

Pre-trained Huge Language Models: Challenges and Opportunities

巨型预训练语言模型的机会和挑战

LIU Qun 刘群

Huawei Noah's Ark Lab 华为诺亚方舟实验室

SPARKS人工智能学术峰会
Online, 2022-01-21



NOAH'S ARK LAB



Content

Introduction

Opportunities

Challenges

Our Work

Content

Introduction

Opportunities

Challenges

Our Work

What are pre-trained huge language models?

- ▶ A new paradigm for NLP (or AI) research
 - ▶ Deep Neural Models
 - ▶ Huge model size & huge training data
 - ▶ Non-task-specific pre-training, can be adapted to various downstream tasks

What are pre-trained huge language models?

► Also called foundation (Language) models

arXiv.org > cs > arXiv:2108.07258

Computer Science > Machine Learning

[Submitted on 15 Aug 2021 ([v1](#)), last revised 16 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 Kaluri, Siddharth Karamchetti, Geoff Keeling, Fereshte Khan, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvrit Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Niforoshan, 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 Tramér, 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]

Emergence and homogenization

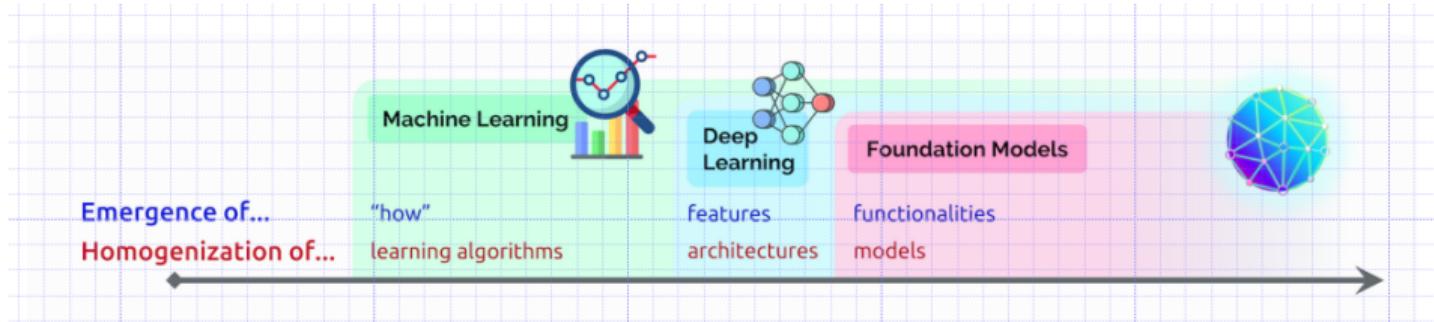


Fig. 1. The story of AI has been one of increasing *emergence* and *homogenization*. With the introduction of machine learning, *how* a task is performed emerges (is inferred automatically) from examples; with deep learning, the high-level features used for prediction emerge; and with foundation models, 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]

Content

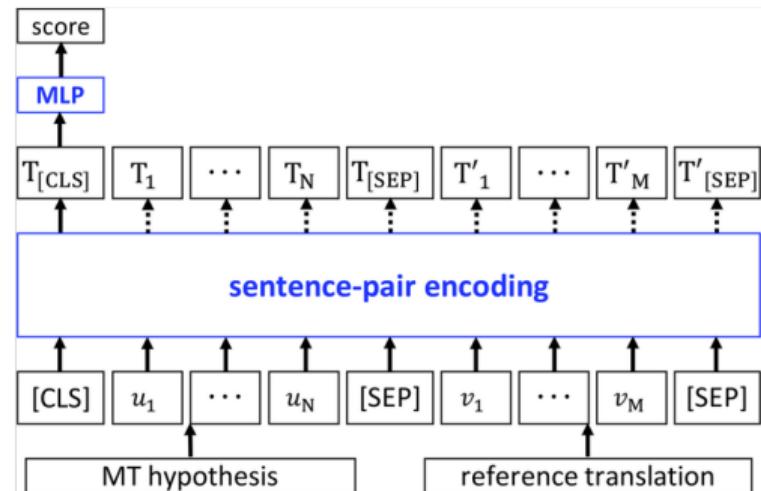
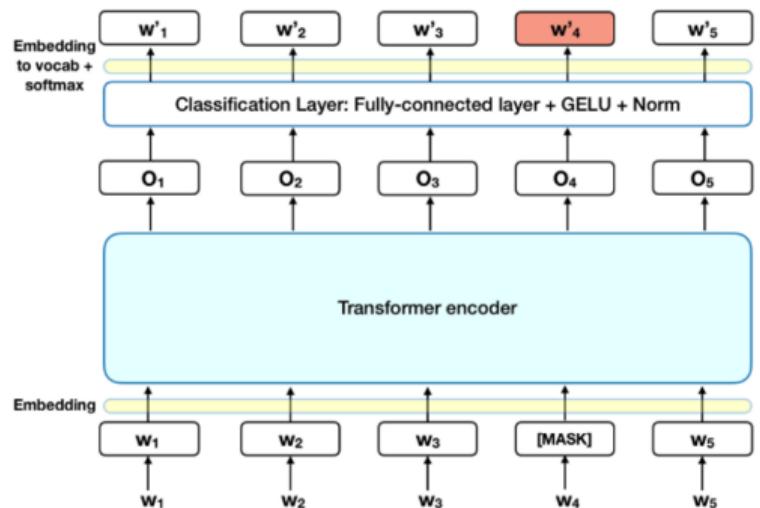
Introduction

Opportunities

Challenges

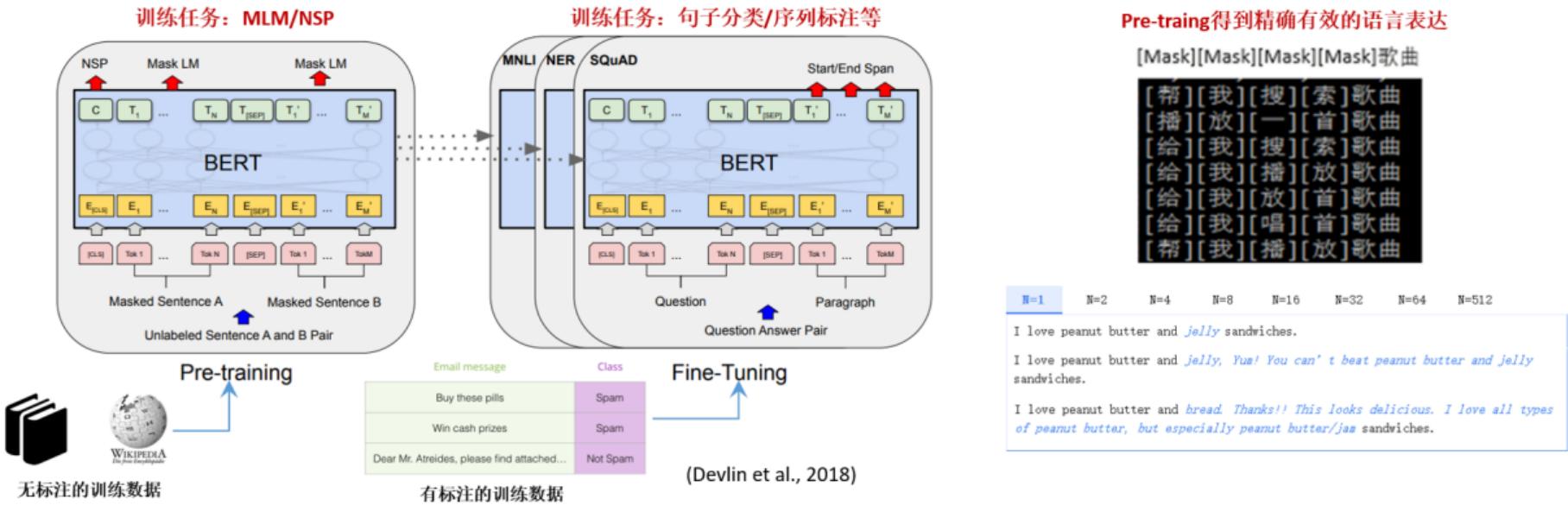
Our Work

Self-supervised Learning



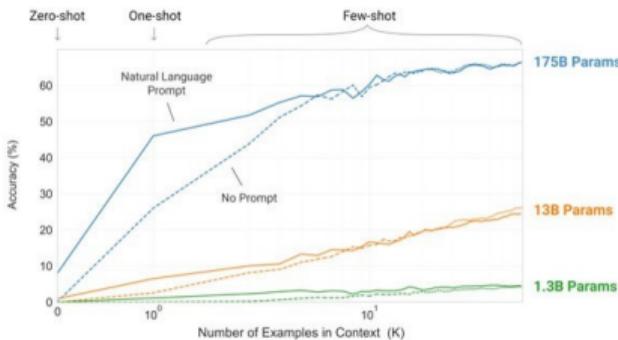
Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805, 2018

Pre-training and fine-tuning framework



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

Few-shot and zero-shot learning



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 description of the task. No gradient updates are performed.



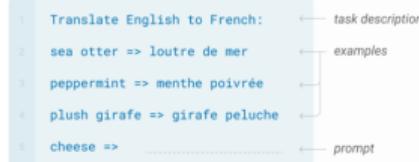
One-shot

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



Few-shot

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



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

Fine-tuning

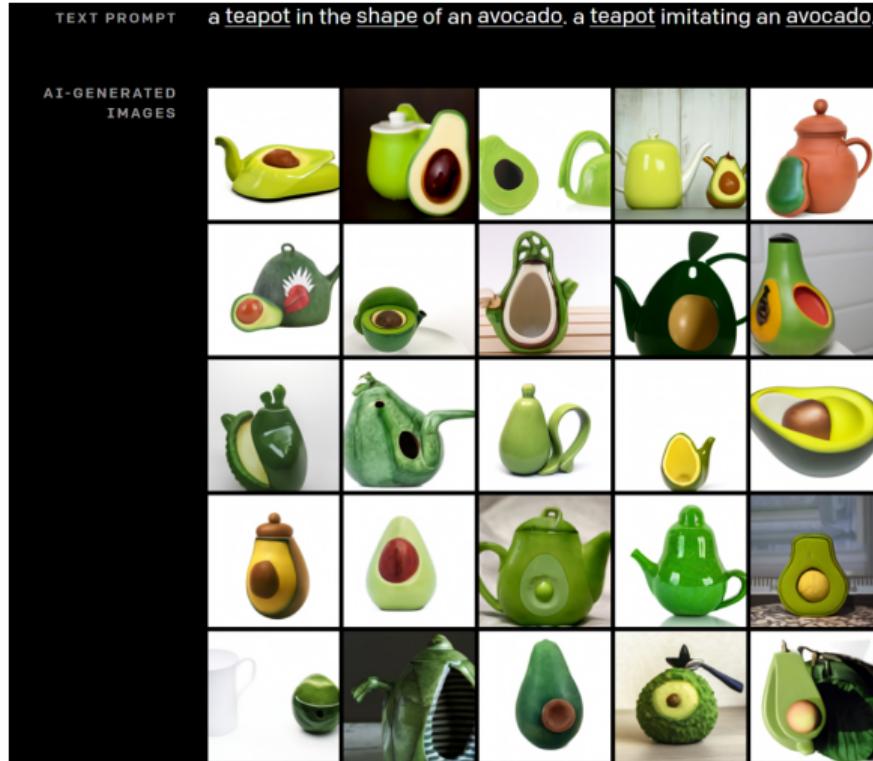
The model is trained via repeated gradient updates using a large corpus of example tasks.



Multilingual representation

The image is a dense, colorful word cloud centered around the word "AHOJ". The word "AHOJ" is repeated in many different languages, including English, Spanish, French, German, Italian, Portuguese, Dutch, Swedish, Danish, Norwegian, and others. The words are arranged in a circular, radiating pattern from the center, with colors ranging from red, orange, yellow, green, blue, and purple. The background is white, and the overall effect is a vibrant, multilingual greeting.

Multimodal interaction



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

New business model

- ▶ Distributed → Centralized, similar to what happened in the history:
 - ▶ Enterprise search, library retrieval system→general search engine (like Google, Baidu)
 - ▶ Enterprise IT system→Cloud computing
- ▶ AI service provider:
 - ▶ Provide centralized AI ability
 - ▶ Models for various domains, application types will flourish
- ▶ AI service customers:
 - ▶ Medium/small size enterprises or persons can order customized AI ability
 - ▶ For example, refrigerator producers can order dialog services for the refrigerators they produce, without develop such systems by themselves
 - ▶ Users can also order small models by using the model compression services provided by the AI service providers

Content

Introduction

Opportunities

Challenges

Our Work

Content

Challenges

Model size challenge: how can we increase the model size?

Competence challenge: doing more and doing better

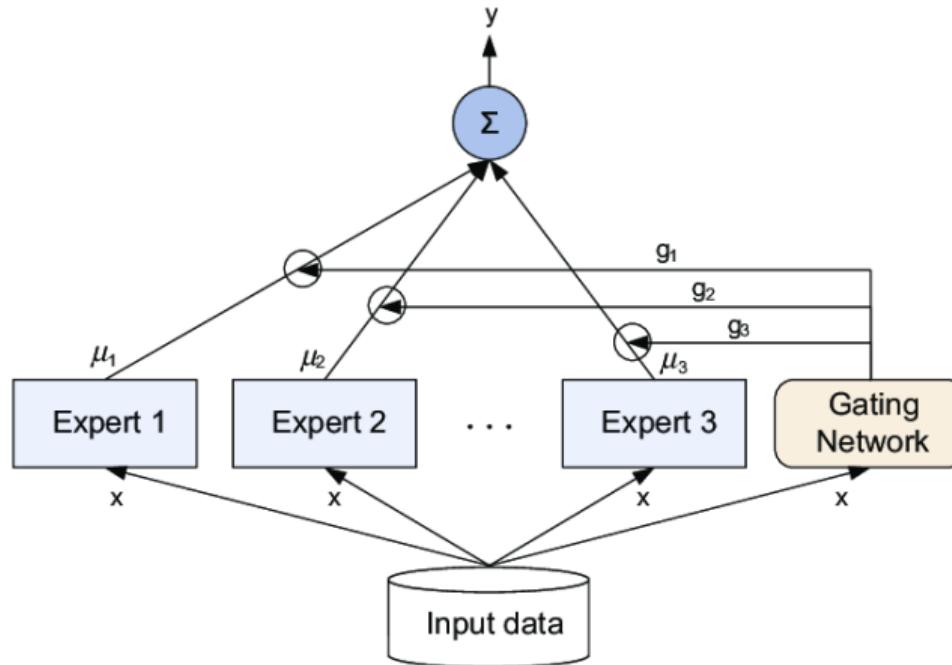
Training challenge: how can we train models more efficiently?

Fine-tuning challenge: how can we fine-tune a huge model?

Challenge of safety and trustworthiness

MoE Transformers

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



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

Content

Challenges

Model size challenge: how can we increase the model size?

Competence challenge: doing more and doing better

Training challenge: how can we train models more efficiently?

Fine-tuning challenge: how can we fine-tune a huge model?

Challenge of safety and trustworthiness

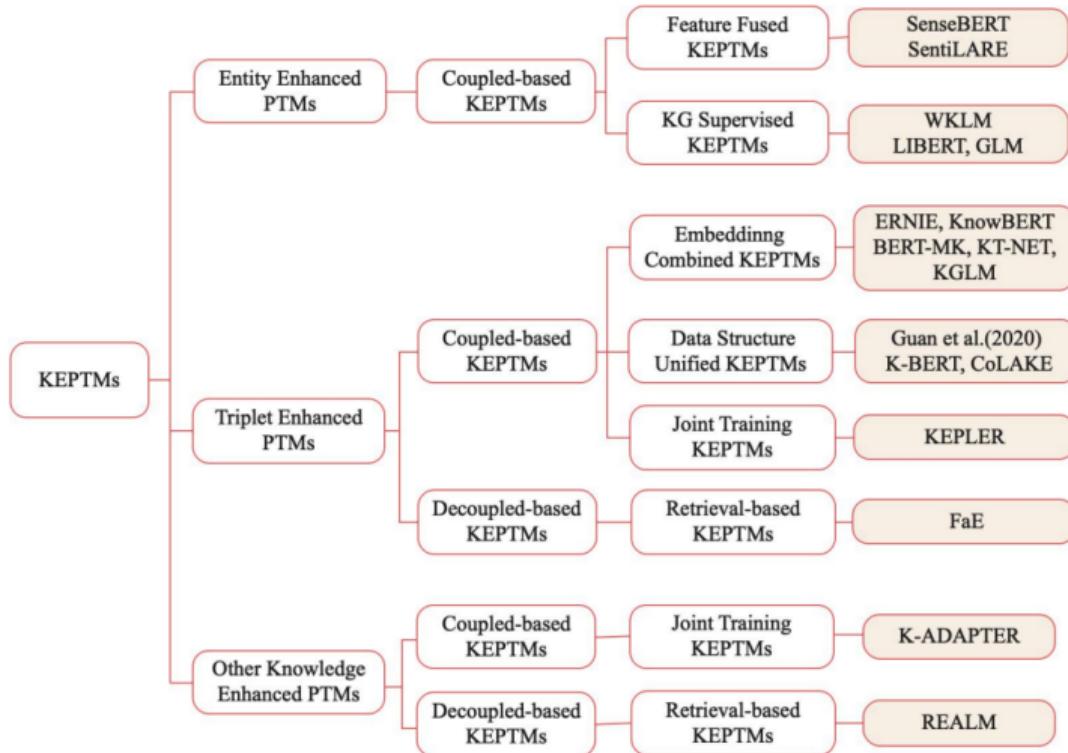
Heterogeneous data training

- ▶ Huge pre-training models are able to modeling data with high diversity
- ▶ Different types of data can mutual enhanced to make a stronger model
- ▶ Heterogeneous data:
 - ▶ Multimodal
 - ▶ Knowledge
 - ▶ External text (enhanced with retrieval)
 - ▶ Codes

Image-text pre-training: How different modalities interact?

- ▶ Two-tower model: model interaction with contrastive loss:
 - ▶ CLIP, ALIGN, WENLAN
- ▶ One-tower model: concatenate the features from images and texts to a single sequence, and feed it to a Transformer model
 - ▶ Encoder: VILT, SOHO
 - ▶ Decoder: DALL-E, Frozen
 - ▶ Mix: M6, OPT
- ▶ Others: 其他: 通过Encoder-decoder结构中Decoder的cross-attention进行模态交互: ALBEF

Combined with knowledge



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

Combined with retrieval

- ▶ 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	✗	✓	✗

Content

Challenges

Model size challenge: how can we increase the model size?

Competence challenge: doing more and doing better

Training challenge: how can we train models more efficiently?

Fine-tuning challenge: how can we fine-tune a huge model?

Challenge of safety and trustworthiness

How can we train models more efficiently?

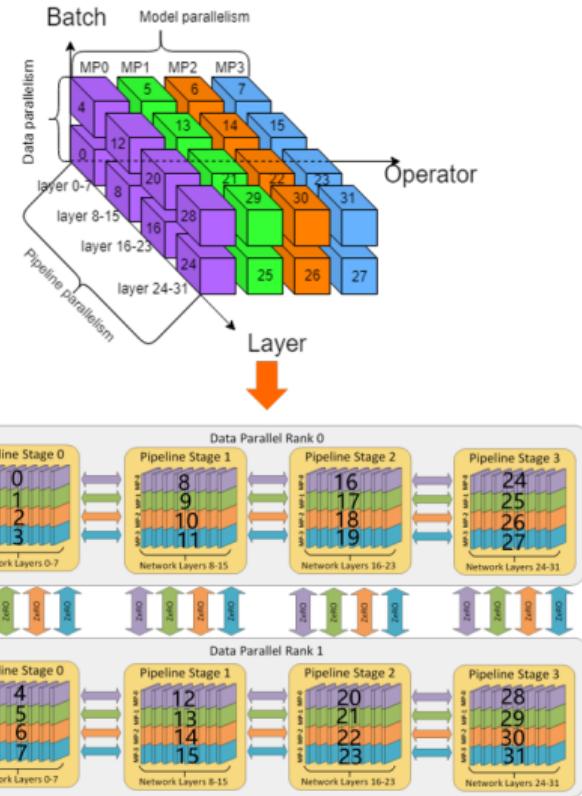
大规模预训练模型因模型巨大，每次训练代价极高，如何节约成本、高效训练，成为必须考虑的重要问题：

- ▶ 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）

3-D parallel training

- ▶ 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



Content

Challenges

Model size challenge: how can we increase the model size?

Competence challenge: doing more and doing better

Training challenge: how can we train models more efficiently?

Fine-tuning challenge: how can we fine-tune a huge model?

Challenge of safety and trustworthiness

How can we fine-tune a huge model?

- ▶ 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	[X1]: An old man with ...		Yes
		[X2]: A man walks ...	[X1]? [Z], [X2]	No ...
Tagging	NER	[X1]: Mike went to Paris.		organization
		[X2]: Paris	[X1] [X2] is a [Z] entity.	location ...
				The victim ... A woman
Text Generation	Summarization	Las Vegas police ...	[X] TL;DR: [Z]	I love you. I fancy you. ...
	Translation	Je vous aime.	French: [X] English: [Z]	

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

Discrete Prompt	Dense Prompt
Tuning-free Prompting Hand-crafted prompt: GPT-3 Automated prompt: AutoPrompt Finetuning on untargeted datasets: instruction tuning	Fixed-prompt, LM Tuning Hand-crafted prompt & Finetuning on target dataset: T5, PET
	Prompt+LM Tuning For better performance: P-Tuning
	Fixed-LM, Prompt Tuning Lightweight finetuning: Prefix-Tuning The scale of PLMs is important: PromptTuning Better initialization for dense prompt: PPT

Content

Challenges

Model size challenge: how can we increase the model size?

Competence challenge: doing more and doing better

Training challenge: how can we train models more efficiently?

Fine-tuning challenge: how can we fine-tune a huge model?

Challenge of safety and trustworthiness

Safety and trustworthiness: social aspects

- ▶ Bias and equality
- ▶ Abuse and misuse
- ▶ Environmental impact
- ▶ Legality
- ▶ Economic impact
- ▶ Ethic problems

Content

Introduction

Opportunities

Challenges

Our Work

Content

Our Work

Pangu- α : A Large-Scale Chinese Pre-trained Language Model

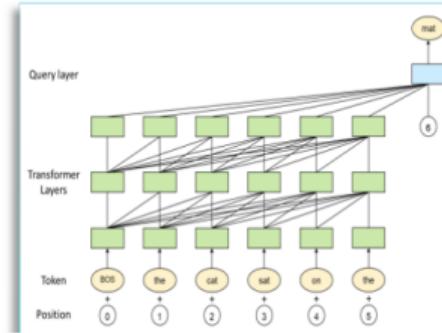
Filip: Fine-Grained Interactive Language-Image Pre-Training

Spiral: Self-Supervised Speech Pre-Training

Pangu- α : Release (May 2021)

联合鹏城实验室发布业界首个两千亿参数量中文预训练语言模型-盘古 α

华为AI全栈: MindSpore + CANN + ModelArts + Atlas 900 集群



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

TECHNICAL REPORT					
Wei Zeng*	Xiaochu Ren*	Teng Su*	Hui Wang*		
Yi Li	Zhiwei Wang	Xin Jiang	Zhenzhang Yang	Kaisheng Wang	Xiaoda Zhang
Chen Li	Ziyao Gong	Yifan Yao	Xiaojing Huang	Jian Wang	Jiufeng Yu
Yue Yu	Yan Zhang	Jin Wang	Hengtan Tang	Daren Yan	Zexuan Yi
Fangqiang Jiang	Han Zhang	Lingfeng Deng	Yehong Zhang	Zhe Lin	Fang Peng
Chao Zhang	Shujie Zhang	Mingyue Guo	Shanchi Gu	Gangjin Fan	Yaowei Wang
Xuefeng Jin	Qun Liu	Yonghong Tian	Wen Gao		

PANGU- α TEAM

<https://arxiv.org/abs/2104.12369>

Pangu- α : Influence



Horace He
@cHilllee

Haven't seen anybody else mention this, but Huawei just announced they trained a 200 BILLION transformer model - PanGu- α . This is bigger than GPT-3, but trained only for 40B tokens.

Moreover, they're trained on an entirely Chinese stack: Huawei chips and Mindspore framework. 1/2

LARGE-SCALE AUTOREGRESSIVE PRE-
LANGUAGE MODELS WITH AUTO-PAR-
COMPUTATION

TECHNICAL REPORT

Table 1: Model size	
#Parameters	#
2.6B	
13.1B	
207.0B	

PANGU- α TEAM

12:45 PM · Apr 26, 2021 · Twitter Web App

151 Retweets 19 Quote Tweets 645 Likes

Facebook研究员Horace He发推，谷歌苹果微软等研究员上百次转发点赞，OpenAI主管Miles Brundage关注转发：
<https://twitter.com/cHilllee/status/1386541907950465028>

The Machine | Special Issue | Become a Member | Sign In

Huawei trained the Chinese-language equivalent of GPT-3

By Kyle Wiggers | @kyle_l_wiggers | April 29, 2021 5:30 AM

This week, a research team at Chinese company Huawei quietly detailed what might be the Chinese-language equivalent of GPT-3. Called [PanGu-Alpha \(stylized PanGu- \$\alpha\$ \)](#), the 750-gigabyte model contains up to 200 billion parameters — 25 million more than GPT-3 — and was trained on 1.1 terabytes of Chinese-language ebooks, encyclopedias, news, social media, and web pages.

科技媒体报道：<https://venturebeat.com/2021/04/29/huawei-trained-the-chinese-language-equivalent-of-gpt-3/>

Posted by [upjenggerman \(20\) · Researcher](#) · 14 days ago
[D] Huawei just announced that they trained a 200 billion transformer model on an entirely Chinese stack

Discussion

My tweet about it: <https://twitter.com/cHilllee/status/1386541907950465028>

They trained a 200 billion parameter decoder-only dense transformer for 40B tokens on 2048 Huawei Ascend 910 chips. Moreover, this was all done using Mindspore, Huawei's ML framework.

In contrast, GPT-3 was a 175B parameter model trained for 300B tokens.

On its own, this is already quite impressive. Even though they've only done 40B tokens, this is the biggest model yet out of China, and represents one of the biggest models yet in the world.

However, the thing that's really impressive to me is that this was done with an all Chinese stack: Huawei Mindspore as the framework, compiling down to Huawei Ascend chips.

I'd known that Huawei was working on AI chips, but I was unaware that they had matured to the point that they could feasibly train a model of this scale.

Code: <https://git.openi.org.cn/PCL-Platform.Intelligence/PanGu-Alpha>

Paper: <https://ieeexplore.ieee.org/document/9450000>

49 Comments 0 Award 0 Share 0 Save 0 Hide Report 90% Upvoted

IMPORT AI

ABOUT

May 3, 2021

Import AI 247: China makes its own GPT3; the AI hackers have arrived; four fallacies in AI research.

Finally, China trains its own GPT3:

...Now the world has two (public) generative models, reflecting two different cultures...

A team of Chinese researchers have created 'PanGu', a large-scale pre-trained language model with around ~200 billion parameters, making it equivalent to GPT3 (175 billion parameters). In terms of parameter complexity, PanGu is trained on 1.1TB of Chinese text (versus 570GB of text for GPT-3), though in the paper they train the 200B model for a lot less time (on way fewer tokens) than OpenAI did for GPT-3. PanGu is the second GPT-3-esque model to come out of China, following the Chinese Pre-trained Language Model (CPM, [Import AI 226](#)), which was trained on 100GB of text and was only a few billion parameters, compared to a couple of hundred!

OpenAI主管Jack Clark撰文：<https://jack-clark.net/2021/05/03/import-ai-247-china-makes-its-own-gpt3-the-ai-hackers-have-arrived-four-fallacies-in-ai-research/>

Reddit论坛讨论：

https://www.reddit.com/r/MachineLearning/comments/myr072/d_huawei_just_announced_that_they_trained_a_200/

Pangu- α : Automatic summarization

Prompt	<p>“据媒体报道，最初华为只决定做车联网。2012年，华为开始进行汽车相关研究，当时在2012实验室下成立车联网实验室，研究电动汽车技术。</p> <p>2013年，华为分别展示过为汽车内部提供3G WiFi热点的DA6810、应用于保险行业及车队管理的汽车在线诊断系统DA3100，以及车载通信模块产品“ME909T”。</p> <p>当时，华为对外的说法是，“华为终端只会聚焦于自己擅长的车联网通信解决方案。华为愿景是做到车联网模块第一，业绩上做到50亿 60亿元的规模”。</p> <p>2019年5月27日，任正非签发华为组织变动文件，批准成立智能汽车解决方案BU部门，隶属于ICT管理委员会管理。智能汽车解决方案BU之下分：智能座舱部门、自动驾驶部门、智能车云服务部门。到这个阶段，华为已经确立了自动驾驶部门。</p> <p>2020年7月12日晚，华为消费者业务CEO余承东宣宣了HUAWEI HiCar车机系统，HiCar系统基于鸿蒙OS。HiCar可以利用移动设备的5G网络能力加强车机系统的数据交互能力，能够接入华为在AI、语音、计算机视觉等方面的能力。同时，能够调用车速、方向盘转角、档位模式、汽车环境光传感器在内的车身数据以及空调、车窗、喇叭等车身控制部件。</p> <p>直到徐直军在2021年的分析师大会透露出一个重点信号，华为将汽车服务的重心定在自动驾驶软件上，才明确了华为的自动驾驶业务。据媒体报道，华为跟百度、小鹏、文远知行、驭势科技等一样，都是选择在乘用车领域重点布局。”的新闻摘要是</p>
Generation	<p>“华为在自动驾驶领域的布局，是基于华为在人工智能、大数据、云计算等方面的技术积累，以及华为在汽车领域的多年经验。”</p>

Pangu- α : Title Generation

Prompt	<p>“据媒体报道，最初华为只决定做车联网。2012年，华为开始进行汽车相关研究，当时在2012实验室下成立车联网实验室，研究电动汽车技术。</p> <p>2013年，华为分别展示过为汽车内部提供3G WiFi热点的DA6810、应用于保险行业及车队管理的汽车在线诊断系统DA3100，以及车载通信模块产品“ME909T”。</p> <p>当时，华为对外的说法是，“华为终端只会聚焦于自己擅长的车联网通信解决方案。华为愿景是做到车联网模块第一，业绩上做到50亿 60亿元的规模”。</p> <p>2019年5月27日，任正非签发华为组织变动文件，批准成立智能汽车解决方案BU部门，隶属于ICT管理委员会管理。智能汽车解决方案BU之下分：智能座舱部门、自动驾驶部门、智能车云服务部门。到这个阶段，华为已经确立了自动驾驶部门。</p> <p>2020年7月12日晚，华为消费者业务CEO余承东官宣了HUAWEI HiCar车机系统，HiCar系统基于鸿蒙OS。HiCar可以利用移动设备的5G网络能力加强车机系统的数据交互能力，能够接入华为在AI、语音、计算机视觉等方面的能力。同时，能够调用车速、方向盘转角、档位模式、汽车环境光传感器在内的车身数据以及空调、车窗、喇叭等车身控制部件。</p> <p>直到徐直军在2021年的分析师大会透露出一个重点信号，华为将汽车服务的重心定在自动驾驶软件上，才明确了华为的自动驾驶业务。据媒体报道，华为跟百度、小鹏、文远知行、驭势科技等一样，都是选择在乘用车领域重点布局。”的新闻标题是《 Generation 华为在车联网业务方面有深厚积累，并将全面发力自动驾驶软件业务》</p>
--------	--

Pangu- α : Open Domain Question Answering

Prompt	阅读文章：株洲北站全称广州铁路（集团）公司株洲北火车站。除站场主体，另外管辖湘潭站、湘潭东站和三个卫星站，田心站、白马垅站、十里冲站，以及原株洲车站货场。车站办理编组、客运、货运业务。车站机关地址：湖南省株洲市石峰区北站路236号，邮编412001。株洲北站位于湖南省株洲市区东北部，地处中南路网，是京广铁路、沪昆铁路两大铁路干线的交汇处，属双向纵列式三级七场路网性编组站。车站等级为特等站，按技术作业性质为编组站，按业务性质为客货运站，是株洲铁路枢纽的主要组成部分，主要办理京广、沪昆两大干线四个方向货物列车的到发、解编作业以及各方向旅客列车的通过作业。每天办理大量的中转车流作业，并有大量的本地车流产生和集散，在路网车流的组织中占有十分重要的地位，是沟通华东、华南、西南和北方的交通要道，任务艰巨，作业繁忙。此外，株洲北站还有连接石峰区喻家坪工业站的专用线。株洲北站的前身是田心车站。请问：株洲北站的前身是哪个车站？回答：
Generation	田心车站
Prompt Generation	中国的首都是哪里? 中国的首都是北京。
Prompt Generation	中国、美国、日本和澳大利亚的首都分别是哪里? 中国的首都是北京，美国的首都是华盛顿，日本的首都是东京，澳大利亚的首都是堪培拉。
Prompt Generation	五一川西行程规划 五一川西旅游活动主要集中在绵阳、成都、雅安、甘孜4市（州），主要景区有大九寨、新都桥、丹巴、海螺沟、泸定、康定、成都新会展中心。

Pangu- α : Dialog Generation

我和妈妈上街的时候淋雨了。

好吧, 那你要注意身体哦!

谢谢关心。

我要上班了, 再见。

再见!

我也想吃饺子, 你还有吗。

我也是, 你想吃什么馅的?

我想吃肉馅的。

那我给你推荐一家饺子馆吧, 叫做
好啊!

这家店的人均价格是多少啊?

人均价格在元左右。

地址在哪儿呢?

地址在北京市朝阳区望京街望京SOHO西区
10号楼(近地铁望京站)。

评分怎么样啊?

评分. 哦哦。

我来给大家读一段吧!

“盘古开天地, 天崩地裂, 日月无光, 山河破碎。”

这是我最喜欢的一首诗。

你喜欢这首诗吗?

我还挺喜欢的。

那你知道这首诗出自谁的笔下吗?

不知道啊! 是谁写的呢?

是王勃的作品。

哦! 原来是王勃啊! 他可是个大才子呢! 他写过很多好
诗呢! 你知道他是哪年出生的吗?

我不清楚啊! 他出生于哪一年啊?

他出生于公元756年, 逝世于公元817年。你知道他有
哪些作品吗?

我只知道《滕王阁序》和《秋日登洪府滕王阁》这
两首诗吧!

Pangu- α : Stylized Document Generation

非常简单地在prompt注入不同场景，不同人物个性化信息，即可实现场景化，个性化，风格化的对话引擎：

Prompt注入职业信息：警察

你好啊！你叫什么名字？
你好，我叫范明明。
好的，明明你是做什么的呢？
我是名警察。
平时上班的主要都干些什么呢？
平时主要是跟警犬一起工作。
警犬？我听我朋友说起过，是个什么类型的呢？
是大型犬类。
大型犬类是不是都是很凶猛的呢？
也不一定，有的就很温顺。

Prompt注入个性化信息：甄嬛体（母子高考前对话）

儿臣不日将高考，能有心仪的学府完成学业真是极好的。
今儿个是开始填志愿了，急切选择学府虽是要紧，却也不能忘了专业二字。
儿臣是真真的不知所措了。
日常暇余之间常听闻五道口有一职业技术学院倒是极不错的。
儿臣倒也听先生提及，谓之清华，其学术氛围浓郁，众学子潜心学业。
整个五道口城楼街道行人驻足皆是热议清华学子，甚是羡慕。
儿臣就想着不日便要面临那番环境，且更想着能早日成为其中一员。可以说，此乃儿臣的荣耀。

Prompt注入个性化场景信息：快递客服

你好啊！你叫什么名字？
你好，我叫范明明。
好的，我是圆通的快递员
好的，有什么事情吗
你的货到了，在这签个字。
我先看看货。
不签收不能验货。
为什么？
我们公司的规定。
好，那我签收了。

Prompt注入职业信息：老师

你好啊！你叫什么名字？
你好，我叫范明明。
好的，明明你是做什么的呢？
我是名老师。
平时上班的主要都干些什么呢？
我平时上班是教小朋友数学的。
哇，是位小朋友的数学老师，真羡慕你，你平时有什么爱好呢？
我平常喜欢看电影。

Prompt注入个性化信息：人话版（母子高考前对话）

儿子马上要高考了，要是能考到心仪的大学就太好了。
今天开始填志愿了，学校很重要，也要选好专业。
我也很迷茫。
我经常听说五道口有一所职业学校非常不错。
我也经常听到老师提到，叫清华，学术氛围很浓，大家学习很刻苦。
我知道，清华是中国的最高学府。
我也知道，清华是中国的骄傲。

Prompt注入个性化场景信息：京东客服

你好啊！你叫什么名字？
你好，我叫范明明。
好的，我是中国京东商城的客户服务代表。请问你有什么事情吗？
我买的木瓜牛奶，牛奶送来的时候，发现没有发票，这个我可以补吗？
请问是这个订单号吗
对的

Pangu- α : Essay Generation

盘古 α 生成效果 – 2021年北京高考作文

论生逢其时

每个人都生活在特定的时代，每个人在特定时代中的人生道路各不相同。

在同一个时代，有人慨叹生不逢时，有人只愿安分随时。有很多人因为种种原因而不能实现自己的理想。他们或因生活所迫、或因性格缺陷、或因家庭变故等等原因而无法实现理想；他们也曾想过要放弃这个目标，但是又没有勇气去面对现实生活中存在着太多不可预知和不可克服困难；他们也曾想过要改变现状、改变命运。当你真正去做了以后才发现自己当初所想与实际相差太远。

每个人都是自己的主人，每个人的命运都把握在自己手里。人生的道路是由自己选择的，每个人都有自己独特的人生道路。在这个世界上，没有谁可以主宰你的命运。但是，你可以选择怎样去走这条路？如何才能使自己不迷失方向、不走错路呢？我们应该正视现实生活中存在着太多太多不可知因素和难以预料因素。只有这样才能使我们更好地把握住人生道路上每个机会。

在人生的道路上，每个人都会遇到各种各样的困难和挫折。面对这些困难和挫折，我们要学会勇敢地面对、积极地战胜它们。只有这样，才能使自己不断地走向成功之路。

在我们这个充满竞争和压力的社会中，每个人都渴望成功，但成功并不是蹴而就、一帆风顺的。当我们遇到困难时，是否能保持一颗平静之心？是否能够坦然面对失败？这一切都取决于个人对待生活、对待工作、对待他人以及自身价值实现等方面所持有怎样种态度和价值观念。

“我不相信命运”这是很多年轻人经常挂在嘴边的话之一。然而事实上命运从来不会眷顾那些没有准备之人！如果你没有足够强大到足以应对各种困难与挑战的话，那么就不要去想着改变自己了！因为改变命运只能靠自己！

Pangu- α : Advertisement Generation

使用Prompt输入推广关键信息，自动写出商业文案：

Prompt注入信息：华为云相关描述

华为坚持开放，合作，共赢的云生态，作为“智能世界”的黑土地，帮助合作伙伴快速融入当地生态。华为云携手合作伙伴联合创新，持续为客户和伙伴创造价值。截至2019年底，华为云已上线210+云服务、210+解决方案，服务全球众多知名企业。

华为云已与国际领先云基础设施提供商合作，在全球数十个国家部署云服务，帮助客户连接国际网络，全球服务网络。利用华为云，全球顶尖跨国企业将平台部署在全球各地，用户能享受到低成本、低风险的云服务，同时还能随时随地安全地访问全球信息资源。

华为云帮助世界级企业搭建云网络、云应用，通过领先的技术和丰富的业务经验，帮助他们将业务快速扩展至5G和物联网领域。我们的客户既有来自全球的顶级航空公司、汽车制造商、半导体公司等行业龙头，也有众多中国的优秀企业。在云服务领域，华为拥有业内最丰富的云服务产品线，拥有独一无二的业务体验，能够满足客户对云服务的更高期望。华为云与合作伙伴的合作有力推动了5G发展，不断促进云服务生态体系的成熟。华为致力于通过开放的云生态系统，与合作伙伴共同帮助企业客户走向成功。

Content

Our Work

Pangu- α : A Large-Scale Chinese Pre-trained Language Model

Filip: Fine-Grained Interactive Language-Image Pre-Training

Spiral: Self-Supervised Speech Pre-Training

Filip: Fine-Grained Interactive Language-Image Pre-Training

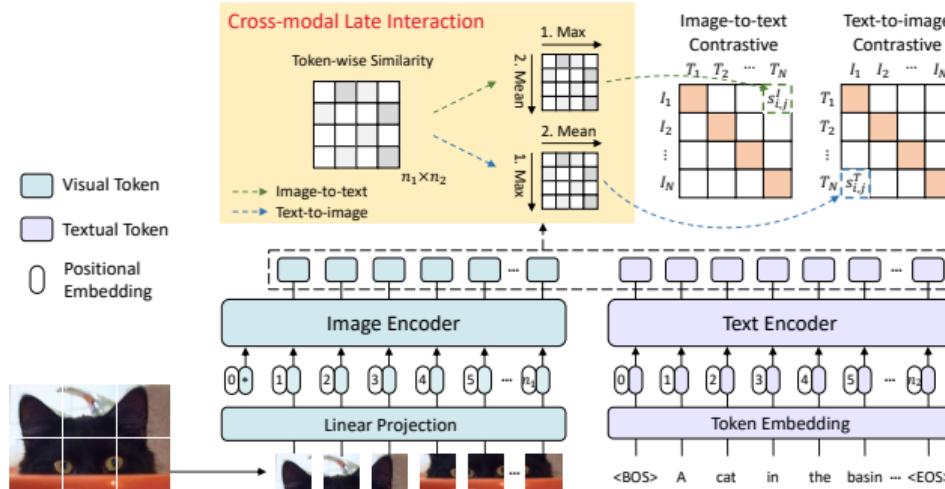


Figure 1: Overall architecture of FILIP, a dual-stream model with Transformer-based image and text encoders. On top of the image and text encoders, the representations of textual tokens and visual tokens are linearly projected to the multi-modal joint space. A novel fine-grained contrastive learning equipped with cross-modal late interaction is proposed, which uses a token-wise maximum similarity between visual and textual tokens.

Yao et al., Filip: Fine-Grained Interactive Language-Image Pre-Training, accepted by ICLR 2022

Filip: Fine-Grained Interactive Language-Image Pre-Training

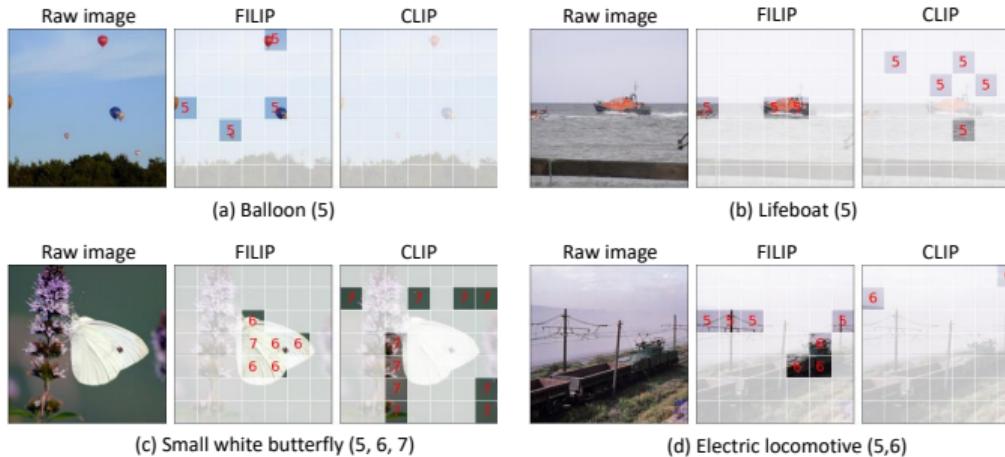


Figure 2: Visualizations of word-patch alignment for 4 classes of the ImageNet dataset and “a photo of a {label}.” is the prompt. Numbers in the parentheses after the class label indicate the location indices of the class label in the tokenized textual sequence. The correct predictions are highlighted by opaque patches with the class label indices in red.

Yao et al., Filip: Fine-Grained Interactive Language-Image Pre-Training, accepted by ICLR 2022

Content

Our Work

Pangu- α : A Large-Scale Chinese Pre-trained Language Model

Filip: Fine-Grained Interactive Language-Image Pre-Training

Spiral: Self-Supervised Speech Pre-Training

Spiral: Self-Supervised Speech Pre-Training

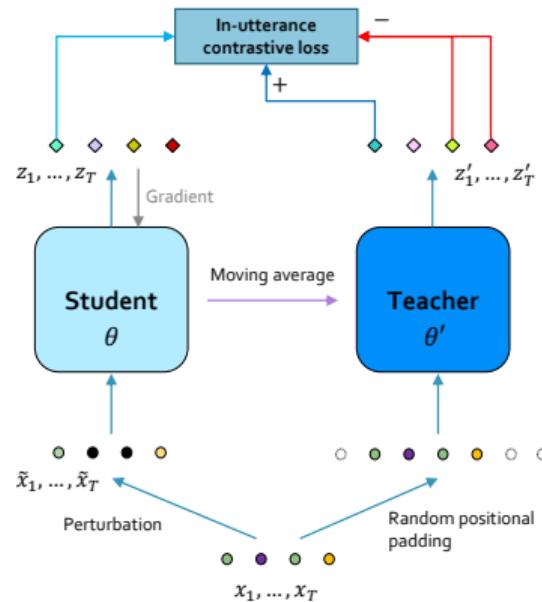


Figure 1: Illustration of SPIRAL architecture for speech pre-training.

Huang et al., Spiral: Self-Supervised Perturbation-Invariant Representation Learning For Speech Pre-Training, accepted by ICLR 2022

Content

Introduction

Opportunities

Challenges

Our Work

Summary

Introduction

Opportunities

Challenges

Our Work

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.

Copyright©2018 Huawei Technologies Co., Ltd.
All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.

