Transformer和预训练模型

邱锡鹏

复旦大学

https://xpqiu.github.io

报告概要

自然语言表示学习

- Transformer模型及改进

预训练模型

- 自监督学习、分类体系

预训练模型的扩展

- 跨语言、跨模态、知识嵌入、模型压缩

迁移到下游任务

未来展望

自然语言表示学习

什么是自然语言处理（NLP）？

- 自然语言≈人类语言，不同于人工语言（比如程序语言）
  - 自然语言处理包括语音识别、自然语言理解、自然语言生成、人机交互以及所涉及的中间阶段。
  - 是人工智能和计算机科学的子学科。

NLP的基础：语言表示

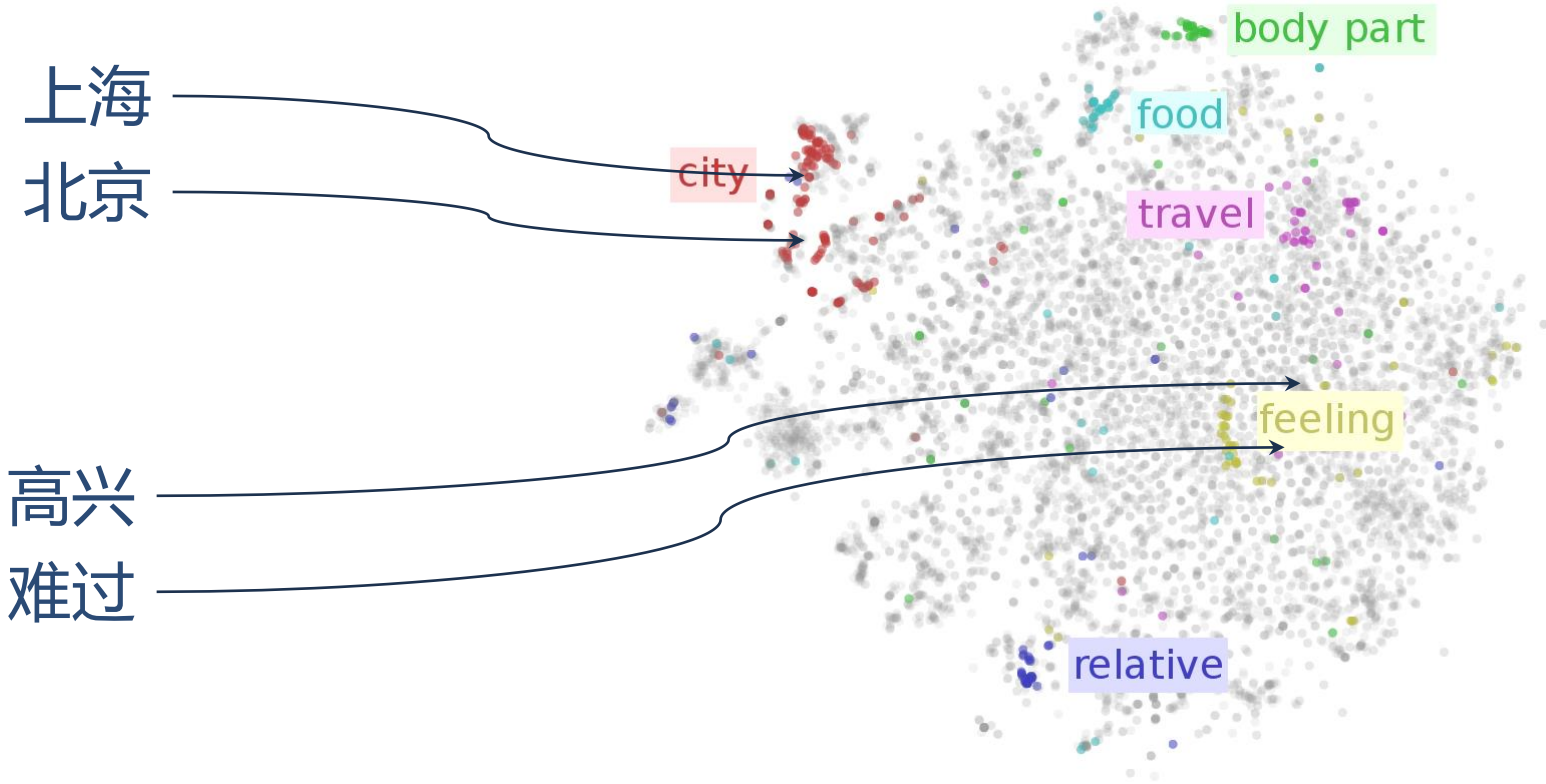
- 语言为离散的符号

- 如何在计算机中表示语言的语义？

- 压缩、低维、稠密向量
- 用O(K)个参数表示 $O(2^k)$区间

- 嵌入（Embeddings）

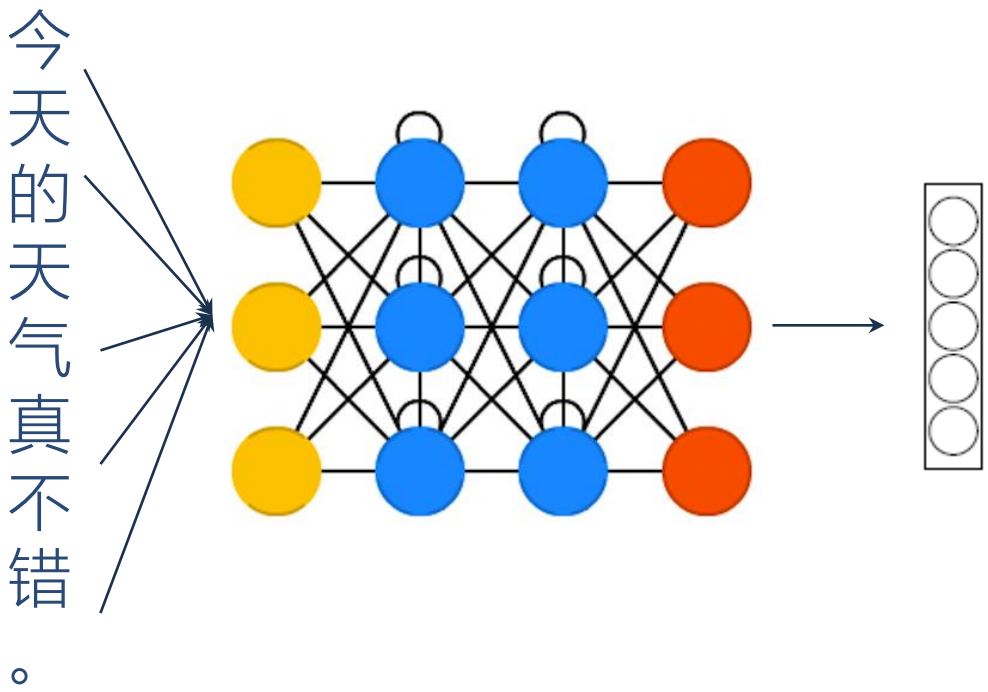
词嵌入（Word Embeddings）



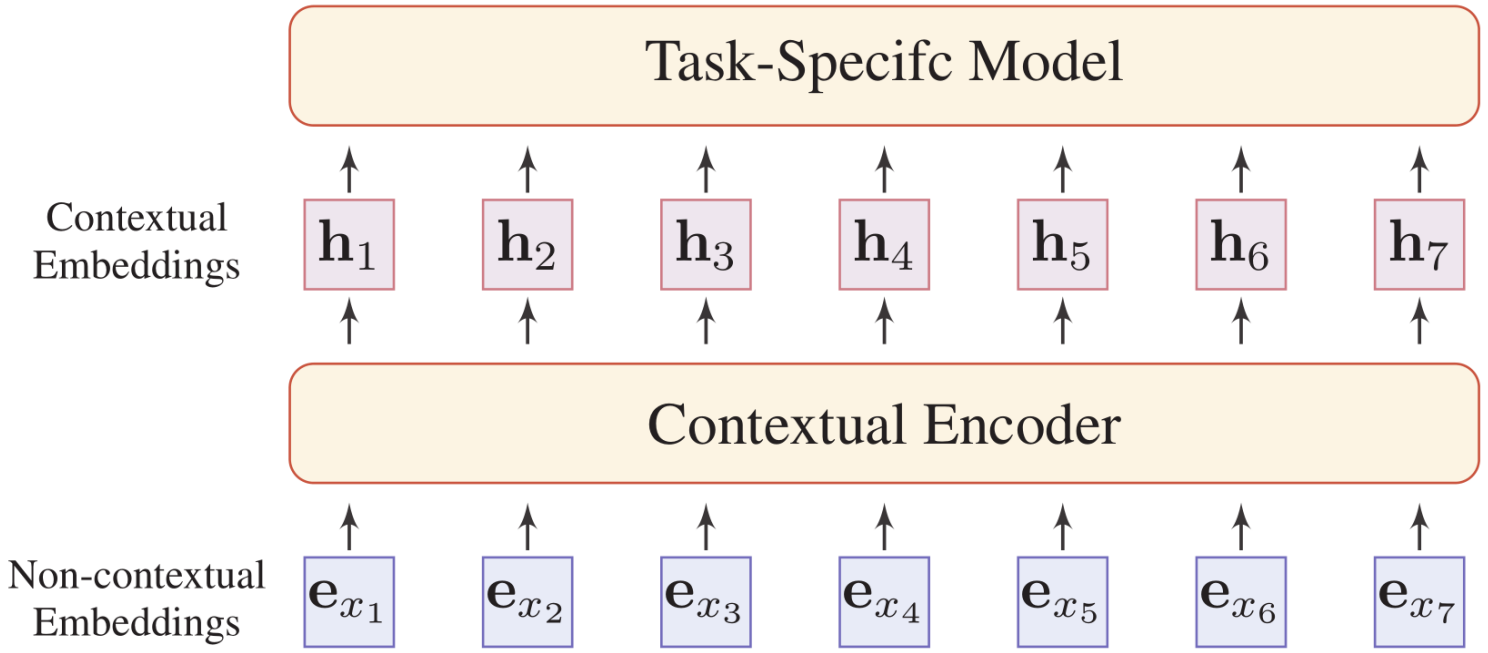
https://indico.io/blog/visualizing-with-t-sne/

语言表示学习

- 词
- 短语
  - 组合语义模型
- 句子
  - 连续词袋模型
  - 序列模型
  - 递归组合模型
  - 卷积模型
- 篇章
  - 层次模型



NLP中神经网络模型的一般架构



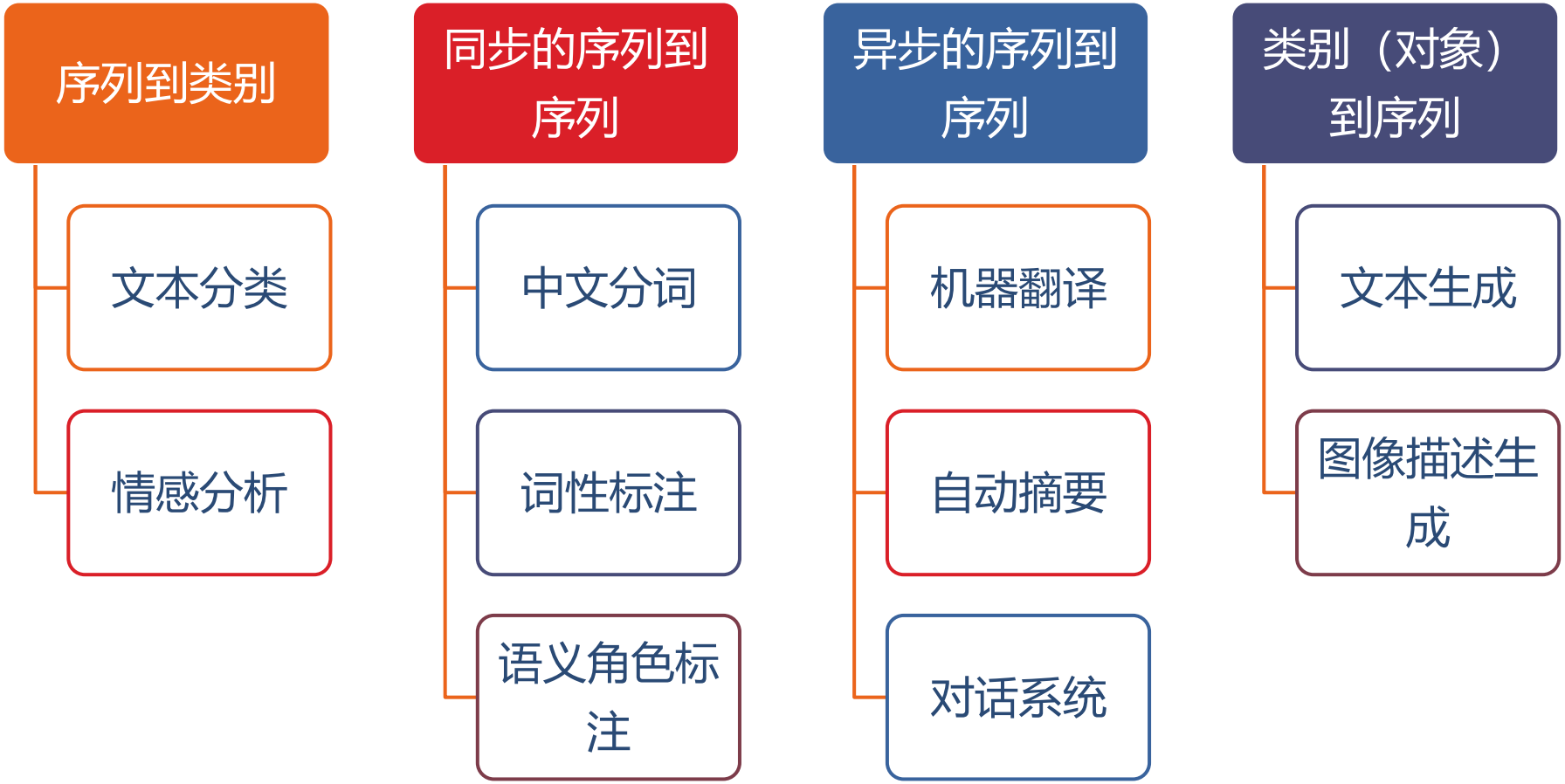
模型驱动

+

数据驱动

自然语言处理任务

- 自然语言处理任务类型划分为：



极大降低了自然语言处理的门槛

序列到类别

带着愉悦的心情看了这部电影

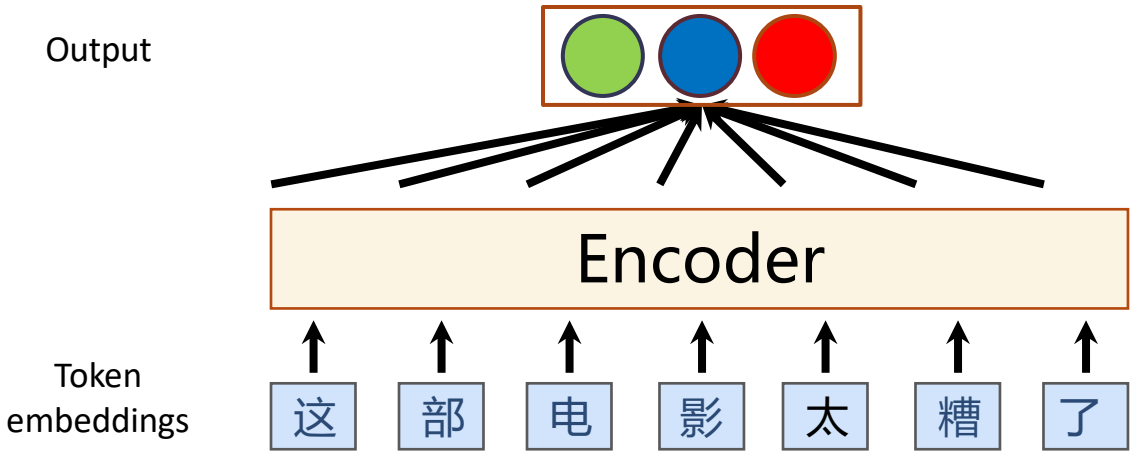
Positive (正面)

这部电影太糟了

Negative (负面)

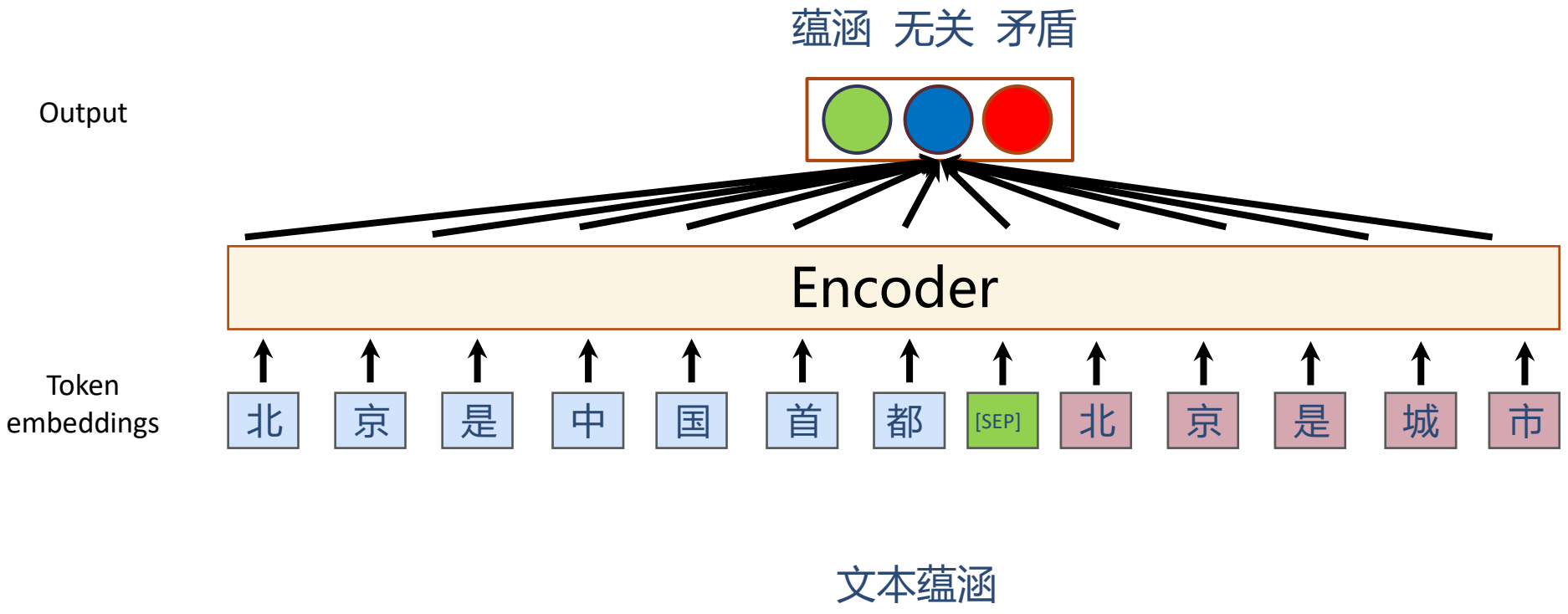
这部电影很棒

Positive (正面)

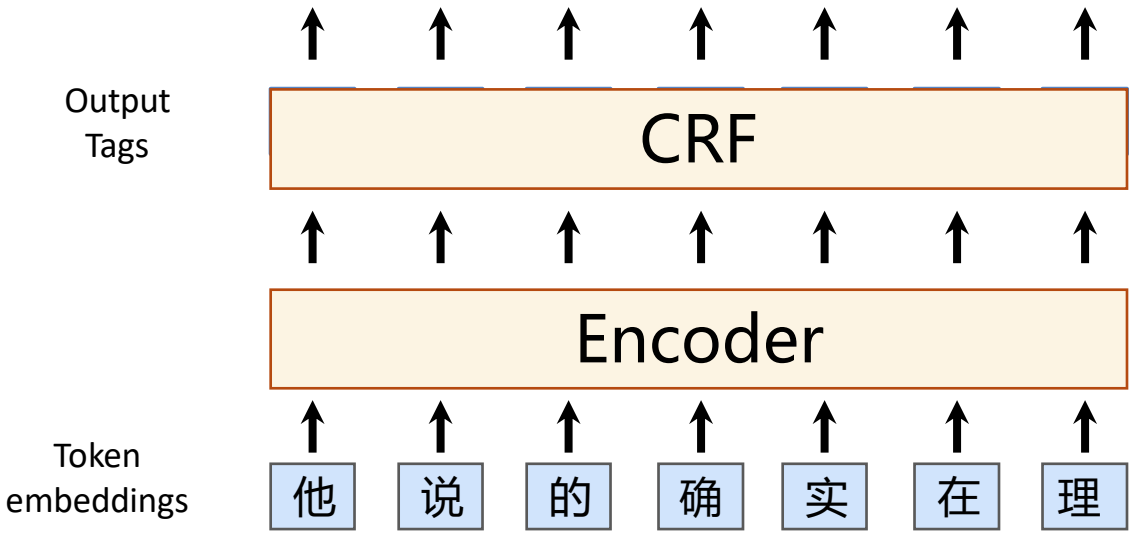


情感分析

序列到类别

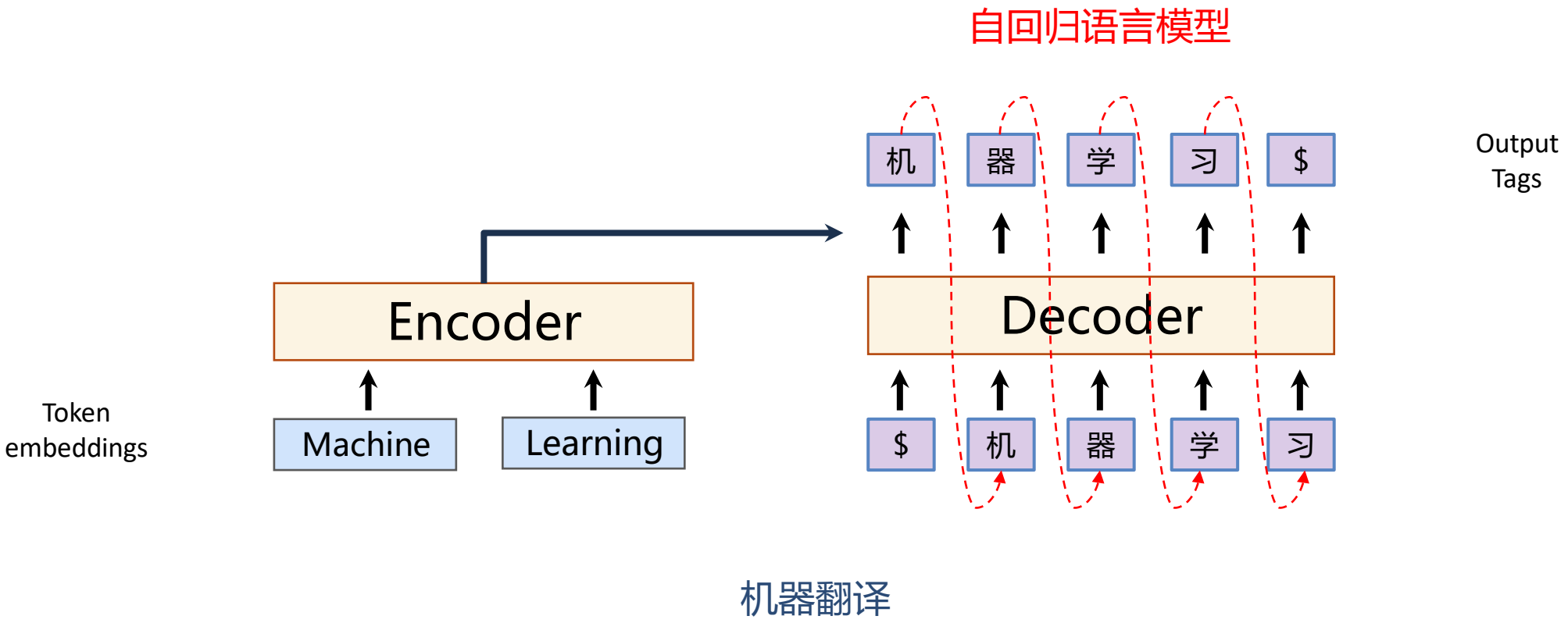

蕴涵 无关 矛盾
Output
Encoder
Token embeddings
北 京 是 中 国 首 都 [SEP] 北 京 是 城 市
文本蕴涵

同步（对齐）的序列到序列



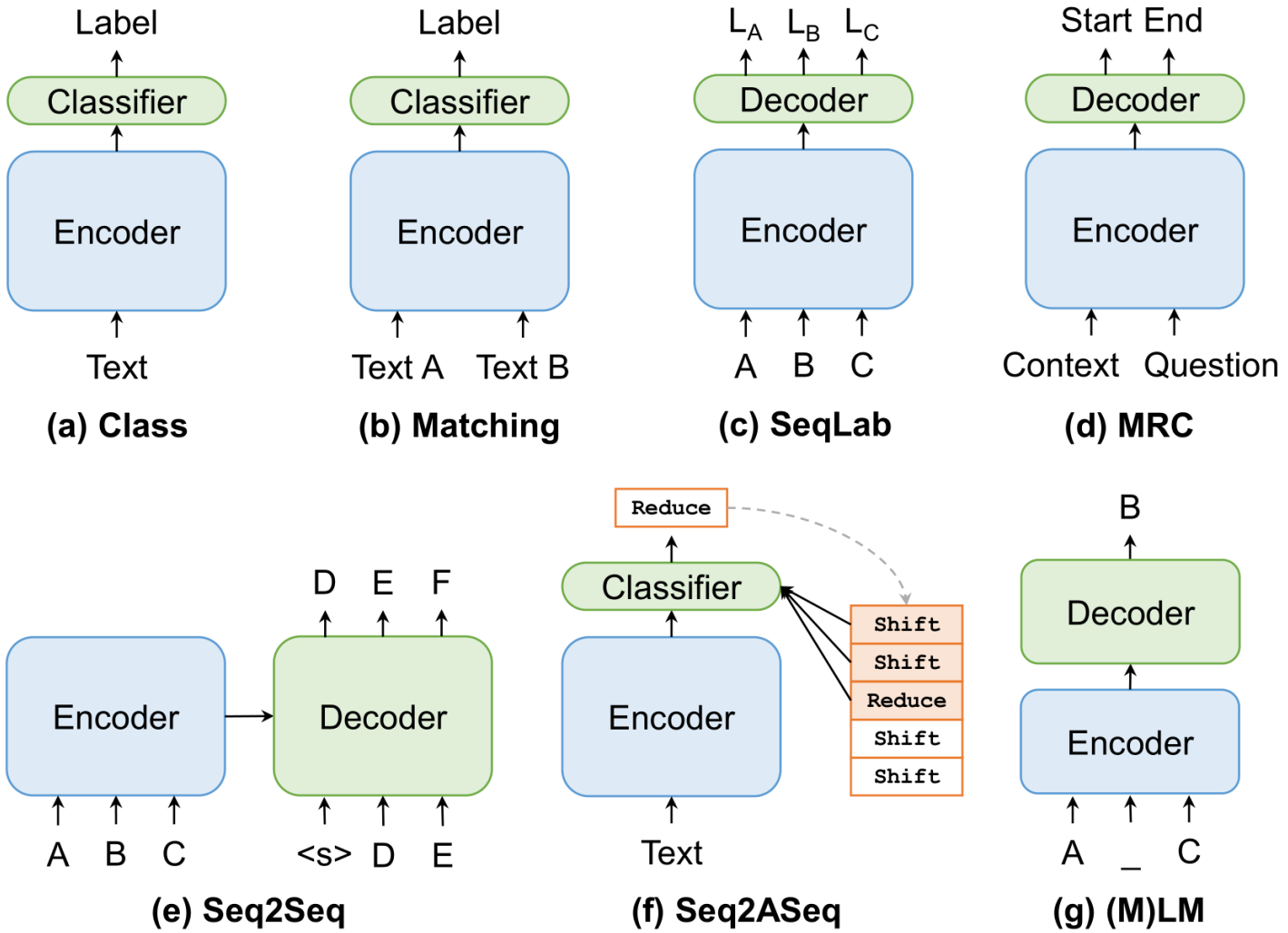
中文分词

异步（非对齐）的序列到序列

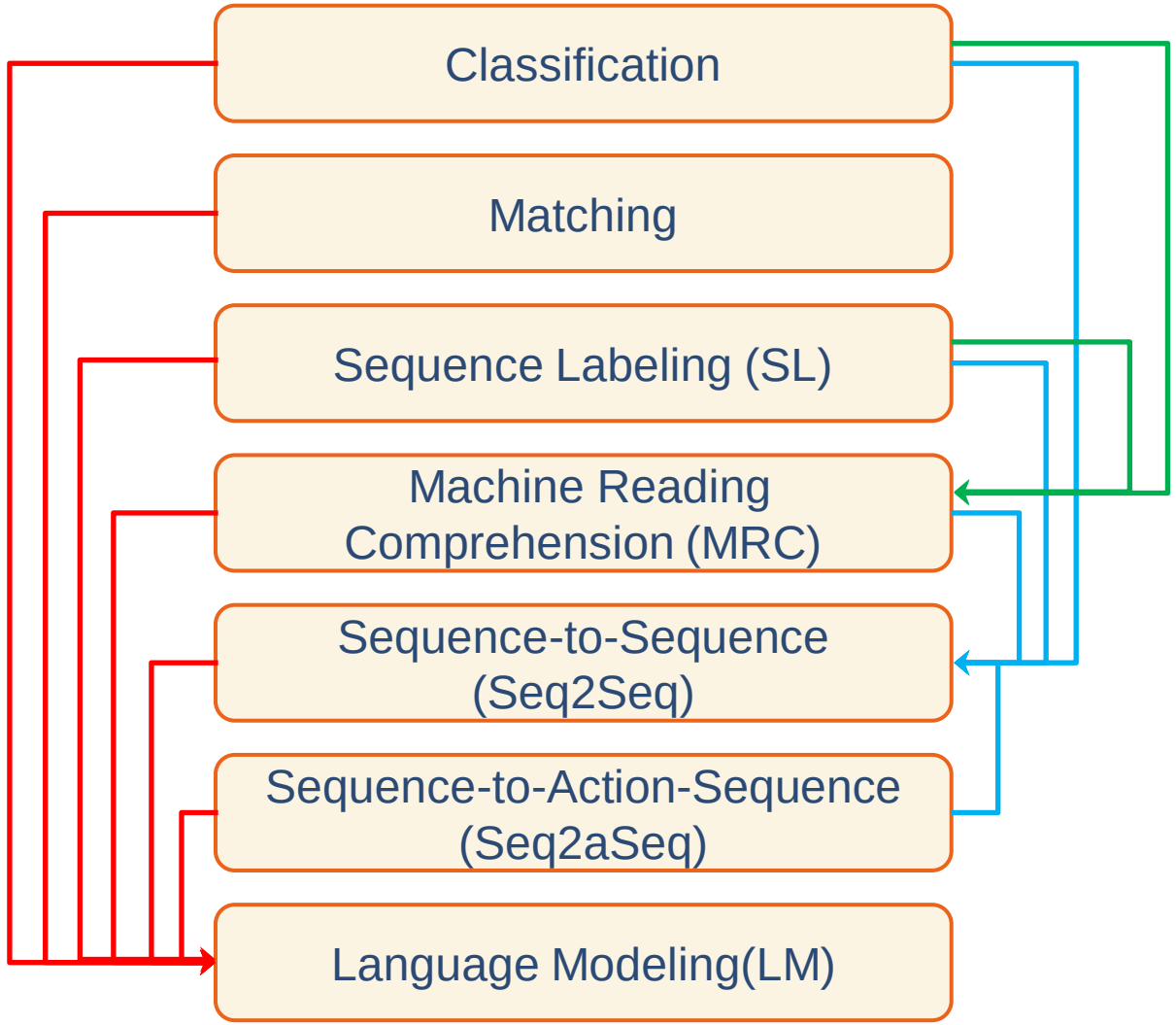


机器翻译

Seven Main Paradigms



Paradigm Shift



Paradigm Shift brings various NLP tasks a unified framework.
The pretrained backbone models accelerate this shift.

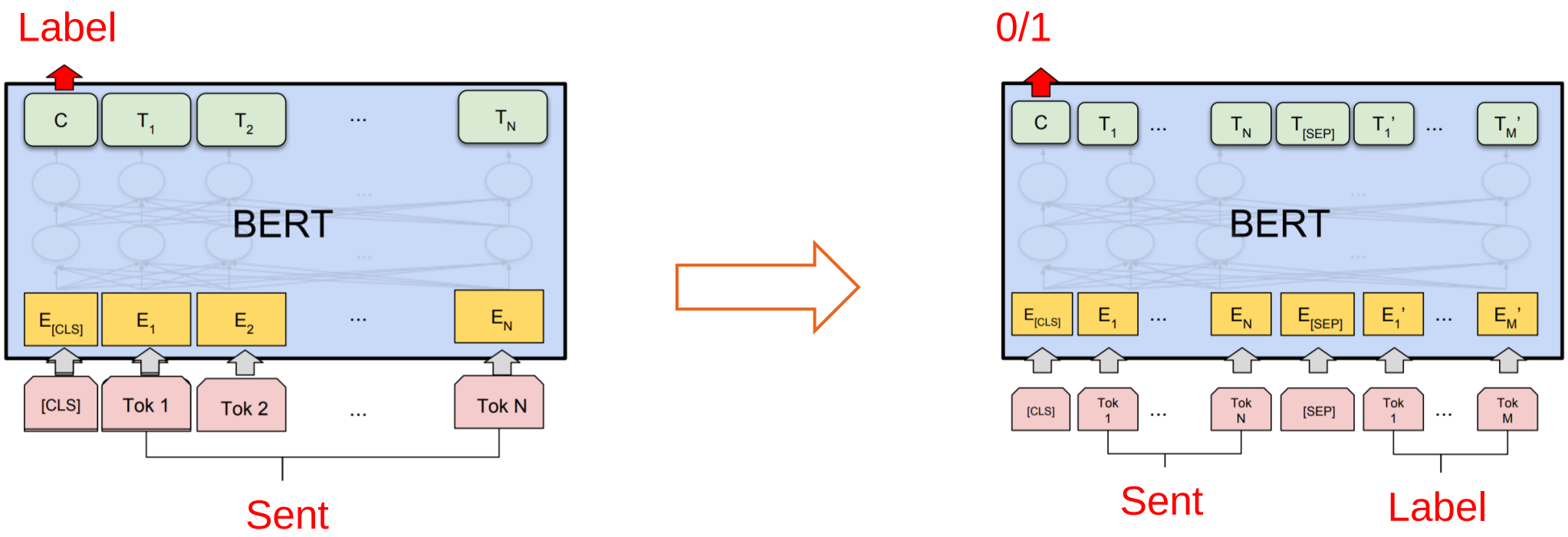
MRC/Seq2Seq/LM are powerful paradigms.

The generative paradigm is more general and flexible.

https://github.com/txsun1997/nlp-paradigm-shift

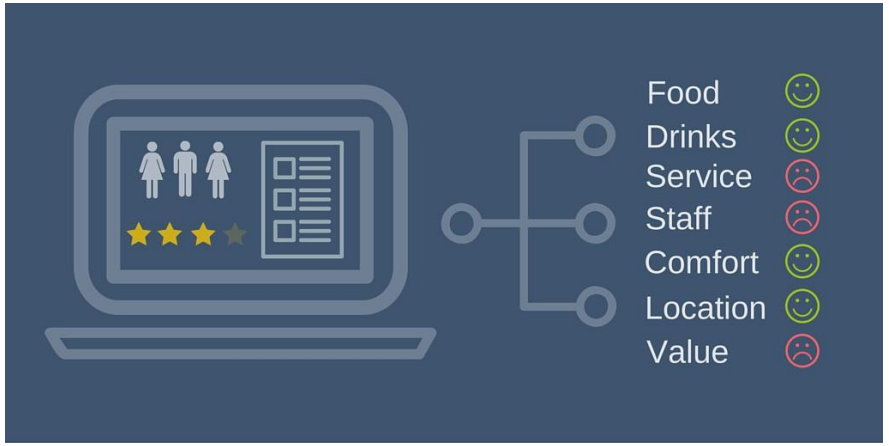
Text Classification as Text Matching

- Constructing Meaningful Labels



Chi Sun, Luyao Huang, Xipeng Qiu, Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence, NAACL 2019, https://arxiv.org/pdf/1903.09588.pdf

Sentiment Analysis



BERT

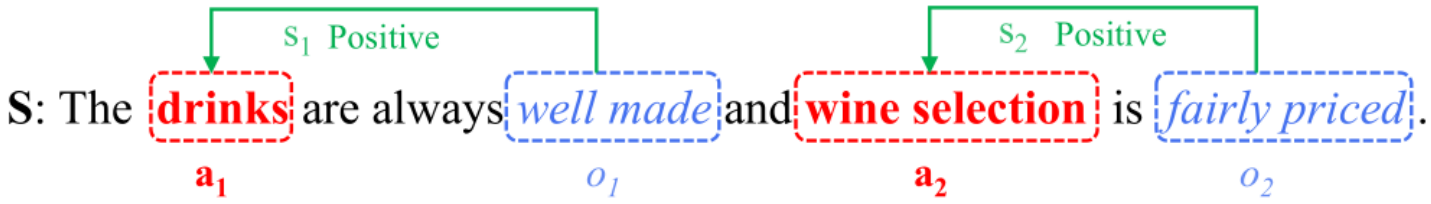
Positive

Constructing Question

what do you think of the safety of location?

MRC

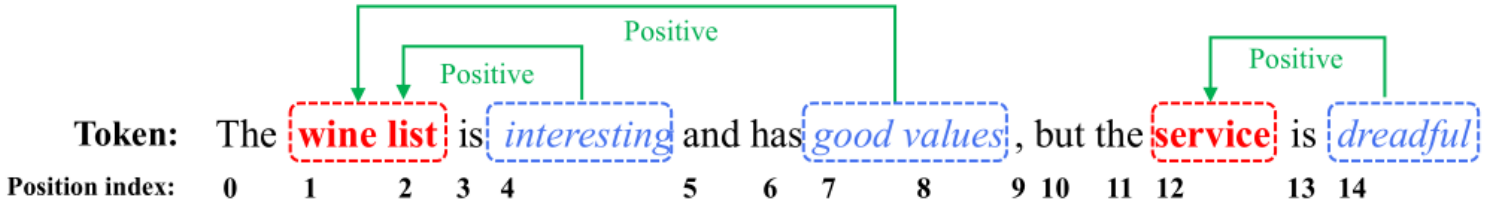
Seven ABSA subtasks



Subtask	Input	Output	Task Type
Aspect Term Extraction(*AE*)	**S**	a_1, a_2	Extraction
Opinion Term Extraction(*OE*)	**S**	o_1, o_2	Extraction
Aspect-level Sentiment Classification(*ALSC*)	**S** + a_1	s_1	Classification
	S + a_2	s_2	
Aspect-oriented Opinion Extraction(*AOE*)	**S** + a_1	o_1	Extraction
	S + a_2	o_2	
Aspect Term Extraction and Sentiment Classification(*AESC*)	**S**	(a_1, s_1), (a_2, s_2)	Extraction & Classification
Pair Extraction(*Pair.*)	**S**	(a_1, o_1), (a_2, o_2)	Extraction
Triplet Extraction(*Triplet.*)	**S**	(a_1, o_1, s_1), (a_2, o_2, s_2)	Extraction & Classification

A Unified Generative Framework for Aspect-based Sentiment Analysis, accepted by ACL 2021

ABSA as Sequence Generation



Subtask	Target Sequence
AE	1, 2, 12, 12, </s>
OE	4, 4, 7, 8, 14, 14, </s>
ALSC	1, 2 , POS, </s>
	12, 12, POS, </s>
AOE	1, 2, 4, 4, 7, 8, </s>
	12, 12, 14, 14, </s>
AESC	1, 2, POS, 12, 12, NEG, </s>
Pair.	1, 2, 4, 4, 1, 2, 7, 8, 12, 12, 14, 14, </s>
Triplet.	1, 2, 4, 4, POS, 1, 2, 7, 8, POS, 12, 12, 14, 14, POS, </s>

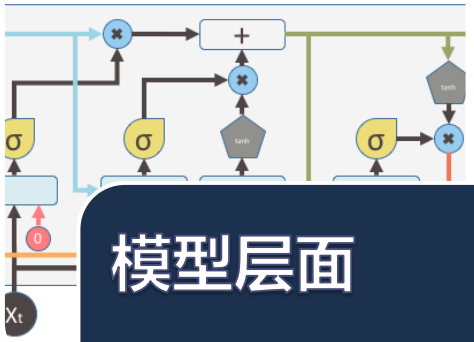
A Unified Generative Framework for Aspect-based Sentiment Analysis, accepted by ACL 2021

语言表示学习的几个问题

认知层面

- 语言知识
- 世界知识
  - 常识

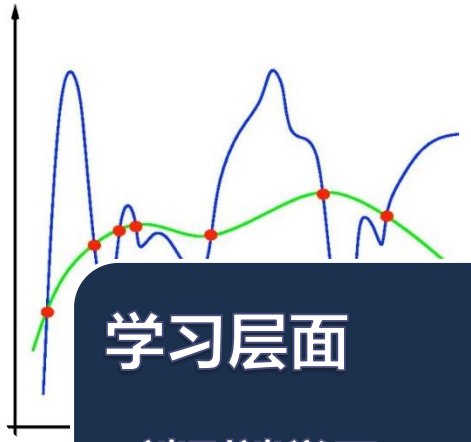
语义表示
先验偏置



模型层面

- 语义组合问题
- 局部VS非局部
- 长期依赖问题

模型驱动



学习层面

- 端到端学习
- 迁移学习
- 多任务学习

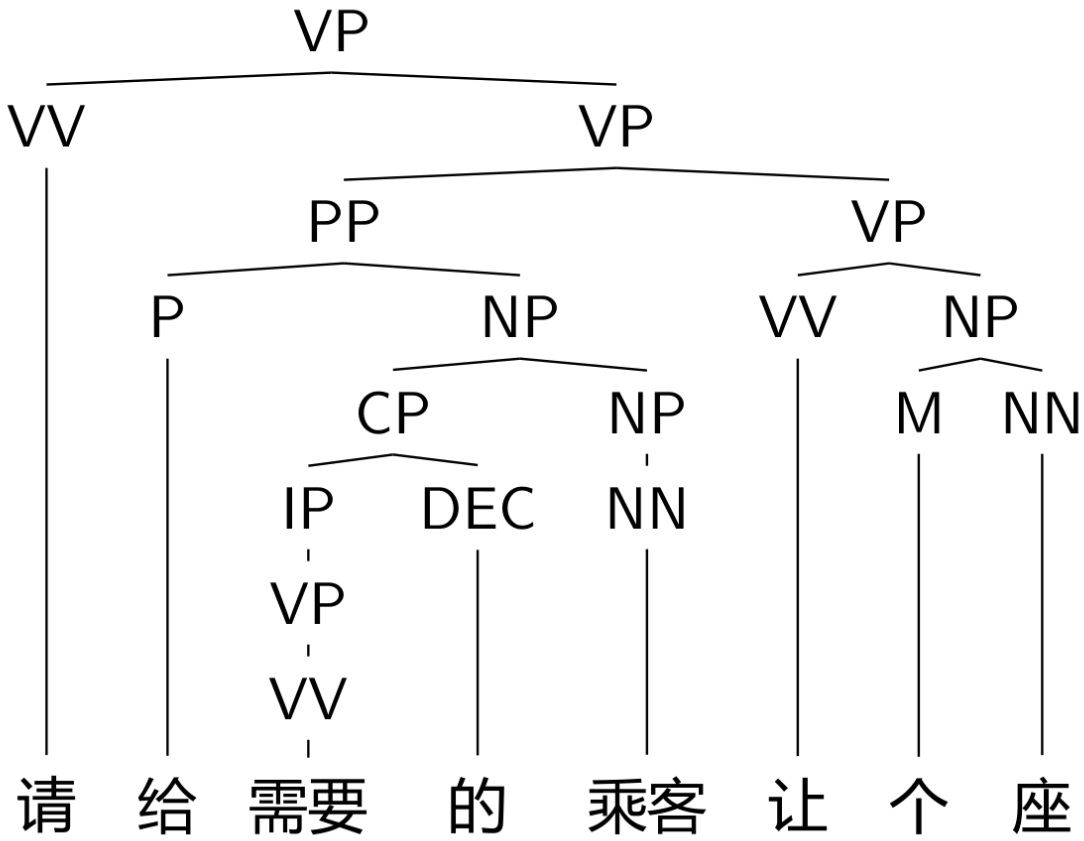
数据驱动

自然语言表示学习

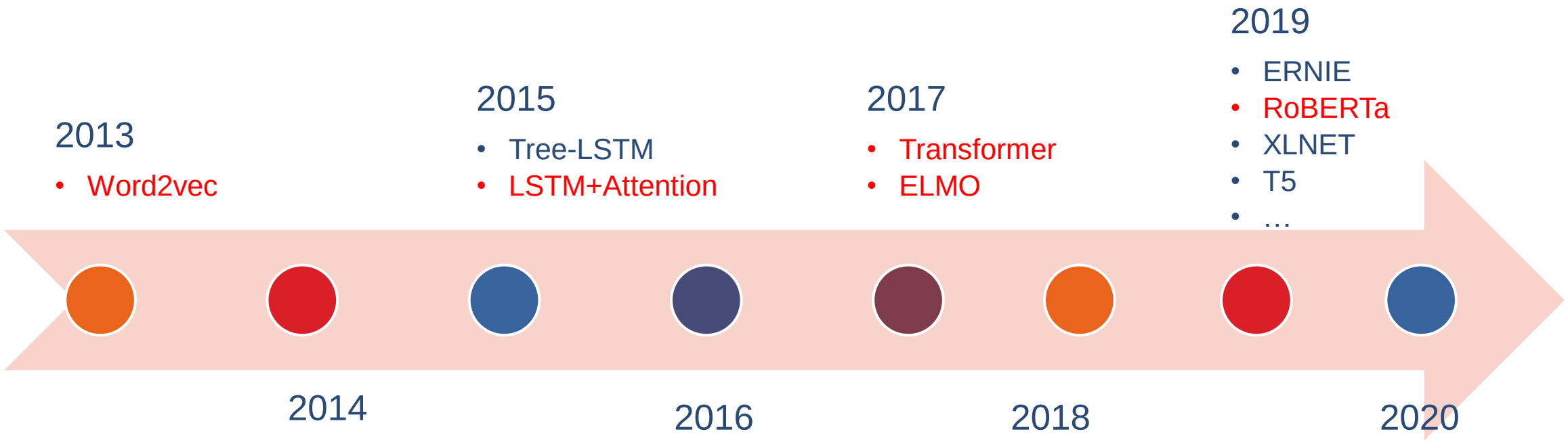
模型驱动 + 数据驱动 = 自然语言表示学习

语义组合

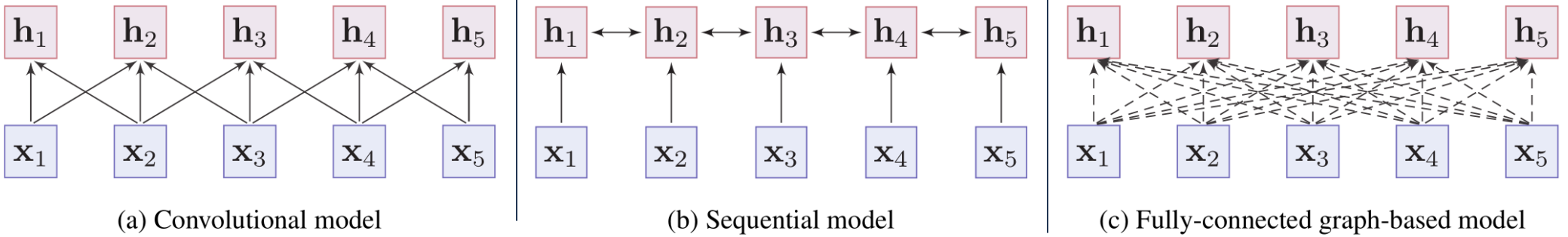
- 语言的性质
  - 层次性
  - 递归性
  - 序列性
- 语义组合
  - 句子的语义可以词组成
  - 长程依赖



寻找最适合NLP的模型

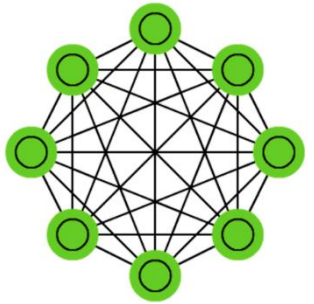


NLP中三大模型



(a) Convolutional model (b) Sequential model (c) Fully-connected graph-based model

隐含先验：局部组合
只建模了输入信息的局部依赖关系



连接权重α_{ij} 由注意力机制动态生成

Transformer模型介绍

注意力机制（Attention Mechanism）

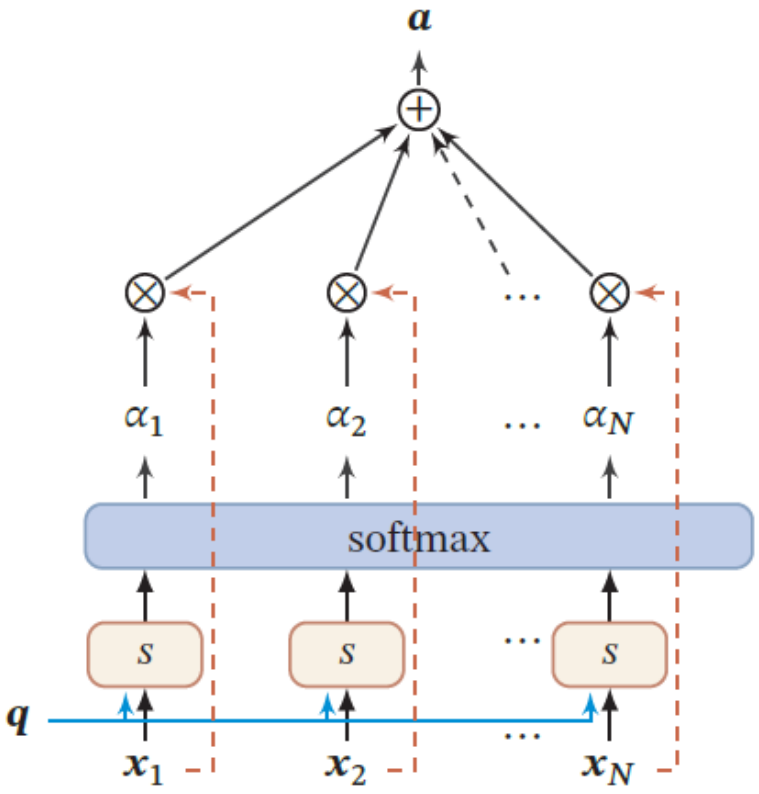
- 注意力机制可以分为两步
  - 计算注意力分布α，

$$\begin{aligned}\alpha_n &= p(z=n|\boldsymbol{X},\boldsymbol{q}) \\ &= \text{softmax}\left(s(\boldsymbol{x}_n,\boldsymbol{q})\right) \\ &= \frac{\exp\left(s(\boldsymbol{x}_n,\boldsymbol{q})\right)}{\sum_{j=1}^{N}\exp\left(s(\boldsymbol{x}_j,\boldsymbol{q})\right)},\end{aligned}$$

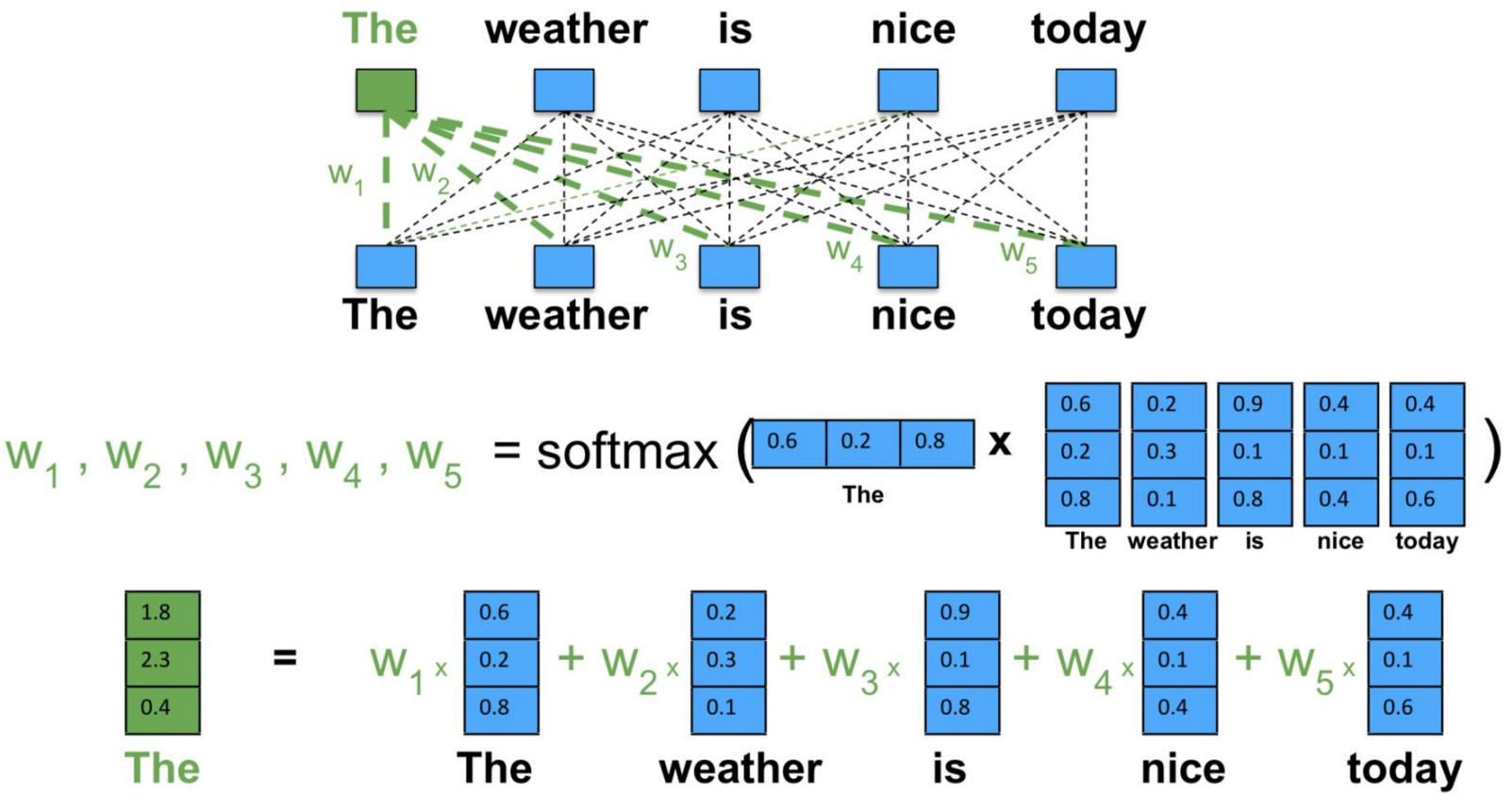
$s(x_i, q)$ 打分函数

  - 根据α来计算输入信息的加权平均。

$$\begin{aligned}\text{att}(\boldsymbol{X},\boldsymbol{q}) &= \sum_{n=1}^{N}\alpha_n\boldsymbol{x}_n, \\ &= \mathbb{E}_{z\sim p(z|\boldsymbol{X},\boldsymbol{q})}[\boldsymbol{x}_z]\end{aligned}$$

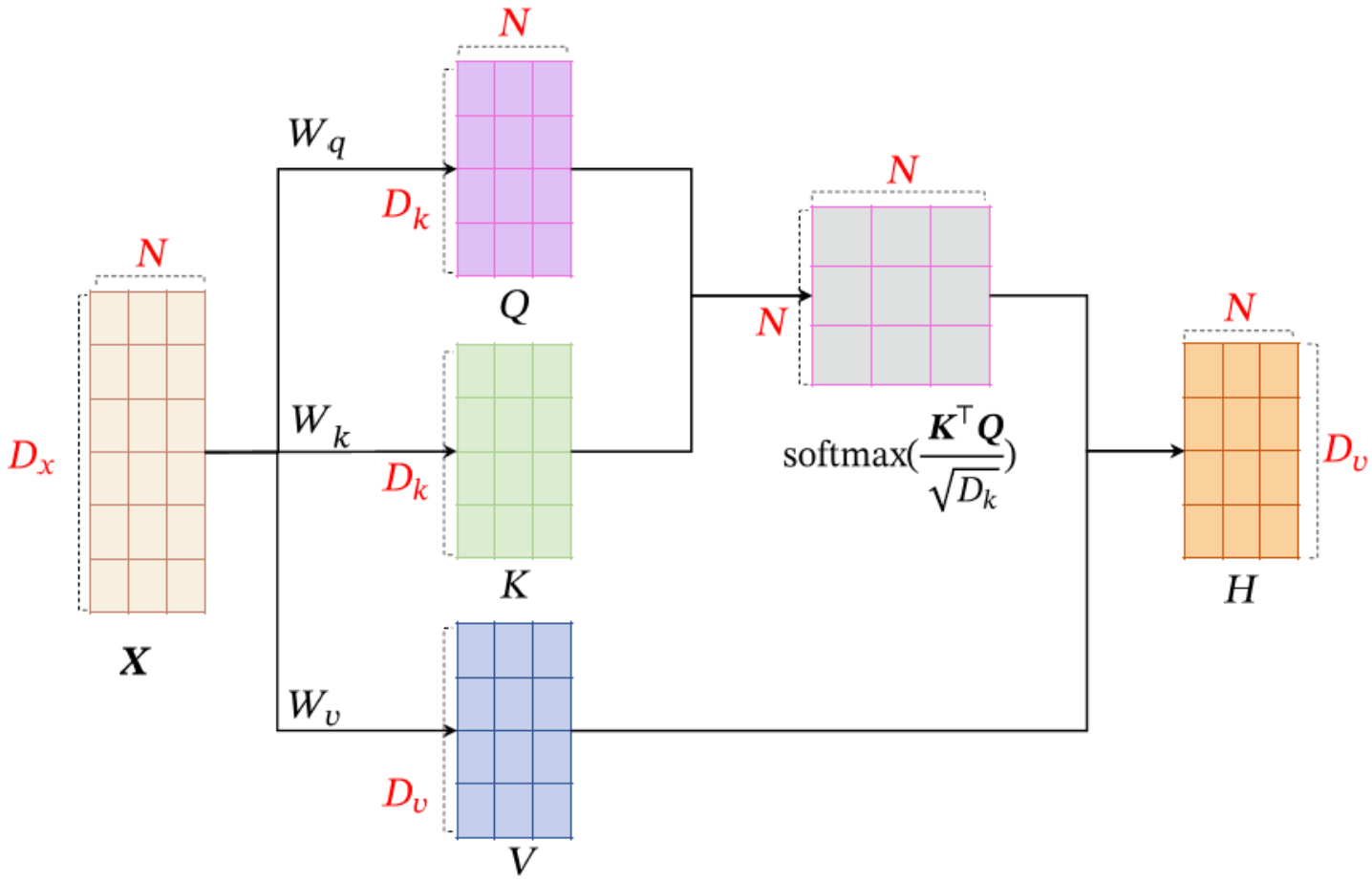


自注意力（Self- Attention ）示例



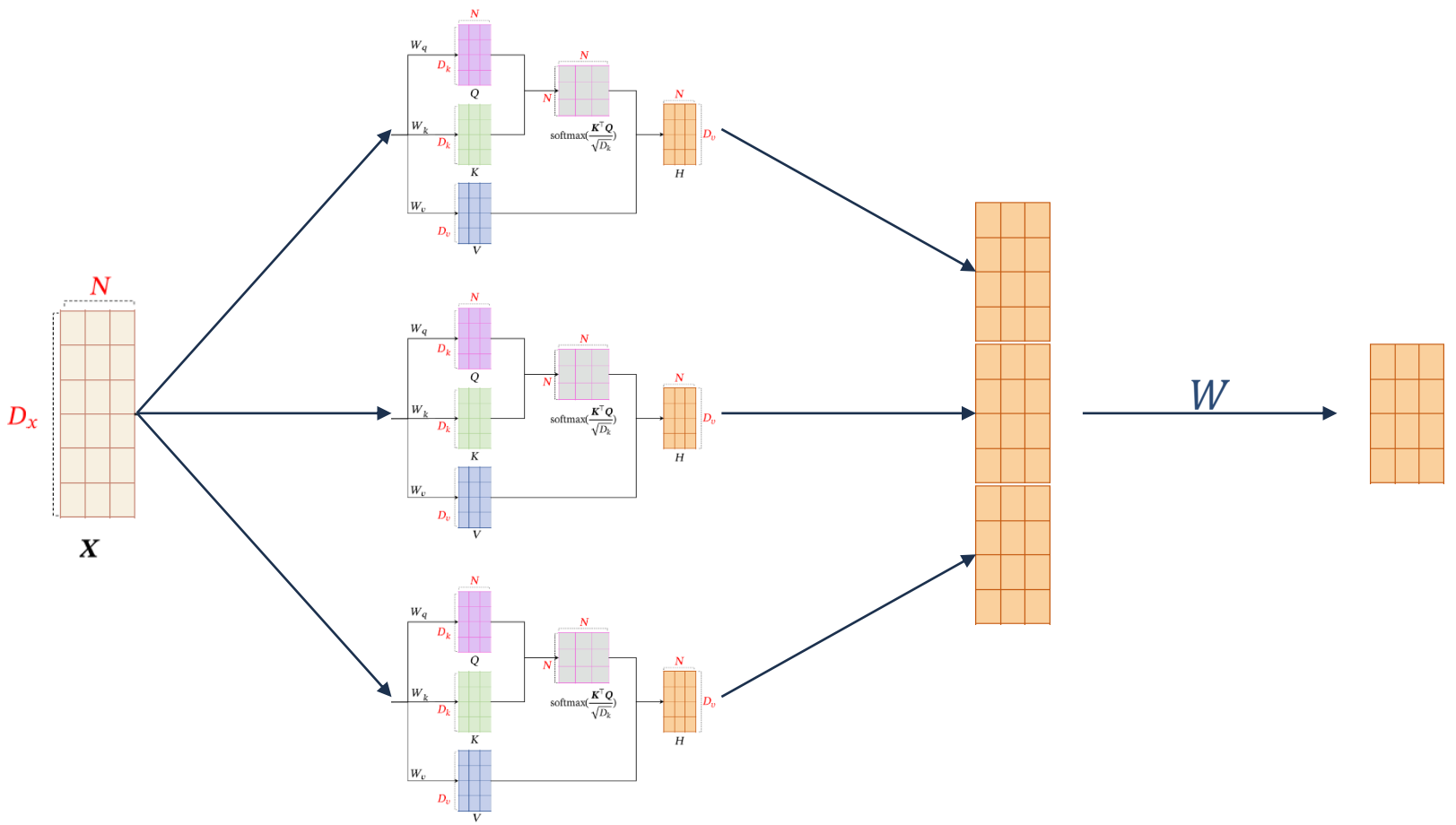
图片来源：http://fuyw.top/NLP_02_QANet/

QKV模式（Query-Key-Value）



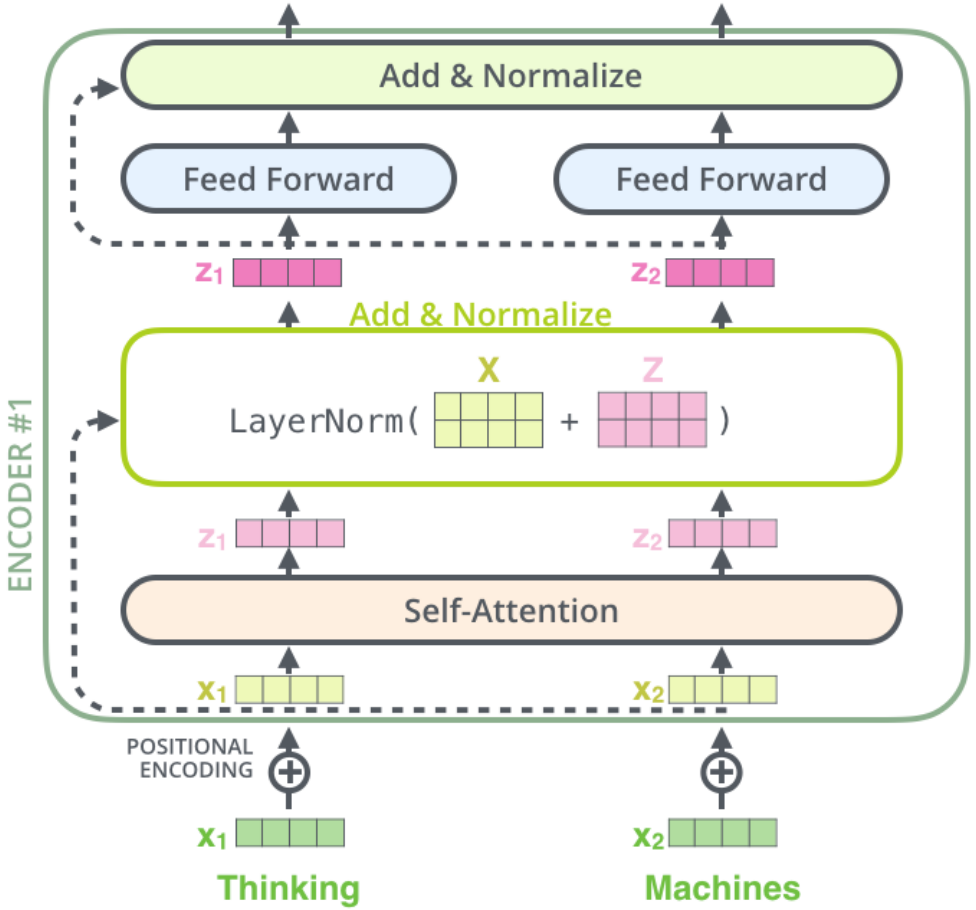
邱锡鹏,神经网络与深度学习,机械工业出版社, 2020. 第八章 “注意力机制与外部记忆”

多头（multi-head）自注意力模型



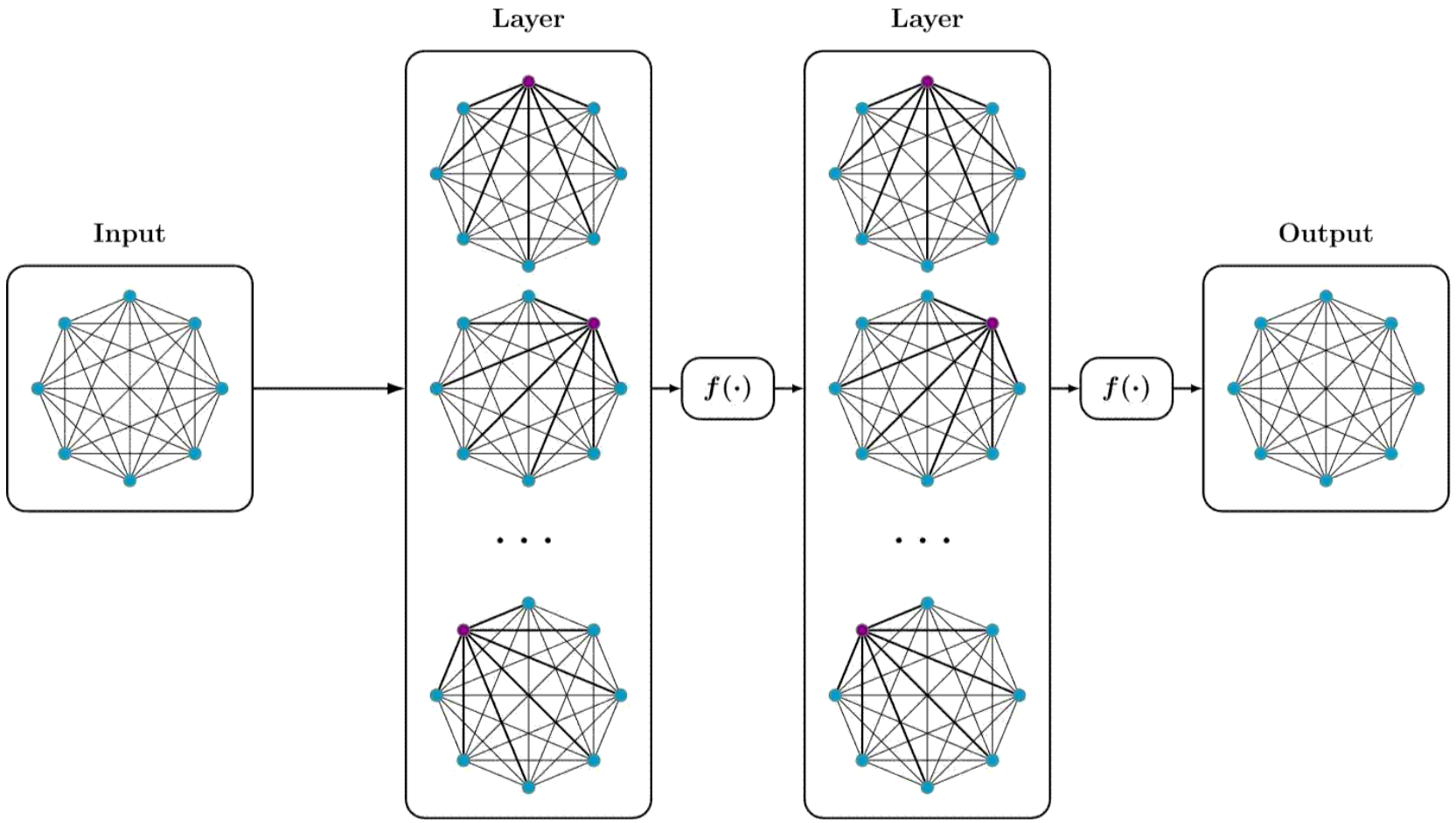
Transformer:可能是目前为止最适合NLP的模型

- 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

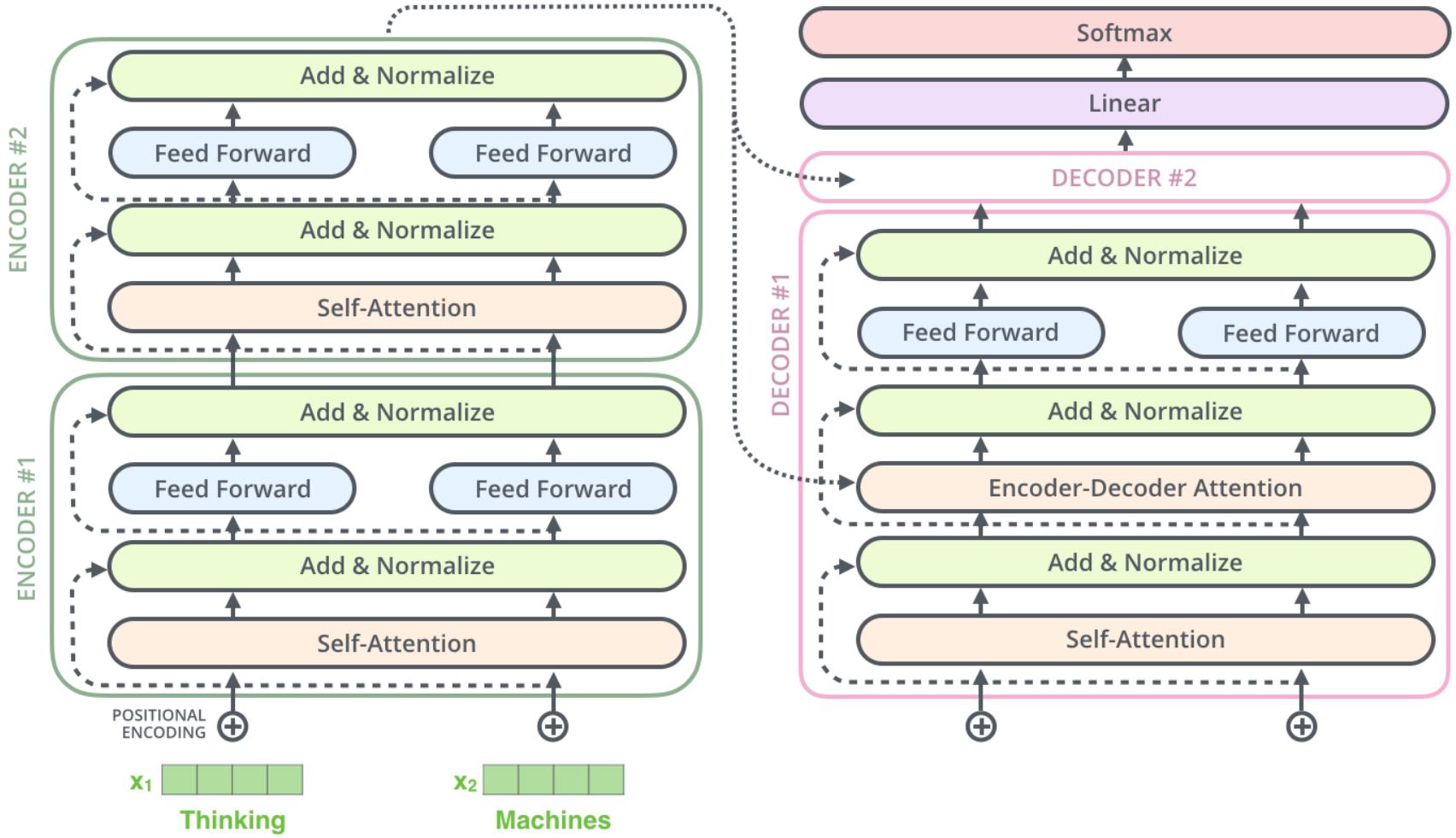


图片来源：http://jalammar.github.io/illustrated-transformer/

Transformer



Transformer完整结构



自注意力模型分析

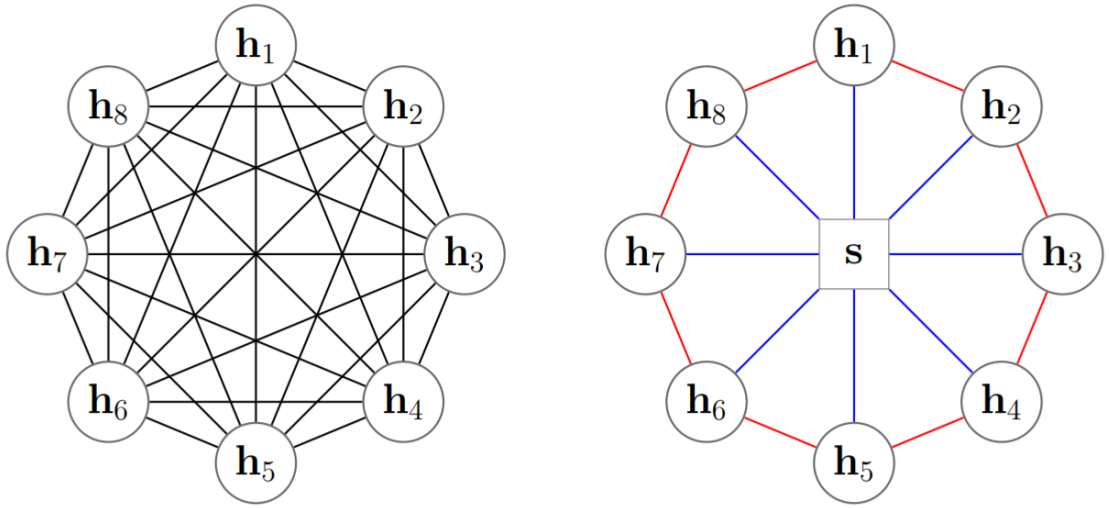
- 模型特点
  - 全连接结构
  - 没有任何先验假设
  - 通过位置编码来建模序列信息

- 复杂度： $O(L^2D)$
  - 无法处理长文档
  - 容易过拟合

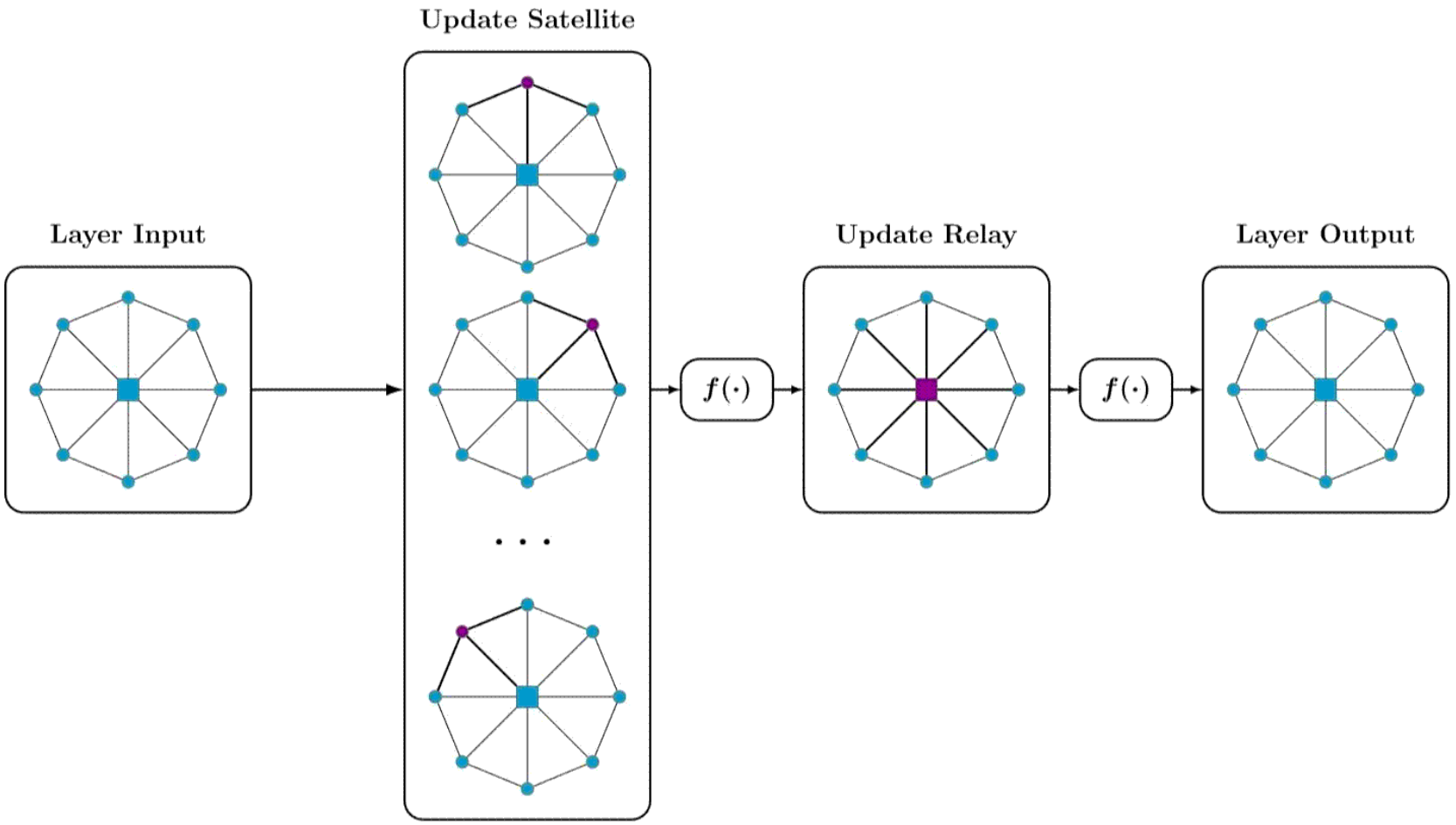
解决思路

Star-Transformer

- Transformer
  - 全连接
  - 复杂度： $O(L^2 d)$
    - 无法处理长文档
  - 容易过拟合
- Star-Transformer
  - 复杂度： $O(2Ld)$
  - 引入局部性先验
    - 不需要Position Embedding
  - 适用于小规模或中等规模数据



Star-Transformer



预训练模型

预训练模型之前

- 深度学习在自然语言处理中的“困境”

What

- 模型并不深，通常为1~2层BiLSTM + Attention
  - 除机器翻译模型外
- 深的模型无法带来更多的收益

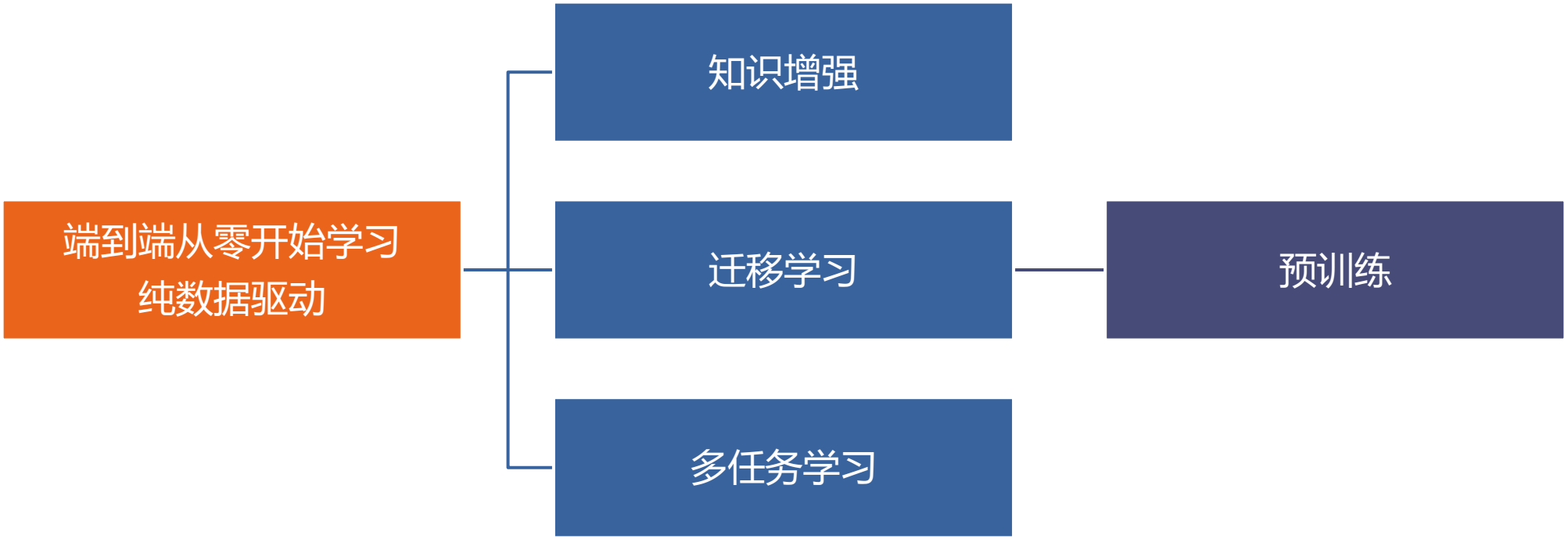
Why

- 缺少大规模的标注数据
- 标注代价太高

How

- 引入知识
- 无监督预训练/迁移学习
- 多任务学习

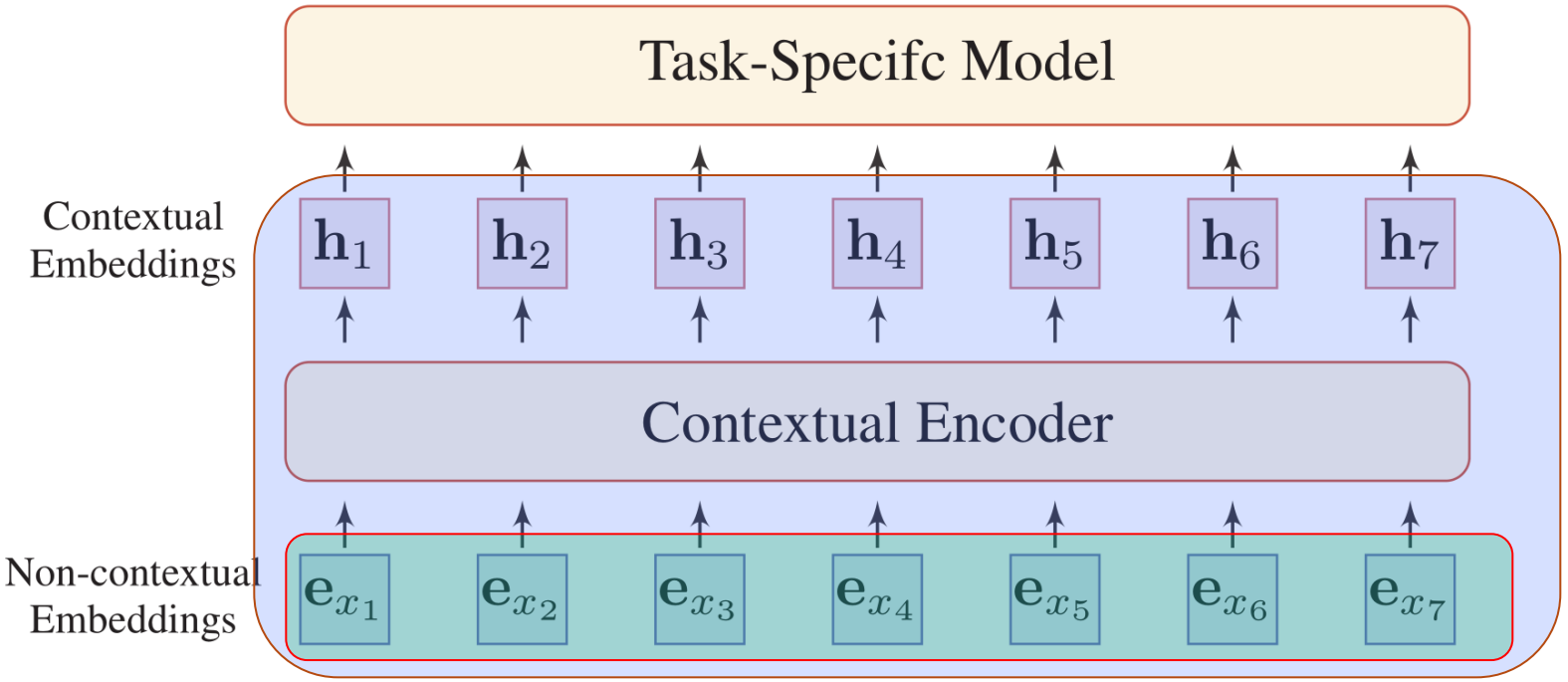
学习方法



为什么要预训练?

- 通用语言表示（universal language representations）
- 获得一个好的初始化
- 预训练可以看作一种正则化方法

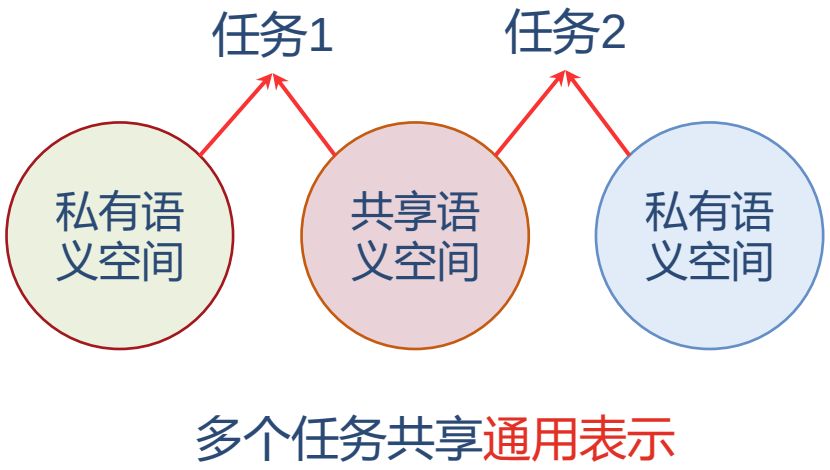
NLP中神经网络模型的一般架构

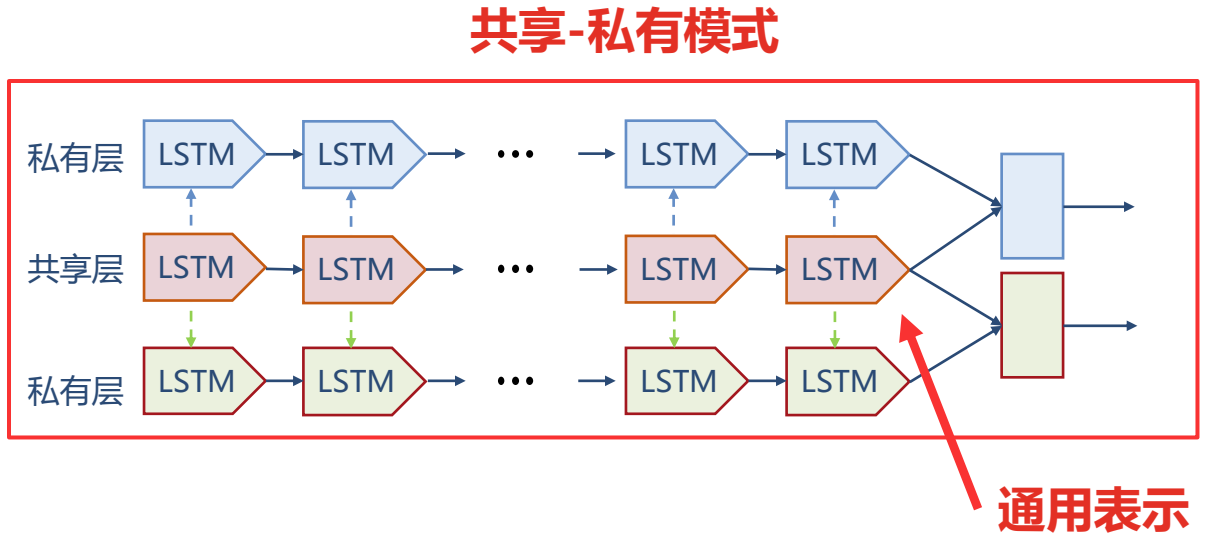


第二代预训练模型

第一代预训练模型

基于多任务的通用表示学习

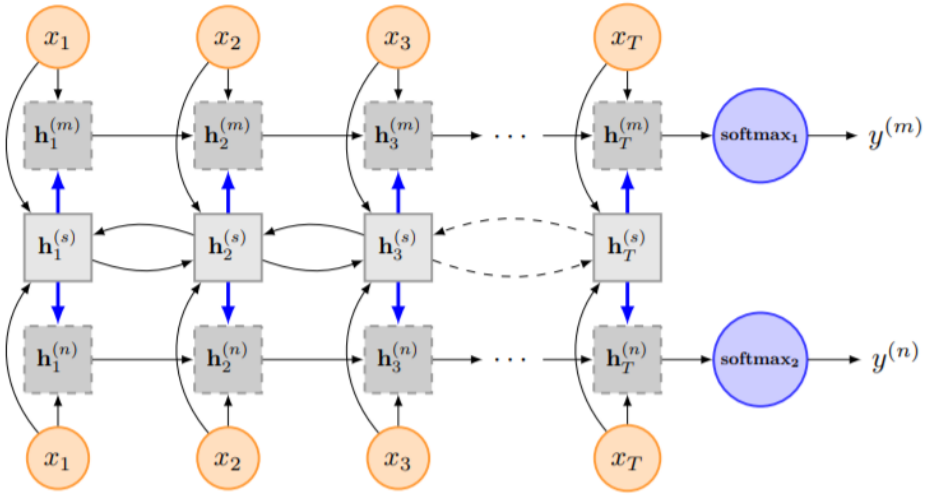




多个任务中，其中一个任务可以设置为语言模型，相当于半监督学习或通用表示的预训练

Pengfei Liu, Xipeng Qiu, Xuanjing Huang, Recurrent Neural Network for Text Classification with Multi-Task Learning, IJCAI 2016, https://arxiv.org/abs/1605.05101

基于多任务的通用表示学习



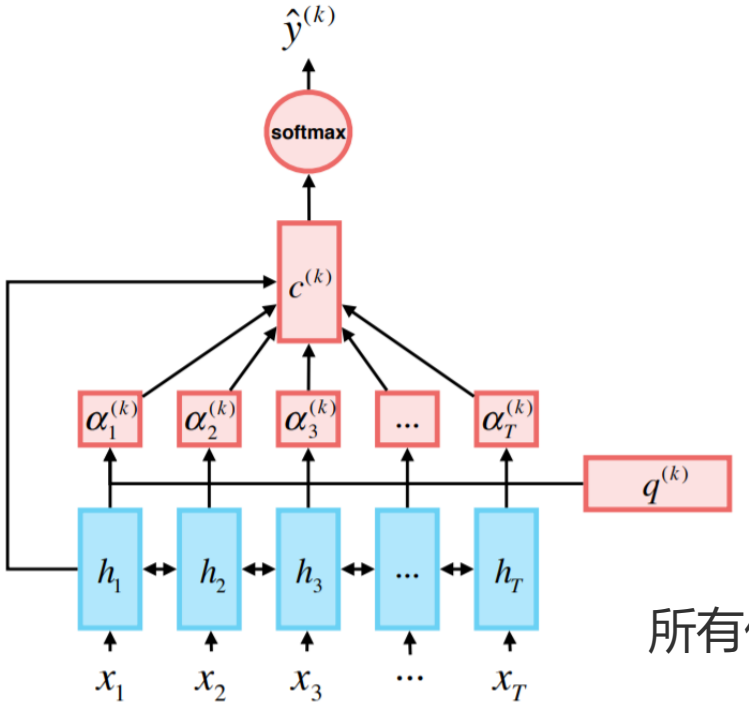
Pre-training of the shared layer with neural language model For model-III, the shared layer can be initialized by an unsupervised pre-training phase. Here, for the shared LSTM layer in Model-III, we initialize it by a language model [Bengio *et al.*, 2007], which is trained on all the four task dataset.

Model	SST-1	SST-2	SUBJ	IMDB	AvgΔ
Single Task	45.9	85.8	91.6	88.5	-
Joint Learning	47.1	87.0	92.5	90.7	+1.4
+ LM	47.9	86.8	93.6	91.0	+1.9
+ Fine Tuning	**49.6**	**87.9**	**94.1**	**91.3**	+2.8

Pengfei Liu, Xipeng Qiu, Xuanjing Huang, Recurrent Neural Network for Text Classification with Multi-Task Learning, IJCAI 2016, https://arxiv.org/abs/1605.05101

基于多任务的通用表示学习

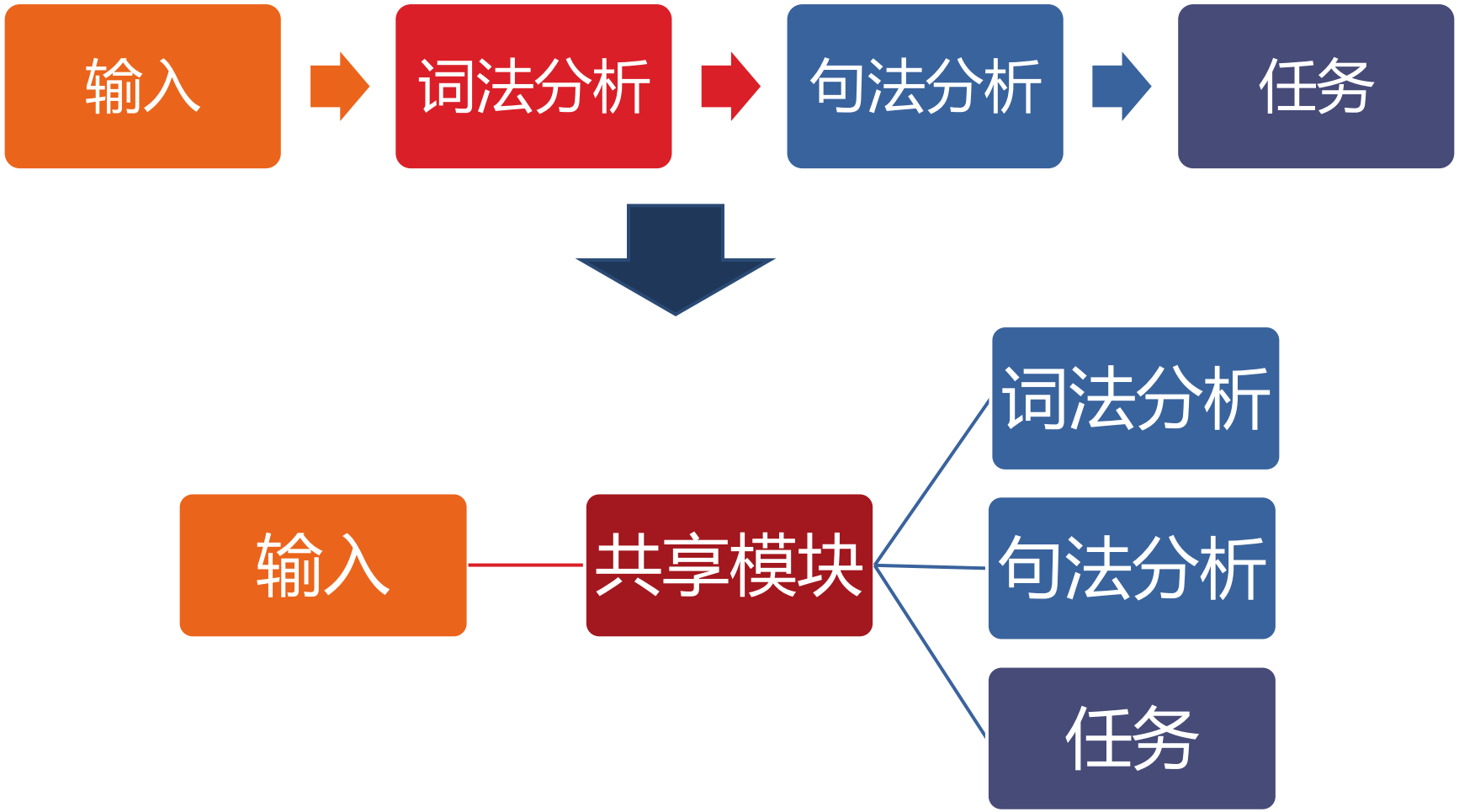
The infantile cart is easy to use.



所有任务共享统一的通用表示

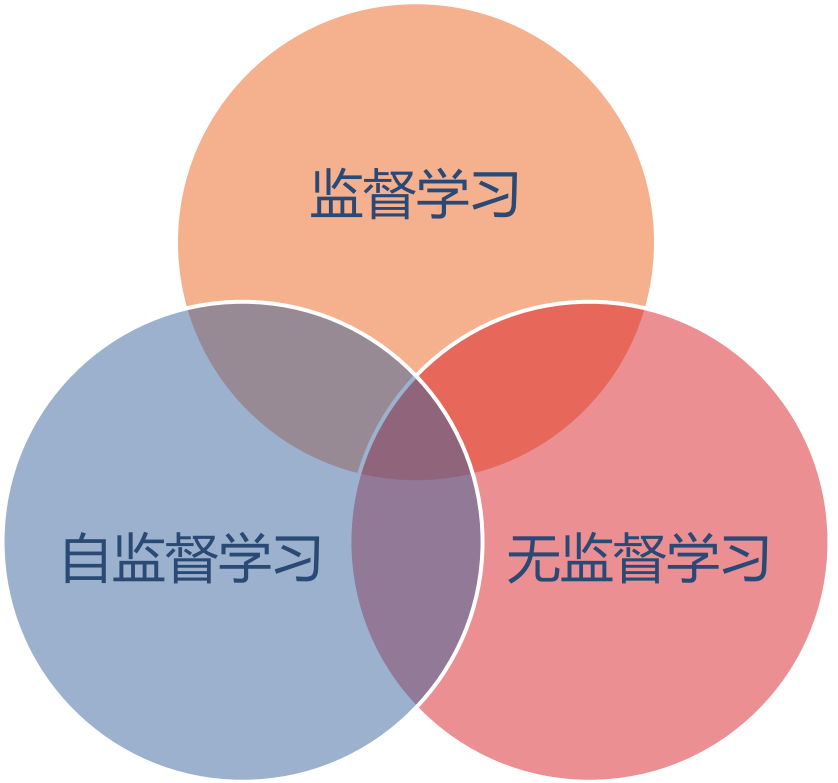
Renjie Zheng, Junkun Chen, Xipeng Qiu, Same Representation, Different Attentions: Shareable Sentence Representation Learning from Multiple Tasks, IJCAI 2018, https://arxiv.org/abs/1804.08139

基于多任务的通用表示学习



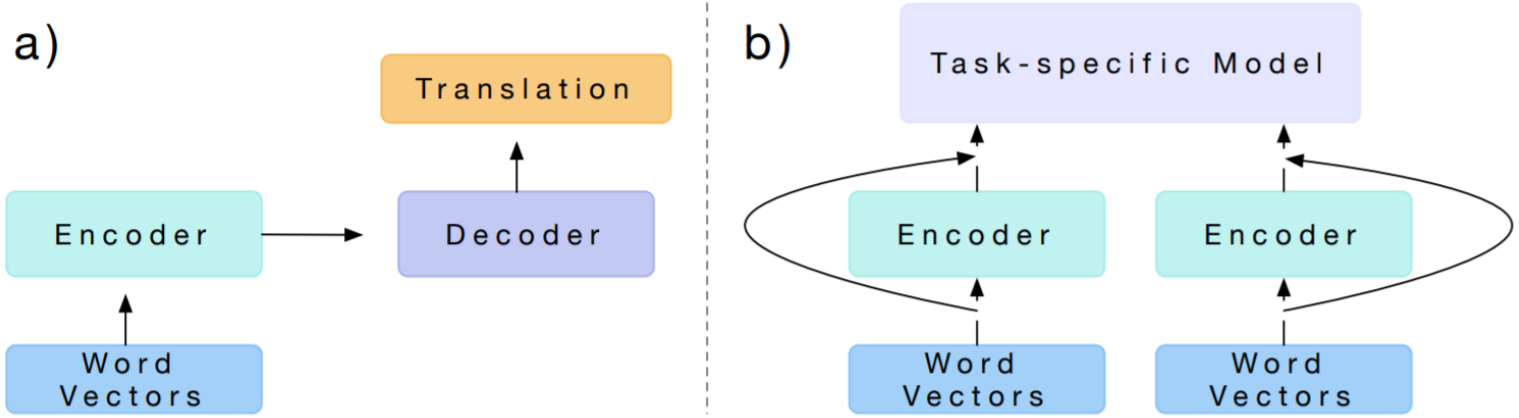
预训练任务

预训练任务



监督学习

- 有大量标注数据的监督任务
  - 机器翻译
  - 机器阅读理解



Context Vectors (CoVe)

McCann B, Bradbury J, Xiong C, et al. Learned in translation: Contextualized word vectors[C]//Advances in Neural Information Processing Systems. 2017.

无监督学习

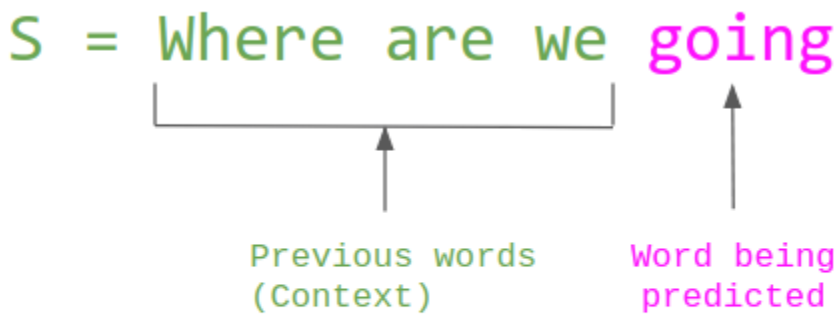
- 密度估计
  - （统计）语言模型
- 重构
  - 自编码器

统计语言模型

- 自然语言理解 → 一个句子的可能性/合理性
  - ！ 在报那猫告做只 
  - 那只猫在作报告！ 
  - 那个人在作报告！ 

- 句子的合理性可以用**概率**来衡量
  - $P(x_1, x_2, \cdots, x_n)$
  - $= \prod_i P(x_i | x_{i-1}, \cdots, x_1)$
  - $\approx \prod_i P(x_i | x_{i-1}, \cdots, x_{i-n+1}) = \mathrm{g}(\mathrm{h_t})$

语言模型

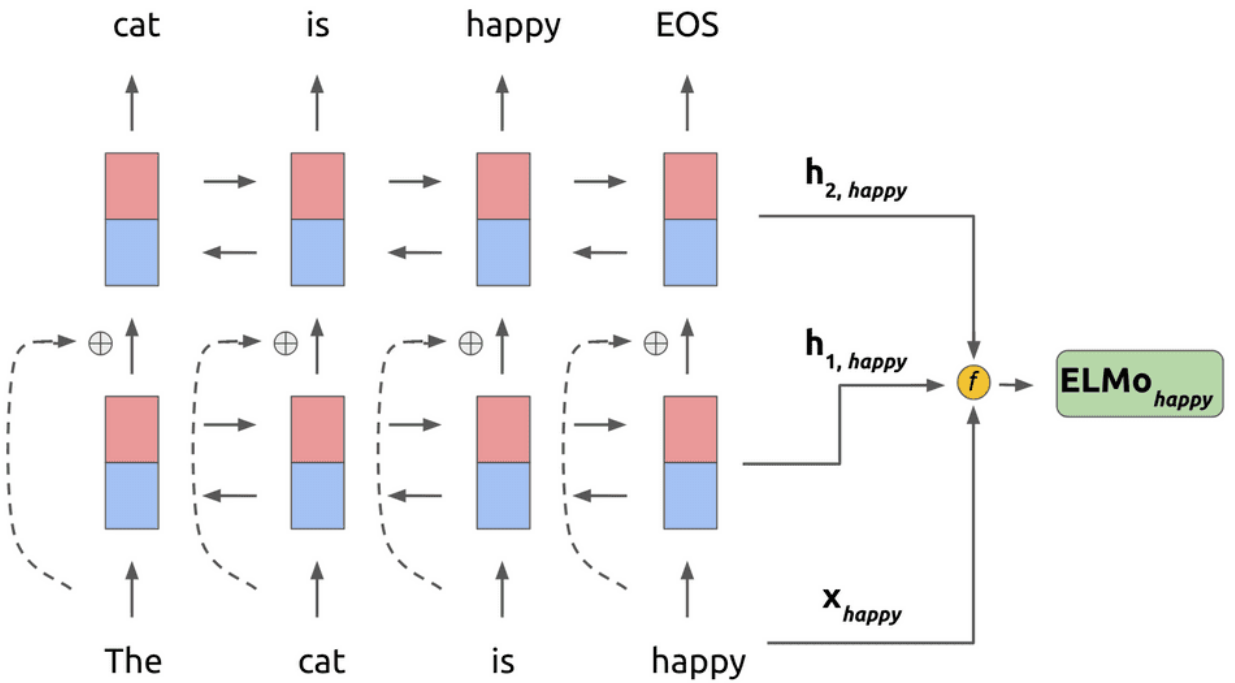


$$P(S) = P(\text{Where}) \times P(\text{are} \mid \text{Where}) \times P(\text{we} \mid \text{Where are}) \times P(\text{going} \mid \text{Where are we})$$

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

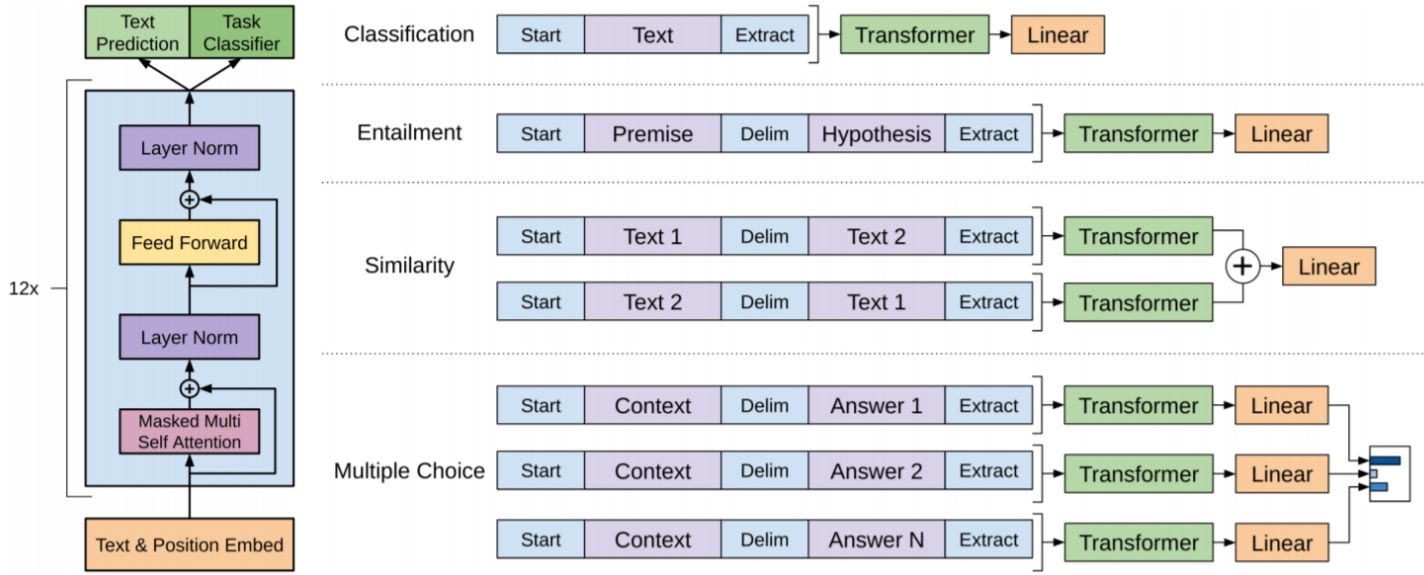


$$ELMo_k^{task} = \gamma_k \cdot (s_0^{task} \cdot x_k + s_1^{task} \cdot h_{1,k} + s_2^{task} \cdot h_{2,k})$$

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

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



$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta)$$

$$U = (u_{-k}, \ldots, u_{-1})$$

$$h_0 = UW_e + W_p$$

$$h_l = \texttt{transformer\_block}(h_{l-1}) \forall i \in [1, n]$$

$$P(u) = \texttt{softmax}(h_n W_e^T)$$

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

GPT-3: Language Models are Few-Shot Learners

https://arxiv.org/abs/2005.14165

The three settings we explore for in-context learning

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  cheese =>                           ← prompt
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

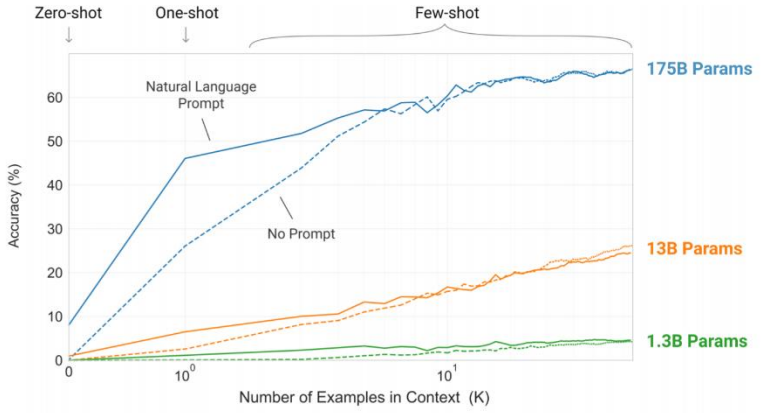
```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← example
3  cheese =>                           ← prompt
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ←
4  plush girafe => girafe peluche      ←
5  cheese =>                           ← prompt
```

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



自监督学习

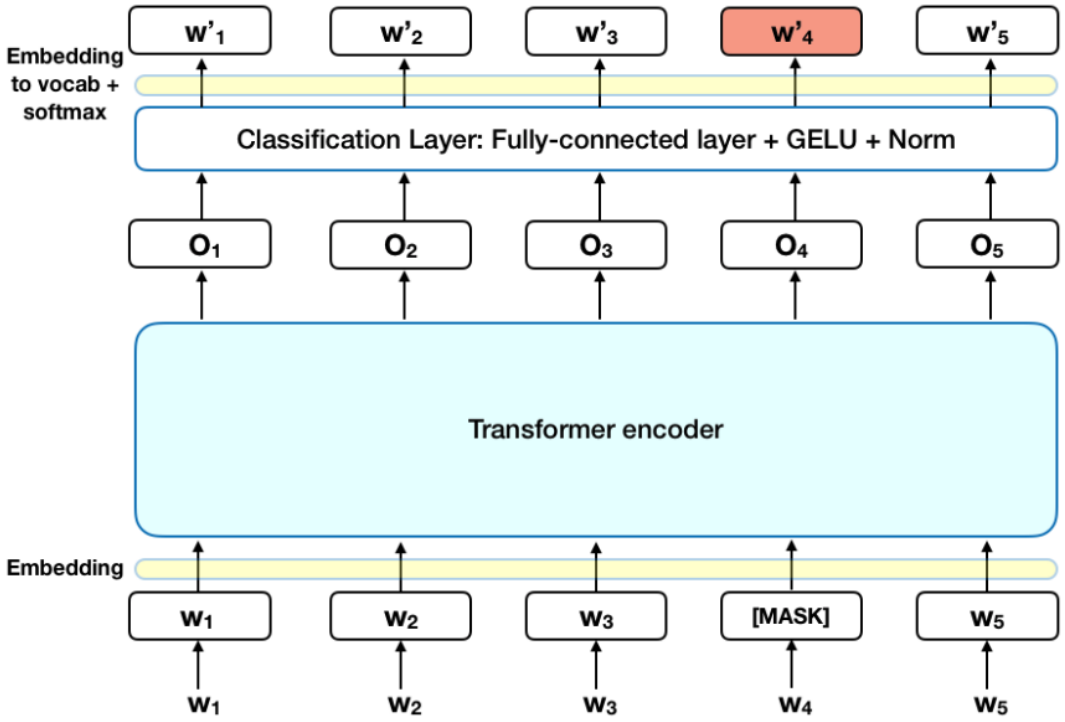
- 一种监督学习和无监督学习的结合
  - 监督式的学习方式
  - 训练数据由无标注数据自动构建
- 典型任务
  - Language Modeling (LM) ?
  - Masked Language Modeling (MLM)
  - Permuted Language Modeling (PLM)
  - Denoising Autoencoder (DAE)
  - Contrastive Learning (CTL)
    - Deep InfoMax (DIM)
    - Replaced Token Detection (RTD)
    - Next Sentence Prediction (NSP)
    - Sentence Order Prediction (SOP)

不依赖标注数据

伪监督任务

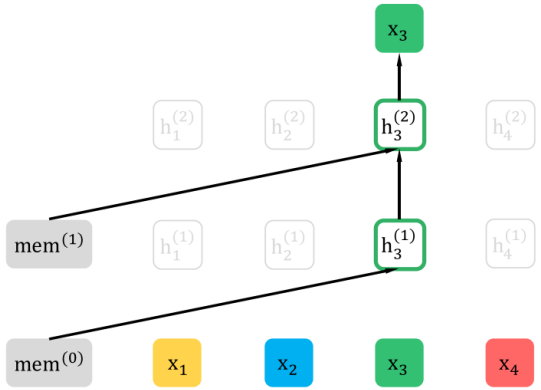
目标：学习数据中的可泛化知识

Masked Language Modeling (MLM)

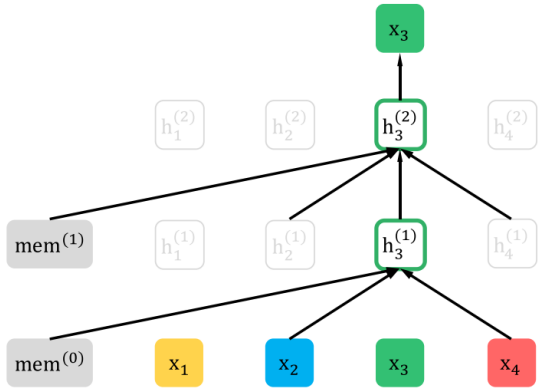


Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

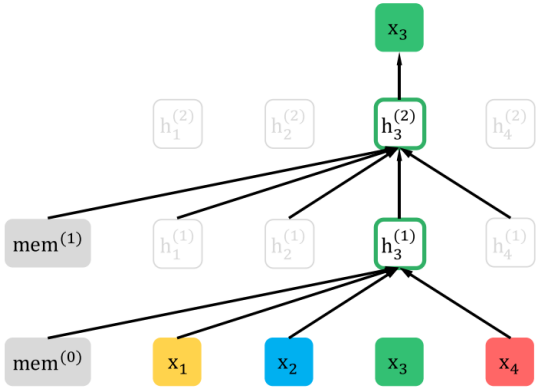
Permutation Language Modeling



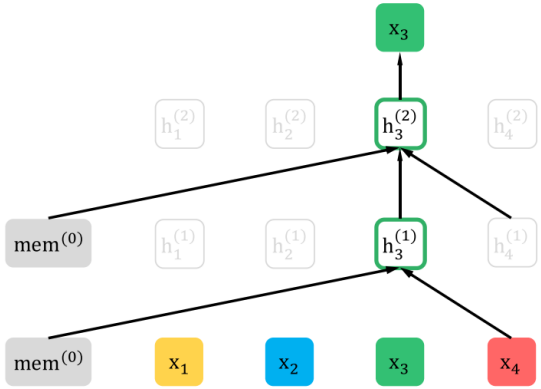
Factorization order: 3 → 2 → 4 → 1



Factorization order: 2 → 4 → 3 → 1



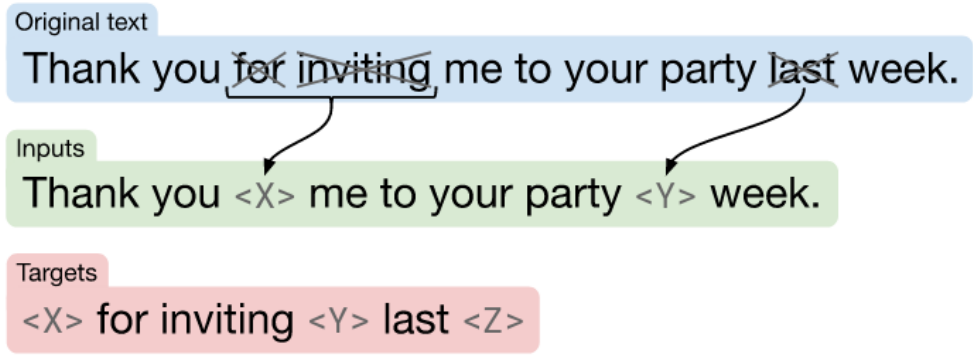
Factorization order: 1 → 4 → 2 → 3



Factorization order: 4 → 3 → 1 → 2

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

Seq2Seq Masked Language Modeling

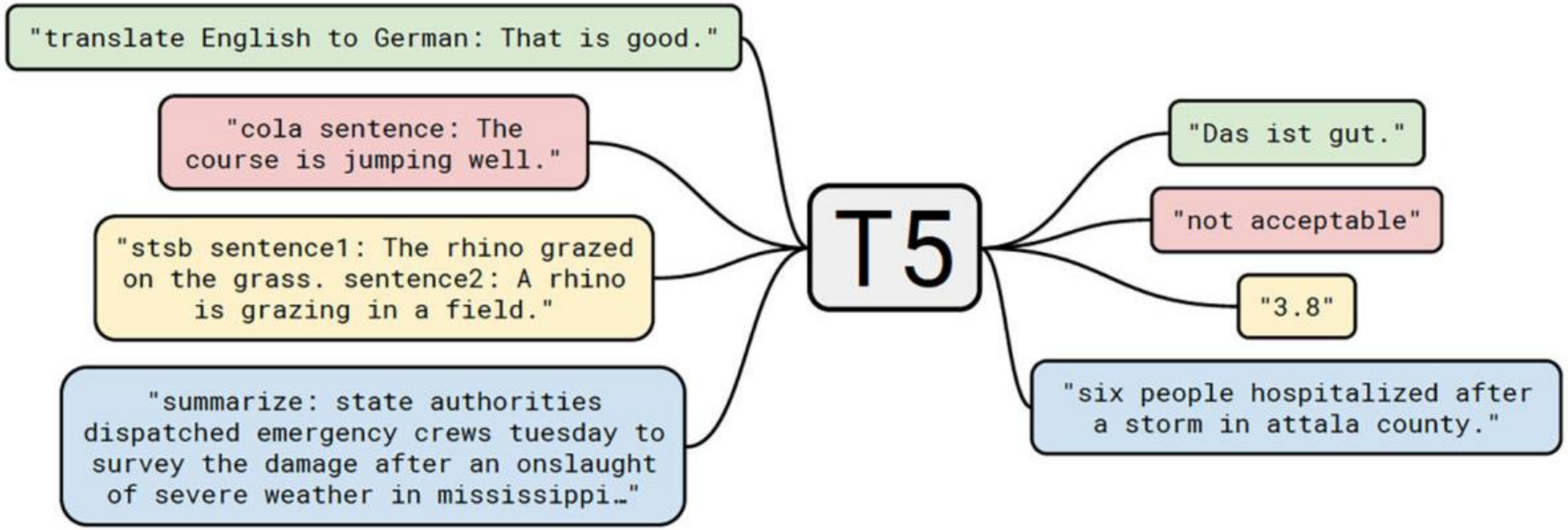


Objective	Inputs	Targets
Prefix language modeling	Thank you for inviting	me to your party last week .
BERT-style	Thank you **<M>** **<M>** me to your party apple week .	*(original text)*
Deshuffling	party me for your to . last fun you inviting week Thank	*(original text)*
I.i.d. noise, mask tokens	Thank you **<M>** **<M>** me to your party **<M>** week .	*(original text)*
I.i.d. noise, replace spans	Thank you **<X>** me to your party **<Y>** week .	**<X>** for inviting **<Y>** last **<Z>**
I.i.d. noise, drop tokens	Thank you me to your party week .	for inviting last
Random spans	Thank you **<X>** to **<Y>** week .	**<X>** for inviting me **<Y>** your party last **<Z>**

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

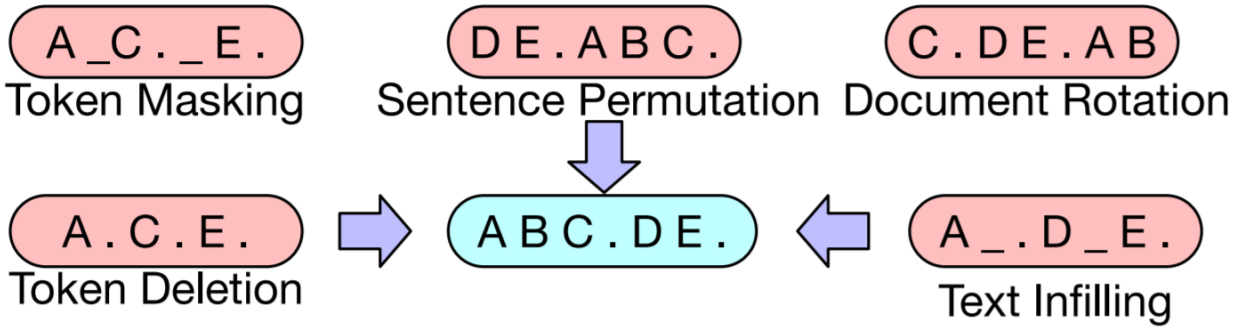
Seq2Seq Masked Language Modeling



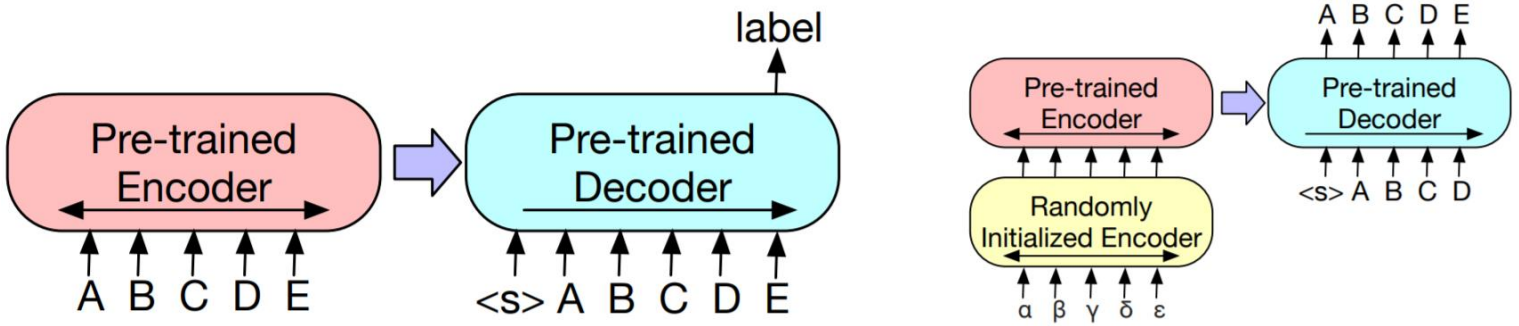
Text-to-Text Transfer Transformer (T5)

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

Denoising Autoencoder (DAE)

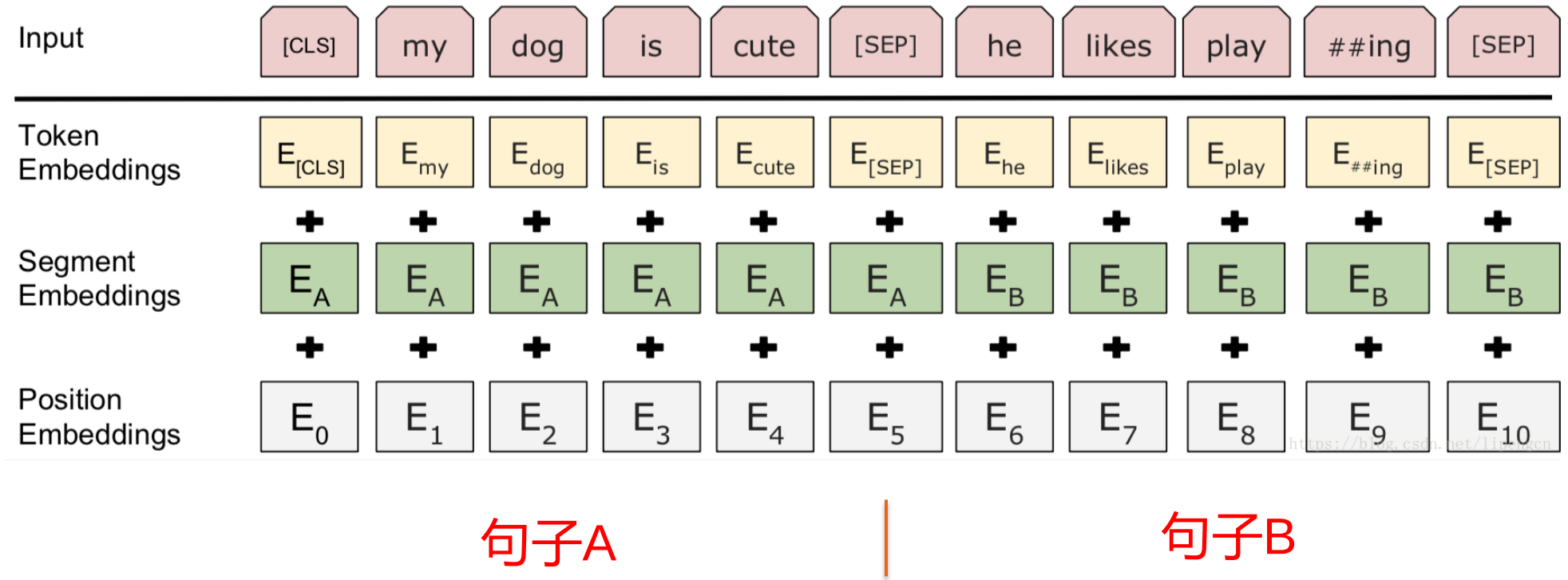


BART: Bidirectional and Auto-Regressive Transformers



Lewis M, Liu Y, Goyal N, et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461, 2019.

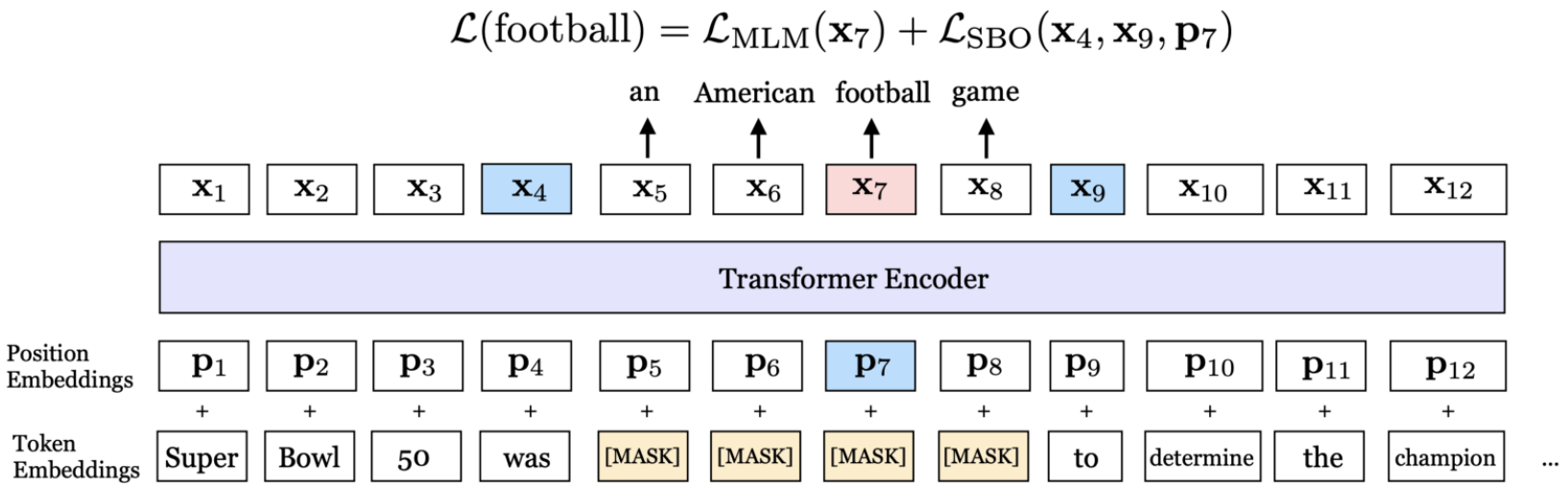
Next Sentence Prediction (NSP)



Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

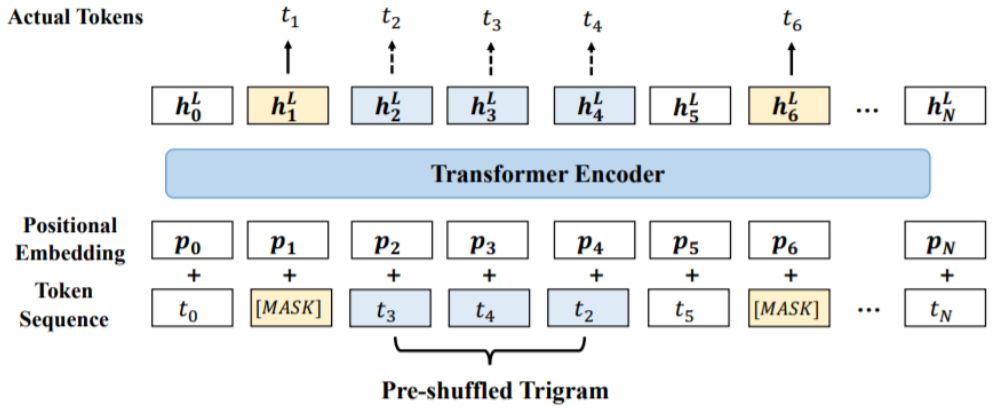
SpanBERT

- 预测一个范围内的所有词
- 去除NSP预训练目标

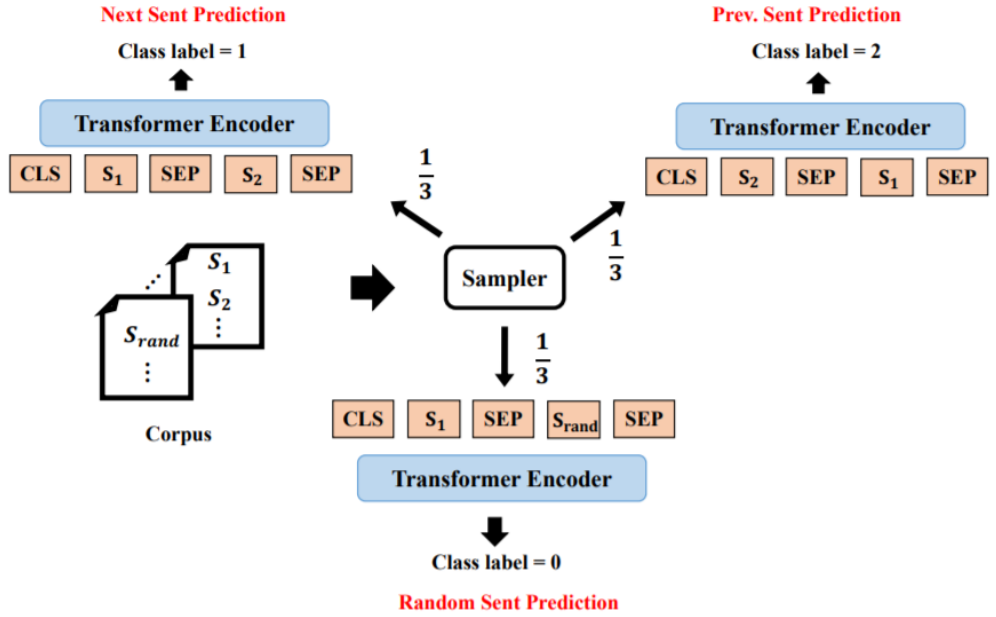


Joshi M, Chen D, Liu Y, et al. Spanbert: Improving pre-training by representing and predicting spans. Transactions of the Association for Computational Linguistics, 2020. https://arxiv.org/abs/1907.10529

StructBERT



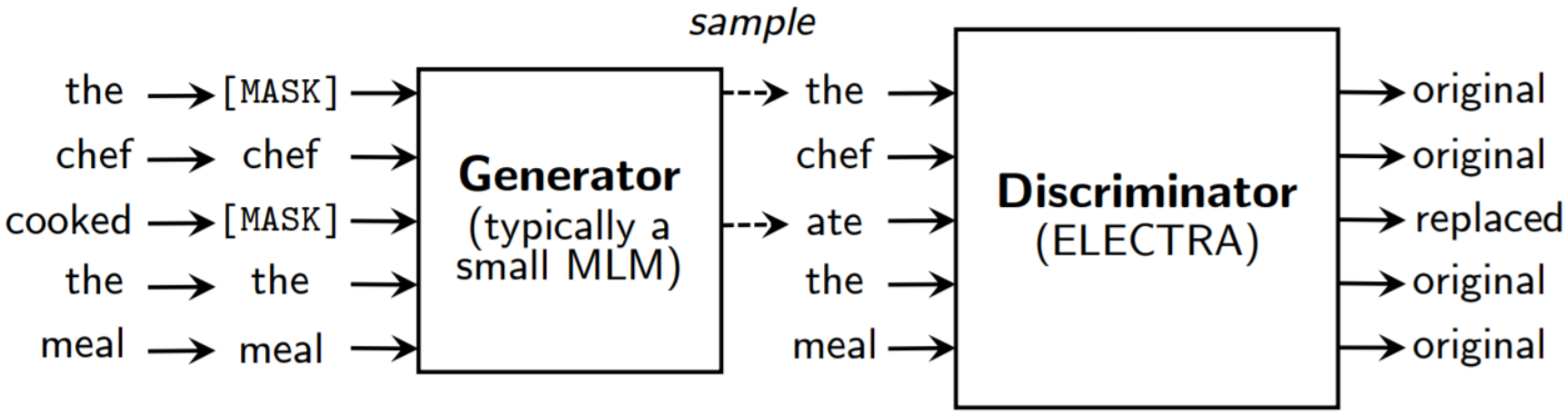
(a) Word Structural Objective



(b) Sentence Structural Objective

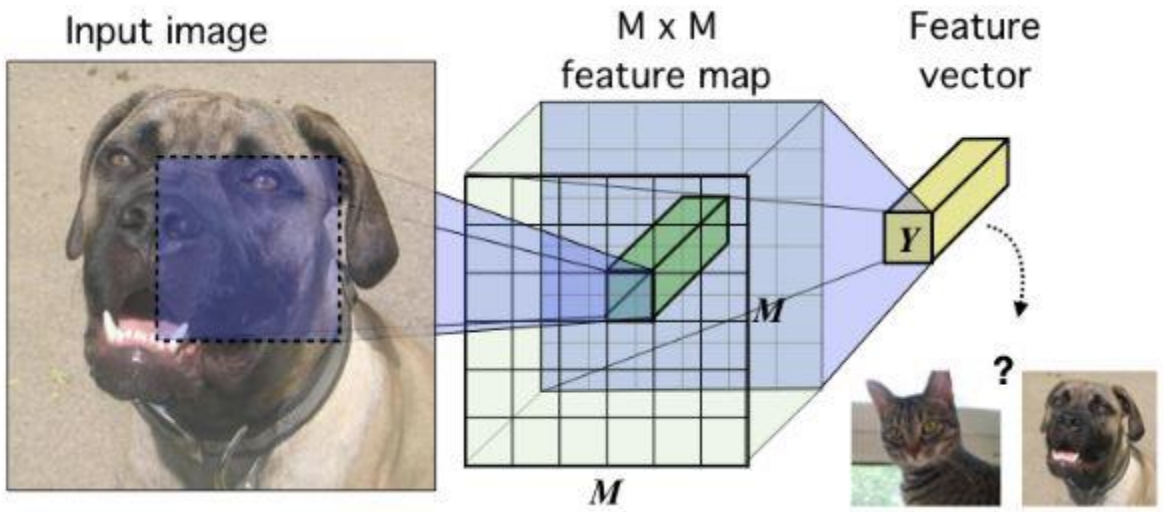
StructBERT: Incorporating Language Structures into Pre-training for Deep Language Understanding, https://arxiv.org/abs/1908.04577

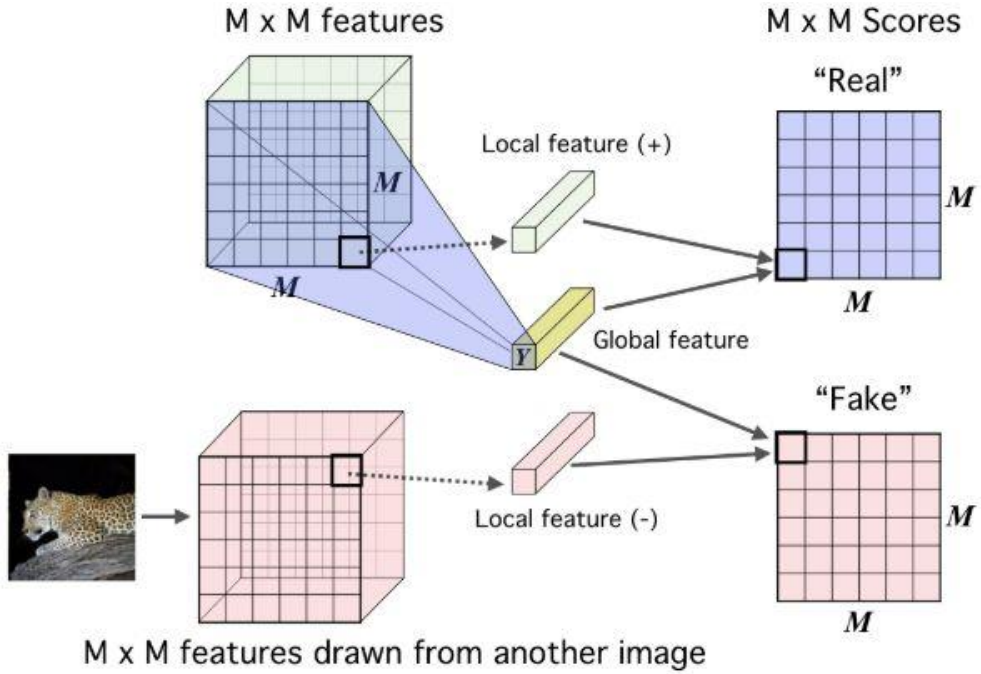
Replaced Token Detection (RTD)



Clark K, Luong M T, Le Q V, et al. Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555, 2020.

Deep InfoMax (DIM)





Hjelm R D, Fedorov A, Lavoie-Marchildon S, et al. Learning deep representations by mutual information estimation and maximization. arXiv preprint arXiv:1808.06670, 2018.

Deep InfoMax (DIM)

INFOWORD

$$\mathcal{J}_{\text{DIM}} = \mathbb{E}_{p(\hat{\boldsymbol{x}}_{i:j}, \boldsymbol{x}_{i:j})} \left[g_{\boldsymbol{\omega}}(\hat{\boldsymbol{x}}_{i:j})^\top g_{\boldsymbol{\omega}}(\boldsymbol{x}_{i:j}) - \log \sum_{\tilde{\boldsymbol{x}}_{i:j} \in \tilde{\mathbb{S}}} \exp(g_{\boldsymbol{\omega}}(\hat{\boldsymbol{x}}_{i:j})^\top g_{\boldsymbol{\omega}}(\tilde{\boldsymbol{x}}_{i:j})) \right]$$

$\boldsymbol{x} = \{x_1, x_2, \ldots, x_T\}$

$\hat{\boldsymbol{x}}_{i:j}$ a masked sequence masked at position i to j

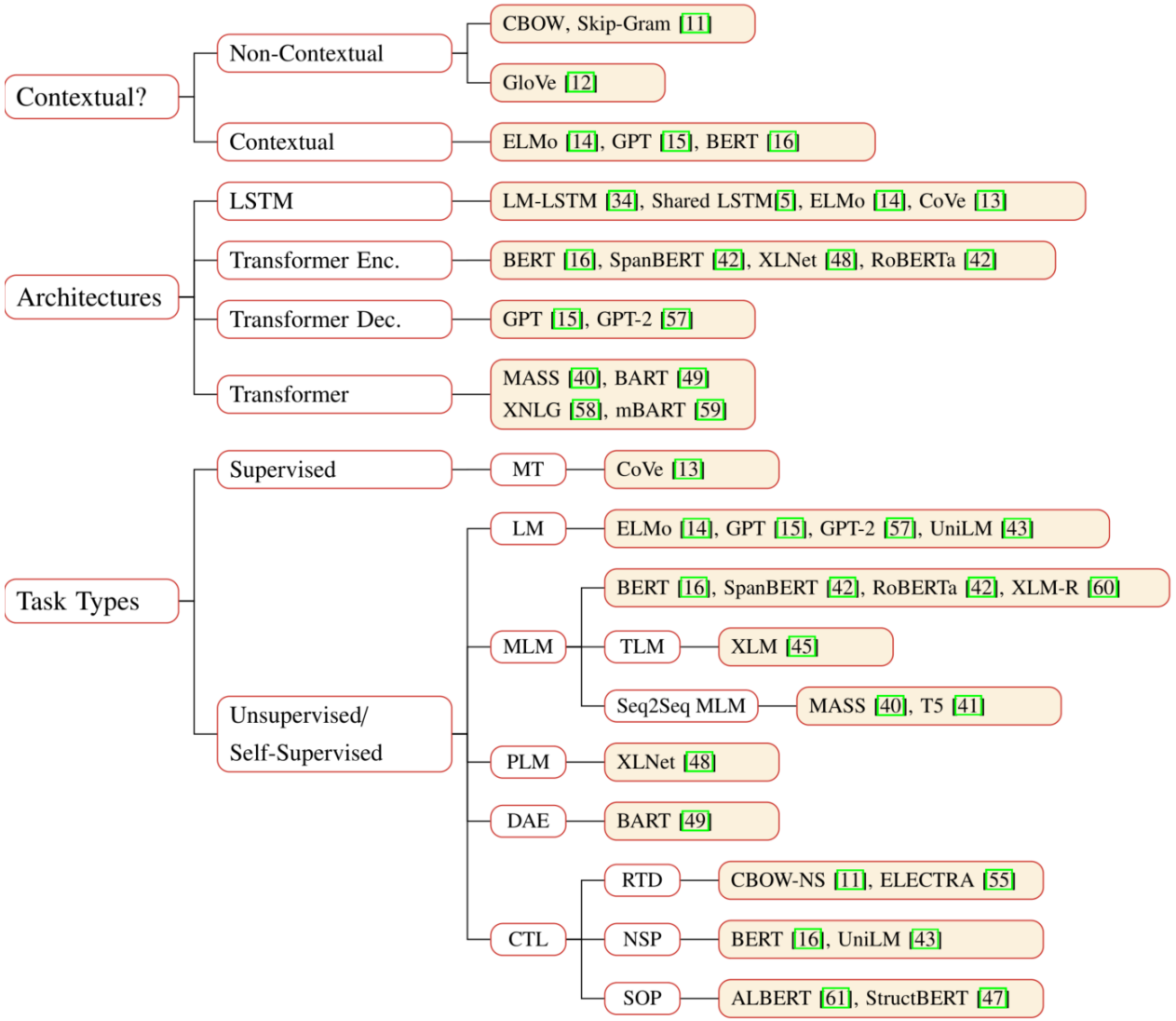
Kong L, d'Autume C M, Ling W, et al. A Mutual Information Maximization Perspective of Language Representation Learning. ICLR 2019.

预训练任务汇总

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

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

小结


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]

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 [57]	Transformer Dec.	Text	LM	WebText	117M ~ 1542M		No
BERT [16]	Transformer Enc.	Text	MLM & NSP	WikiEn+BookCorpus	110M ~ 340M	81.9*	Yes
InfoWord [54]	Transformer Enc.	Text	DIM+MLM	WikiEn+BookCorpus	=BERT	81.1*	Yes
RoBERTa [42]	Transformer Enc.	Text	MLM	BookCorpus+CC-News+OpenWebText+ STORIES	355M	88.5	Yes
XLNet [48]	Two-Stream Transformer Enc.	Text	PLM	WikiEn+ BookCorpus+Giga5 +ClueWeb+Common Crawl	≈BERT	90.5§	Yes
ELECTRA [55]	Transformer Enc.	Text	RTD+MLM	same to XLNet	335M	88.6	Yes
UniLM [43]	Transformer Enc.	Text	MLM§+ NSP	WikiEn+BookCorpus	340M	80.8	Yes
MASS [40]	Transformer	Text	Seq2Seq MLM	*Task-dependent			Yes
BART [49]	Transformer	Text	DAE	same to RoBERTa	110% of BERT	88.4*	Yes
T5 [41]	Transformer	Text	Seq2Seq MLM	Colossal Clean Crawled Corpus (C4)	220M ~ 11B	89.7*	Yes

预训练模型的扩展

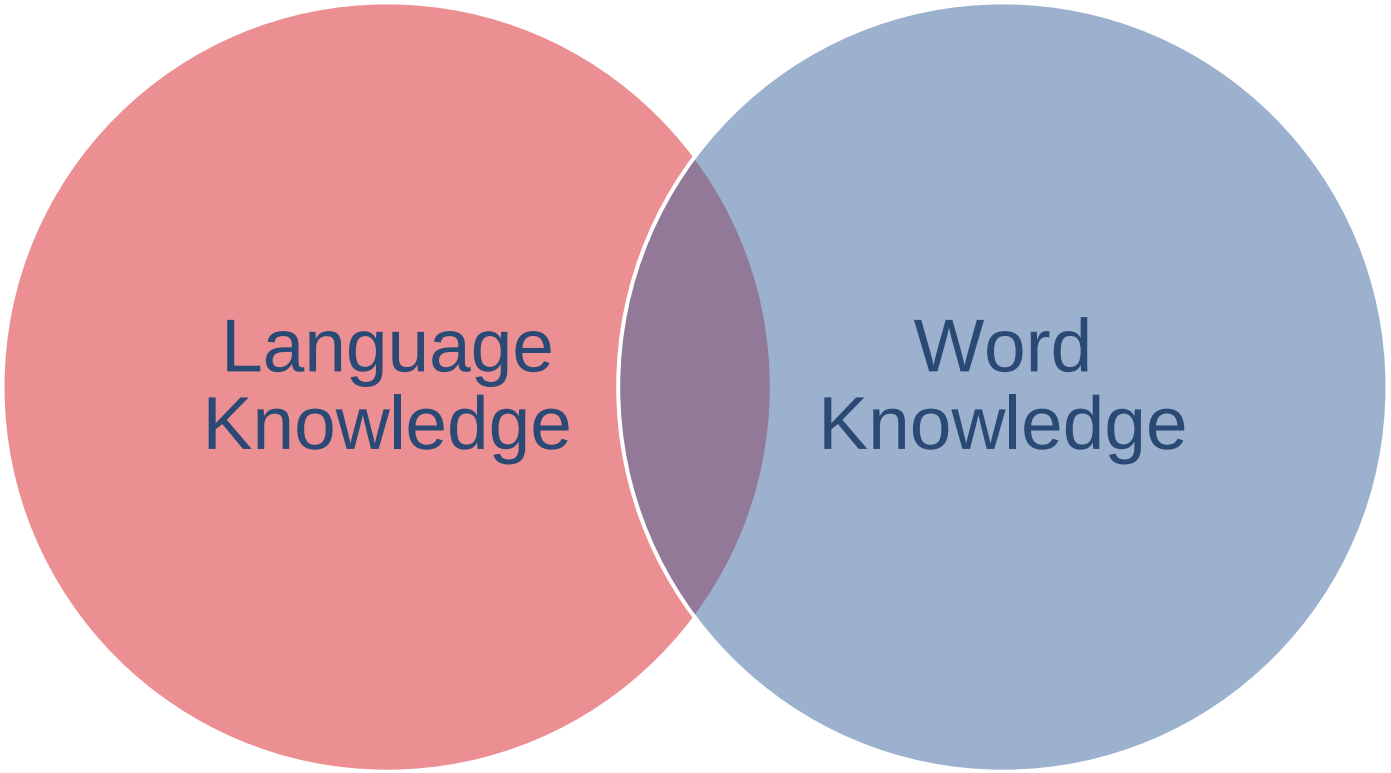
预训练模型的扩展

- Knowledge-Enriched PTMs
- Multilingual and Language-Specific PTMs
- Multi-Modal PTMs
- Domain-Specific and Task-Specific PTMs
- Model Compression

知识增强

预训练模型中的知识

What type of knowledge brings the improvements?



Language Models Need Knowledge

- PLMs perform poorly on entity recognition
  - Contextualized PLMs achieved small improvements on entity & semantic related tasks compared with non-contextualized methods. (Tenney et al.)
  - BPE tokenization breaks entities

```
The native language of Jean Mara ##is is French.
```

  - Surface form-based reasoning

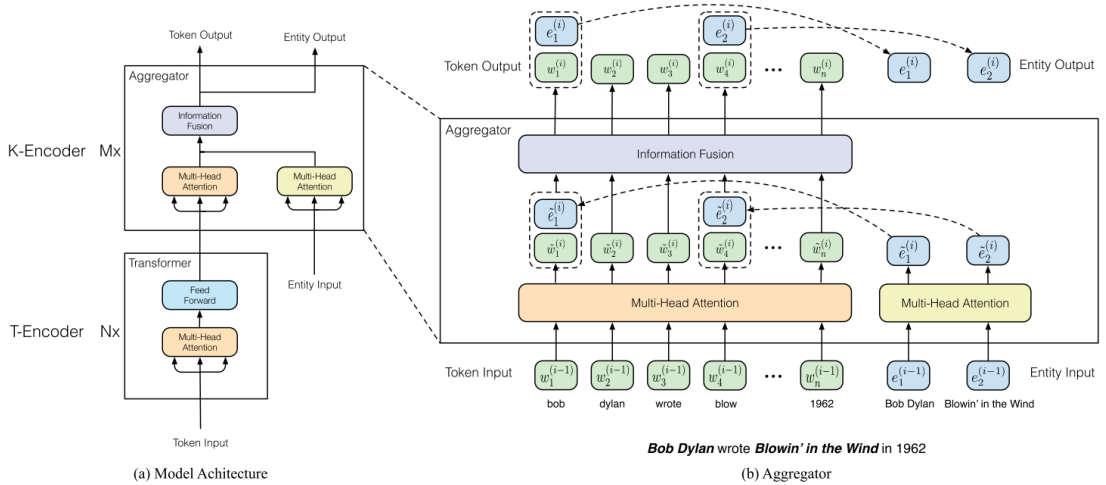
	original BERT	E-BERT-replace	E-BERT-concat	ERNIE	Know-Bert
Jean Marais	French	French	French	french	french
Daniel Ceccaldi	Italian	French	French	french	italian
Orane Demazis	Albanian	French	French	french	french
Sylvia Lopez	Spanish	French	Spanish	spanish	spanish
Annick Alane	English	French	French	english	english

BERT does not know *Daniel Ceccaldi* (as an entity) at all. It just think *Daniel Ceccaldi* looks like an Italian name.

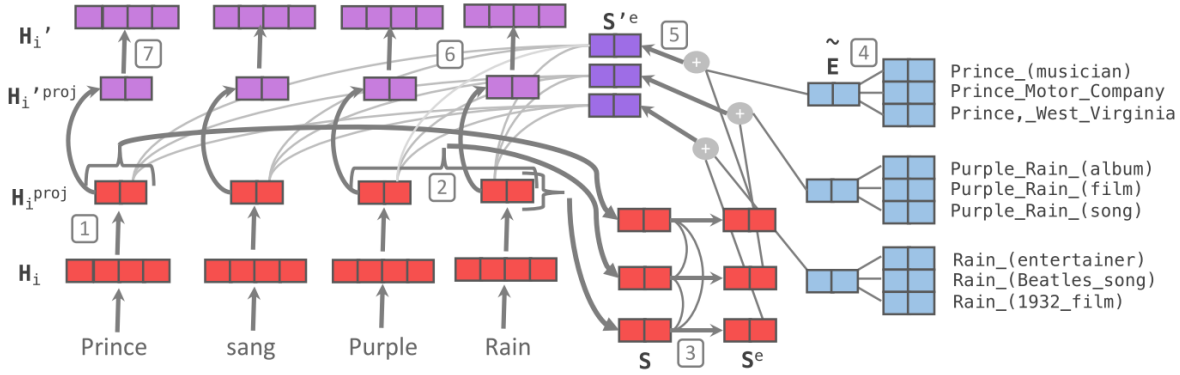
E-BERT: Efficient-Yet-Effective Entity Embeddings for BERT. https://arxiv.org/abs/1911.03681

Injecting Knowledge into Pre-Training

- Injecting entity embeddings
  - ERNIE, KnowBERT, K-BERT
  - The entity embeddings are NOT
    - Jointly learned along with PLM
    - Contextualized
- Knowledge as supervision
  - WKLM



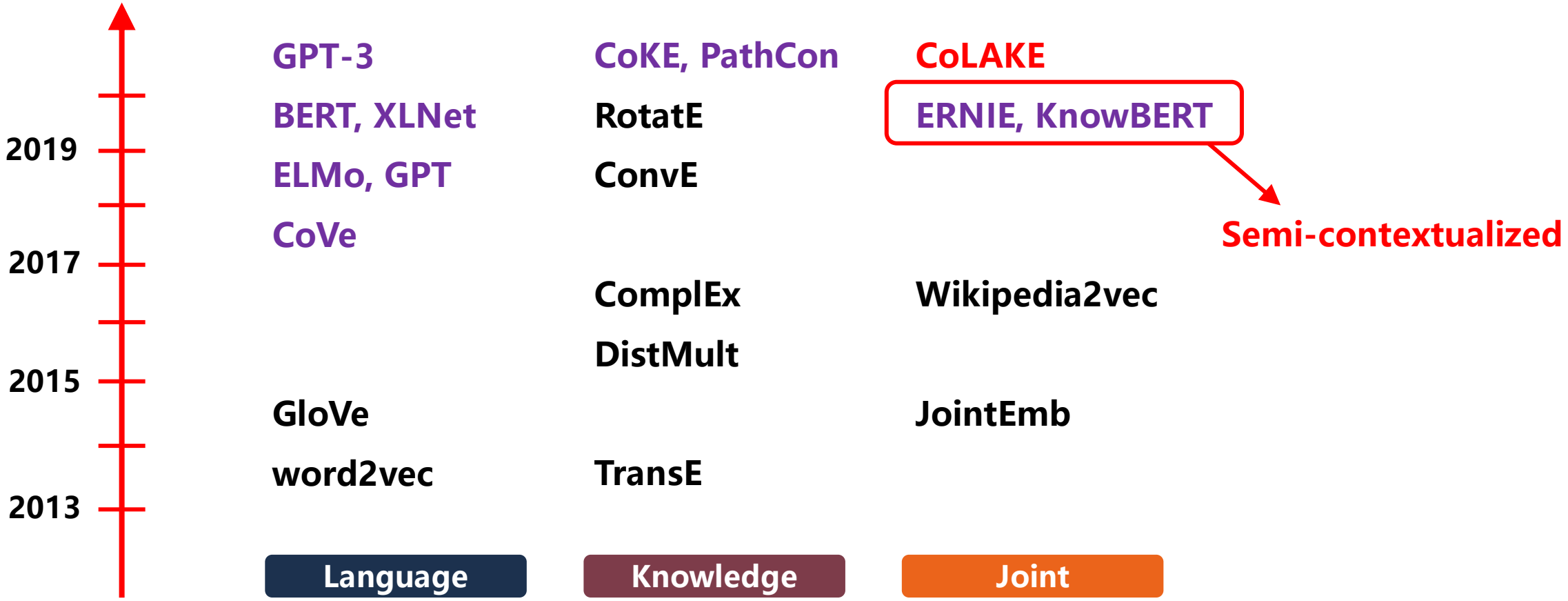
ERNIE: Enhanced Language Representation with Informative Entities
https://arxiv.org/abs/1905.07129



Knowledge Enhanced Contextual Word Representations
https://arxiv.org/abs/1909.04164

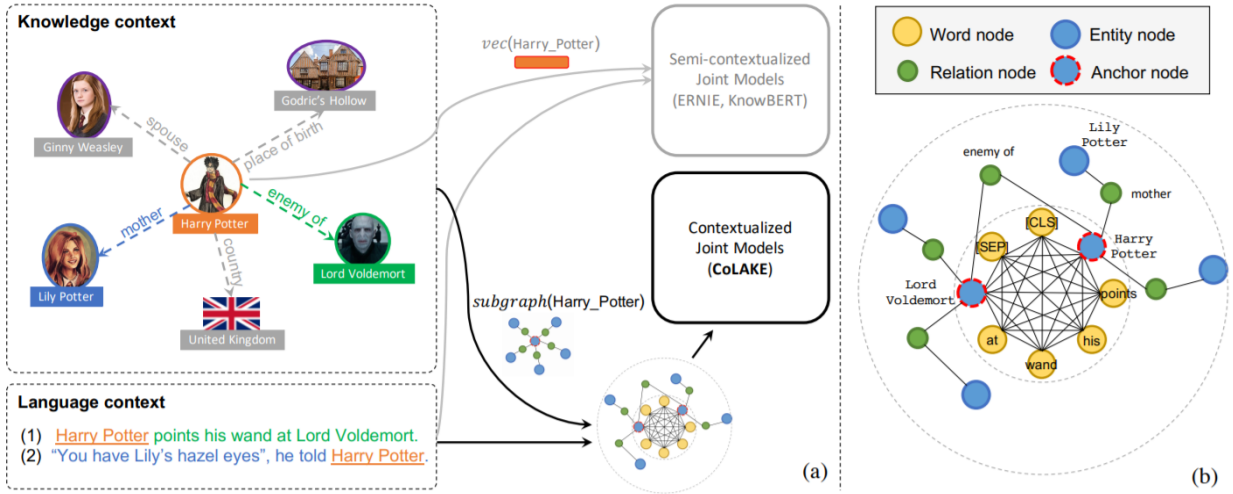
Representation in Language and Knowledge

Combine the success of both sides -- CoLAKE



CoLAKE：预训练的“语言-知识”联合上下文表示

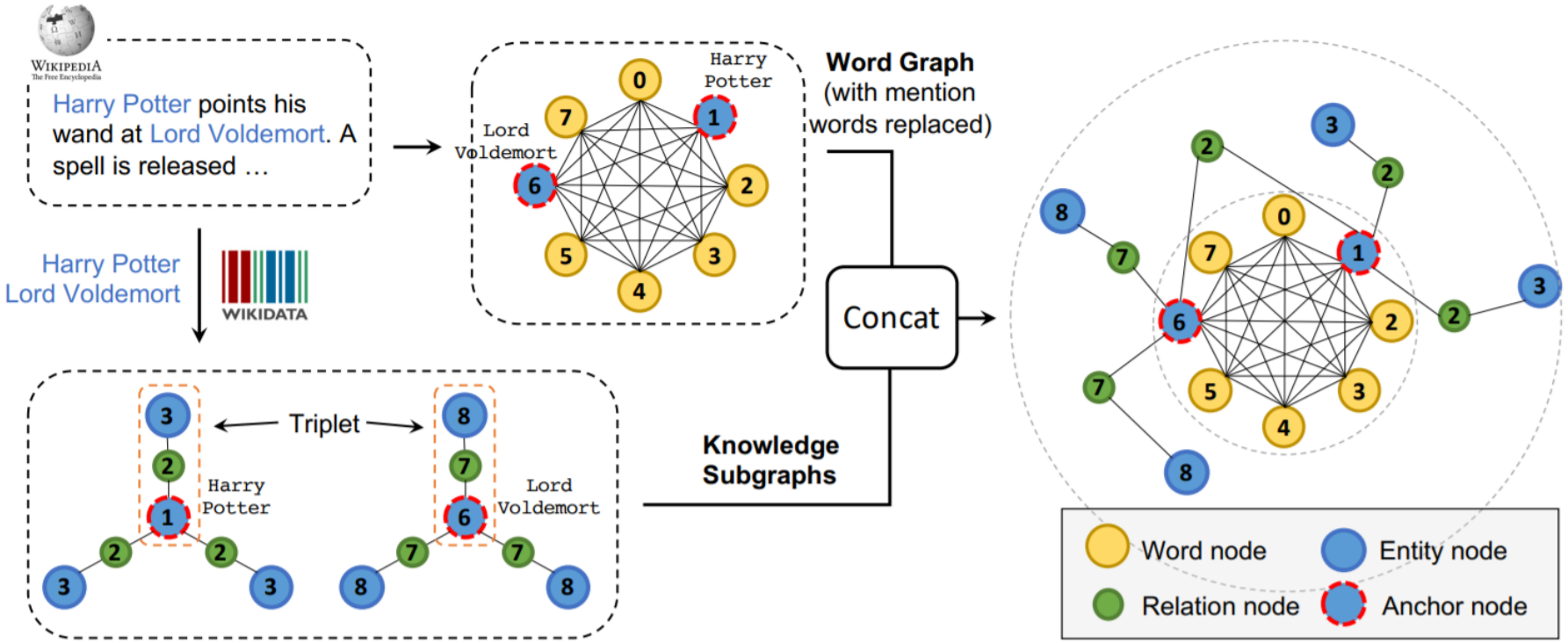
- Why?
  - Different facts should be accessed to help understand different sentences.
- What?
  - Word-knowledge graph is a unified data structure to integrate language context and knowledge context.



Tianxiang Sun, Yunfan Shao, Xipeng Qiu, Qipeng Guo, Yaru Hu, Xuanjing Huang, Zheng Zhang, CoLAKE: Contextualized Language and Knowledge Embedding, COLING 2020, https://arxiv.org/abs/2010.00309

Word-knowledge Graph

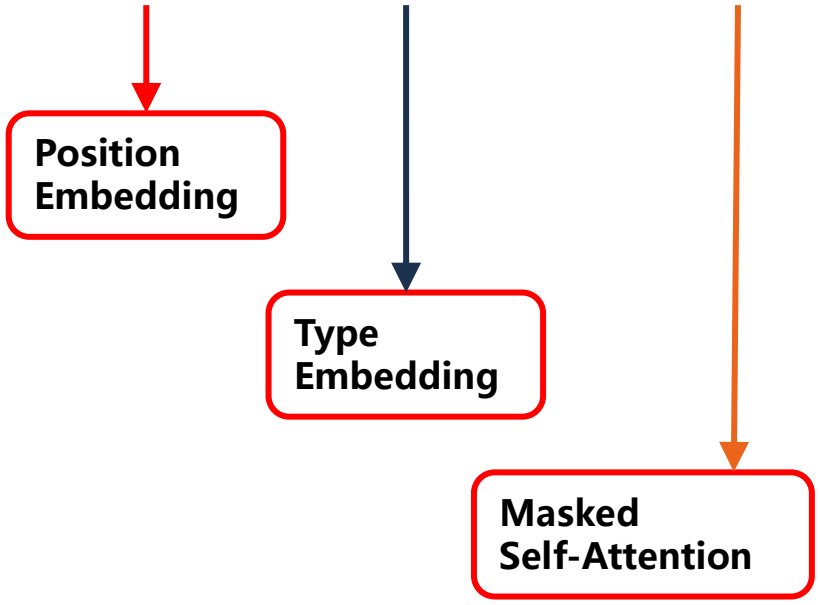
- How?
  - Word graph + Knowledge subgraph

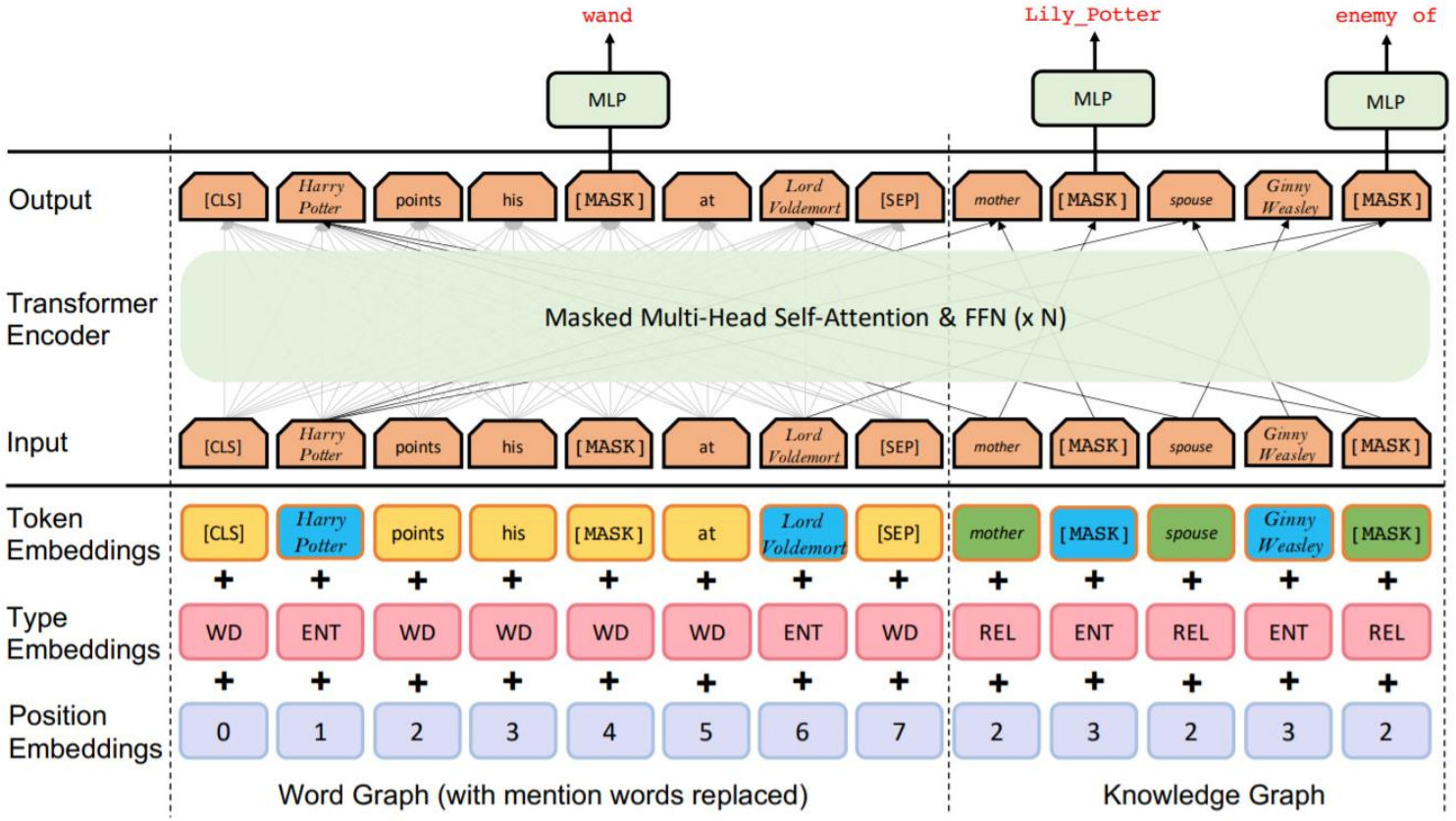


Model Architecture

- Modify Transformer for word-knowledge graph

Word-knowledge graph is a positional heterogeneous graph.



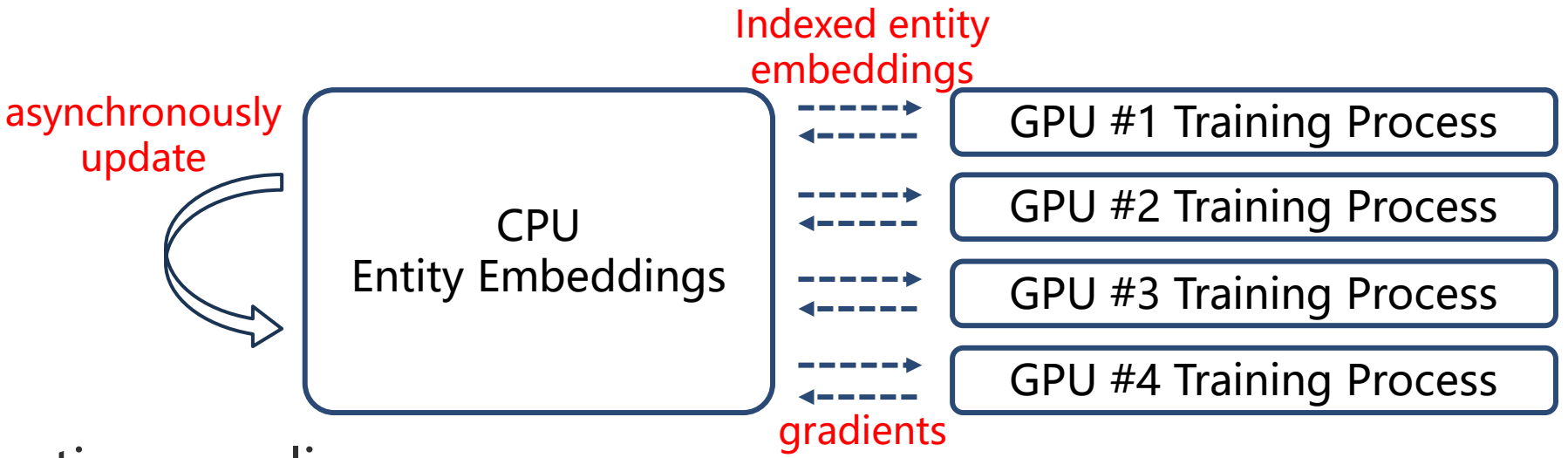


Pre-Training Objective

- Masked Language Model (MLM) on word-knowledge graph
  - Masking word nodes
    - Learn linguistic knowledge
  - Masking entity nodes
    - Anchor nodes masked – Learn to align the two spaces
    - Entity nodes masked – Learn contextualized entity embeddings
  - Masking relation nodes
    - Relation between anchor nodes – Learn to do relation extraction
    - Otherwise – Learn contextualized relation embeddings

Some Details

- Mixed CPU-GPU training



- Negative sampling
  - Sample negative entities from the 3/4 powered entity distribution.
- Alignment of the two spaces
  - Discard neighbors of anchor nodes in 50% of time.
  - Replace mention words with anchor nodes.

Contextualized Representation

	Joint Models	Language		Knowledge	
		Objective	Contextualized?	Objective	Contextualized?
Non-contextual	Wang et al. (2014)	Skip-Gram	✗	TransE	✗
	Yamada et al. (2016)	Skip-Gram	✗	Skip-Gram	✗
Semi-contextualized	ERNIE (Zhang et al., 2019)	MLM	✓	TransE*	✗
	KnowBERT (Peters et al., 2019)	MLM	✓	-	✗
	KEPLER (Wang et al., 2019c)	MLM	✓	TransE	✗
Contextualized	CoLAKE (Ours)	MLM	✓	MLM	✓

Experiments

- Knowledge-driven tasks
  - Entity typing
  - Relation extraction
- Knowledge probing tasks
  - LAMA
  - LAMA-UHN
- Language understanding tasks
  - GLUE
- Synthetic graph task
  - Word-knowledge graph completion

Experiments

- Knowledge-driven tasks

Model	Open Entity			FewRel		
	P	R	F	P	R	F
BERT (Devlin et al., 2019)	76.4	71.0	73.6	85.0	85.1	84.9
RoBERTa (Liu et al., 2019)	77.4	73.6	75.4	85.4	85.4	85.3
ERNIE (Zhang et al., 2019)	78.4	72.9	75.6	88.5	88.4	88.3
KnowBERT (Peters et al., 2019)	**78.6**	73.7	76.1	-	-	-
KEPLER (Wang et al., 2019c)	77.8	74.6	76.2	-	-	-
E-BERT (Pörner et al., 2019)	-	-	-	88.6	88.5	88.5
CoLAKE (Ours)	77.0	**75.7**	**76.4**	**90.6**	**90.6**	**90.5**

- Knowledge probing tasks

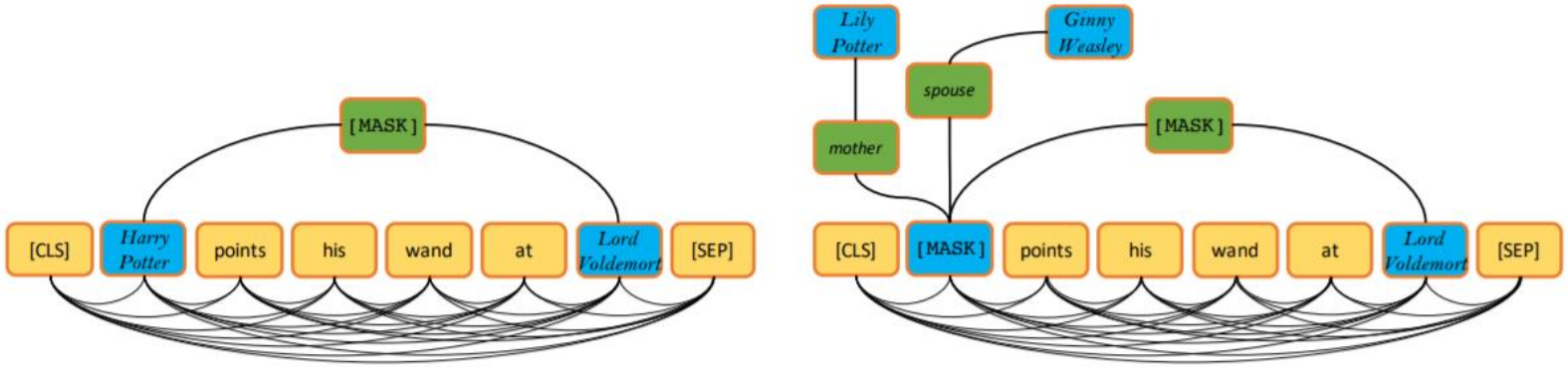
Corpus	Pre-trained Models					
	ELMo	ELMo5.5B	BERT	RoBERTa	CoLAKE	K-Adapter*
LAMA-Google-RE	2.2	3.1	11.4	5.3	9.5	7.0
LAMA-UHN-Google-RE	2.3	2.7	5.7	2.2	4.9	3.7
LAMA-T-REx	0.2	0.3	32.5	24.7	28.8	29.1
LAMA-UHN-T-REx	0.2	0.2	23.3	17.0	20.4	23.0

Experiments

- Language understanding tasks

Model	MNLI (m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	AVG.
RoBERTa	87.5 / 87.3	91.9	92.8	94.8	63.6	91.2	90.2	78.7	86.4
KEPLER	87.2 / 86.5	91.5	92.4	94.4	62.3	89.4	89.3	70.8	84.9
CoLAKE	87.4 / 87.2	92.0	92.4	94.6	63.4	90.8	90.9	77.9	86.3

- Synthetic graph task
  - Word-knowledge graph completion



(a) Transductive setting. (b) Inductive setting.

Experiments

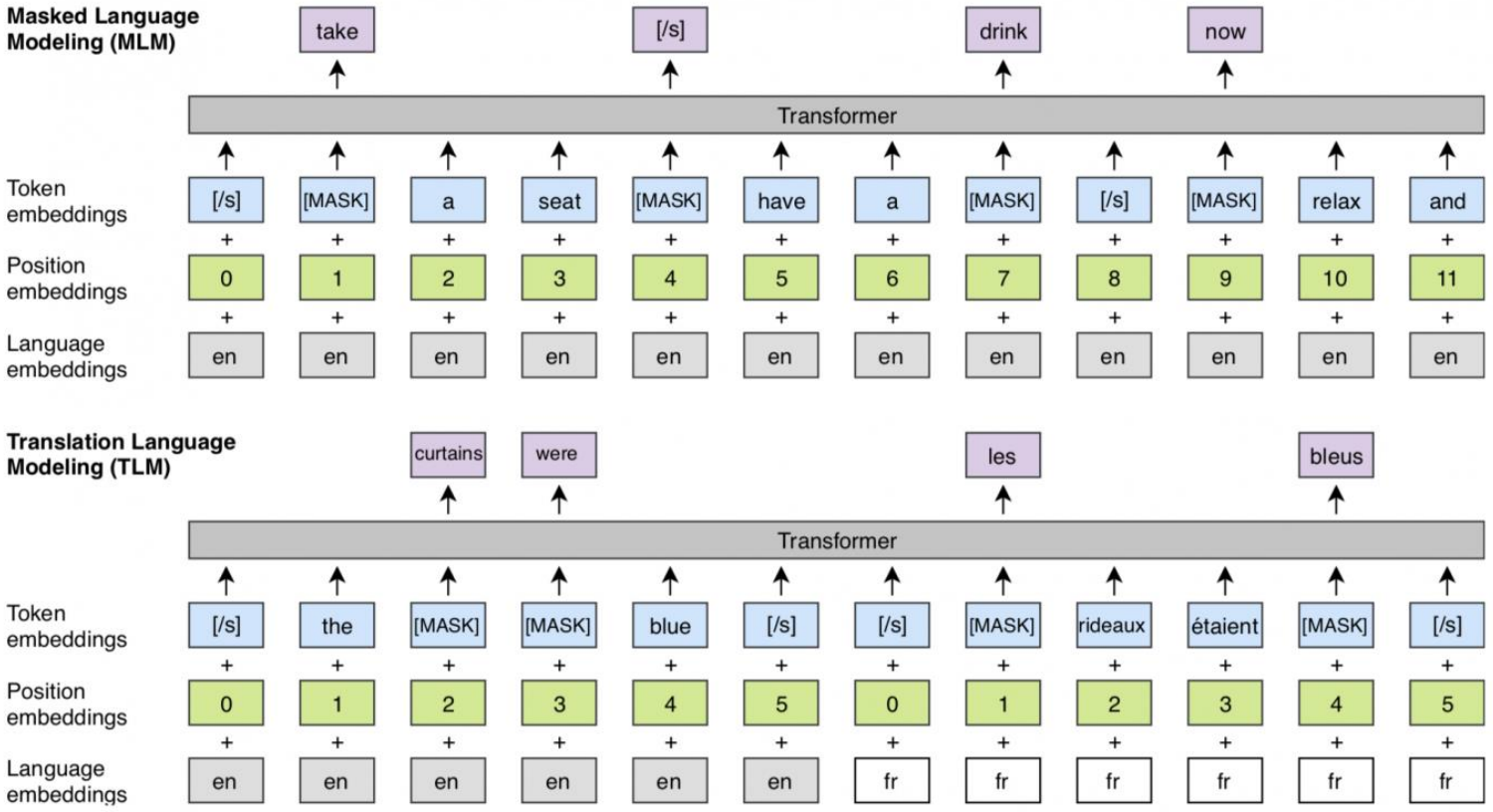
- Results on word-knowledge graph completion
  - CoLAKE is essentially a pre-trained inductive GNN which simultaneously models structural knowledge and text semantics.

Model	MR ↓	MRR	HITS@1	HITS@3	HITS@10
Transductive setting					
TransE (Bordes et al., 2013)	15.97	67.30	60.28	70.96	79.75
DistMult (Yang et al., 2015)	27.09	60.56	48.66	69.69	79.61
ComplEx (Trouillon et al., 2016)	26.73	61.09	49.80	70.64	79.78
RotatE (Sun et al., 2019)	30.36	70.90	64.74	74.89	81.05
CoLAKE	2.03	82.48	72.14	92.19	98.58
Inductive setting					
DKRL (Xie et al., 2016)	168.21	8.18	5.03	7.28	14.13
CoLAKE	31.01	28.10	15.69	30.28	58.05

针对跨语言、特定语言的预训练模型

Multilingual and Language-Specific PTMs

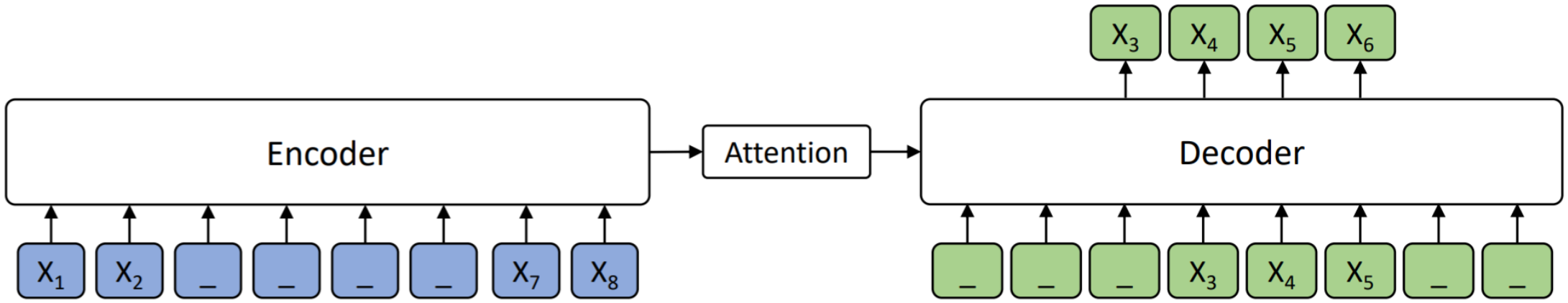
- Cross-Lingual Language Understanding (XLU)



XLM – Enhancing BERT for Cross-lingual Language Model

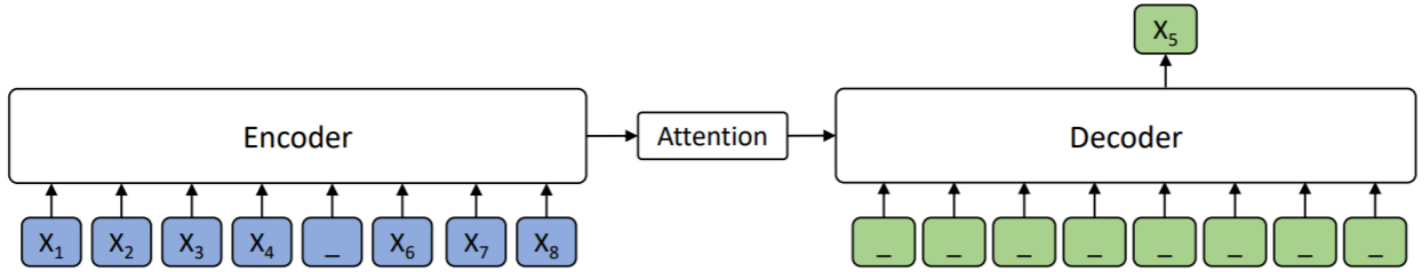
Multilingual and Language-Specific PTMs

- Cross-Lingual Language Generation (XLG)

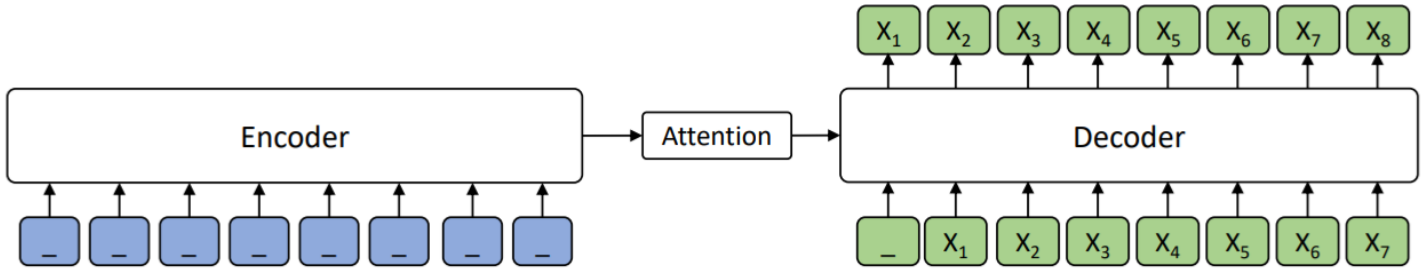


MASS: Masked Sequence to Sequence Pre-training for Language Generation

Multilingual and Language-Specific PTMs



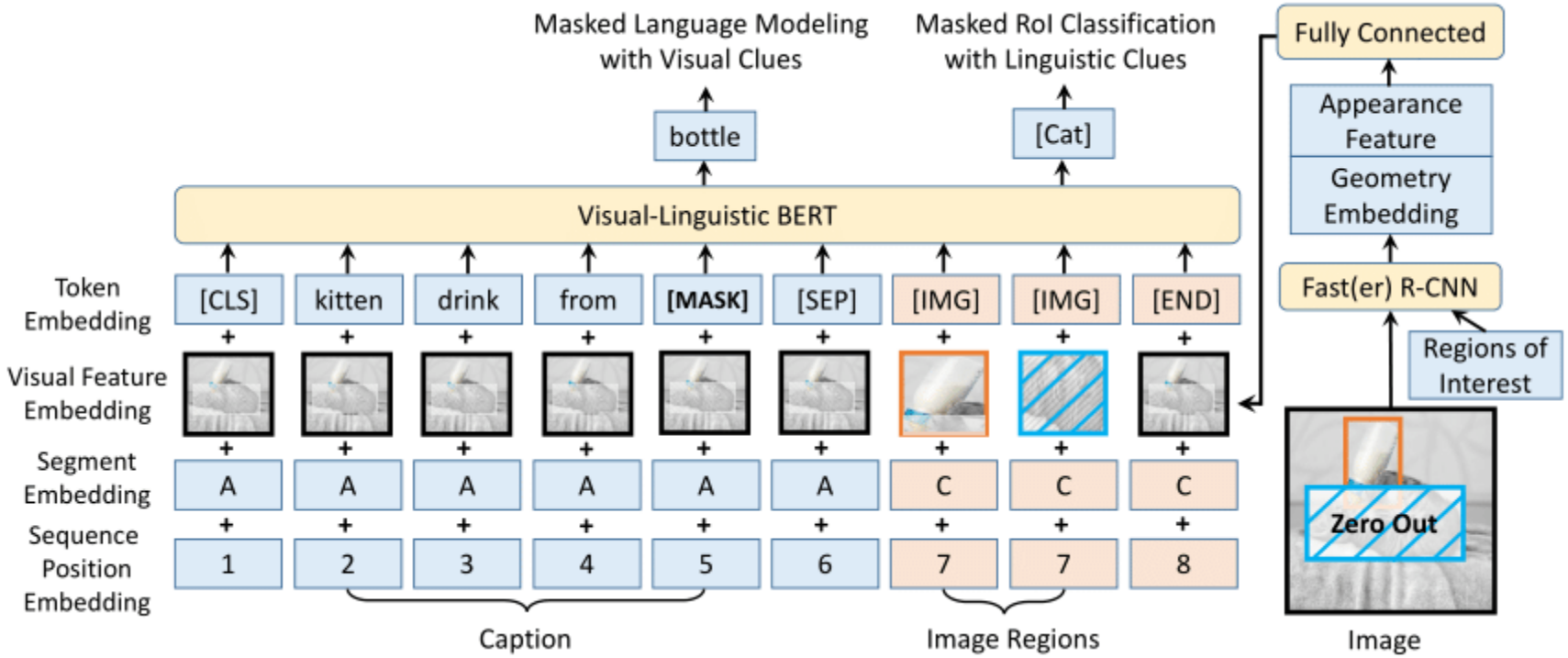
(a) Masked language modeling in BERT ($k = 1$)



(b) Standard language modeling ($k = m$)

跨模态语言模型

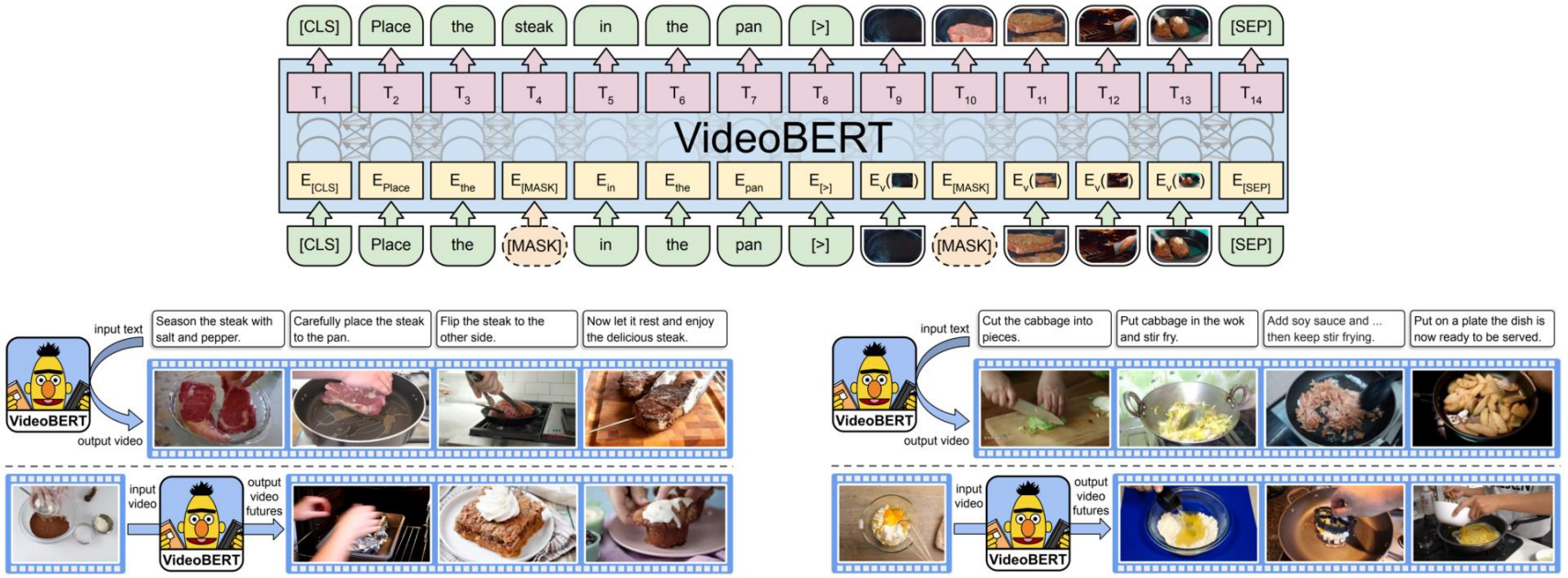
VL-BERT



Su W, Zhu X, Cao Y, et al. VL-BERT: Pre-training of generic visual-linguistic representations. 2019. https://arxiv.org/abs/1908.08530

VideoBERT

将文本和视频对作为BERT的输入，同时Mask词以及图像块



Sun C, Myers A, Vondrick C, et al. VideoBERT: A joint model for video and language representation learning. Proceedings of the IEEE International Conference on Computer Vision. 2019.

OpenAI DALL·E

(a) a tapir made of accordion. a tapir with the texture of an accordion.

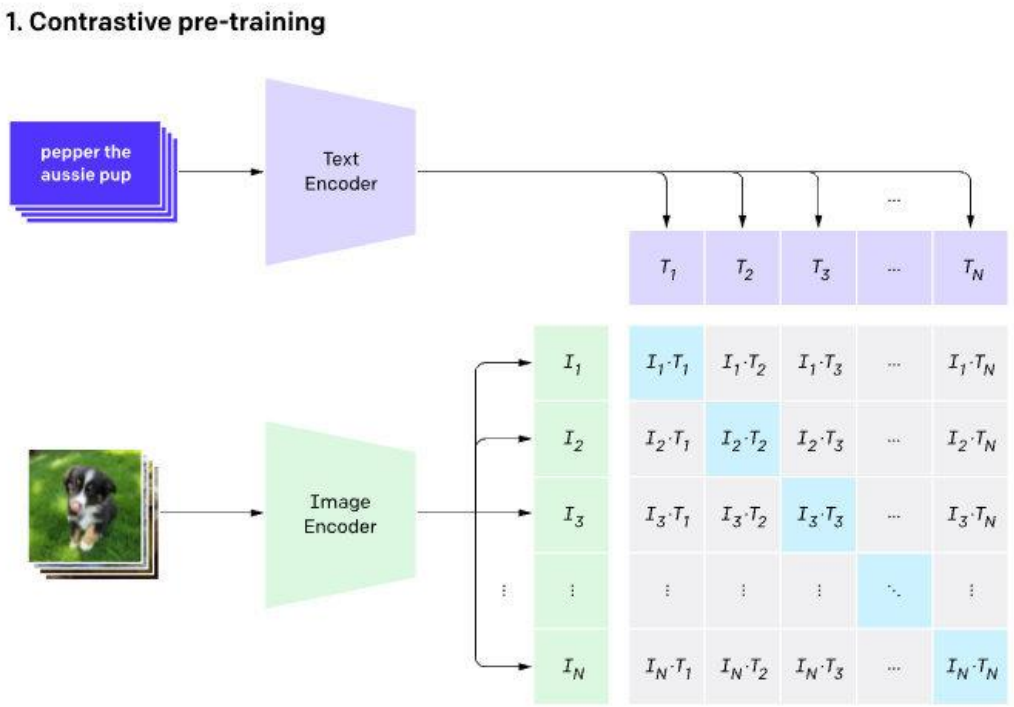
(b) an illustration of a baby hedgehog in a christmas sweater walking a dog

(c) a neon sign that reads "backprop". a neon sign that reads "backprop". backprop neon sign

(d) the exact same cat on the top as a sketch on the bottom

Zero-Shot Text-to-Image Generation

OpenAI CLIP

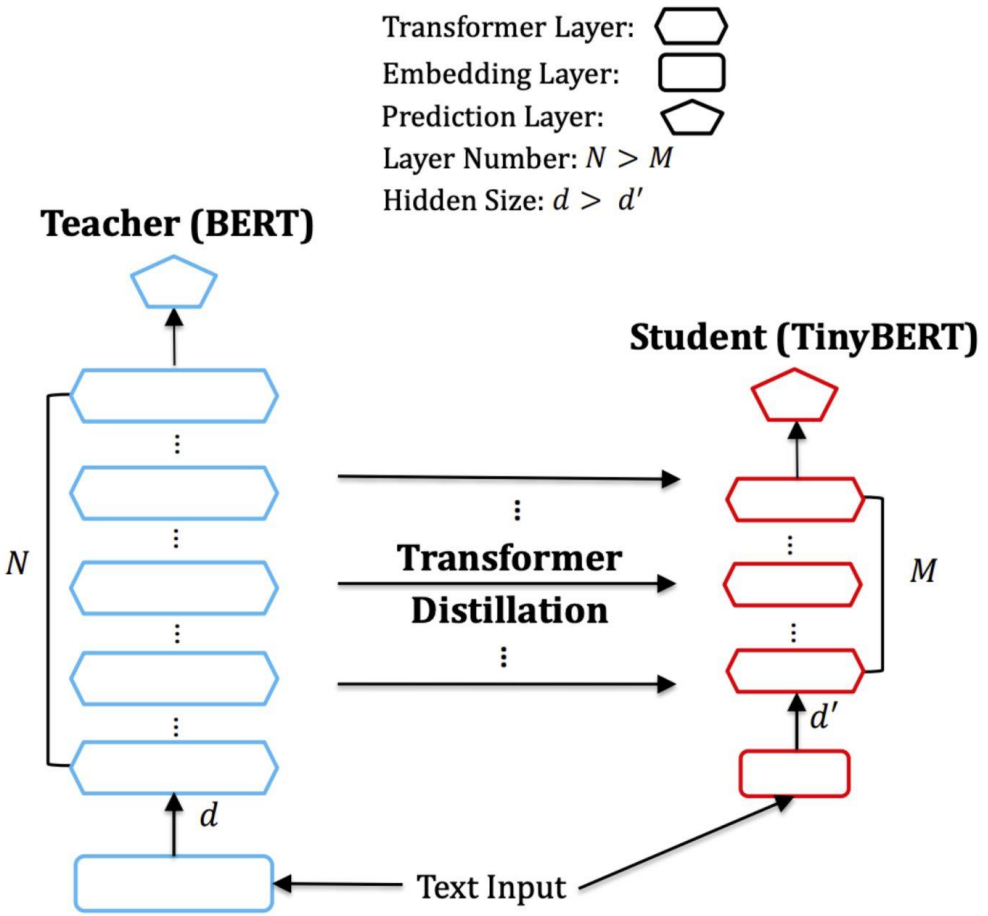


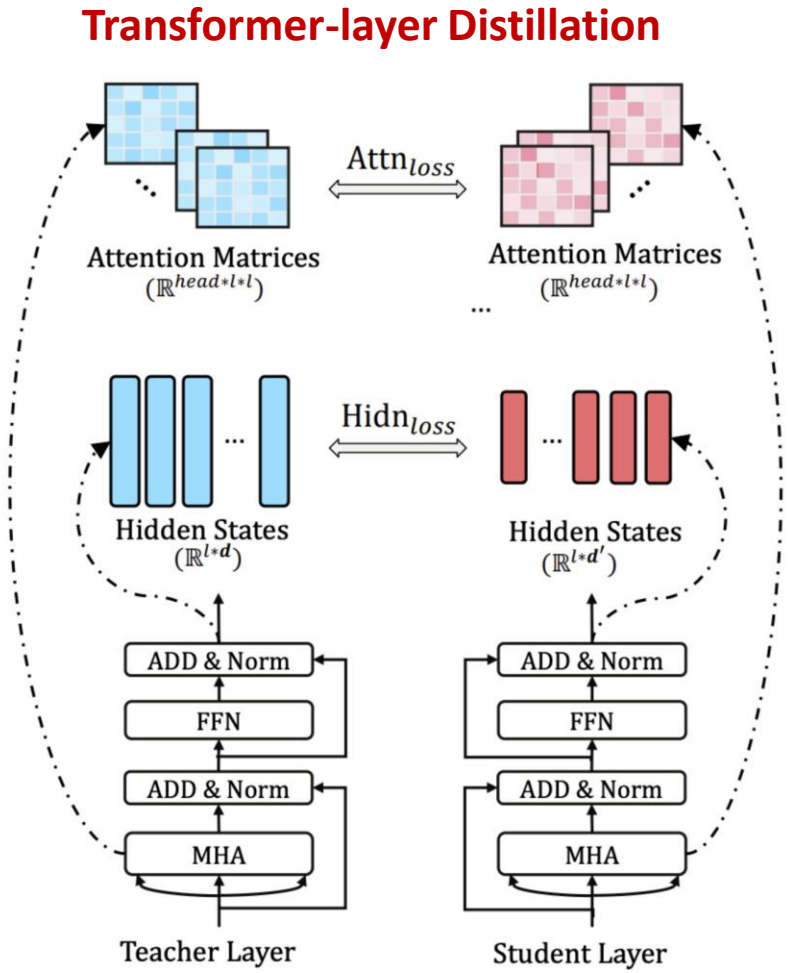
Zero-Shot Text-to-Image Generation

模型压缩

- Model pruning
  - removes less important parameters
- Weight Quantization
  - uses fewer bits to represent the parameters
- Parameter sharing
- Knowledge distillation
  - trains a smaller student model that learns from intermediate outputs from the original model
- Module replacing

TinyBERT



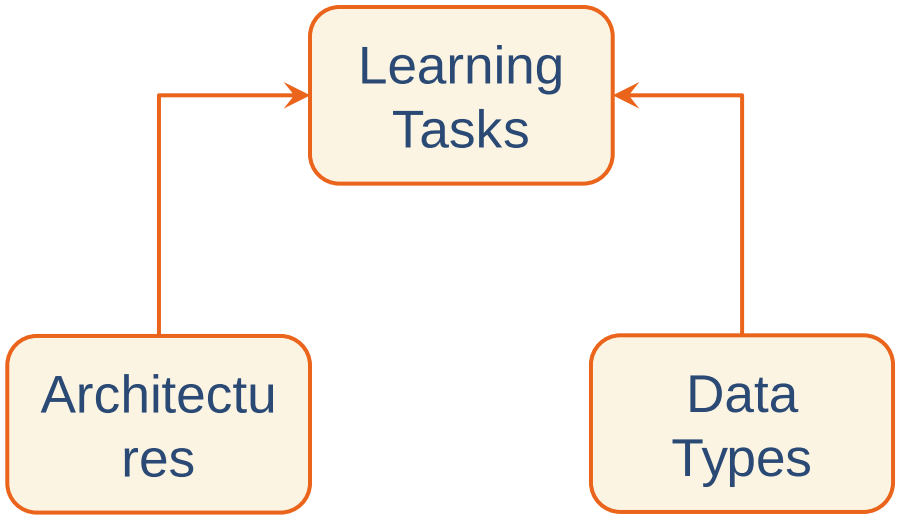


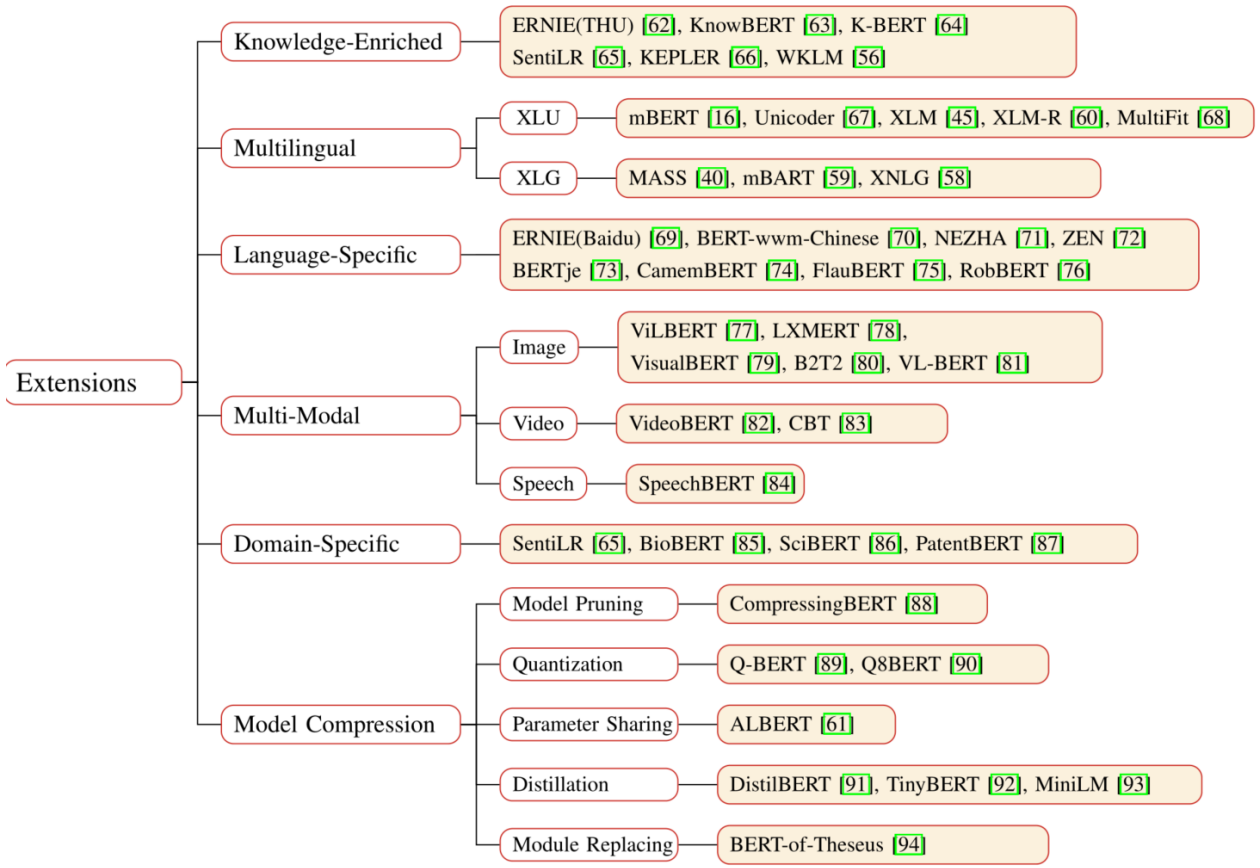
Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, Qun Liu, TinyBERT: Distilling BERT for Natural Language Understanding, https://arxiv.org/abs/1909.10351

小结

Method	Type	#Layer	Loss Function*	Speed Up	Params	Source PTM	GLUE‡
$\text{BERT}_{\text{BASE}}$ [16]	Baseline	12	$\mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{NSP}}$		110M		79.6
$\text{BERT}_{\text{LARGE}}$ [16]		24	$\mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{NSP}}$		340M		81.9
Q-BERT [89]	Quantization	12	HAWQ + GWQ	-		$\text{BERT}_{\text{BASE}}$	≈ 99% BERT◇
Q8BERT [90]		12	DQ + QAT	-		$\text{BERT}_{\text{BASE}}$	≈ 99% BERT
ALBERT§ [61]	Param. Sharing	12	$\mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{SOP}}$	×5.6 ~ 0.3	12 ~ 235M		89.4 (ensemble)
DistilBERT [91]	Distillation	6	$\mathcal{L}_{\text{KD-CE}}+\text{Cos}_{\text{KD}}+\mathcal{L}_{\text{MLM}}$	×1.63	66M	$\text{BERT}_{\text{BASE}}$	77.0 (dev)
TinyBERT§ † [92]		4	$\text{MSE}_{\text{embed}}+\text{MSE}_{\text{attn}}+\text{MSE}_{\text{hidn}}+\mathcal{L}_{\text{KD-CE}}$	×9.4	14.5M	$\text{BERT}_{\text{BASE}}$	76.5
BERT-PKD [147]		3 ~ 6	$\mathcal{L}_{\text{KD-CE}}+\text{PT}_{\text{KD}}+\mathcal{L}_{\text{Task}}$	×3.73 ~ 1.64	45.7 ~ 67 M	$\text{BERT}_{\text{BASE}}$	76.0 ~ 80.6♯
PD [148]		6	$\mathcal{L}_{\text{KD-CE}}+\mathcal{L}_{\text{Task}}+\mathcal{L}_{\text{MLM}}$	×2.0	67.5M	$\text{BERT}_{\text{BASE}}$	81.2♯
MobileBERT§ [149]		24	$\text{FMT}+\text{AT}+\text{PKT}+\mathcal{L}_{\text{KD-CE}}+\mathcal{L}_{\text{MLM}}$	×4.0	25.3M	$\text{BERT}_{\text{LARGE}}$	79.7
MiniLM [93]		6	AT+AR	×1.99	66M	$\text{BERT}_{\text{BASE}}$	81.0♭
DualTrain§ † [150]		12	Dual Projection+$\mathcal{L}_{\text{MLM}}$	-	1.8 ~ 19.2M	$\text{BERT}_{\text{BASE}}$	75.8 ~ 81.9♮
BERT-of-Theseus [94]	Module Replacing	6	$\mathcal{L}_{\text{Task}}$	×1.94	66M	$\text{BERT}_{\text{BASE}}$	78.6

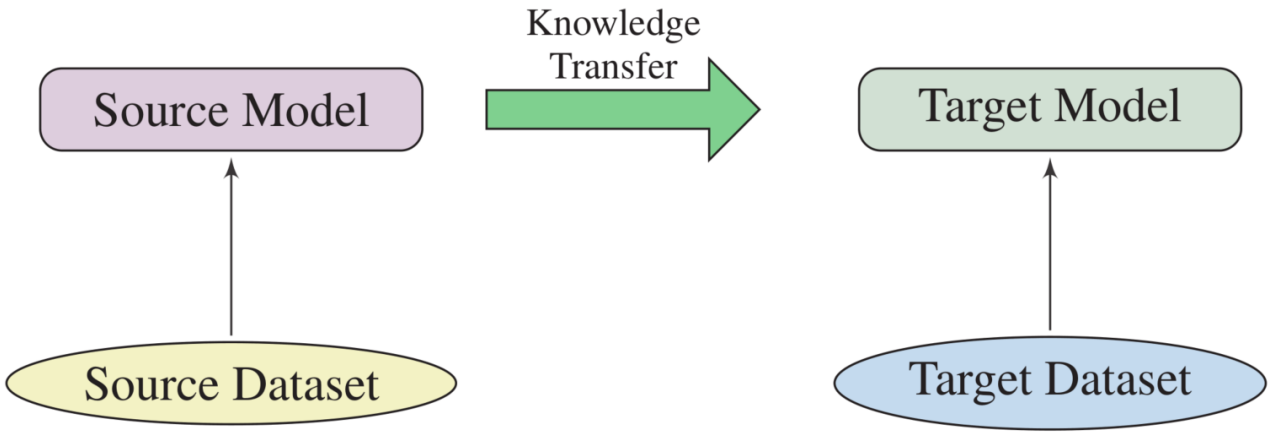
小结


Learning Tasks
Architectures
Data Types


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]

迁移到下游任务

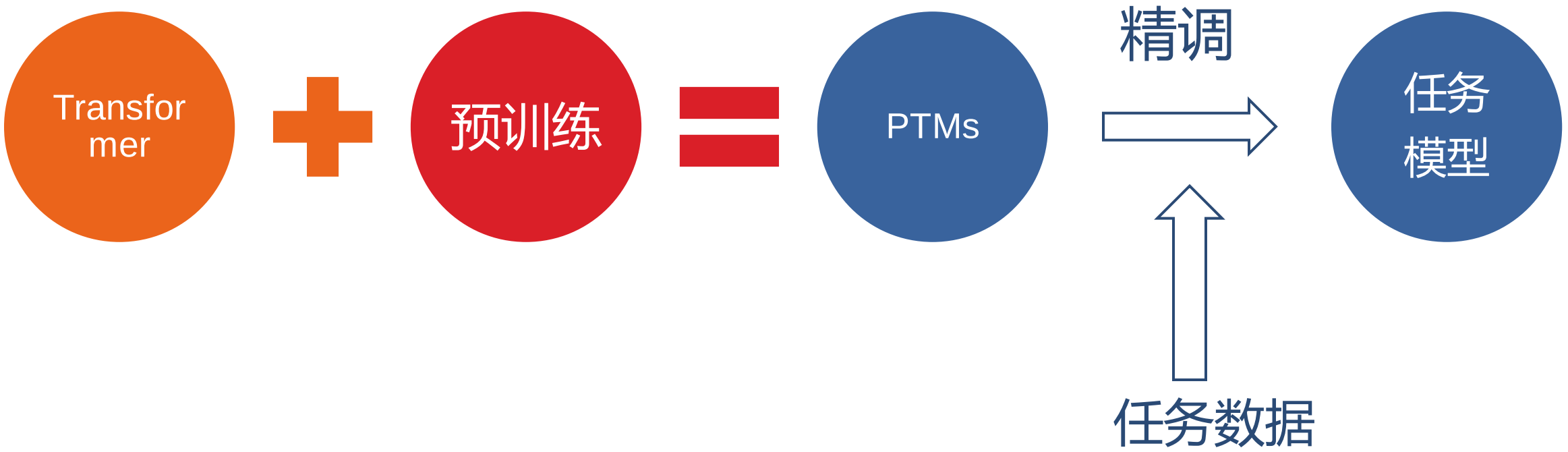
Transfer Learning



sequential transfer learning

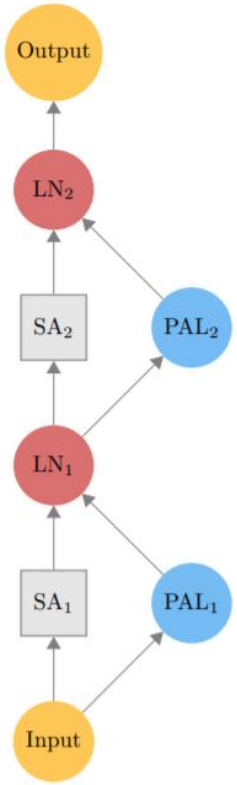
预训练+精调

迁移学习



引入额外自适应模块

- 多任务场景下，精调是参数低效的
  - Fine-tuning with extra adaptation modules

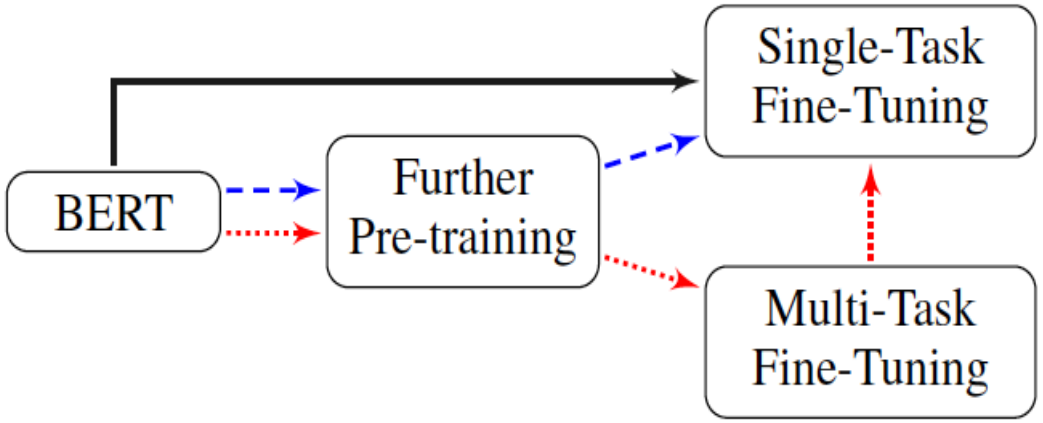


METHOD	PARAMS	MNLI-(M/MM) 392K	QQP 363K	QNLI 108K	SST-2 67K	COLA 8.5K	STS-B 5.7K	MRPC 3.5K	RTE 2.5K	AV.
BERT-BASE	8×	84.6/83.4	89.2/71.2	90.1	93.5	52.1	85.8	84.8/88.9	66.4	79.6
SHARED	1.00×	84.0/83.4	88.9/70.8	89.3	93.4	51.2	83.6	81.3/86.7	76.6	79.9
TOP PROJ. ATTN.	1.10×	84.0/83.2	88.8/71.2	89.7	93.2	47.1	85.3	83.1/87.5	75.5	79.6
PALS (204)	1.13×	84.3/83.5	89.2/71.5	90.0	92.6	51.2	85.8	84.6/88.7	76.0	**80.4**

BERT and PALs: Projected Attention Layers for Efficient Adaptation in Multi-Task Learning, https://arxiv.org/abs/1902.02671

精调策略

- 多阶段迁移：继续预训练+精调



Multi-stage Transfer

Chi Sun, Xipeng Qiu, Yige Xu, Xuanjing Huang, How to Fine-Tune BERT for Text Classification?, CCL 2019, Best Paper Award, https://arxiv.org/abs/1905.05583

多阶段迁移

- In-Domain and Cross-Domain Further Pre-Training

Domain	sentiment			question		topic	
Dataset	IMDb	Yelp P.	Yelp F.	TREC	Yah. A.	AG's News	DBPedia
IMDb	**4.37**	2.18	29.60	2.60	22.39	5.24	0.68
Yelp P.	5.24	1.92	29.37	2.00	22.38	5.14	**0.65**
Yelp F.	5.18	1.94	29.42	2.40	22.33	5.43	**0.65**
all sentiment	4.88	**1.87**	29.25	3.00	22.35	5.34	0.67
TREC	5.65	2.09	29.35	3.20	22.17	5.12	0.66
Yah. A.	5.52	2.08	29.31	**1.80**	22.38	5.16	0.67
all question	5.68	2.14	29.52	2.20	**21.86**	5.21	0.68
AG's News	5.97	2.15	29.38	2.00	22.32	**4.80**	0.68
DBPedia	5.80	2.13	29.47	2.60	22.30	5.13	0.68
all topic	5.85	2.20	29.68	2.60	22.28	4.88	**0.65**
all	5.18	1.97	**29.20**	2.80	21.94	5.08	0.67
w/o pretrain	5.40	2.28	30.06	2.80	22.42	5.25	0.71

Further Pre-Training on BERT Large

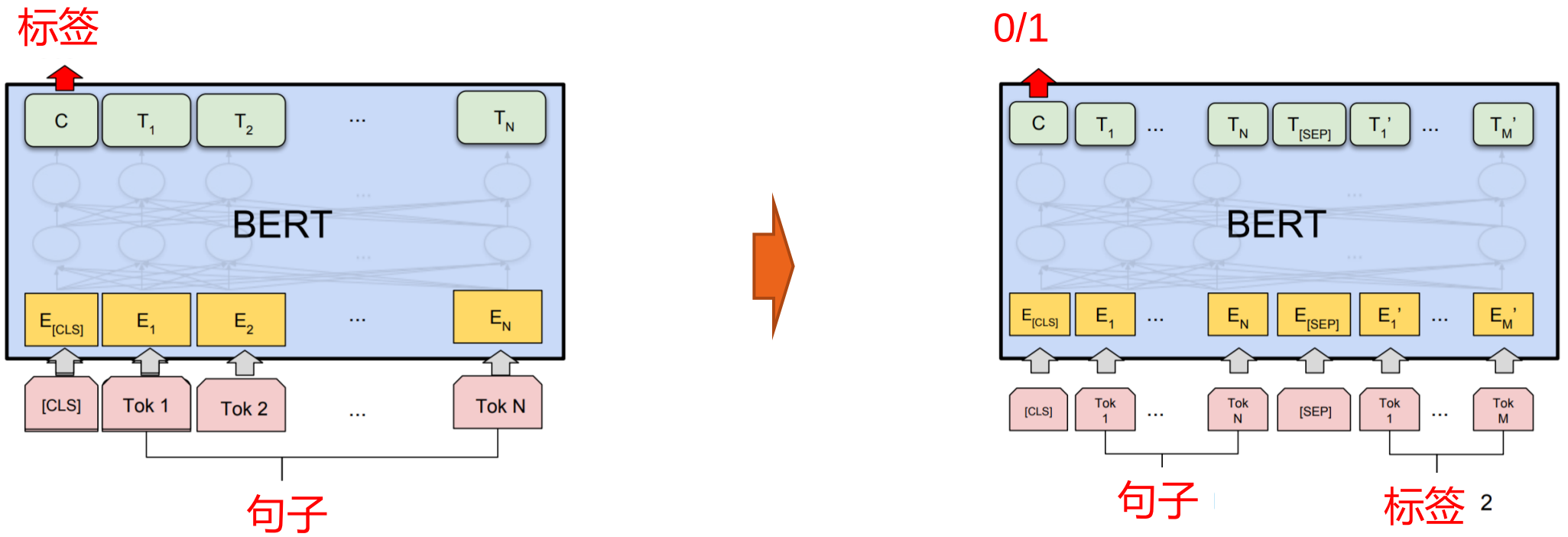
Test error rates (%) on five text classification datasets

Model	IMDb	Yelp P.	Yelp F.	AG	DBP
ULMFiT	4.60	2.16	29.98	5.01	0.80
$\text{BERT}_{\text{BASE}}$	5.40	2.28	30.06	5.25	0.71
+ ITPT	4.37	1.92	29.42	4.80	0.68
$\text{BERT}_{\text{LARGE}}$	4.86	2.04	29.25	4.86	0.62
+ ITPT	**4.21**	**1.81**	**28.62**	**4.66**	**0.61**

Current SOTA results!

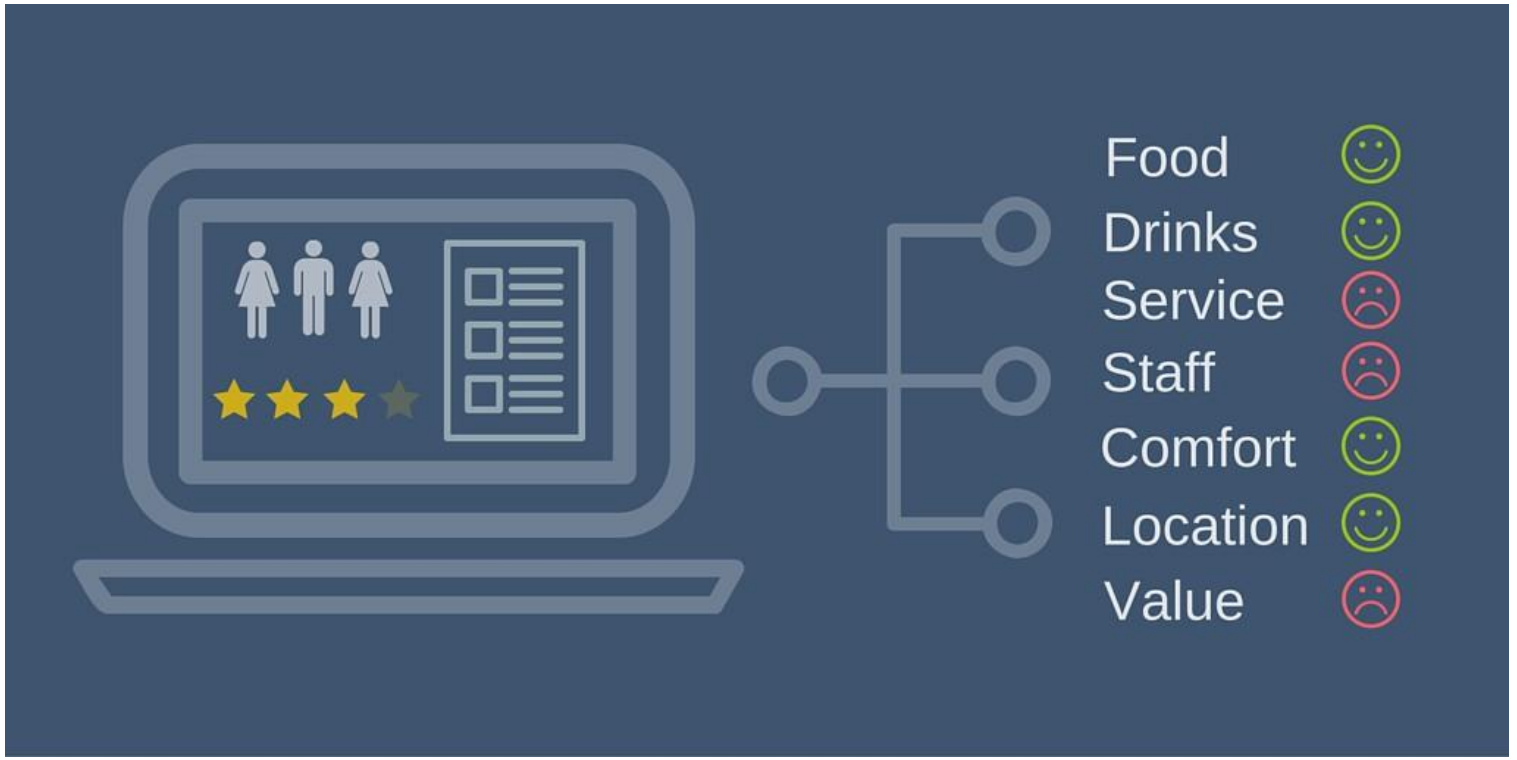
任务转换

- 当标签（label）含有语义信息时，将单句(single sentence)分类问题转换为句对(sentence-pair)分类。
  - 标签不含语义时，可以人工构造辅助句子



Chi Sun, Luyao Huang, Xipeng Qiu, Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence, NAACL 2019, https://arxiv.org/pdf/1903.09588.pdf

TABSA :Aspect-Based Sentiment Analysis



方法：构建辅助句子的四种方法

(location, safety)

Target Aspect

1. QA-M:

what do you think of the safety of location?

2. NLI-M:

location - safety

3. QA-B:

the polarity of the aspect safety of location is positive

the polarity of the aspect safety of location is negative

the polarity of the aspect safety of location is none

4. NLI-B:

location - safety - positive

location - safety - negative

location - safety - none

实验结果

- Sentihood数据集：

Model	Aspect			Sentiment	
	Acc.	F_1	AUC	*Acc.*	AUC
LR (Saeidi et al., 2016)	-	39.3	92.4	87.5	90.5
LSTM-Final (Saeidi et al., 2016)	-	68.9	89.8	82.0	85.4
LSTM-Loc (Saeidi et al., 2016)	-	69.3	89.7	81.9	83.9
LSTM+TA+SA (Ma et al., 2018)	66.4	76.7	-	86.8	-
SenticLSTM (Ma et al., 2018)	67.4	78.2	-	89.3	-
Dmu-Entnet (Liu et al., 2018)	73.5	78.5	94.4	91.0	94.8
BERT-single	73.7	81.0	96.4	85.5	84.2
BERT-pair-QA-M	79.4	86.4	97.0	**93.6**	96.4
BERT-pair-NLI-M	78.3	87.0	**97.5**	92.1	96.5
BERT-pair-QA-B	79.2	**87.9**	97.1	93.3	**97.0**
BERT-pair-NLI-B	**79.8**	87.5	96.6	92.8	96.9

实验结果

SemEval 2014 任务4的子任务3和4：

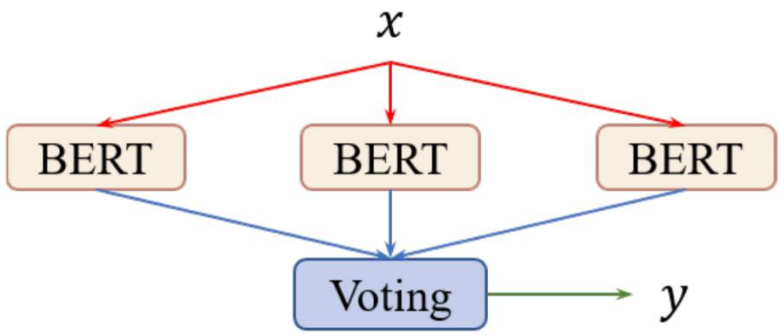
Models	P	R	F1
XRCE	83.23	81.37	82.29
NRC-Canada	91.04	86.24	88.58
BERT-single	92.78	89.07	90.89
BERT-pair-QA-M	92.87	90.24	91.54
BERT-pair-NLI-M	93.15	90.24	91.67
BERT-pair-QA-B	93.04	89.95	91.47
BERT-pair-NLI-B	93.57	90.83	**92.18**

方面类别检测子任务

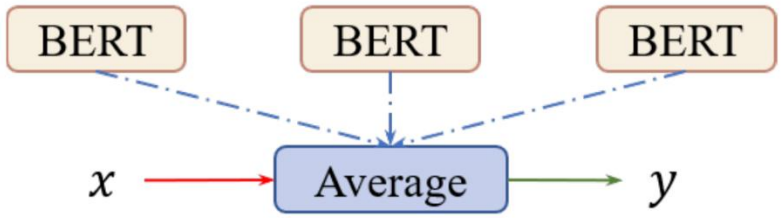
Models	4-way	3-way	Binary
XRCE	78.1	-	-
NRC-Canada	82.9	-	-
LSTM	-	82.0	88.3
ATAE-LSTM	-	84.0	89.9
BERT-single	83.7	86.9	93.3
BERT-pair-QA-M	85.2	89.3	95.4
BERT-pair-NLI-M	85.1	88.7	94.4
BERT-pair-QA-B	**85.9**	**89.9**	**95.6**
BERT-pair-NLI-B	84.6	88.7	95.1

方面情感极性分类子任务

精调方法优化



(a) Voted BERT



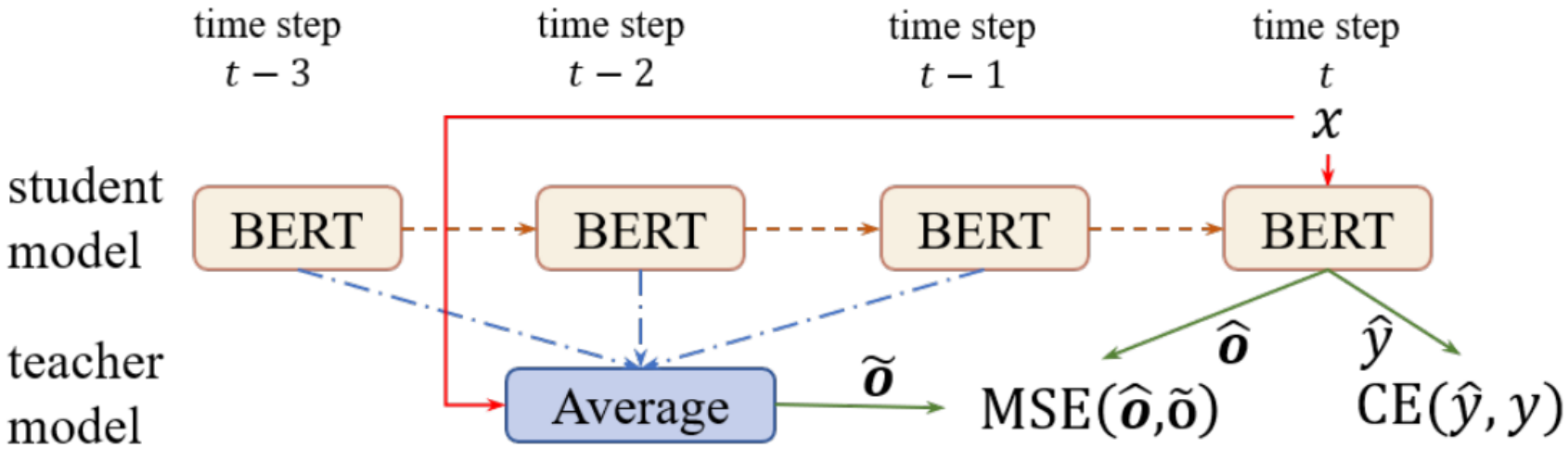
(b) Averaged BERT

Ensemble BERT is very successful in practical applications.

$$\mathrm{BERT}_{\mathrm{VOTE}}(x;\Theta) = \sum_{k=1}^{K} \mathrm{BERT}(x;\theta_k)$$

$$\mathrm{BERT}_{\mathrm{AVG}}(x;\bar{\theta}) = \mathrm{BERT}(x;\frac{1}{K}\sum_{k=1}^{K}\theta_k)$$

Self-Ensemble and Self-Distillation



Self-Distillation-Averaged (SDA)

Yige Xu, Xipeng Qiu, Ligao Zhou, Xuanjing Huang. Improving BERT Fine-Tuning via Self-Ensemble and Self-Distillation, https://arxiv.org/abs/2002.10345

实验结果

Model	IMDb	AG's News	DBPedia	Yelp P.	Yelp F.	Avg. Δ	SNLI	MNLI (m/mm)	Avg. Δ
	Test Error Rate (%)						Accuracy (%)		
ULMFiT [Howard and Ruder, 2018]	4.60	5.01	0.80	2.16	29.98	/	/	/	/
$\text{BERT}_{\text{BASE}}$ [Sun *et al.*, 2019]*	5.40	5.25	0.71	2.28	30.06	/	/	/	/
$\text{BERT}_{\text{BASE}}$	5.80	5.71	0.71	2.25	30.37	-	90.7	84.6/83.3	-
$\text{BERT}_{\text{VOTE}}$ ($K = 4$)	5.60	5.41	0.67	2.03	29.44	5.44%	91.2	85.3/84.4	5.50%
BERT_{AVG} ($K = 4$)	5.68	5.53	0.68	2.03	30.03	4.07%	90.8	85.1/84.2	3.24%
BERT_{SE} (ours)	5.82	5.59	0.65	2.19	30.48	2.50%	90.8	84.2/83.3	-0.51%
BERT_{SDV} (ours)	5.35	5.38	**0.68**	2.05	**29.88**	5.65%	**91.2**	**85.3/84.3**	**5.30%**
BERT_{SDA} (ours)	**5.29**	**5.29**	**0.68**	**2.04**	**29.88**	**6.26%**	91.2	85.0/84.3	4.65%

资源、系统

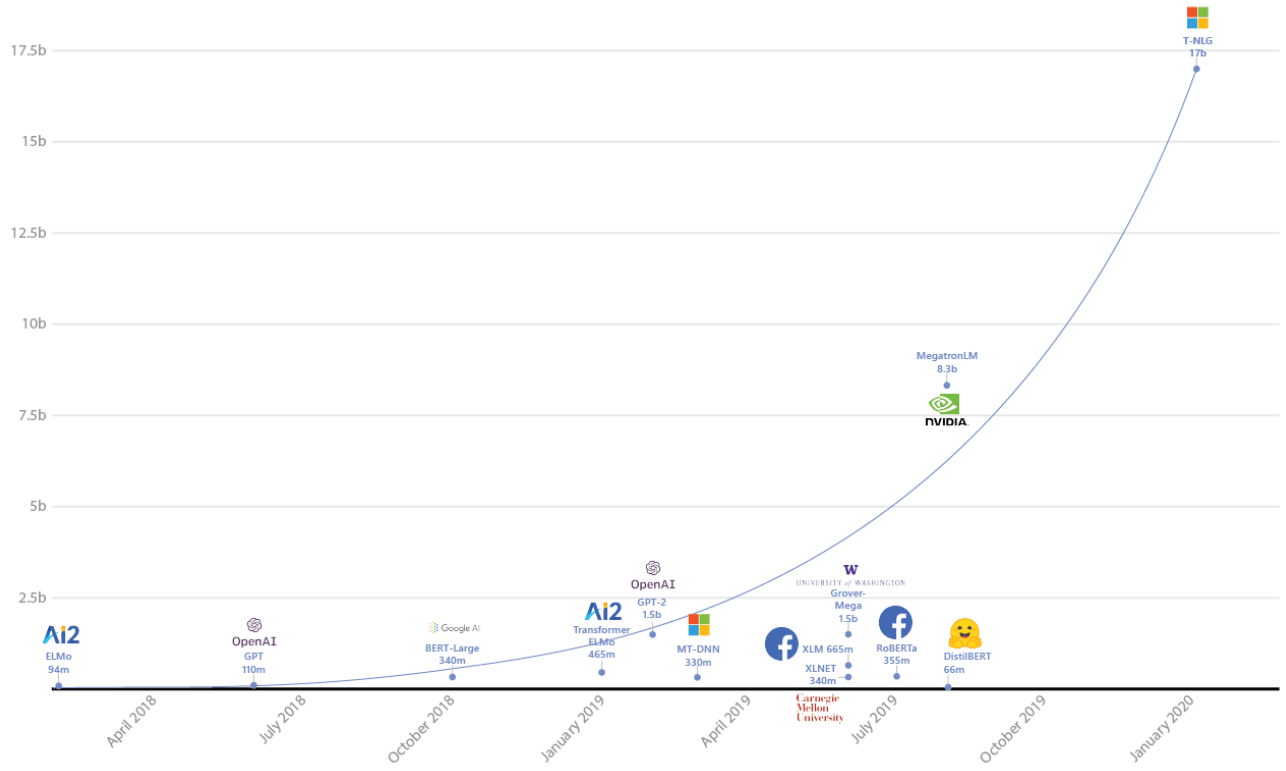
资源、系统

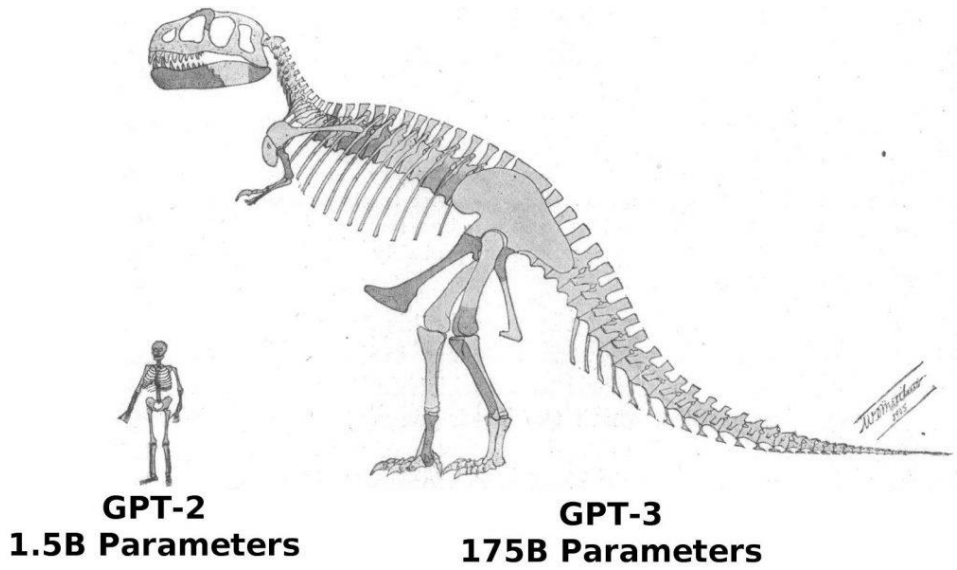
- Pre-trained Languge Model (PLM) 论文汇总
  - https://github.com/thunlp/PLMpapers
- HuggingFace开源Transformers
  - https://github.com/huggingface/transformers

未来展望

Upper Bound of PTMs

GPT-3
175b





The Bitter Lesson

Rich Sutton
March 13, 2019

The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin.

This is a big lesson. As a field, we still have not thoroughly learned it, as we are continuing to make the same kind of mistakes. To see this, and to effectively resist it, we have to understand the appeal of these mistakes. We have to learn the bitter lesson that building in how we think we think does not work in the long run. The bitter lesson is based on the historical observations that

1) AI researchers have often tried to build knowledge into their agents,
2) this always helps in the short term, and is personally satisfying to the researcher, but
3) in the long run it plateaus and even inhibits further progress, and
4) breakthrough progress eventually arrives by an opposing approach based on scaling computation by search and learning.

The eventual success is tinged with bitterness, and often incompletely digested, because it is success over a favored, human-centric approach.

http://www.incompleteideas.net/IncIdeas/BitterLesson.html

Architecture of PTMs

- The transformer has been proved to be an effective architecture for pre-training.
  - computation complexity
    - most of current PTMs cannot deal with the sequence longer than 512 tokens.

- Searching for more efficient model architecture for PTMs is important to capture longer-range contextual information.
  - neural architecture search (NAS)

Knowledge Transfer Beyond Fine-tuning

- Currently, fine-tuning is the dominant method to transfer PTMs' knowledge to downstream tasks.
  - parameter inefficiency

- Mining knowledge from PTMs can be more flexible, such as
  - feature extraction
  - knowledge distillation
  - data augmentation
  - using PTMs as external knowledge
  - Retrieval Augmented Generation

Interpretability and Reliability of PTMs

- Explainable artificial intelligence (XAI)

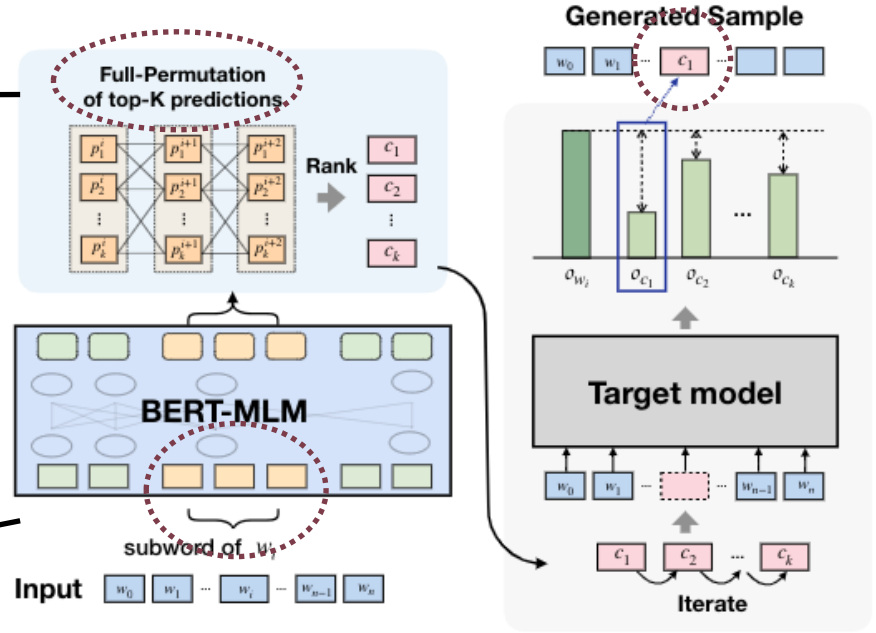
- Attack
  - PTMs are also vulnerable to adversarial attacks.
  - The studies of adversarial attacks against PTMs help us understand their capabilities by fully exposing their vulnerabilities.

- Defense
  - Adversarial defenses for PTMs are also promising, which improve the robustness of PTMs and make them immune against adversarial attack.

BERT-Attack

deal with sub-words

not mask the word



find one substitute

Advantages:

- effective
- context-aware generation of substitutes

Figure 1: One step of our replacement strategy.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, Xipeng Qiu, BERT-ATTACK: Adversarial Attack Against BERT Using BERT, EMNLP2020. https://arxiv.org/abs/2004.09984

Summary

- Architectures
- Learning Tasks
- Knowledge Transfer
- Reliability
- Model Compression
- Beyond Text