future
- XLNet: Generalized Autoregressive Pretraining for Language Understanding
- Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer
- ERNIE: Enhanced Representation through Knowledge Integration
- A Robustly Optimized BERT Pretraining Approach
- Cross-lingual Language Model Pretraining
- ERNIE: Enhanced Language Representation with Informative Entities
- K-BERT: Enabling Language Representation with Knowledge Graph
- CROSS-MODAL PRE-TRAINING WITH LARGE-SCALE WEAK-SUPERVISED IMAGE-TEXT DATA
- Language Models are Few-Shot Learners
- PANGU-α: LARGE-SCALE AUTOREGRESSIVE PRETRAINED CHINESE LANGUAGE MODELS WITH AUTO-PARALLEL COMPUTATION
- ALBERT: A Lite BERT for Self-supervised Learning of Language Representations

欢迎加入DL4NLP！

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS