关于预训练大模型发展前景的一些思考和探讨

Qun Liu 刘群

Huawei Noah's Ark Lab 华为诺亚方舟实验室

全国人工智能大会，自然语言处理论坛
成都，2021-10-13

NOAH'S ARK LAB

HUAWEI

Content

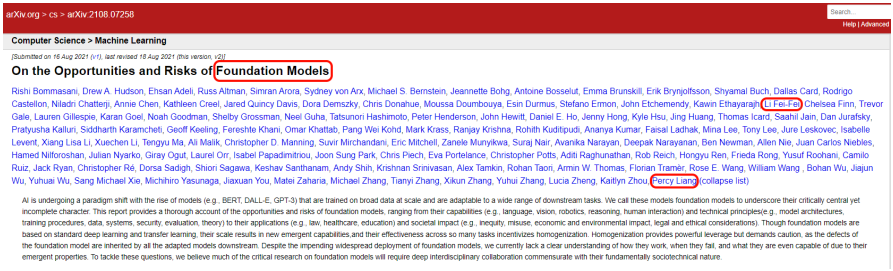
Content

什么是预训练大模型？

- 人工智能研究的新范式
  - 深度学习模型
  - 参数规模大、训练数据大
  - 非特定任务预训练，可以应用于广泛的下游任务
- 又被称为基础模型（Foundation Models）

什么是预训练大模型？

- 人工智能研究的新范式
  - 深度学习模型
  - 参数规模大、训练数据大
  - 非特定任务预训练，可以应用于广泛的下游任务
- 又被称为基础模型（Foundation Models）



arXiv.org > cs > arXiv:2108.07258

Search...

Help | Advanced

Computer Science > Machine Learning

[Submitted on 16 Aug 2021 (v1), last revised 18 Aug 2021 (this version, v2)]

On the Opportunities and Risks of Foundation Models

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Kohd, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang , Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, Percy Liang (collapse list)

AI is undergoing a paradigm shift with the rise of models (e.g., BERT, DALL-E, GPT-3) that are trained on broad data at scale and are adaptable to a wide range of downstream tasks. We call these models foundation models to underscore their critically central yet incomplete character. This report provides a thorough account of the opportunities and risks of foundation models, ranging from their capabilities (e.g., language, vision, robotics, reasoning, human interaction) and technical principles(e.g., model architectures, training procedures, data, systems, security, evaluation, theory) to their applications (e.g., law, healthcare, education) and societal impact (e.g., inequity, misuse, economic and environmental impact, legal and ethical considerations). Though foundation models are based on standard deep learning and transfer learning, their scale results in new emergent capabilities,and their effectiveness across so many tasks incentivizes homogenization. Homogenization provides powerful leverage but demands caution, as the defects of the foundation model are inherited by all the adapted models downstream. Despite the impending widespread deployment of foundation models, we currently lack a clear understanding of how they work, when they fail, and what they are even capable of due to their emergent properties. To tackle these questions, we believe much of the critical research on foundation models will require deep interdisciplinary collaboration commensurate with their fundamentally sociotechnical nature.

Bommasani et al., On the Opportunities and Risks of Foundation Models, arXiv:2108.07258 [cs.LG]

涌现和同质化

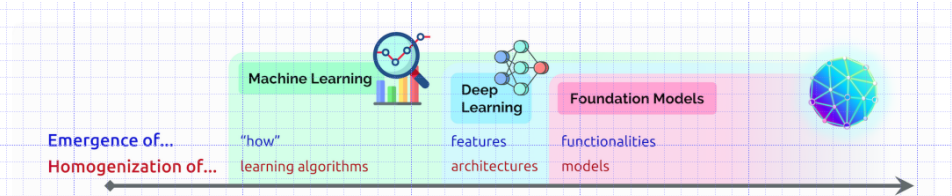
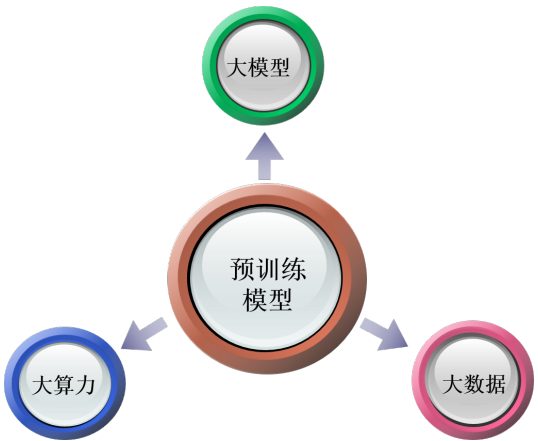


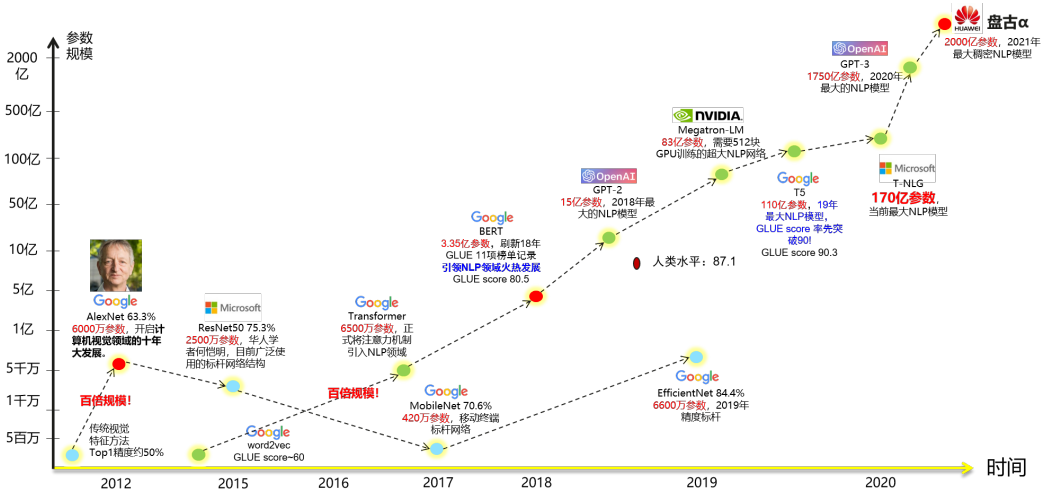
Fig. 1. The story of AI has been one of increasing *emergence* and *homogenization*. With the introduction of machine learning, *how* a task is performed emerges (is inferred automatically) from examples; with deep learning, the high-level features used for prediction emerge; and with foundation models, even advanced functionalities such as in-context learning emerge. At the same time, machine learning homogenizes learning algorithms (e.g., logistic regression), deep learning homogenizes model architectures (e.g., Convolutional Neural Networks), and foundation models homogenizes the model itself (e.g., GPT-3).

Bommasani et al., On the Opportunities and Risks of Foundation Models, arXiv:2108.07258 [cs.LG]

预训练大模型的特点

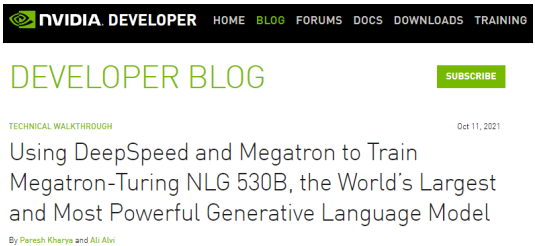


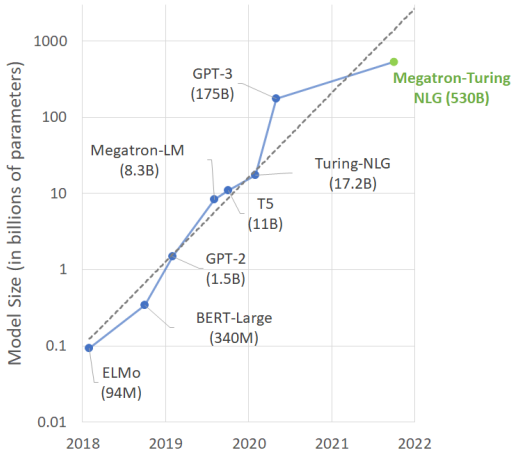
预训练大模型的参数规模



预训练大模型的参数规模

2021-10-11 刚刚发布



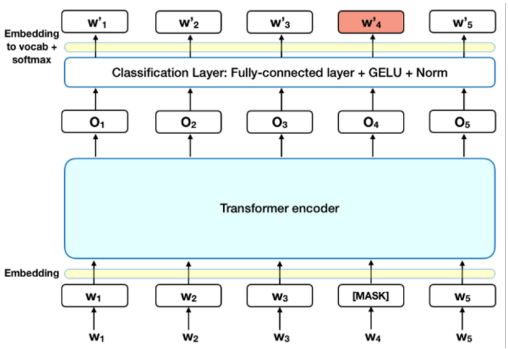


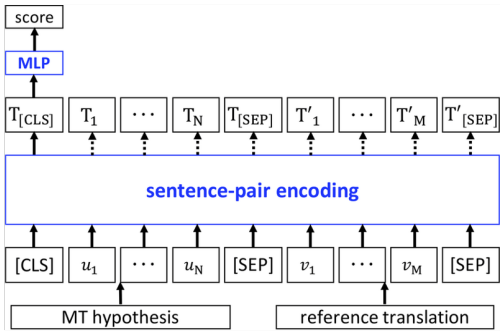
HUAWEI NOAH'S ARK LAB

预训练大模型给我们带来了什么？

- 海量无标注或弱标注数据的利用（自监督学习）
- 预训练+微调框架：下游任务模型结构的简化+性能的普遍提高
- 少样本和零样本的学习
- 多语言表达能力
- 多模态交互

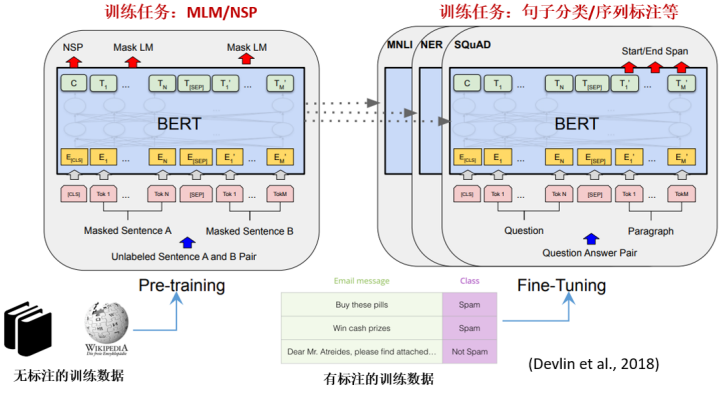
海量无标注或弱标注数据的利用（自监督学习）





Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805, 2018

预训练+微调框架：下游任务模型结构的简化 / 性能的普遍提高



Pre-traing得到精确有效的语言表达

[Mask][Mask][Mask][Mask]歌曲

[帮][我][搜][索]歌曲
[播][放][一][首]歌曲
[给][我][搜][索]歌曲
[给][我][播][放]歌曲
[给][我][放][首]歌曲
[给][我][唱][首]歌曲
[帮][我][播][放]歌曲

N=1 N=2 N=4 N=8 N=16 N=32 N=64 N=512

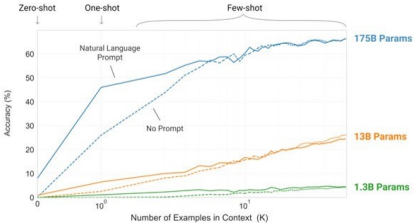
I love peanut butter and *jelly* sandwiches.

I love peanut butter and *jelly, Yum! You can't beat peanut butter and jelly* sandwiches.

I love peanut butter and *bread. Thanks!! This looks delicious. I love all types of peanut butter, but especially peanut butter/jam* sandwiches.

Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805, 2018

少样本和零样本的学习



Brown et al., Language Models are Few-Shot Learners, arXiv:2005.14165, 2021

The three settings we explore for in-context learning

Zero-shot

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

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

One-shot

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

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

Few-shot

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

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

Traditional fine-tuning (not used for GPT-3)

Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.

```
sea otter => loutre de mer      ← example #1
        ↓
   gradient update
        ↓
peppermint => menthe poivrée    ← example #2
        ↓
   gradient update
        ↓
        ...
        ↓
plush giraffe => girafe peluche ← example #N
        ↓
   gradient update
        ↓
cheese =>                       ← prompt
```

多语言表达能力

多语言表达能力

Models

There are two multilingual models currently available. We do not plan to release more single-language models, but we may release `BERT-Large` versions of these two in the future:

- **`BERT-Base, Multilingual Cased (New, recommended)`**: 104 languages, 12-layer, 768-hidden, 12-heads, 110M parameters
- **`BERT-Base, Multilingual Uncased (Orig, not recommended)`**: 102 languages, 12-layer, 768-hidden, 12-heads, 110M parameters
- **`BERT-Base, Chinese`**: Chinese Simplified and Traditional, 12-layer, 768-hidden, 12-heads, 110M parameters

Data Source and Sampling

The languages chosen were the top 100 languages with the largest Wikipedias. The entire Wikipedia dump for each language (excluding user and talk pages) was taken as the training data for each language

Model	D	#M	#lg	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Avg
Fine-tune multilingual model on English training set (Cross-lingual Transfer)																			
mBERT	Wiki	N	102	82.1	73.8	74.3	71.1	66.4	68.9	69.0	61.6	64.9	69.5	55.8	69.3	60.0	50.4	58.0	66.3
XLM (MLM+TLM)	Wiki+MT	N	15	85.0	78.7	78.9	77.8	76.6	77.4	75.3	72.5	73.1	76.1	73.2	76.5	69.6	68.4	67.3	75.1
XLM-R	CC	1	100	**88.8**	**83.6**	**84.2**	**82.7**	**82.3**	**83.1**	**80.1**	**79.0**	**78.8**	**79.7**	**78.6**	**80.2**	**75.8**	**72.0**	**71.7**	**80.1**
Translate everything to English and use English-only model (TRANSLATE-TEST)																			
BERT-en	Wiki	1	1	88.8	81.4	82.3	80.1	80.3	80.9	76.2	76.0	75.4	72.0	71.9	75.6	70.0	65.8	65.8	76.2
RoBERTa	CC	1	1	**91.3**	82.9	84.3	81.2	81.7	83.1	78.3	76.8	76.6	74.2	74.1	77.5	70.9	66.7	66.8	77.8
Fine-tune multilingual model on each training set (TRANSLATE-TRAIN)																			
XLM (MLM)	Wiki	N	100	82.9	77.6	77.9	77.9	77.1	75.7	75.5	72.6	71.2	75.8	73.1	76.2	70.4	66.5	62.4	74.2
Fine-tune multilingual model on all training sets (TRANSLATE-TRAIN-ALL)																			
XLM (MLM+TLM)	Wiki+MT	1	15	85.0	80.8	81.3	80.3	79.1	80.9	78.3	75.6	77.6	78.5	76.0	79.5	72.9	72.8	68.5	77.8
XLM (MLM)	Wiki	1	100	84.5	80.1	81.3	79.3	78.6	79.4	77.5	75.2	75.6	78.3	75.7	78.3	72.1	69.2	67.7	76.9
XLM-R	CC	1	100	**88.7**	**85.2**	**85.6**	**84.6**	**83.6**	**85.5**	**82.4**	**81.6**	**80.9**	**83.4**	**80.9**	**83.3**	**79.8**	**75.9**	**74.3**	**82.4**

https://github.com/google-research/bert/blob/master/multilingual.md

多模态交互

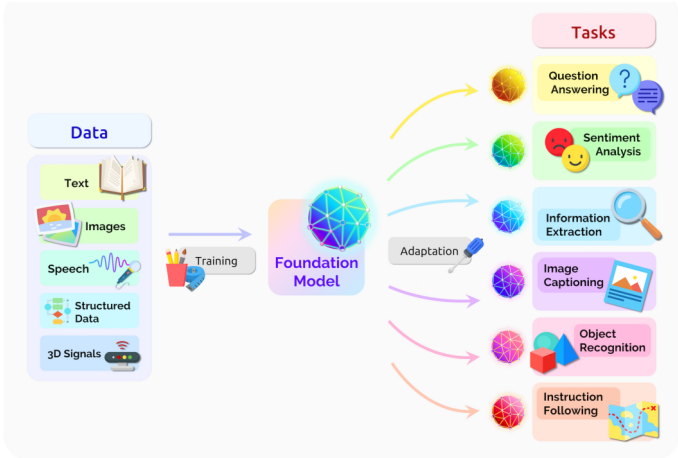
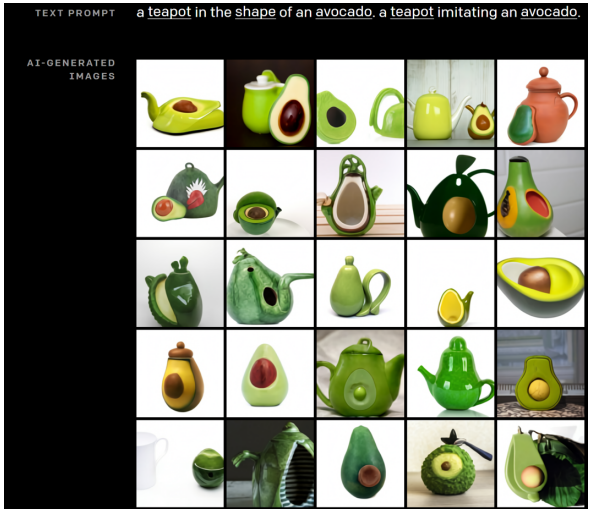


Fig. 2. A foundation model can centralize the information from all the data from various modalities. This one model can then be adapted to a wide range of downstream tasks.

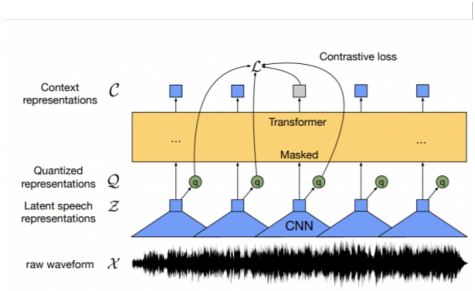
Bommasani et al., On the Opportunities and Risks of Foundation Models, arXiv:2108.07258 [cs.LG]

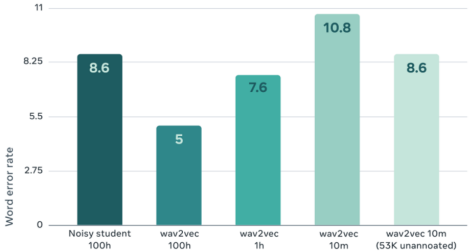
多模态交互



OpenAI DALL-E demo, source: https://openai.com/blog/dall-e/

多模态交互





WER for Noisy Student self-training with 100 hours of labeled data. Wav2vec 2.0 with 100 hours, 1 hour, and only 10 minutes of labeled data. All models use the remainder of the LibriSpeech corpus (total 960 hours) as unannotated data, except for the last result, which uses 53K hours from LibriVox.

Facebook AI Wav2Vec 2.0 https://ai.facebook.com/blog/wav2vec-20-learning-the-structure-of-speech-from-raw-audio/

Content

Content

预训练大模型的研究现状和发展趋势

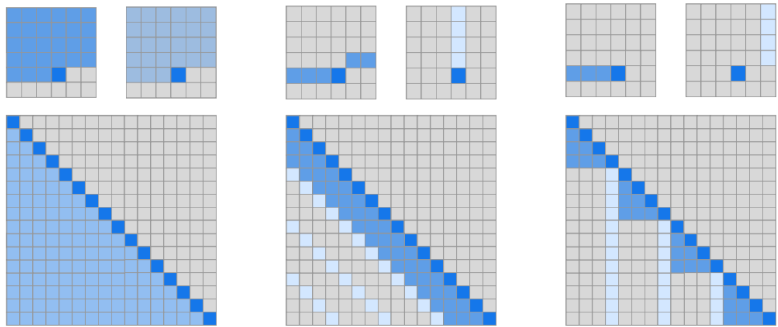
- 如何做得更大？
- 如何更能干？
- 如何更有效地训练？
- 如何应用和推理？
- 如何确保安全可信？

如何做得更大？

- 数据：
  - 更大：数据规模的增长没有止境
  - 更多样：更多样化、多模态的数据将融入预训练大模型
- 模型：
  - 模型参数数量反应了模型的容量（Capacity）
  - 模型的参数规模还可以大幅度增长
- 算力：
  - 单一集中式模型所使用的算力几乎达到极限，除非出现新的计算模式（如量子计算）
  - 稀疏模型和分布式模型将是发展趋势：采用合适的稀疏模型（如MoE，Mixture-of-Experts），可以在不增加算力消耗的情况下，大幅度增加模型参数和模型表达能力（Capacity）

稀疏模型 Sparse Transformers

- Sparse factorizations of the attention matrix which reduce this to $O(n\sqrt{n})$:



(a) Transformer (b) Sparse Transformer (strided) (c) Sparse Transformer (fixed)

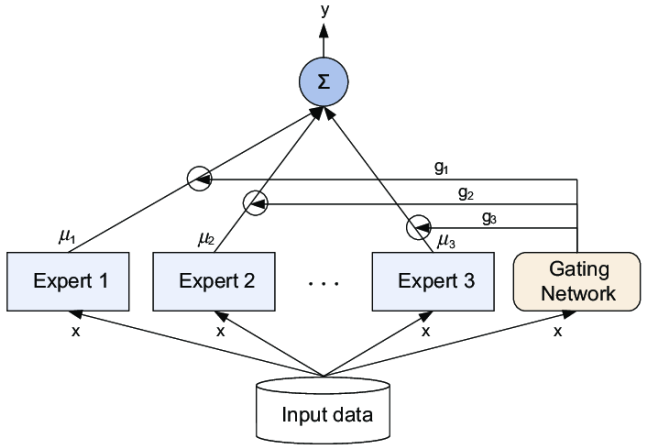
Child et al., Generating Long Sequences with Sparse Transformers, arXiv:1904.10509

稀疏模型 Sparse Transformers

- Related work:
  - Big Bird (Zaheer et al. 2020, NeurIPS),
  - Longformer (Beltagy et al. 2020),
  - Reformer (Kitaev et al. 2020, ICLR),
  - Routing Transformer (Roy et al. 2021, ACL),

稀疏模型 MoE Transformers

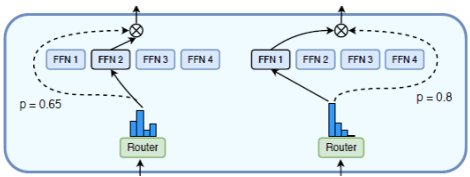
- Introduces Mixture-of-Experts (MoE) in Transformer components

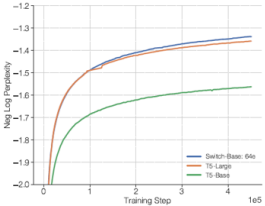


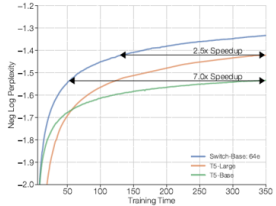
Jason Brownlee, A Gentle Introduction to Mixture of Experts Ensembles (blog)

稀疏模型 MoE Transformers

- Switch Transformers (Google, 2021.01)
  - Backbone: T5
  - Parameters: 1571B, 15 layers, 2048 experts
  - Dataset: C4 (180B tokens)
  - Router: switch routing (top-1)







Fedus et al., Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. arXiv:2101.03961, 2021

稀疏模型 MoE Transformers

- Top-1 routing (Google)
  - 单个expert可以减少通信代价，提升训练速度
  - Fedus et al., Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. arXiv:2101.03961, 2021
- Top-K routing (Google)
  - 通常从N个expert中选择2个进行稀疏路由
  - Shazeer et al., Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. arXiv:1701.06538, 2017
- Hash routing (Facebook)
  - 不需要router的学习，通过设定token-expert的映射来指导路由
  - Roller et al., Hash Layers For Large Sparse Models. arXiv:2106.04426, 2021
- Domain routing (AI2 & Facebook)
  - 对不同领域数据设置不同的expert，根据领域进行路由
  - Gururangan et al., DEMix Layers: Disentangling Domains for Modular Language Modeling. arXiv:2108.05036. 2021

Content

预训练大模型的研究现状和发展趋势

- 如何做得更大？
- 如何更能干？
- 如何更有效地训练？
- 如何应用和推理？
- 如何确保安全可信？

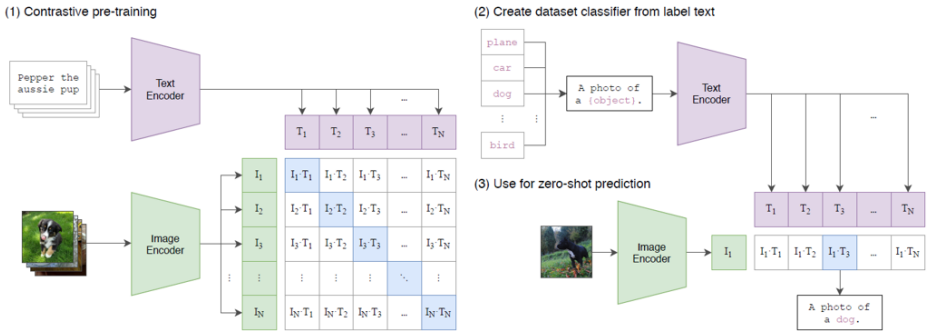
异构数据引入

- 大规模预训练语言模型强大的能力可以从更多样的数据中吸收知识
- 同时，更多样的知识来源可以互相增强，使得大规模预训练语言模型更加强大
- 异构知识来源：
  - 多模态融入
  - 知识融入：不同模式
  - 文本融入：与检索技术融合
  - 程序代码

图文预训练模型: 关键在于怎么进行模态之间的交互

- 双塔模型：通过对比学习Loss进行模态交互: CLIP, ALIGN, WENLAN
- 单塔模型：将图文特征拼接成一个序列通过Transformer模型Encoder或（和）Decoder的self-attention进行模态交互:
  - Encoder: VILT, SOHO
  - Decoder: DALL-E, Frozen
  - Mix: M6, OPT
- 其他：通过Encoder-decoder结构中Decoder的cross-attention进行模态交互: ALBEF

图文预训练模型: CLIP：典型双塔模型



Connecting Text and Images by Contrastive Language-Image Pre-training, OpenAI 2021

图文预训练模型: CLIP：典型双塔模型

- 技术创新：多模态对比学习
  - 使用图像和文本的global feature来进行对比学习
- 大规模数据库构建：
  - OpenAI 4亿单语言数据库
- 大规模训练：
  - CLIP_SMALL: VIT-B/32 + GPT(12L-8head-emb512)
  - CLIP_LARGE: VIT-L/14 + GPT-BASE(12L-12head-emb768)
- 多种下游任务：包括zero-shot图像分类，image-text 检索

- Zero-shot Image Classification

		Food101	CIFAR10	CIFAR100	Birdsnap	SUN397	Stanford Cars	FGVC Aircraft	VOC2007	DTD	Oxford Pets	Caltech101	Flowers102	MNIST	FER2013	STL10	EuroSAT	RESISC45	GTSRB	KITTI	Country211	PCam	UCF101	Kinetics700	CLEVR	HatefulMemes	Rendered SST2	ImageNet
CLIP-ResNet	RN50	81.1	75.6	41.6	32.6	59.6	55.8	19.3	82.1	41.7	85.4	82.1	65.9	66.6	42.2	94.3	41.1	54.2	35.2	42.2	16.1	57.6	63.6	43.5	20.3	59.7	56.9	59.6
	RN101	83.9	81.0	49.0	37.2	59.9	62.3	19.5	82.4	43.9	86.2	85.1	65.7	59.3	45.6	96.7	33.1	58.5	38.3	33.3	16.9	55.2	62.2	46.7	28.1	61.1	64.2	62.2
	RN50x4	86.8	79.2	48.9	41.6	62.7	67.9	24.6	83.0	49.3	88.1	86.0	68.0	75.2	51.1	96.4	35.0	59.2	35.7	26.0	20.2	57.5	65.5	49.0	17.0	58.3	66.6	65.8
	RN50x16	90.5	82.2	54.2	45.9	65.0	72.3	30.3	82.9	52.8	89.7	87.6	71.9	80.0	56.0	97.8	40.3	64.4	39.6	33.9	24.0	62.5	68.7	53.4	17.6	58.9	67.6	70.5
	RN50x64	91.8	86.8	61.3	48.9	66.9	76.0	35.6	83.8	53.4	93.4	90.6	77.3	90.8	61.0	98.3	59.4	69.7	47.9	33.2	29.6	65.0	74.1	56.8	27.5	62.1	70.7	73.6
CLIP-ViT	B/32	84.4	91.3	65.1	37.8	63.2	59.4	21.2	83.1	44.5	87.0	87.9	66.7	51.9	47.3	97.2	49.4	60.3	32.2	39.4	17.8	58.4	64.5	47.8	24.8	57.6	59.6	63.2
	B/16	89.2	91.6	68.7	39.1	65.2	65.6	27.1	83.9	46.0	88.9	89.3	70.4	56.0	52.7	98.2	54.1	65.5	43.3	44.0	23.3	48.1	69.8	52.4	23.4	61.7	59.8	68.6
	L/14	92.9	96.2	77.9	48.3	67.7	77.3	36.1	84.1	55.3	93.5	92.6	78.7	87.2	57.5	99.3	59.9	71.6	50.3	23.1	32.7	58.8	76.2	60.3	24.3	63.3	64.0	75.3
	L/14-336px	93.8	95.7	77.5	49.5	68.4	78.8	37.2	84.3	55.7	93.5	92.8	78.3	88.3	57.7	99.4	59.6	71.7	52.3	21.9	34.9	63.0	76.9	61.3	24.8	63.3	67.9	76.2

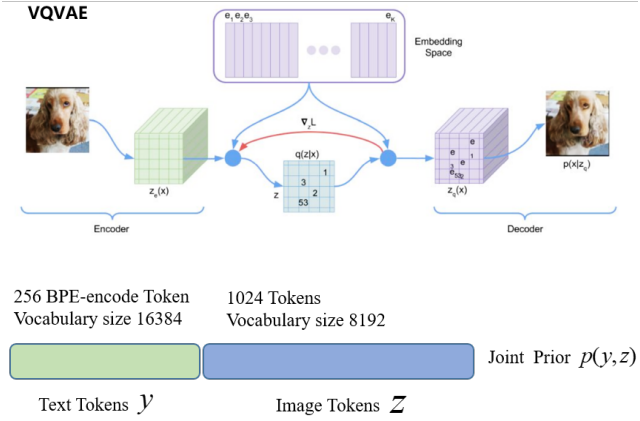
- Image-text retrieval

		Text Retrieval						Image Retrieval					
		Flickr30k			MSCOCO			Flickr30k			MSCOCO		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Finetune	Unicoder-VL[a]	86.2	96.3	99.0	62.3	87.1	92.8	71.5	90.9	94.9	46.7	76.0	85.3
	Uniter[b]	87.3	98.0	99.2	65.7	88.6	93.8	75.6	94.1	**96.8**	52.9	79.9	88.0
	VILLA[c]	87.9	97.5	98.8	-	-	-	76.3	**94.2**	**96.8**	-	-	-
	Oscar[d]	-	-	-	**73.5**	**92.2**	**96.0**	-	-	-	**57.5**	**82.8**	**89.8**
	ERNIE-ViL[e]	**88.7**	98.0	99.2	-	-	-	**76.7**	93.6	96.4	-	-	-

Connecting Text and Images by Contrastive Language-Image Pre-training, OpenAI 2021

图文预训练模型: Dall-E：典型单塔模型

- 视觉模态（numeric data）：用VQVAE等模型的encoder当成某个模态的contextualized tokenizer, decoder作为generator恢复到原本模态
- 文本(symbolic data)：本身就是离散的，普通的文本tokenizer
- 将视觉token和文本token连接成一个序列，用Autoregressive LM进行训练（类似GPT）



Zero-Shot Text-to-Image Generation. OpenAI, 2021

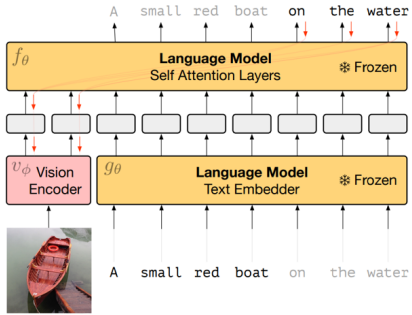
图文预训练模型: Dall-E：典型单塔模型

- Text-grounded Image Generation

Zero-Shot Text-to-Image Generation. OpenAI, 2021

图文预训练模型: Frozen：典型小样本模型

- 预训练
  - 模型：
    - 固定住7B纯文本预训练模型GPT，训练vision prefix (prompt)
  - 目标：
    - 使用Image Caption为训练目标在CC12M数据集上面fine-tune 一个NF-ResNet-50模型
  - 拥有类似gpt-3的跨模态few-shot(in-context) learning能力
- 和NLP的Pre-fix tuning的异同
  - 同：
    - 固定住纯文本预训练大模型，只fientune可学习的Prefix
  - 异：
    - 跨模态Prefix
    - 这个Prefix是sample-dependent的，不同的图片会产生不同的Prefix



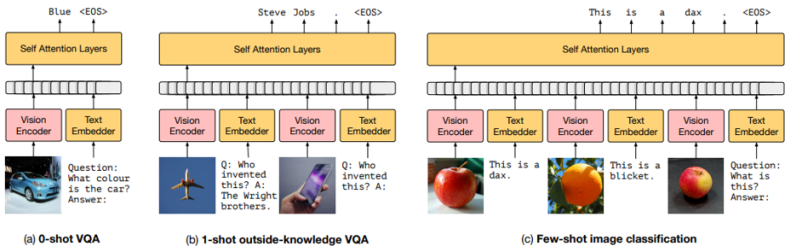
Multimodal Few-Shot Learning with Frozen Language Models, Deepmind 2021

图文预训练模型: Frozen：典型小样本模型

- VQA:

n-shot Acc.	n=0	n=1	n=4	τ
Frozen	29.5	35.7	38.2	✗
Frozen scratch	0.0	0.0	0.0	✗
Frozen finetuned	24.0	28.2	29.2	✗
Frozen train-blind	26.2	33.5	33.3	✗
Frozen VQA	48.4	–	–	✓
Frozen VQA-blind	39.1	–	–	✓
Oscar [23]	73.8	–	–	✓

- Inference:

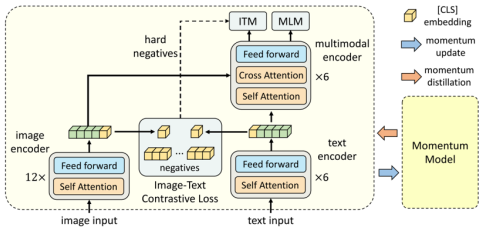


Multimodal Few-Shot Learning with Frozen Language Models, Deepmind 2021

图文预训练模型: ALBEF: 典型cross-attention模型

- 预训练
  - 模型：
    - 图像pre-trained VIT, 文本pre-trained BERT
  - 目标：
    - 图像VIT最后一层[CLS]特征和文本BERT第六层[CLS]特征做image-text contrastive learning
    - 文本部分做masked language modeling
    - 图像特征输入到文本decoder（BERT后六层）以cross-attention作了多模态交互之后做image-text matching
  - 拥有类似gpt-3的跨模态few-shot(in-context) learning能力
- 多种下游任务
  - 包括image-text 检索、VQA、VE、

Align before Fuse: Vision and Language Representation Learning with Momentum Distillation, Salesforce 2021



图文预训练模型: ALBEF: 典型cross-attention模型

- Image-text Retrieval:

Method	# Pre-train Images	Flickr30K (1K test set)						MSCOCO (5K test set)					
		TR			IR			TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
UNITER	4M	87.3	98.0	99.2	75.6	94.1	96.8	65.7	88.6	93.8	52.9	79.9	88.0
VILLA	4M	87.9	97.5	98.8	76.3	94.2	96.8	-	-	-	-	-	-
OSCAR	4M	-	-	-	-	-	-	70.0	91.1	95.5	54.0	80.8	88.5
ALIGN	1.2B	95.3	99.8	100.0	84.9	97.4	98.6	77.0	93.5	96.9	59.9	83.3	89.8
ALBEF	4M	94.3	99.4	99.8	82.8	96.7	98.4	73.1	91.4	96.0	56.8	81.5	89.2
ALBEF	14M	**95.9**	**99.8**	**100.0**	**85.6**	**97.5**	**98.9**	**77.6**	**94.3**	**97.2**	**60.7**	**84.3**	**90.5**

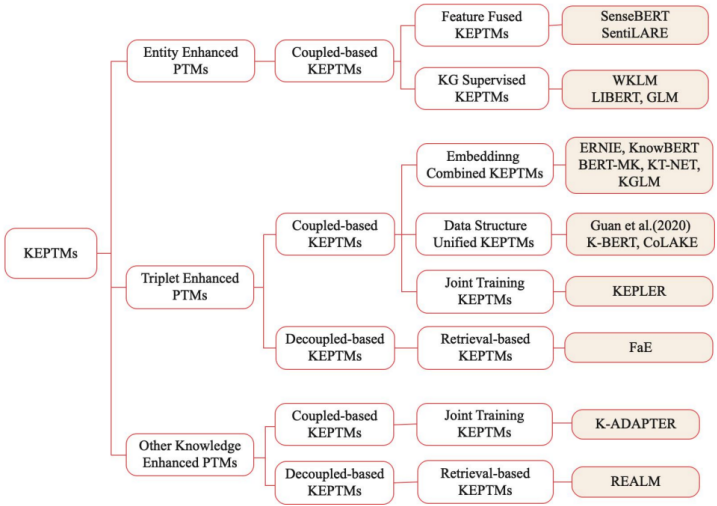
Table 2: Fine-tuned image-text retrieval results on Flickr30K and COCO datasets.

- Visual Grounding:

Figure 6: Grad-CAM visualizations on the cross-attention maps corresponding to individual words.

Align before Fuse: Vision and Language Representation Learning with Momentum Distillation, Salesforce 2021

知识融入



Yang et al., A Survey of Knowledge Enhanced Pre-trained Models, arXiv:2110.00269

知识融入：Triplet-Enhanced PLMs

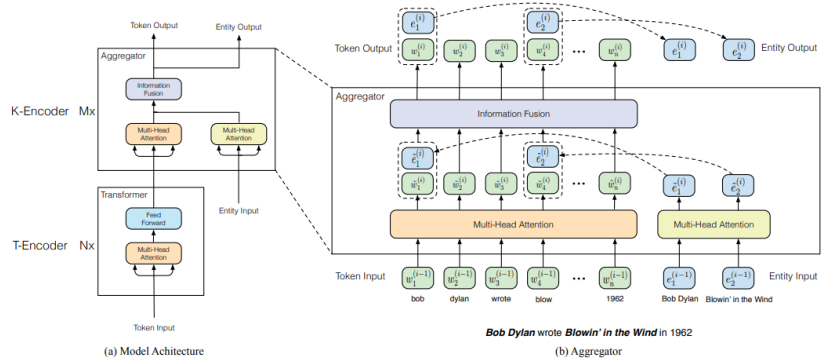


Figure 2: The left part is the architecture of ERNIE. The right part is the aggregator for the mutual integration of the input of tokens and entities. Information fusion layer takes two kinds of input: one is the token embedding, and the other one is the concatenation of the token embedding and entity embedding. After information fusion, it outputs new token embeddings and entity embeddings for the next layer.

Zhang et al., ERNIE: Enhanced Language Representation with Informative Entities, ACL 2019

知识融入：Entity-Enhanced PLMs

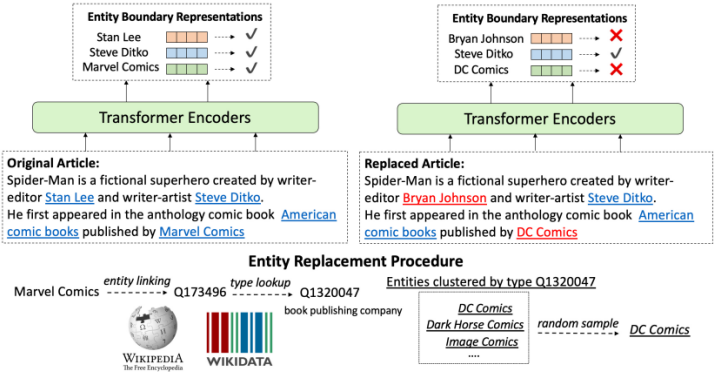


Figure 1: Type-Constrained Entity Replacements for Knowledge Learning.

Xiong et al., Pretrained Encyclopedia: Weakly Supervised Knowledge-Pretrained Language Model, ICLR 2020

加入检索

- 为什么PLMs需要Retrieval
  - 更忠实于客观事实的文本生成
  - 适配高速动态变化的客观世界知识
- 对于Retrieval augmented我们需要关注
  - 在Pre-training还是Fine-tuning阶段做retrieval
  - Retrieval到的(多个)文档如何建模
  - Retriever与Generator(Predictor)是否端到端训练

	Backbone model	Downstream tasks	Retrieval in pre-training	Retrieval in fine-tuning	End2End training
REALM[1]	BERT	ODQA	✔	✔	✔
RAG[2]	BART	ODQA/Generative QA/Dialogue generation	✘	✔	✔
FiD[3]	T5/BART	ODQA/Generative QA/Dialogue generation/Multi docs summarization	✘	✔	✘

加入检索：REALM(Retrieval-augmented Pre-training)

- Retrieval Augmented的预训练
  - 预训练阶段同时训练Retriever和Generator
  - 从原始BERT单纯的模式记忆->检索+记忆

$$p(y \mid x) = \sum_{z \in \mathcal{Z}} p_\phi(y|x,z)\, p_\theta(z|x)$$

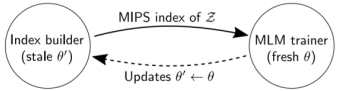
Retriever　Generator　Z=Top(K) passages

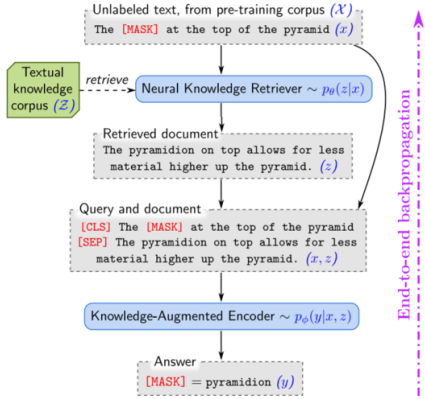
- Knowledge Retriever
  - MLM object可提供远程监督信号训练Retriever

$$p(z \mid x) = \frac{\exp f(x,z)}{\sum_{z'} \exp f(x,z')},$$

$$f(x,z) = \mathtt{Embed}_{\mathtt{input}}(x)^\top \mathtt{Embed}_{\mathtt{doc}}(z)$$

- End2End训练的最大挑战: Document Index update
  - 异步MIPS更新





Guu, Kelvin, et al. "Realm: Retrieval-augmented language model pre-training."

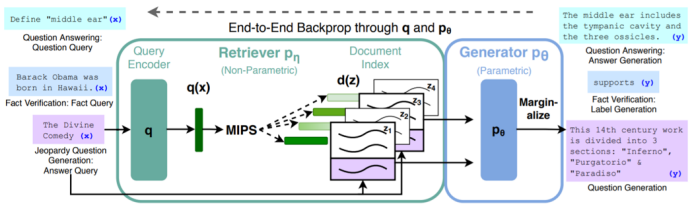
加入检索：RAG(Retrieval-augmented Generation)

- 在fine-tuning阶段使用retriever
  - 和REALM虽同为End2end training，但RAG并不更新document索引
  - 和REALM类似，直接将检索文档和query拼接建模，都会受制于encoder的max-seq-length
- 优化目标
  - RAG-Sequence Model:

$$p_{\text{RAG-Sequence}}(y|x) \approx \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) p_\theta(y|x,z) = \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) \prod_i^N p_\theta(y_i|x,z,y_{1:i-1})$$

  - RAG-Token Model:

$$p_{\text{RAG-Token}}(y|x) \approx \prod_i^N \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) p_\theta(y_i|x,z,y_{1:i-1})$$



Lewis, Patrick, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks."

加入检索：FiD(Fusion in Decoder)

- FiD给出一种在Decoder端进行信息融合的方式
  - Encoder端文档独立编码
  - 文档间的交互通过decoder端的Cross-Attention实现
- FiD可以更加高效地利用多文档信息
  - Generator与Retriever解耦，使用上较REALM和RAG更加灵活
  - Cross Attention Score具备一定的可解释性
  - 在问答生成，对话生成等任务上均取得了SOTA的效果

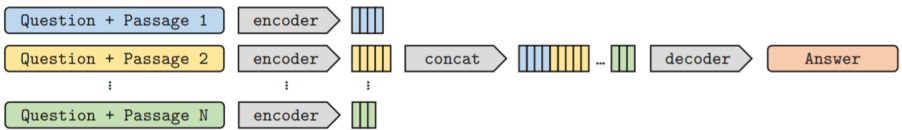


Figure 2: Architecture of the Fusion-in-Decoder method.

Izacard, Gautier, and Edouard Grave. "Leveraging passage retrieval with generative models for open domain question answering."

Content

预训练大模型的研究现状和发展趋势

- 如何做得更大？
- 如何更能干？
- 如何更有效地训练？
- 如何应用和推理？
- 如何确保安全可信？

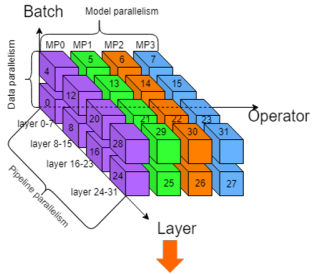
如何更有效地训练？

大规模预训练模型因模型巨大，每次训练代价极高，如何节约成本、高效训练，成为必须考虑的重要问题：

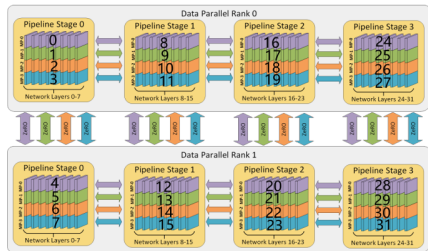
- 分布式并行训练
- 迁移学习（尽量复用已有大模型参数）
- 持续训练（增量式训练、终身学习，避免灾难性遗忘）

三维并行训练

- 三维混合并行策略：数据并行+Pipeline并行+模型并行
  - 数据并行： Batch维度的切分
  - Pipeline并行： Layer维度的切分
  - 模型并行：算子维度的切分
- 通过建立三维并行坐标和物理设备之间的映射，可自由扩展，高效训练如盘古α、GPT3等千亿参数级别的模型

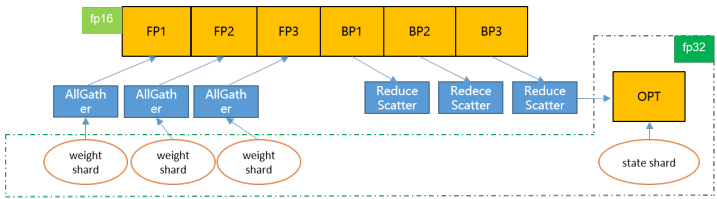


Coordinate	RANK	Coordinate	RANK	Coordinate	RANK	Coordinate	RANK
(0, 0, 0)	0	(1, 0, 0)	8	(2, 0, 0)	16	(3, 0, 0)	24
(0, 0, 1)	1	(1, 0, 1)	9	(2, 0, 1)	17	(3, 0, 1)	25
(0, 0, 2)	2	(1, 0, 2)	10	(2, 0, 2)	18	(3, 0, 2)	26
(0, 0, 3)	3	(1, 0, 3)	11	(2, 0, 3)	19	(3, 0, 3)	27
(0, 1, 0)	4	(1, 1, 0)	12	(2, 1, 0)	20	(3, 1, 0)	28
(0, 1, 1)	5	(1, 1, 1)	13	(2, 1, 1)	21	(3, 1, 1)	29
(0, 1, 2)	6	(1, 1, 2)	14	(2, 1, 2)	22	(3, 1, 2)	30
(0, 1, 3)	7	(1, 1, 3)	15	(2, 1, 3)	23	(3, 1, 3)	31



https://www.microsoft.com/en-us/research/blog/deepspeed-extreme-scale-model-training-for-everyone/

训练状态并行



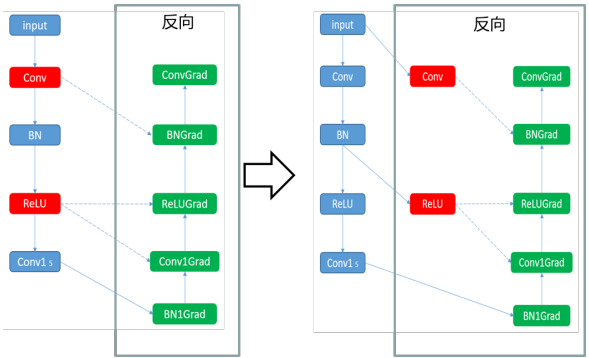
内存开销

Optimizer切分前

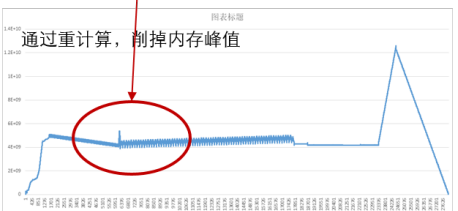
Optimizer切分后

- Feature：
  - inner-layer切分：在参数、优化器状态、梯度最高维切分。
  - 通信分组并行：allgather和reduce-catter分别和正、反向运算并行。
  - 混合精度：正反向传播和通信采用fp16运算，优化器及参数采用fp32。

重计算



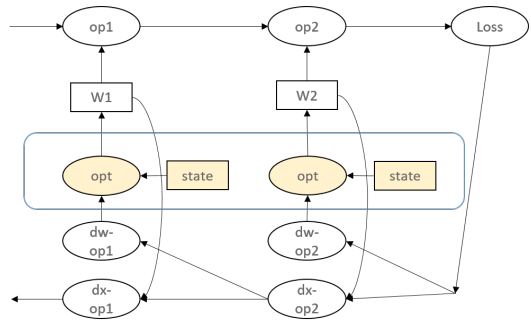
正向activation累积，出现内存峰值



- 正向activation不存，反向计算时，重新计算正向activation，时间换空间。

异构计算

- 过去几年，模型规模增大了1000倍以上，但是并行计算设备的内存只增大了5倍（GPU 16G 到 80G）
- 将一部分训练计算转移到CPU上进行，将一部分存储放在Host内存。有代表性的是优化器异构计算：
  - Adam Optimizer State是Weight的2倍；175B参数量的GPT3模型，就有350B的Optimizer State；
  - Adam Optimizer调度到Host CPU执行，Optimizer State存储到Host内存；
  - 极大节省GPU，NPU等计算内存空间。



Optimizer CPU执行

Content

预训练大模型的研究现状和发展趋势

- 如何做得更大？
- 如何更能干？
- 如何更有效地训练？
- 如何应用和推理？
- 如何确保安全可信？

如何应用和推理？

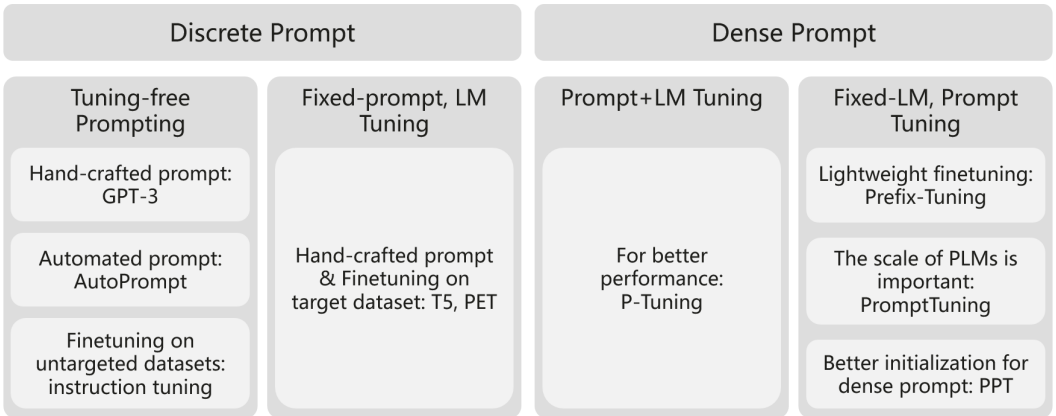
- 因模型规模巨大，传统的预训练模型的微调模式很难被采用：全量参数更新代价太高
  - 基于Prompt的微调模式受到广泛关注
  - 基于Adapter的微调模式也可以适用于大规模预训练模型，但近期进展不大
- 传统的模型蒸馏也变得代价极高，因为蒸馏过程需要在大量的数据上进行推理
- 其他模型压缩算法如量化、剪枝等等都面临新的问题

Prompting Methods for Downstream Tasks

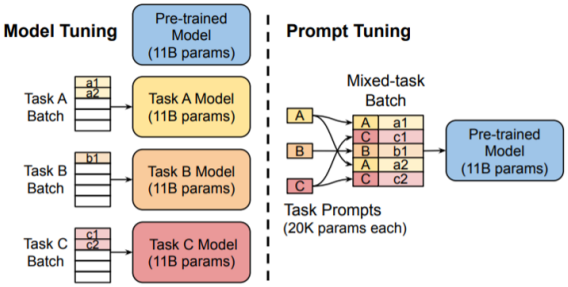
Type	Task	Input ([X])	Template	Answer ([Z])
Text CLS	Sentiment	I love this movie.	[X] The movie is [Z].	great fantastic ...
	Topics	He prompted the LM.	[X] The text is about [Z].	sports science ...
	Intention	What is taxi fare to Denver?	[X] The question is about [Z].	quantity city ...
Text-span CLS	Aspect Sentiment	Poor service but good food.	[X] What about service? [Z].	Bad Terrible ...
Text-pair CLS	NLI	[X1]: An old man with ... [X2]: A man walks ...	[X1]? [Z], [X2]	Yes No ...
Tagging	NER	[X1]: Mike went to Paris. [X2]: Paris	[X1][X2] is a [Z] entity.	organization location ...
Text Generation	Summarization	Las Vegas police ...	[X] TL;DR: [Z]	The victim ... A woman ... ...
	Translation	Je vous aime.	French: [X] English: [Z]	I love you. I fancy you. ...

Liu et al., Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing, arXiv:2107.13586, 2021

Prompt-based Learning for Large-scale PLMs



PromptTuning



Lester et al., The Power of Scale for Parameter-Efficient Prompt Tuning, arXiv:2104.08691, 2021

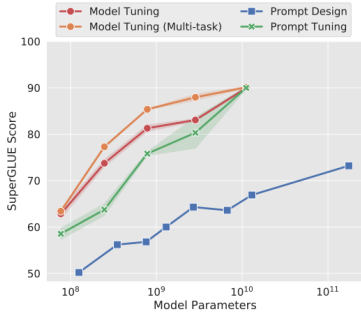
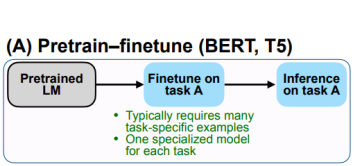
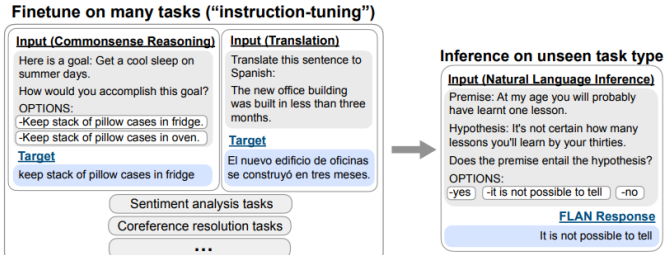
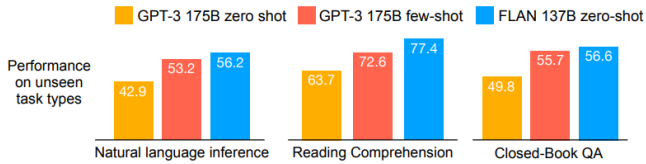


Figure 1: Standard **model tuning** of T5 achieves strong performance, but requires storing separate copies of the model for each end task. Our **prompt tuning** of T5 matches the quality of model tuning as size increases, while enabling the reuse of a single frozen model for all tasks. Our approach significantly outperforms few-shot **prompt design** using GPT-3. We show mean and standard deviation across 3 runs for tuning methods.

Instruction Tuning







Wei et al., Finetuned Language Models Are Zero-Shot Learners, arXiv:2109.01652, 2021 (Google)

Content

预训练大模型的研究现状和发展趋势

- 如何做得更大？
- 如何更能干？
- 如何更有效地训练？
- 如何应用和推理？
- 如何确保安全可信？

安全可信：预训练大模型的社会因素

- 偏见和公平性
- 滥用和误用
- 环境影响
- 合法性
- 经济影响
- 伦理问题

Content

如何变现？（商业模式）

- 分散式→集中式，类似于：
  - 企业搜索、图书检索系统→通用搜索引擎
  - 企业IT→云计算
- AI服务提供商：
  - 提供集中式的AI能力
  - 面向不同领域、不同应用的大模型会百花齐放
- AI服务客户/消费者：
  - 中小企业甚至个人可以轻松为自己的设备定制AI能力（比如微波炉可以跟人对话），无需自己开发
  - 客户也可以通过模型压缩、蒸馏、量化等方式定制自己的小模型

Content

总结

- 介绍了预训练大模型的定义和特点
  - 两个主要特点：大、预训练
  - 与传统模型比具有全新的特点（浮现和同质化）
- 介绍了预训练大模型近期的研究进展和面临的问题
  - 越来越大
  - 吸收异构数据
  - 需要全新的训练算法和应用模式
  - 安全可信问题需要更多关注
- 讨论了预训练大模型为了的应用前景
  - 集中的预训练模型提供随处可得的AI能力和服务

Thank you!

把数字世界带入每个人、每个家庭、
每个组织，构建万物互联的智能世界。

Bring digital to every person, home and organization for a fully connected, intelligent world.

HUAWEI