Security Level:

预训练语言模型的研究与应用

www.huawei.com

刘群

华为诺亚方舟实验室

北京智源大会NLP论坛

2019-10-31

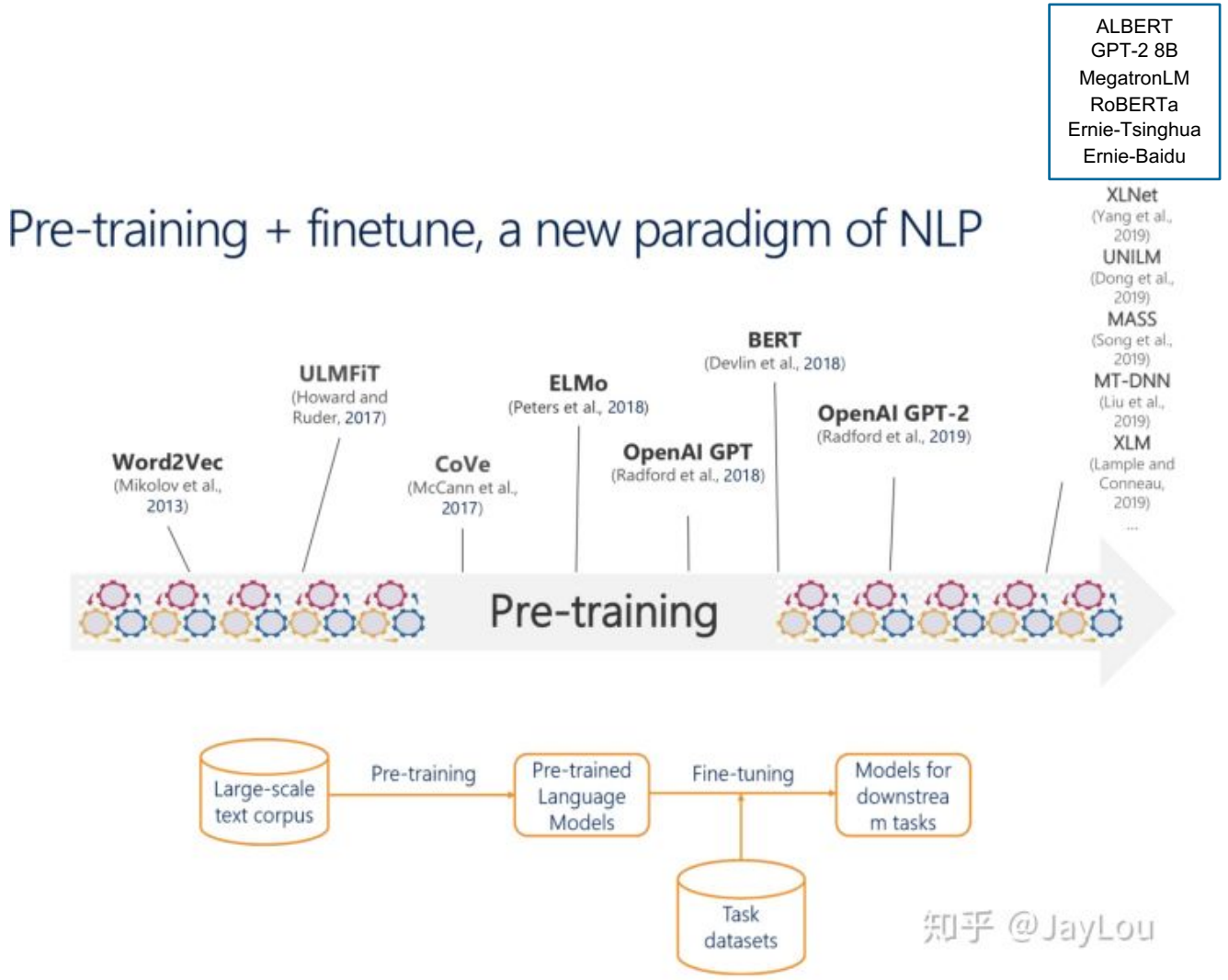
HUAWEI

内容概要

预训练语言模型概述

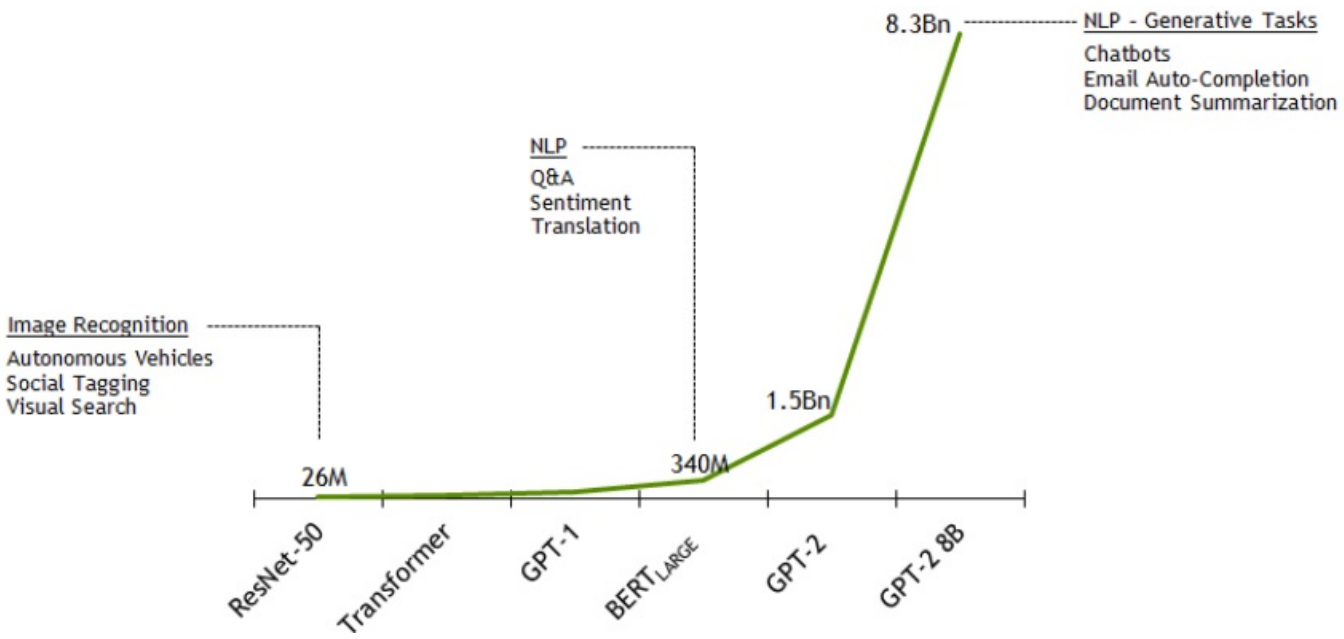
- 神经网络语言模型
- 使用大规模无标注纯文本语料进行训练
- 可以用于各类下游nlp任务，各项性能指标均获得大幅度提高，并可以将各类下游任务的解决方案统一简化为几种固定的fine-tune框架
- 两大类型：
  - Encoder：用于自然语言理解，ELMo，BERT
  - Decoder：用于自然语言生成，GPT

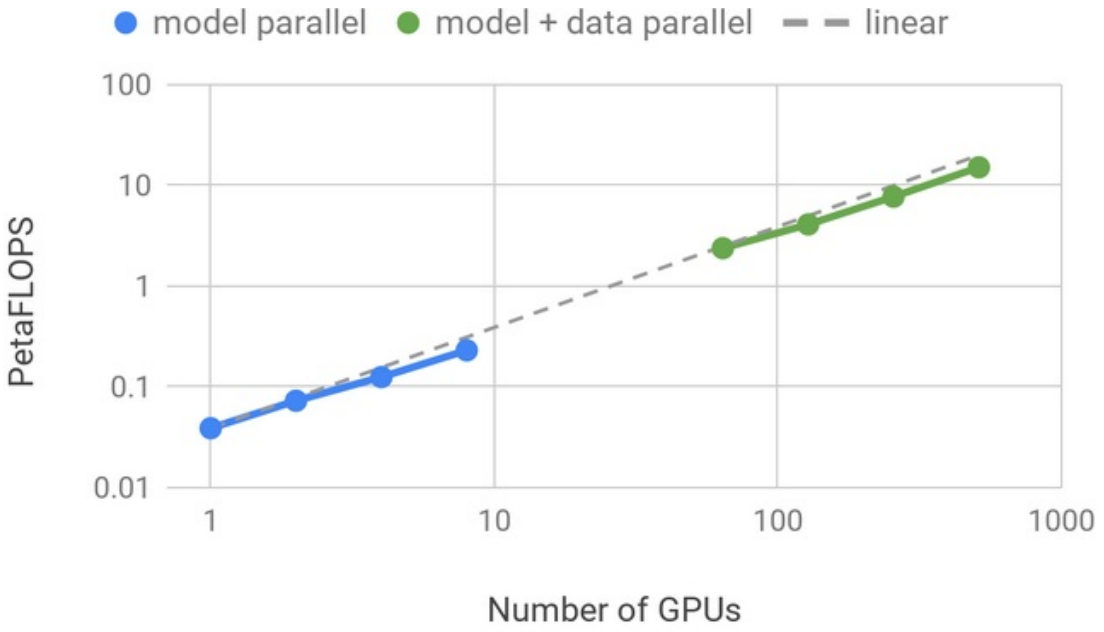
预训练语言模型的发展



图片来自：https://zhuanlan.zhihu.com/p/76912493

预训练语言模型概述





图片来自：https://devblogs.nvidia.com/training-bert-with-gpus/#targetText=GPT%2D2%208B%20is%20the,on%20a%20single%20V100%20GPU.

诺亚方舟实验室的预训练语言模型研究

- 我们在内部的服务器上重现了Google BERT-base和BERT-large的实验，并尝试了各种改进的方法
- 利用BERT的代码，实现了OpenAI GPT-2模型
- 实现了基于GPU的多卡多机并行训练，并对训练过程进行了优化，提高了训练效率
- 对模型的细节进行了多方面的改进，提高了系统性能
- 研究了多种模型压缩优化方案

预训练语言模型的应用

- 中文古诗生成
- 对话生成

- 对话理解
- 多标签分类
- 推荐和搜索

内容概要

哪吒：诺亚方舟实验室的中文预训练语言模型

NEZHA: NEURAL CONTEXTUALIZED REPRESENTATION FOR CHINESE LANGUAGE UNDERSTANDING

TECHNICAL REPORT

Junqiu Wei, Xiaozhe Ren, Xiaoguang Li, Wenyong Huang, Yi Liao,
Yasheng Wang, Jiashu Lin*, Xin Jiang, Xiao Chen, Qun Liu
Noah's Ark Lab, *HiSilicon, Huawei Technologies
{wei.junqiu1, renxiaozhe, lixiaoguang11, wenyong.huang, liao.yi,
wangyasheng, linjiashu, jiang.xin, chen.xiao2, qun.liu}@huawei.com

September 4, 2019

技术报告下载地址： https://arxiv.org/abs/1909.00204
即将开放源代码和模型数据（公司流程申请中）

哪吒：诺亚方舟实验室的中文预训练语言模型

- 首先，我们在内部的服务器上重现了Google BERT-base和BERT-large的实验
  - 在华为云上训练和运行成功
  - 英文预训练数据与BERT相同：Wikipedia + BookCorpus
  - 中文预训练数据包括：Wikipedia + Baike + News

哪吒：模型的训练和优化

- 基于华为云
- 多卡多机的数据并行
- 混合精度训练
- LAMB优化器

哪吒：模型的改进

- 函数式相对位置编码
- 全词覆盖

函数式相对位置编码

$$a_{ij}[2k] = \sin((j-i)/(10000^{\frac{2\cdot k}{d_z}})),$$
$$a_{ij}[2k+1] = \cos((j-i)/(10000^{\frac{2\cdot k}{d_z}})).$$

相关工作性能对比

Table 3: Results of pre-trained models on downstream Chinese NLU tasks.

Model	CMRC		XNLI		LCQMC		PD-NER		ChnSenti	
	EM	F1	Dev	Test	Dev	Test	Dev	Test	Dev	Test
BASE MODELS										
$\text{BERT}_{\text{BASE}}$	64.06	85.01	78.75	77.27	89.04	87.61	96.53	98.58	94.91	95.42
$\text{BERT}_{\text{BASE}}$-WWM	64.96	85.79	78.79	78.44	89.19	87.16	96.86	98.58	94.67	94.58
$\text{BERT}_{\text{BASE}}$-WWM (in [8])	66.30	85.60	79.00	78.20	89.40	87.00	95.30	65.10	95.10	95.40
ERNIE-$\text{Baidu}_{\text{BASE}}$ 1.0 (in [3])	65.1	85.1	79.9	78.4	89.70	87.40	-	-	95.20	95.40
ERNIE-$\text{Baidu}_{\text{BASE}}$ 2.0 (in [4])	**69.10**	**88.60**	81.20	**79.70**	**90.90**	87.90	-	-	**95.70**	95.50
$\text{NEZHA}_{\text{BASE}}$ (ours)	67.07	86.35	**81.37**	79.32	89.98	87.41	97.22	**98.58**	94.74	95.17
$\text{NEZHA}_{\text{BASE}}$-WWM (ours)	67.82	86.25	81.25	79.11	89.85	87.10	**97.41**	98.35	94.75	**95.84**
LARGE MODELS										
ERNIE-$\text{Baidu}_{\text{LARGE}}$ 2.0 (in [4])	**71.50**	**89.90**	**82.60**	81.00	**90.90**	87.90	-	-	**96.10**	95.80
$\text{NEZHA}_{\text{LARGE}}$ (ours)	68.10	87.20	81.53	80.44	90.18	87.20	**97.51**	**97.87**	95.92	95.83
$\text{NEZHA}_{\text{LARGE}}$-WWM (ours)	67.32	86.62	82.21	**81.17**	90.87	**87.94**	97.26	97.63	95.75	**96.00**

相关工作技术比较

Table 2: Pre-training Techniques Adopted in Chinese pre-trained language models (MLM: Masked Language Modeling, NSP: Next Sentence Prediction, WWM: Whole Word Masking, KM: Knowledge Masking, SR: Sentence Reordering, SD: Sentence Distance, DR: Discourse Relation, IR: IR Relevance, PAPE: Parametric Absolute Position Encoding, FRPE: Functional Relative Position Encoding)

Model	Pre-Training Tasks			Training Precision	Optimizer	Position Encoding
	Word-Aware	Sentence-Aware	Semantic-Aware			
$\text{BERT}_{\text{BASE}}$ $\text{BERT}_{\text{BASE}}$-WWM	MLM MLM (WWM)	NSP NSP	-	Single Precision (FP32)	ADAM LAMB	PAPE
ERNIE-$\text{Baidu}_{\text{BASE}}$ 1.0	MLM (KM)	NSP	-	Single Precision (FP32)	ADAM	PAPE
ERNIE-$\text{Baidu}_{\text{BASE}}$ 2.0 ERNIE-$\text{Baidu}_{\text{LARGE}}$ 2.0	MLM (KM)	SR & SD	DR & IR	Mixed Precision	ADAM	PAPE
$\text{NEZHA}_{\text{BASE}}$ $\text{NEZHA}_{\text{LARGE}}$	MLM (WWM)	NSP	-	Mixed Precision	LAMB	FRPE

各项改进的效果比较

Table 4: Ablation studies. (APE: absolute position encoding; PRPE: parametric relative position encoding; FRPE: functional relative position encoding; WWM: whole word masking; SL: sequence length.)

Model	CMRC		XNLI		LCQMC		PD-NER		ChnSenti	
	EM	F1	Dev	Test	Dev	Test	Dev	Test	Dev	Test
$\text{NEZHA}_{\text{BASE}}$										
News, APE, SL:128	37.96	58.40	78.79	77.72	89.31	86.74	94.87	98.10	94.17	95.67
News, PRPE, SL:128	65.26	86.17	79.18	77.98	89.21	86.92	96.93	98.12	94.67	95.08
News, FRPE, SL:128	65.95	86.46	79.96	78.32	89.40	87.23	96.69	98.10	95.58	95.75
News, FRPE, SL:512	67.79	86.60	80.57	79.52	90.06	86.73	97.04	97.62	95.09	95.08
News+Wiki+Baike, FRPE, SL:128	66.95	86.41	81.25	79.06	89.83	87.13	97.21	97.41	95.25	94.42
News+Wiki+Baike, FRPE, WWM, SL:128	67.82	86.25	81.25	79.11	89.85	87.10	97.41	98.35	94.75	95.84
News+Wiki+Baike, FRPE, WWM, SL:512	66.45	86.16	80.96	79.86	89.64	86.18	96.79	98.10	95.08	95.42

内容概要

ERNIE：实体表示增强的预训练语言模型

ERNIE: Enhanced Language Representation with Informative Entities

Zhengyan Zhang[1,2,3*], Xu Han[1,2,3*], Zhiyuan Liu[1,2,3†], Xin Jiang[4], Maosong Sun[1,2,3], Qun Liu[4]

[1]Department of Computer Science and Technology, Tsinghua University, Beijing, China

[2]Institute for Artificial Intelligence, Tsinghua University, Beijing, China

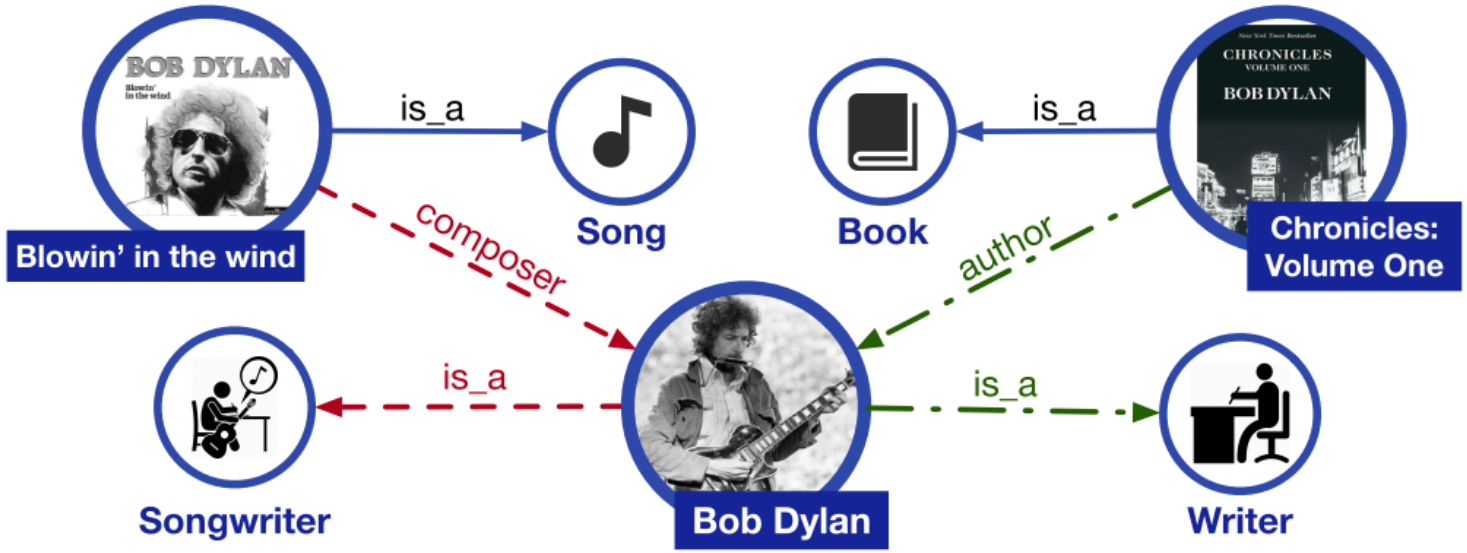
[3]State Key Lab on Intelligent Technology and Systems, Tsinghua University, Beijing, China

[4]Huawei Noah's Ark Lab

{zhangzhengyan14,hanxu17}@mails.tsinghua.edu.cn

- 论文发表在：ACL 2019
- 论文下载：https://arxiv.org/abs/1905.07129
- 代码下载：https://github.com/thunlp/ERNIE

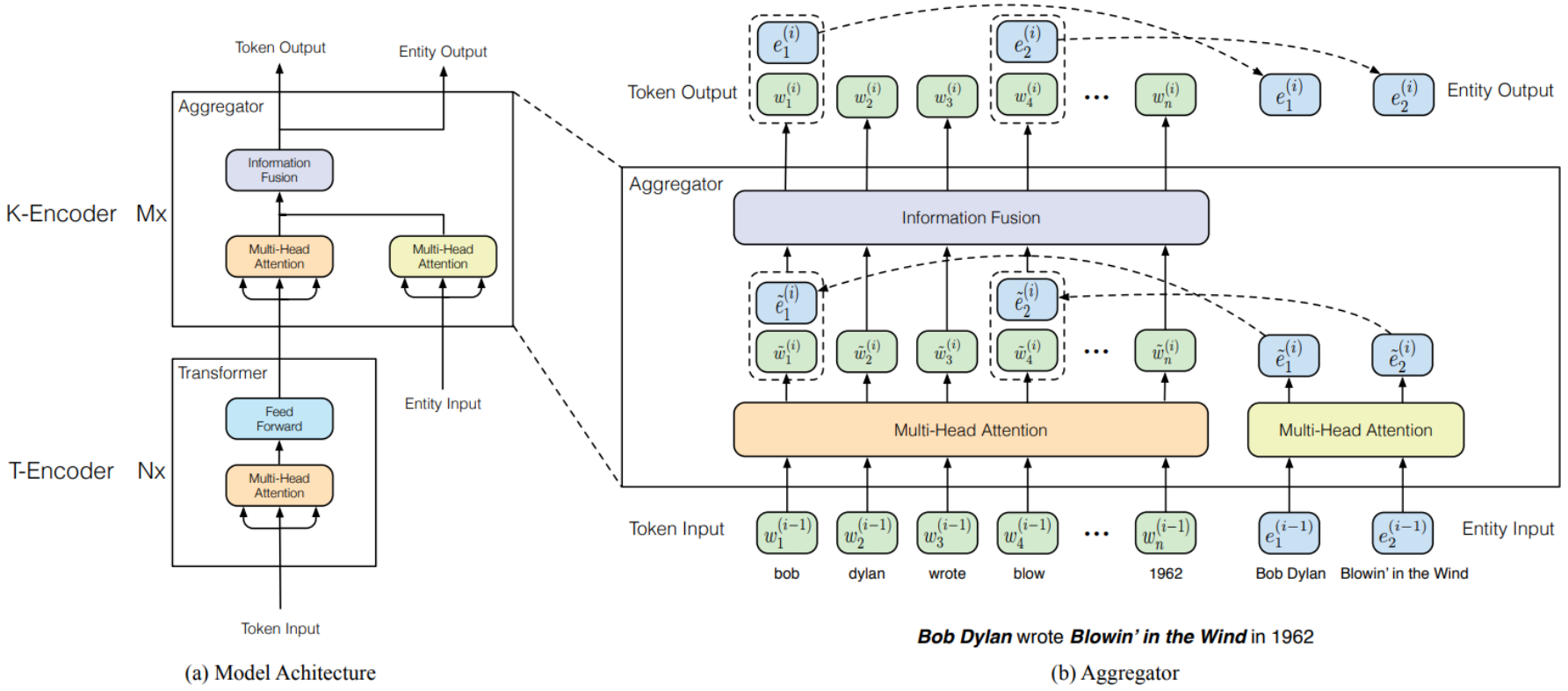
为语言理解注入外部知识



Bob Dylan wrote ***Blowin' in the Wind*** in 1962, and wrote ***Chronicles: Volume One*** in 2004.

向BERT注入外部知识：Ernie

- 采用TransE预训练的实体嵌入
- 知识编码 + 实体预测任务



(Credit: Zhang et al., 2019, ERNIE: Enhanced Language Representation with Informative Entities)

向BERT注入外部知识：Ernie

- 普通NLU任务上取得与BERT可比的结果
- 知识驱动的NLU任务上显著超过BERT

Entity Typing tasks:

Model	P	R	F1
NFGEC (LSTM)	68.80	53.30	60.10
UFET	77.40	60.60	68.00
BERT	76.37	70.96	73.56
ERNIE	**78.42**	**72.90**	**75.56**

Table 3: Results of various models on Open Entity (%).

Relation Classification tasks:

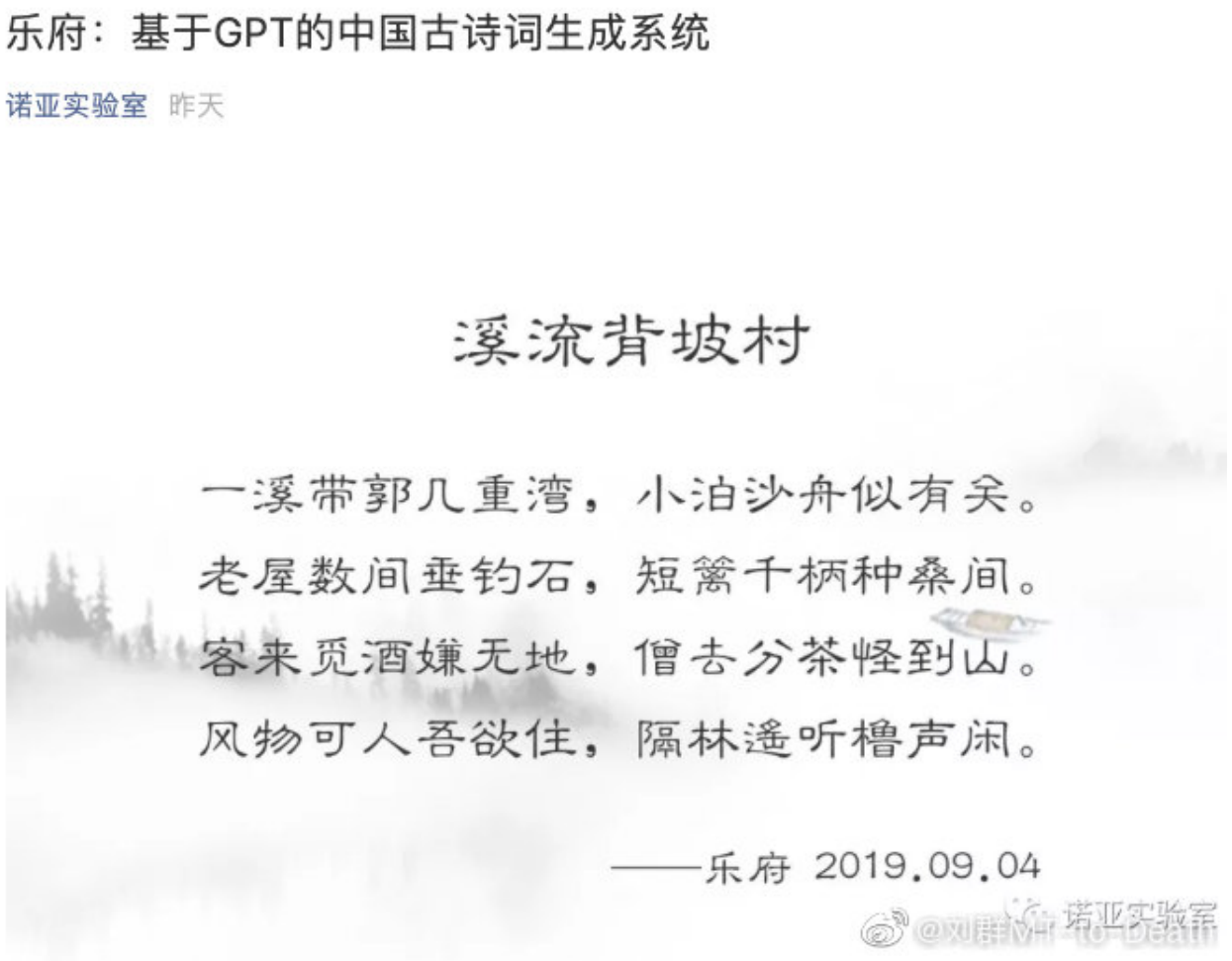
Model	FewRel			TACRED		
	P	R	F1	P	R	F1
CNN	69.51	69.64	69.35	70.30	54.20	61.20
PA-LSTM	-	-	-	65.70	64.50	65.10
C-GCN	-	-	-	69.90	63.30	66.40
BERT	85.05	85.11	84.89	67.23	64.81	66.00
ERNIE	88.49	88.44	**88.32**	69.97	66.08	**67.97**

Table 5: Results of various models on FewRel and TACRED (%).

(Credit: Zhang et al., 2019, ERNIE: Enhanced Language Representation with Informative Entities)

内容概要

乐府：基于GPT的中国古诗词生成系统



微信小程序：

论文下载：https://arxiv.org/abs/1907.00151

乐府：基于GPT的中国古诗词生成系统

- 系统：基于BERT代码开发的GPT-2系统
- 预训练：新闻语料（2.35亿句子）
- Fine-tune：
  - 绝句和律诗：25万
  - 词：2万
  - 对联：70万
- 生成策略：Top-K随机取样
- 我们没有使用任何韵律、平仄、句子长度的规则信息，完全通过语言模型训练自动学习这些作诗规则

(Credit: Liao et al., 2019, GPT-based Generation for Classical Chinese Poetry)

乐府：基于GPT的中国古诗词生成系统

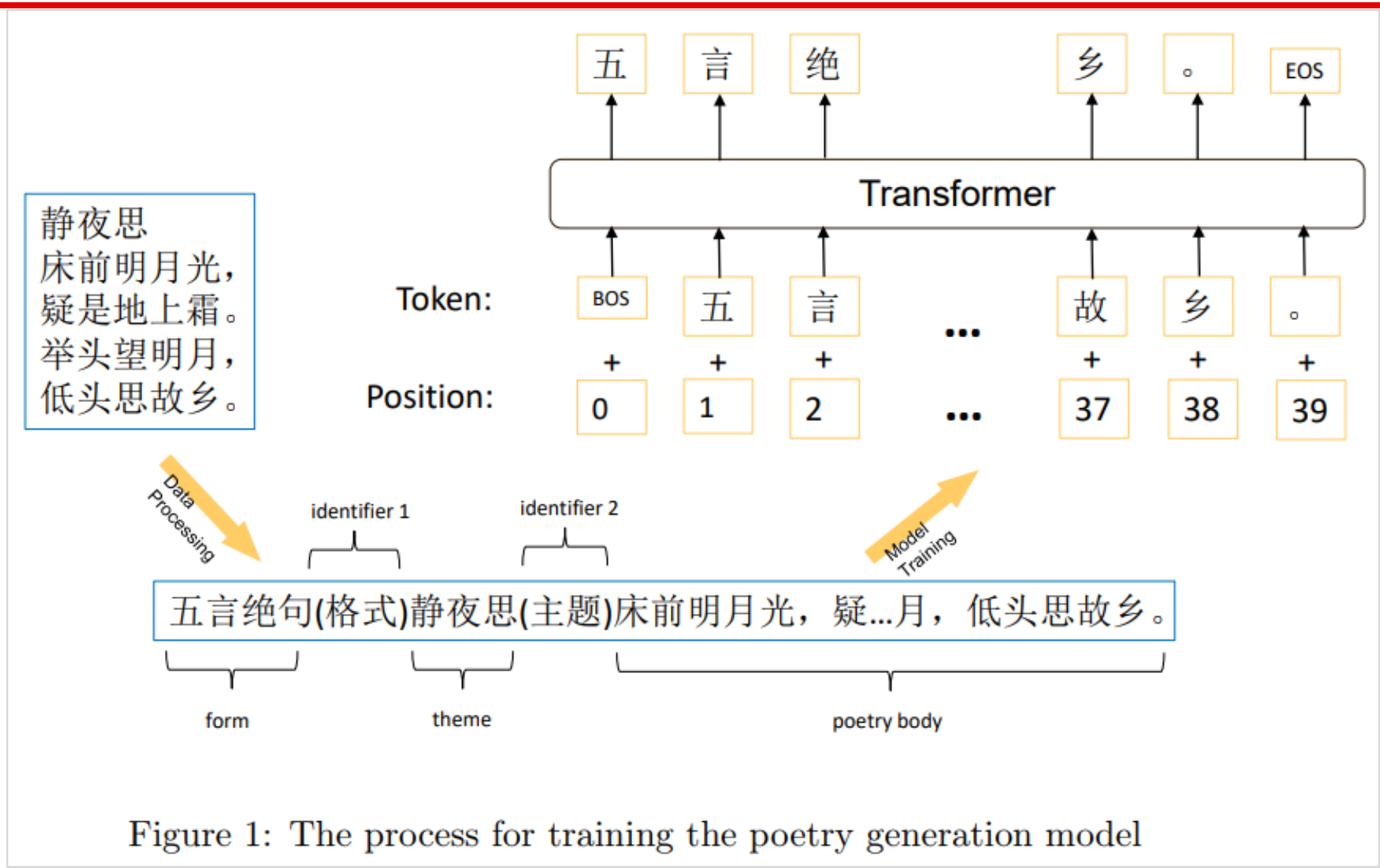
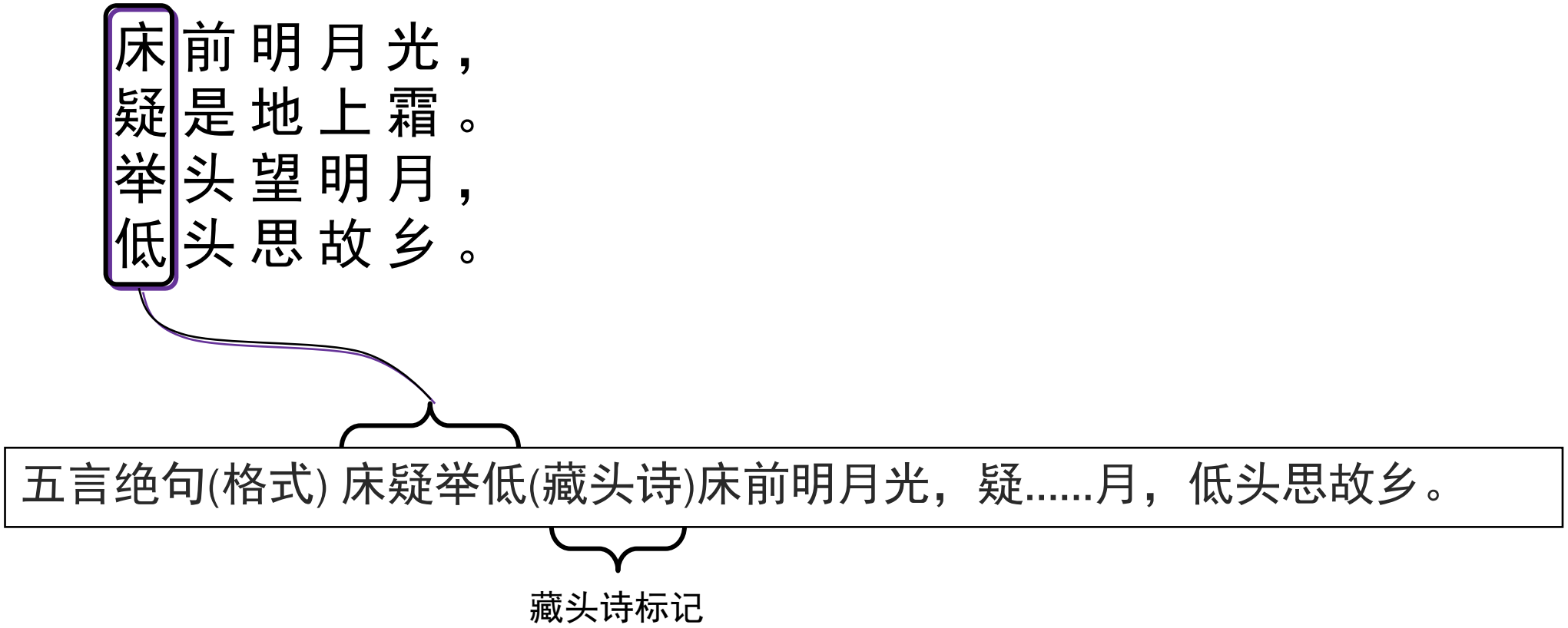


Figure 1: The process for training the poetry generation model

(Credit: Liao et al., 2019, GPT-based Generation for Classical Chinese Poetry)

乐府：基于GPT的中国古诗词生成系统



藏头诗训练语料的处理

【乐府】生成对联样例

上联(First line)	下联(Second line)
一句相思吟岁月	几分寂寞醉诗词 三杯落泪诉年华 三分眷恋到天涯
风弦未拨心先乱	诗卷未题笔已枯 月盏虽清梦已酣 草色初青眼欲穿
海上飞燕飞上海	城外环山环外城 江中落叶落中江 山前雾气雾前山
水墨丹青，烟雨江南春纵笔	花红柳绿，桃源粤地燕裁云 诗词歌赋，风花塞外夏抒怀 天光风韵，云霞岭上日倾杯

叠字联

(Credit: Liao et al., 2019, GPT-based Generation for Classical Chinese Poetry)

【乐府】作诗机作品赏析

这四首诗只有一首是真的唐诗，其他三首是【乐府】的作品：

江上田家

村南喧鸟雀，江北梦悠悠。
桑熟蚕三眠，人家半依楼。
一身千万里，何处得穷愁。
日暮歌明月，长河满斛秋。

江上田家

江边田舍好，茅屋远相迎。
竹里开门入，芦中引水行。
犬来沙上吠，鸥去岸间鸣。
不是无吟兴，谁知乐此生。

江上田家

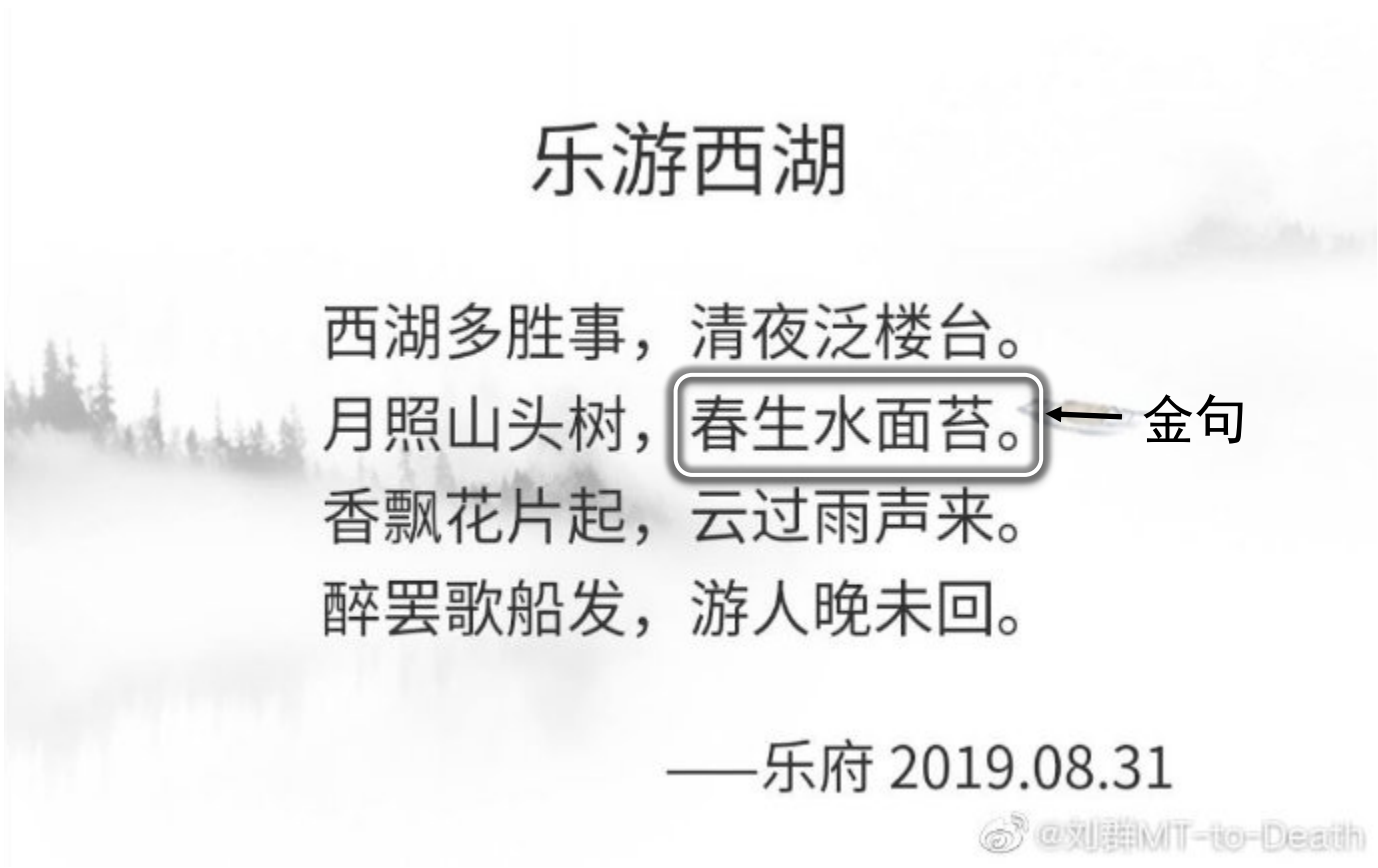
近海川原薄，人家本自稀。
黍苗期腊酒，霜叶是寒衣。
市井谁相识，渔樵夜始归。
不须骑马问，恐畏狎鸥飞。

江上田家

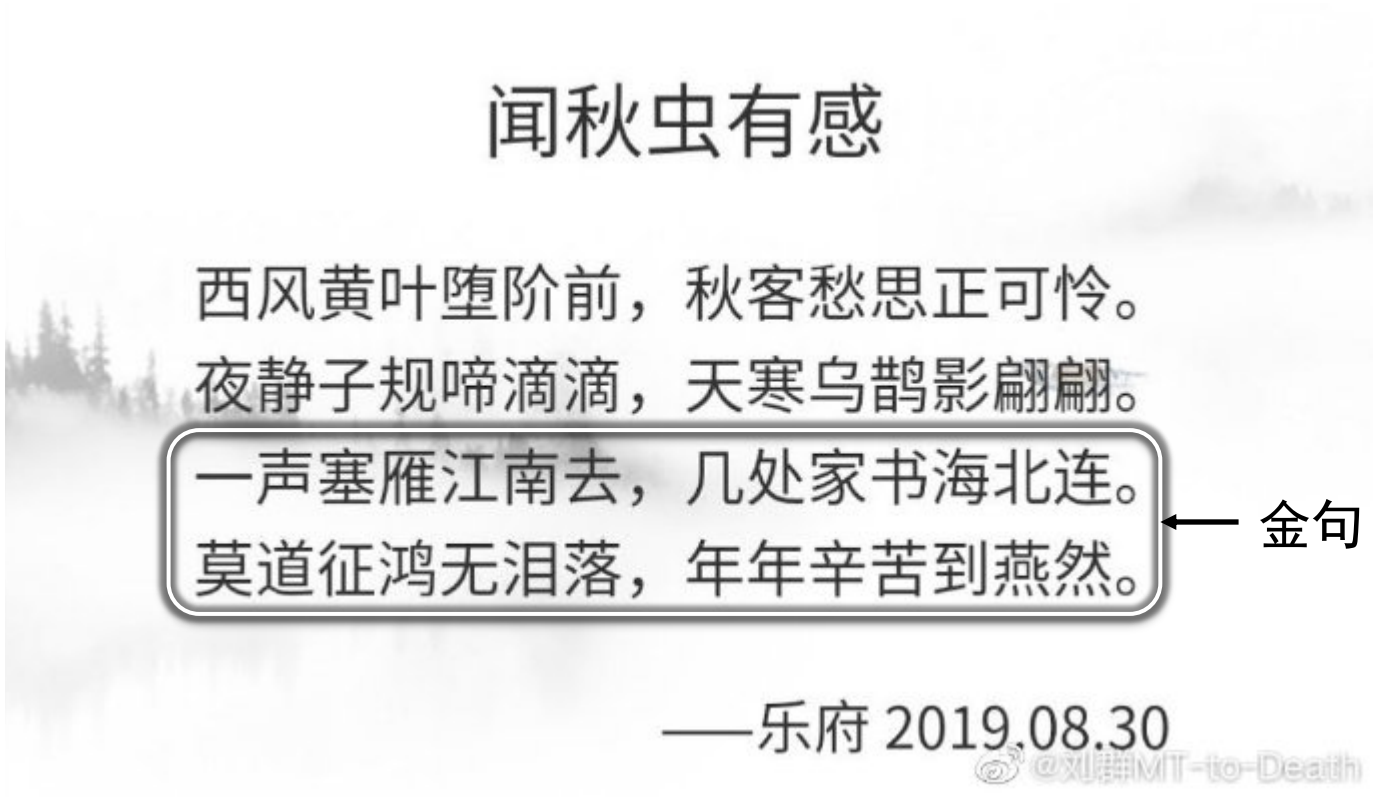
野水通渔路，江村带夕阳。
数家深竹里，一树隔芦塘。
牧去牛羊下，人行果橘旁。
相逢皆贺岁，还有醉眠乡。

受过一定古诗训练的人还是能分辨出来的，但普通人不太容易分辨

【乐府】作诗机作品赏析


乐游西湖
西湖多胜事，清夜泛楼台。
月照山头树，春生水面苔。
香飘花片起，云过雨声来。
醉罢歌船发，游人晚未回。
——乐府 2019.08.31
金句

【乐府】作诗机作品赏析


闻秋虫有感
西风黄叶堕阶前，秋客愁思正可怜。
夜静子规啼滴滴，天寒乌鹊影翩翩。
一声塞雁江南去，几处家书海北连。
莫道征鸿无泪落，年年辛苦到燕然。
——乐府 2019.08.30
金句

【乐府】作诗机：一些观察到的现象

- 总体上句子长度、押韵、平仄都不错，但严格检查起来还是有一些不符合的地方。
- 扣题不错，能够找到跟主题相关的词语写进诗里，使人感觉总体上形成一定的意境
- 有一定的起承转合，最后一句通常会点题。
- 偶尔会出现一些非常好的金句。
- 七言诗比五言诗效果好。
- 作词成功率比较低（70%多），可能是语料比较少造成的。
- 会犯一些逻辑错误和常识错误。
- 作出的诗的质量跟标题关系比较大：
  - 在传统古诗中比较常见的主题容易作出好诗，如写景抒情类的题目；
  - 纯现代主题作出的诗相对质量较低，例如与科学相关的标题（如机器翻译）有时会写成佛道主题的诗。

相关工作

CTRL: A Conditional Transformer Language Model for Controllable Generation

Nitish Shirish Keskar[*], **Bryan McCann**[*], **Lav R. Varshney**, **Caiming Xiong**, **Richard Socher**
Salesforce Research[†]

Abstract

Large-scale language models show promising text generation capabilities, but users cannot easily control particular aspects of the generated text. We release CTRL, a 1.63 billion-parameter conditional transformer language model, trained to condition on control codes that govern style, content, and task-specific behavior. Control codes were derived from structure that naturally co-occurs with raw text, preserving the advantages of unsupervised learning while providing more explicit control over text generation. These codes also allow CTRL to predict which parts of the training data are most likely given a sequence. This provides a potential method for analyzing large amounts of data via model-based source attribution. We have released multiple full-sized, pretrained versions of CTRL at https://github.com/salesforce/ctrl.

arXiv:1909.05858v2 [cs.CL] 20 Sep 2019

相关工作

Wikipedia *Anarchism is* a political philosophy that advocates the abolition of all forms of hierarchy

Books *Anarchism is* the \n only true and practical form of Socialism. It has been said that Socialism

Horror *A knife* handle pulled through the open hole in the front. I jumped when the knife

Reviews *A knife* is a tool and this one does the job well.\n\nRating: 4.0\n\nI bought these for my

Relationships *My neighbor is* a jerk and I don't know what to do\n\nText: So my neighbors

Legal *My neighbor is* threatening to sue me for not letting him use my pool\n\nText: I live in a

Science Title: Scientists have discovered a new type of bacteria that

Politics Title: The US is the only country in history to have a national debt of more than

Running Text: I have been running for about a year and a half now but never really got into it.\n\n

Reviews Rating: 5.0\n\n I have been using this product for a few years and it is the best thing

内容概要

高效的BERT压缩模型：TinyBERT

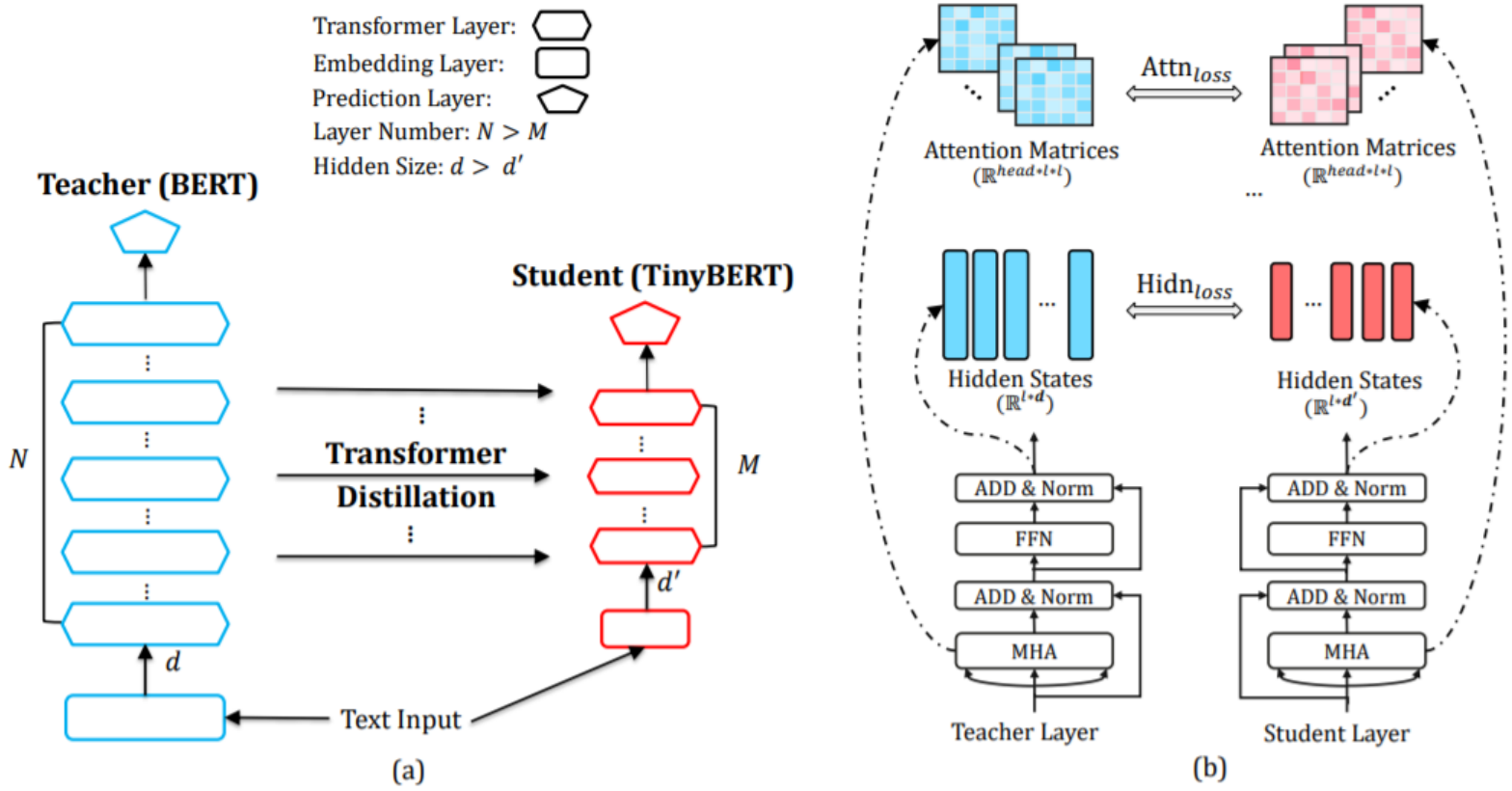
- BERT性能强大，但不便于部署到算力、内存有限的设备上
- 提出一种专为Transformer模型设计的知识蒸馏方法，以BERT作为老师蒸馏出一个小模型—TinyBERT
- TinyBERT参数量为 BERT 的 1/7，预测速度是 BERT 的 9 倍，在GLUE评测上相比BERT下降3个百分点

System	MNLI-m	MNLI-mm	QQP	SST-2	QNLI	MRPC	RTE	CoLA	STS-B	Average
$\text{BERT}_{\text{BASE}}$ (Google)	84.6	83.4	71.2	93.5	90.5	88.9	66.4	52.1	85.8	79.6
$\text{BERT}_{\text{BASE}}$ (Teacher)	83.9	83.4	71.1	93.4	90.9	87.5	67.0	52.8	85.2	79.5
$\text{BERT}_{\text{SMALL}}$	75.4	74.9	66.5	87.6	84.8	83.2	62.6	19.5	77.1	70.2
Distilled $\text{BiLSTM}_{\text{SOFT}}$	73.0	72.6	68.2	90.7	-	-	-	-	-	-
BERT-PKD	79.9	79.3	70.2	89.4	85.1	82.6	62.3	24.8	79.8	72.6
DistilBERT	78.9	78.0	68.5	91.4	85.2	82.4	54.1	32.8	76.1	71.9
TinyBERT	82.5	81.8	71.3	92.6	87.7	86.4	62.9	43.3	79.9	76.5

System	Layers	Hidden Size	Feed-forward Size	Model Size	Inference Time
$\text{BERT}_{\text{BASE}}$ (Teacher)	12	768	3072	109M(×1.0)	188s(×1.0)
Distilled $\text{BiLSTM}_{\text{SOFT}}$	1	300	400	10.1M(×10.8)	24.8s(×7.6)
BERT-PKD/DistilBERT	4	768	3072	52.2M(×2.1)	63.7s(×3.0)
TinyBERT	4	312	1200	14.5M(×7.5)	19.9s(×9.4)

高效的BERT压缩模型：TinyBERT

- Transformer蒸馏



深度语义理解：TinyBERT

- TinyBERT蒸馏：预训练蒸馏+下游任务蒸馏+数据增强

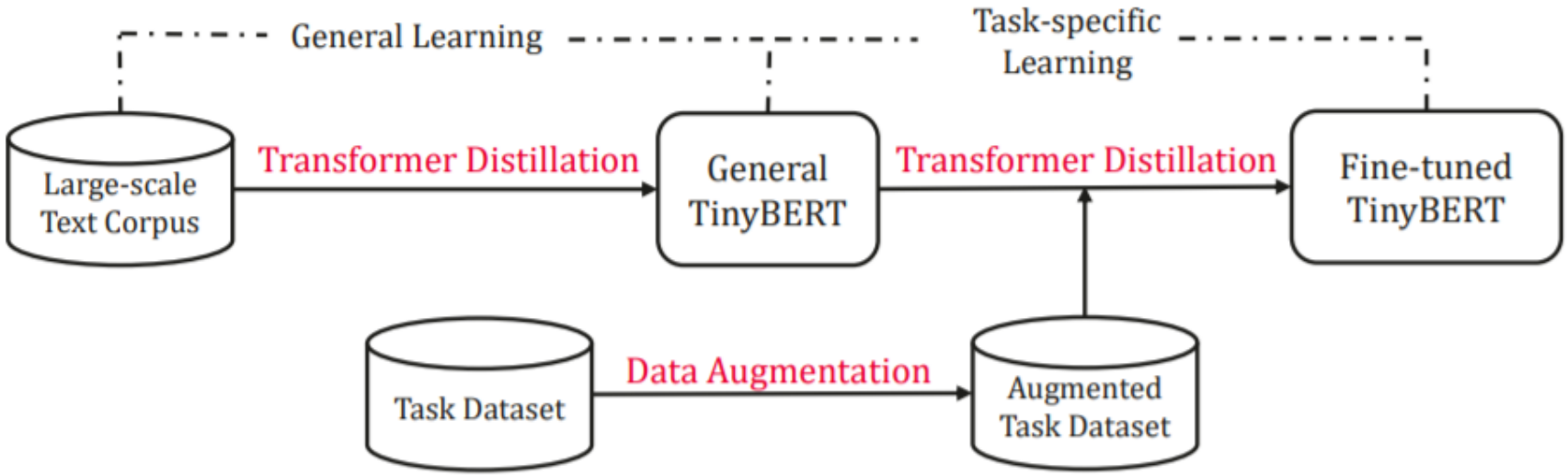


Figure 2: The illustration of TinyBERT learning

System	MNLI-m	MNLI-mm	MRPC	CoLA	Average
TinyBERT	82.8	82.9	85.8	49.7	75.3
No GD	82.5	82.6	84.1	40.8	72.5
No TD	80.6	81.2	83.8	28.5	68.5
No DA	80.5	81.0	82.4	29.8	68.4

相关工作

ALBERT: A LITE BERT FOR SELF-SUPERVISED LEARNING OF LANGUAGE REPRESENTATIONS

- Google新推出ALBERT模型
- 主要改进:
  - Factorized embedding parameterization
  - Cross-layer parameter sharing
  - Inter-sentence coherence loss
- 优点：
  - 大幅度减少模型参数，并加快训练速度
  - 通过加深模型，可以在参数减少的情况下获得更好的性能
- 缺点：
  - 模型如果不加深性能会有较多的下降
  - 模型加深后推理时间增加了

内容概要

小结

- 我们在预训练语言模型方向开展了一系列工作
- 重现了主流的预训练语言模型
- 在华为云上实现了大规模并行化训练并做了针对性优化
- 对预训练语言模型的技术细节进行了深入的研究，并推出了中文预训练语言模型【哪吒】
- 与清华大学合作尝试将实体知识融入BERT开发了【ERNIE】
- 研究了基于GPT模型的中国古诗词生成方法，推出了【乐府】作诗机
- 提出来一种高效的BERT压缩算法：TinyBERT
- 探索了预训练语言模型在多标签分类、对话生成、搜索推荐等领域的应用

展望

- 研究更好、更强大的预训练语言模型
  - 更好地融入知识
  - 跟语音、图像结合
- 更好地应用预训练语言模型
  - 应用在更多领域
- 模型的压缩和优化：在终端落地
- 在华为自研的AI芯片上实现和优化

Security Level:

谢谢关注

www.huawei.com