Large Language Models: Research and Practice

刘群 LIU Qun

华为诺亚方舟实验室 Huawei Noah's Ark Lab

FST Symposium on Science and Technology and Graduation Ceremony

2023-05-25, University of Macao

Content

Content

Large Language Models (LLMs): Background

Pangu Models

LLM Research in Huawei Noah's Ark Lab

Future Work

What is a Large Language Model?

From Wikipedia, the free encyclopedia

- A large language model (LLM) is a language model consisting of a neural network with many parameters (typically billions of weights or more), trained on large quantities of unlabeled text using self-supervised learning or semi-supervised learning. LLMs emerged around 2018 and perform well at a wide variety of tasks. This has shifted the focus of natural language processing research away from the previous paradigm of training specialized supervised models for specific tasks.
- Though the term large language model has no formal definition, it often refers to deep learning models having a parameter count on the order of billions or more. LLMs are general purpose models which excel at a wide range of tasks, as opposed to being trained for one specific task (such as sentiment analysis, named entity recognition, or mathematical reasoning). The skill with which they accomplish tasks, and the range of tasks at which they are capable, seems to be a function of the amount of resources (data, parameter-size, computing power) devoted to them, in a way that is not dependent on additional breakthroughs in design.
- Though trained on simple tasks along the lines of predicting the next word in a sentence, neural language models with sufficient training and parameter counts are found to capture much of the syntax and semantics of human language. In addition, large language models demonstrate considerable general knowledge about the world, and are able to "memorize" a great quantity of facts during training.

From Pre-trained Language Models (PLMs) to LLMs

	Pre-trained Language Models (PLMs)	Large Language Models (LLMs)
Typical Models	ELMo, BERT, GPT	GPT-2, GPT-3
Model Structure	BiLSTM, Transformer	Transformer
Model Framework	Encoder, Encoder-decoder, Decoder	Decoder
Attention Mechanism	Bidirectional、Unidirectional	Unidirectional
Tranining Method	Mask & Predict Autoregressive Generation	Autoregressive Generation
Downs. Task Types	NLU	NLU & NLG
Model Size	0.1-1B parameters	1-1000B parameters
Downs. Tasks Adapt.	Fine-tuning	Prompting & Fine-tuning & RLHF
Emergence Abilities	Inductive Transfer Learning	Zero-shot Learning Few-shot/In-context Learning Chain-of-Thought

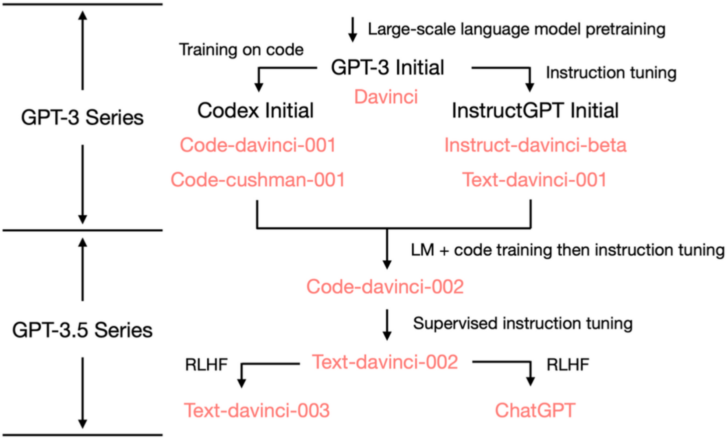
List of typical LLMs

Name	Release date	Developer	Number of parameters	Corpus size
GPT-2	2019-02-14	OpenAI	1.5 billion	40GB (~10 billion tokens)
GPT-3	2020-06-11	OpenAI	175 billion	499 billion tokens
GPT-Neo	2021-03-01	EleutherAI	2.7 billion	825 GiB
PanGu-α	2021-04-26	Pengcheng Lab and Huawei	200 billion	40 billion tokens
GPT-J	2021-06-01	EleutherAI	6 billion	825 GiB
Megatron-Turing NLG	2021-10-01	Microsoft and Nvidia	530 billion	338.6 billion tokens
Gopher	2021-12-01	DeepMind	280 billion	300 billion tokens
GLaM (Generalist Language Model)	2021-12-01	Google	1.2 trillion (sparse)	1.6 trillion tokens
Ernie 3.0 Titan	2021-12-01	Baidu	260 billion	4 Tb
Claude	2021-12-01	Anthropic	52 billion	400 billion tokens
LaMDA (Language Models for Dialog Applications)	2022-01-01	Google	137 billion	1.56T words, 168 billion tokens

List of typical LLMs

Name	Release date	Developer	Number of parameters	Corpus size
GPT-NeoX	2022-02-01	EleutherAI	20 billion	825 GiB
Chinchilla	2022-03-01	DeepMind	70 billion	1.4 trillion tokens
PaLM (Pathways Language Model)	2022-04-01	Google	540 billion	768 billion tokens
OPT (Open Pretrained Transformer)	2022-05-01	Meta	175 billion	180 billion tokens
YaLM 100B	2022-06-01	Yandex	100 billion	1.7TB
Minerva	2022-06-01	Google	540 billion	38.5B tokens from webpages filtered for mathematical content and from papers submitted to the arXiv preprint server
BLOOM	2022-07-01	Large collaboration led by Hugging Face	175 billion	350 billion tokens (1.6TB)
AlexaTM (Teacher Models)	2022-11-01	Amazon	20 billion	1.3 trillion
LLaMA (Large Language Model Meta AI)	2023-02-01	Meta	65 billion	1.4 trillion
GPT-4	2023-03-01	OpenAI	Unknown	Unknown
PanGu-Σ	2023-03-20	Huawei	1 trillion (sparse)	300 billion tokens

The road map of GPT-3 families



Yao Fu, How does GPT Obtain its Ability? Tracing Emergent Abilities of Language Models to their Sources (Blog)

Pros and Cons of LLMs

- Pros:
  - Language Ability
  - Simple Reasoning Ability
  - Human-like Behaviour
- Cons:
  - Halluciation
  - Math, Logic and Complex Reasoning Ablilties
  - Security: Bias, Offence, Discrimination...

HUAWEI NOAH'S ARK LAB

Content

Large Language Models (LLMs): Background

Pangu Models

LLM Research in Huawei Noah's Ark Lab

Future Work

Content

Pangu Models

PanGu-α: A Chinese 200-billion-parameters dense language lodel

Pangu-Σ series: a multi-domian one-trillion-parameters sparse language model

Pangu- α : A Large-scale Autoregressive Pretrained Chinese Language Model

PanGu-α: Large-scale Autoregressive Pretrained Chinese Language Models with Auto-parallel Computation

Technical Report

Wei Zeng* Xiaozhe Ren* Teng Su* Hui Wang*

Yi Liao Zhiwei Wang Xin Jiang Zhenzhang Yang Kaisheng Wang Xiaoda Zhang

Chen Li Ziyan Gong Yifan Yao Xinjing Huang Jun Wang Jianfeng Yu Qi Guo

Yue Yu Yan Zhang Jin Wang Hengtao Tao Dasen Yan Zexuan Yi Fang Peng

Fangqing Jiang Han Zhang Lingfeng Deng Yehong Zhang Zhe Lin

Chao Zhang Shaojie Zhang Mingyue Guo Shanzhi Gu Gaojun Fan Yaowei Wang

Xuefeng Jin Qun Liu Yonghong Tian

PanGu-α Team

Technical report: http://arxiv.org/abs/2104.12369

Pangu- α : A Large-scale Autoregressive Pretrained Chinese Language Model

- The first Chinese autoregressive dense LM with 200B parameters
- State-of-the-art performance in few-shot Chinese NLP tasks
- Code and model open-sourced
- Fully based on Huawei technology stack (MindSpore+CANN+Ascend910)
- Collaboration with Pengcheng Lab, Peking University and Huawei CSL

Technical report: http://arxiv.org/abs/2104.12369

Pangu- α : Model architecture

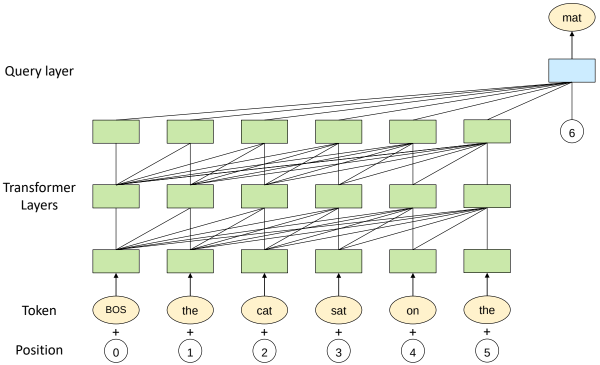


Figure 1: The architecture of PanGu-α. The model is based on a uni-directional Transformer decoder. A query layer is stacked on top of Transformer layers with the position embedding as the query in the attention mechanism to generate the token at the next position.

Pangu- α : Model sizes and data collection and filtering

Table 1: Model sizes and hyperparameters of PanGu-α models.

Model	#Parameters	#Layers (L)	Hidden size (d)	FFN size (d_{ff})	#Heads (N_h)
PanGu-α 2.6B	2.6B	32	2560	10240	40
PanGu-α 13B	13.1B	40	5120	20480	40
PanGu-α 200B	207.0B	64	16384	65536	128

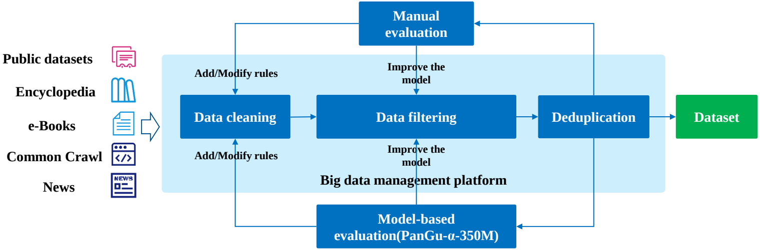


Figure 2: The data sources and the process of constructing pretraining data for PanGu-α.

HUAWEI NOAH'S ARK LAB

Pangu- α : Data composition and sampling strategy

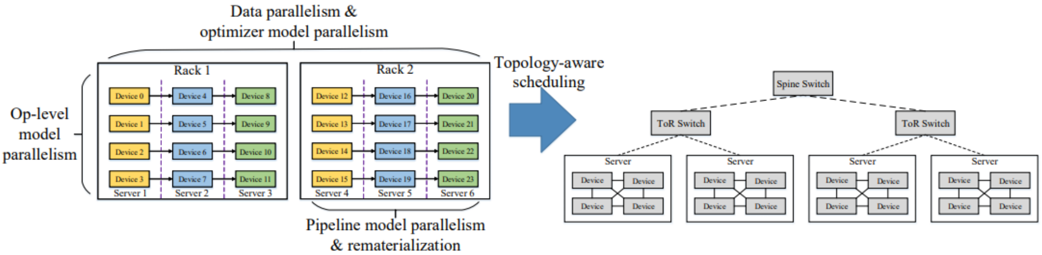
Table 3: Data composition of the 1.1TB Chinese text corpus.

	Size (GB)	Data source	Processing steps
Public datasets	27.9	15 public datasets including DuReader, BaiDuQA, CAIL2018, Sogou-CA, etc.	Format conversion[11] and text deduplication
Encyclopedia	22	Baidu Baike, Sogou Baike, etc.	Text deduplication
e-Books	299	e-Books on various topics (e,g., novels, history, poetry, ancient prose, etc.).	Sensitive word and model-based spam filtering
Common Crawl	714.9	Web data from January 2018 to December 2020 from Common Crawl.	All steps
News	35.5	News data from 1992 to 2011.	Text deduplication

Table 4: Sampling strategy of the corpora in training PanGu-α models.

	PanGu-α 200B			PanGu-α 2.6B&13B	
	Quantity (tokens)	Weight in training mix	Epochs elapsed when training	Quantity (tokens)	Weight in training mix
Public datasets	25.8B	10.23%	3.65	7B	27.99%
e-Books	30.9B	12.23%	0.41	5.6B	18%
Common Crawl	176.2B	62.81%	0.85	2.5B	10%
News	19.8B	7.83%	2.2	5.6B	22%
Encyclopedia data	5.8B	6.9%	3	5.8B	23%

PanGu- α : Training techniques - Model Parallelization



(a) How the partitioned model and data are mapped onto the hardware

(b) A brief example of hardware topology

Technical Report: https://arxiv.org/pdf/2104.12369.pdf

HUAWEI NOAH'S ARK LAB

Pangu- α : Training techniques - Parallelization strategy

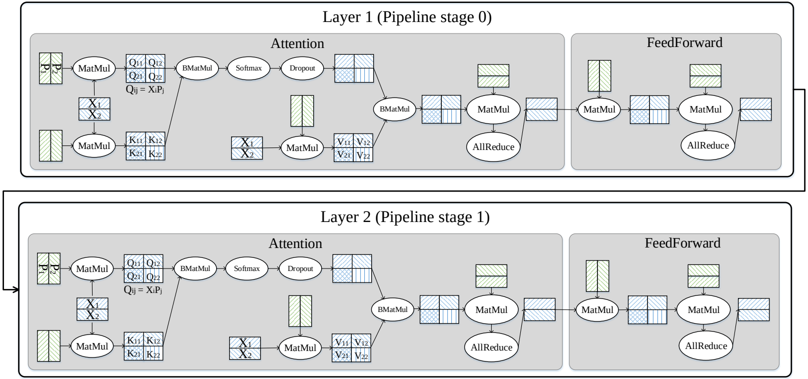
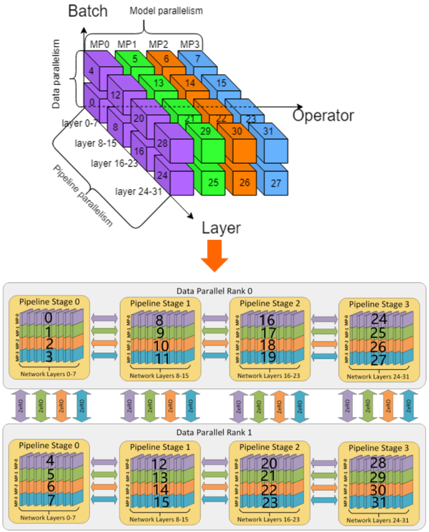


Figure 6: A simplified PanGu-α's parallelization strategy. The ellipsoids stand for the operators, blue rectangles represent tensors, and green rectangles represent trainable parameters. Parameters are partitioned along the row and column dimension respectively, and the input tensor is partitioned along the row dimension. And, two layers are assigned to different pipeline stages.

PanGu- α : Training techniques - 3-D parallel training

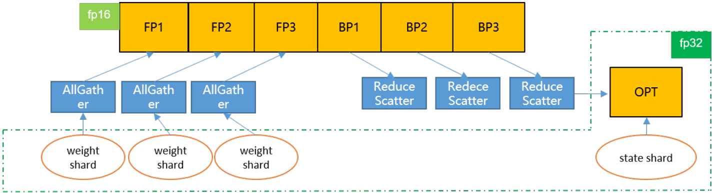
- 3-D mixture paralle: data parallel + pipeline parallel + model parallel
  - Data parallel: partition in batche dimension
  - Pipeline parallel: partition in layer dimension
  - Model parallel: partition in operator dimension
- By mapping 3-D coordinates to physical devices, we can train the huge models like GPT-3 efficiently.

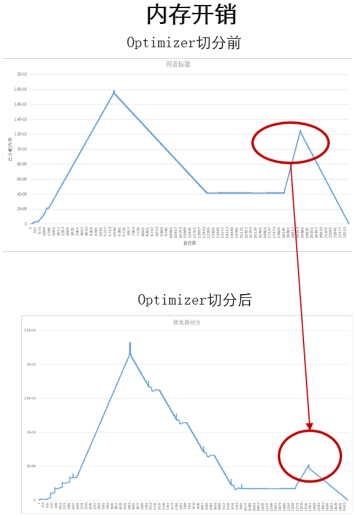


Coordinate	RANK	Coordinate	RANK	Coordinate	RANK	Coordinate	RANK
(0, 0, 0)	0	(1, 0, 0)	8	(2, 0, 0)	16	(3, 0, 0)	24
(0, 0, 1)	1	(1, 0, 1)	9	(2, 0, 1)	17	(3, 0, 1)	25
(0, 0, 2)	2	(1, 0, 2)	10	(2, 0, 2)	18	(3, 0, 2)	26
(0, 0, 3)	3	(1, 0, 3)	11	(2, 0, 3)	19	(3, 0, 3)	27
(0, 1, 0)	4	(1, 1, 0)	12	(2, 1, 0)	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/

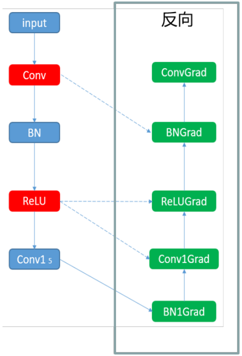
PanGu- α : Training techniques - Optimizer state parallel

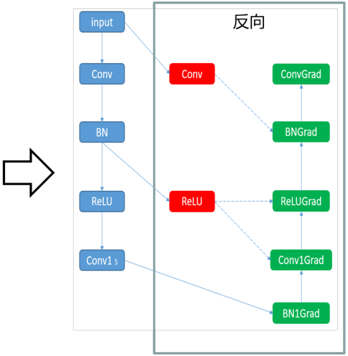




- Feature:
  - inner-layer partition: partition in dimensions of parameters, optimizer states and gradients
  - communication grouping parallel: allgather and reduce-catter, forward and backword computing
  - mixture precision: use fp16 for forward-backword propogation and communication, use fp32 for optimizer parameters

PanGu- α : Training techniques - Re-computing



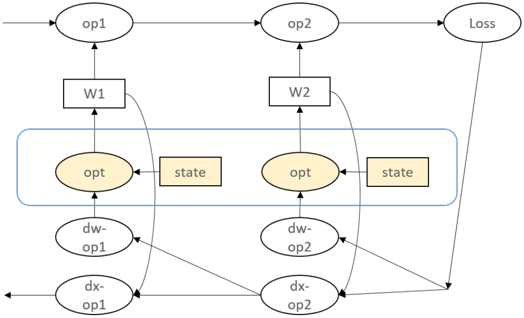


- Abandon activitions in forward computing, and re-computing them in backward propogation. Trade time for spaces.

HUAWEI NOAH'S ARK LAB

PanGu- α : Training techniques - Heterogeneous computing

- In the past few years, the model sizes increased by 1000 times, while the memory of parallel computing devices only increased by 5 times (GPU memory: 16G to 80G)
- Move parts of computing of training to Host CPUs and Host memories. A typical solution is optimizer heterogeneous computing.
  - The number of Adam Optimizer states is twice of the number of model weights: A 175B GPT-3 model has 350B optimizer states
  - Move the adam optimizer computing to Host CPU, and optimizer states to Host memory.
  - This can greatly reduce the memory cost in GPU/NPUs.



Optimizer CPU execution

HUAWEI NOAH'S ARK LAB

Pangu- α : Training curves

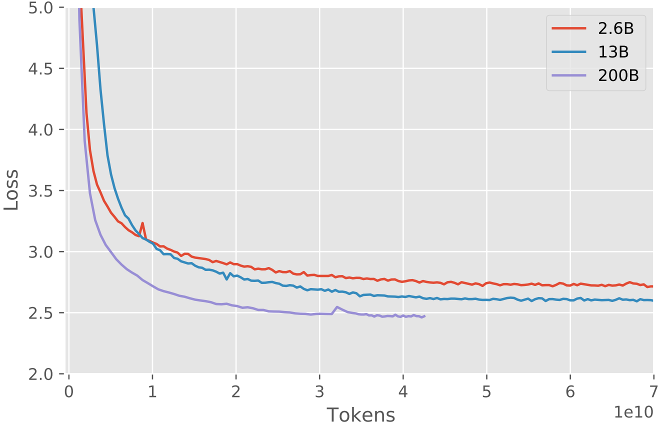


Figure 8: Training curves of three PanGu-α models with different model sizes. The x-axis denotes the number of training tokens, which is measured as $training\_steps * batch\_size * sequence\_length$. The y-axis denotes the training loss.

HUAWEI NOAH'S ARK LAB

Pangu- α : Experimental results

Table 9: Performance comparison of CPM 2.6B v.s. PanGu-α 2.6B on few-shot NLP tasks.

				Zero-Shot		One-Shot		Few-Shot		
Dataset	Method	Metrics	Task Types	CPM 2.6B	PanGu-α 2.6B	CPM 2.6B	PanGu-α 2.6B	#Shot(K)	CPM 2.6B	PanGu-α 2.6B
CMRC2018	Generation	Em/F1	Read Comprehension	0.59/10.12	**1.21/16.647**	1.71/11.29	**2.49/18.57**	Dynamic	3.11/14.64	**5.68/23.22**
DRCD	Generation	Em/F1	Read Comprehension	0/4.62	**0.8/9.99**	0.22/5.17	**2.47/12.48**	Dynamic	0.15/7.14	**5.31/18.29**
DuReader	Generation	Rouge-1	Read Comprehension	16.63	**21.07**	16.42	**20.18**	6,6	17.85	**21.43**
WebQA	Generation	Em/f1	Closed-Book QA	6/12.59	**6/16.32**	6/11.82	**12/23.39**	8,8	4/12.23	**24/33.94**
PD-CFT	Generation	Acc	Cloze(without choices)	35.73/38.99	**38.47/42.39**	33.3/39.73	**38.8/41.61**	3,3	32.03/39.84	**39.07/42.05**
CMRC2017	Generation	Acc	Cloze(without choices)	24.60	**37.83**	25.40	**38.00**	3,3	23.50	**36.33**
CHID	PPL	Acc	Cloze(multi-choices)	68.62	**68.73**	67.91	**68.16**	3,3	**66.82**	66.56
CMRC2019	PPL	Acc	Cloze (multi-choices)	47.69	**61.93**	47.99	**61.54**	2,2	47.20	**62.42**
CMNLI	PPL	Acc	Natural Language Inference	49.10	**50.20**	47.56	**49.54**	6,12	49.29	**51.17**
OCNLI	PPL	Acc	Natural Language Inference	**44.20**	42.61	**44.30**	44.00	3,6	44.00	**46.78**
TNEWS	PPL	Acc	Text classification	**65.44**	60.95	**69.50**	57.95	6,6	**70.17**	63.62
IFLYTEK	PPL	Acc	Text classification	68.91	**74.26**	**79.84**	79.03	3,3	**83.99**	80.15
AFQMC	PPL	Acc	Sentence Pair Similarity	**66.34**	59.29	39.70	**64.62**	4,4	38.29	**69.00**
CSL	PPL	Acc	Keyword Recognition	**52.30**	50.50	**51.20**	50.90	10,10	50.50	**52.00**
CLUEWSC2020	PPL	Acc	WSC	**73.684**	73.36	73.684	**75.33**	14,14	70.065	**72.70**
C^3	PPL	Acc	Common Sense Reasoning	49.81	**53.42**	51.43	**52.82**	3,3	51.60	**53.64**

Pangu- α : Experimental results

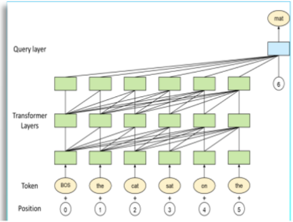
Table 10: Performance comparison of PanGu-α 2.6B v.s. PanGu-α 13B on few-shot NLP tasks.

Dataset	Method	Metrics	Task Types	Zero-Shot PanGu-α 2.6B	Zero-Shot PanGu-α 13B	One-Shot PanGu-α 2.6B	One-Shot PanGu-α 13B	Few-Shot #Shot(K)	Few-Shot PanGu-α 2.6B	Few-Shot PanGu-α 13B
CMRC2018	Generation	Em/F1	Read Comprehension	1.21/16.65	**1.46/19.28**	2.49/18.57	**3.76/21.46**	Dynamic	5.68/23.22	**9.76/29.23**
DRCD	Generation	Em/F1	Read Comprehension	**0.8**/9.99	0.66/**10.55**	2.47/12.48	**4.22/15.01**	Dynamic	5.31/18.29	**9.09/23.46**
DuReader	Generation	Rouge-1	Read Comprehension	21.07	**24.46**	20.18	**25.99**	6,6	21.43	**27.67**
WebQA	Generation	Em/f1	Closed-Book QA	4.43/13.71	**5.13/14.47**	10.22/20.56	**13.43/24.52**	8,8	23.71/33.81	**31.18/41.21**
PD-CFT	Generation	Acc	Cloze(without choices)	38.47/42.39	**43.86/46.60**	38.8/41.61	**40.97/45.42**	3,3	39.07/42.05	**41.13/45.86**
CMRC2017	Generation	Acc	Cloze(without choices)	37.83	**38.90**	38.00	**38.40**	3,3	36.33	**37.86**
CHID	PPL	Acc	Cloze(multi-choices)	68.73	**70.64**	68.16	**70.05**	3,3	66.56	**70.91**
CMRC2019	PPL	Acc	Cloze (multi-choices)	68.22	**70.54**	68.05	**70.02**	2,2	66.26	**71.28**
CMNLI	PPL	Acc	Natural Language Inference	**50.20**	48.44	**49.54**	46.81	6,12	**51.17**	46.18
OCNLI	PPL	Acc	Natural Language Inference	**42.61**	41.53	44.00	**44.10**	3,6	**46.78**	46.44
TNEWS	PPL	Acc	Text classification	**60.95**	60.26	57.95	**63.83**	6,6	63.62	**65.17**
IFLYTEK	PPL	Acc	Text classification	**74.26**	73.80	**79.03**	78.95	3,3	80.15	**80.34**
AFQMC	PPL	Acc	Sentence Pair Similarity	59.29	**65.76**	**64.62**	63.55	4,4	**69.00**	68.91
CSL	PPL	Acc	Keyword Recognition	**50.50**	49.30	**50.90**	50.20	10,10	52.00	**55.70**
CLUEWSC2020	PPL	Acc	WSC	73.36	**75.00**	**75.33**	75.00	14,14	72.70	**78.62**
C^3	PPL	Acc	Common Sense Reasoning	53.42	**54.47**	52.82	**53.92**	3,3	53.64	**54.58**
WPLC	PPL	ppl	Chinese WPLC	16.70	**19.18**	-	-	-	-	-

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* Xiaozhe Ren* Teng Su* Hui Wang*
Yi Liao Zhiwei Wang Xin Jiang Zhenzhang Yang Kaisheng Wang Xiaoda Zhang
Chen Li Ziyan Gong Yifan Yao Xinjing Huang Jun Wang Jianfeng Yu Qi Guo
Yue Yu Yan Zhang Jin Wang Hengtao Tao Dasen Yan Zexuan Yi Fang Peng
Fangqing Jiang Han Zhang Lingfeng Deng Yehong Zhang Zhe Lin
Chao Zhang Shaojie Zhang Mingyue Guo Shanzhi Gu Gaojun Fan Yaowei Wang
Xuefeng Jin Qun Liu Yonghong Tian Wen Gao

PANGU-α TEAM

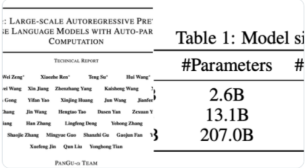
https://arxiv.org/abs/2104.12369

Pangu- α : Influence

Horace He
@cHHillee
...

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
E LANGUAGE MODELS WITH AUTO-PAR
COMPUTATION

TECHNICAL REPORT

Wei Zeng* Xiaozhe Ren* Teng Su* Hui Wang*
ei Wang Xin Jiang Zhenzhang Yang Kaisheng Wang
Gong Yifan Yao Xinjing Huang Jun Wang Jianfe
hang Jin Wang Hengtao Tao Dasen Yan Zexuan Y
iang Han Zhang Lingfeng Deng Yehong Zhang
Shaojie Zhang Mingyue Guo Shanzhi Gu Gaojun Fan
Xuefeng Jin Qun Liu Yonghong Tian

PANGU-α TEAM

Table 1: Model si

#Parameters	#
2.6B	
13.1B	
207.0B	

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人员上百次转发点赞，OpenAI主管Miles Brundage关注转发：https://twitter.com/cHHillee/status/1386541907950465028

VB The Machine GamesBeat Jobs Special Issue Become a Member | Sign In

The Machine
Making sense of AI

Huawei trained the Chinese-language equivalent of GPT-3

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-α), 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 u/programmerChilli Researcher 14 days ago

86

[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/cHHillee/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://t.co/BwQepOVIYq?amp=1

49 Comments Award Share Save Hide Report 90% Upvoted

Reddit论坛讨论：https://www.reddit.com/r/MachineLearning/comments/myr072/d_huawei_just_announced_that_they_trained_a_200/

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/

HUAWEI NOAH'S ARK LAB

Pangu- α : Examples - Text summarization

Prompt	"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。 2013年，华为分别展示过为汽车内部提供3GWifi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"。 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"。 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。 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。 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
Generation	"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。"

Pangu- α : Examples - Title generation

Prompt	"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。 2013年，华为分别展示过为汽车内部提供3GWifi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"。 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"。 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。 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。 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》

Pangu- α : Examples - Open domain QA

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，以及原株洲车站货房。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	中国的首都是哪里?
Generation	中国的首都是北京。
Prompt	中国、美国、日本和澳大利亚的首都分别是哪里?
Generation	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	五一川西行程规划
Generation	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 : Examples - 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 : Examples - Stylized text 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 : Examples - Gaokao 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 : Examples - 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

Pangu Models

PanGu- α : A Chinese 200-billion-parameters dense language lodel

Pangu- Σ series: a multi-domian one-trillion-parameters sparse language model

Pangu- Σ :
A multi-domian one-trillion-parameters sparse language model

PANGU-Σ: TOWARDS TRILLION PARAMETER LANGUAGE MODEL WITH SPARSE HETEROGENEOUS COMPUTING

TECHNICAL REPORT

Xiaozhe Ren[1*] Pingyi Zhou[1*] Xinfan Meng[1*] Xinjing Huang[2*] Yadao Wang[1*]

Weichao Wang[1] Pengfei Li[1] Xiaoda Zhang[2] Alexander Podolskiy[1] Grigory Arshinov [1]

Andrey Bout [1] Irina Piontkovskaya [1] Jiansheng Wei[1] Xin Jiang[1]

Teng Su[2] Qun Liu[1] Jun Yao[1]

[1]Noah's Ark Lab, Huawei Technologies
[2]Distributed and Parallel Software Lab, Huawei Technologies

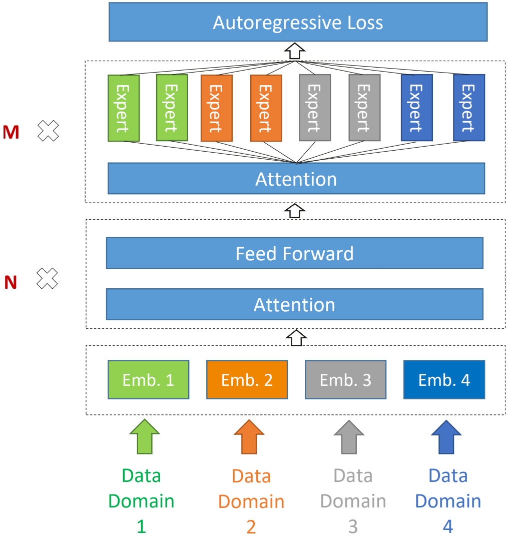
Ren et al. “PanGu- Σ : Towards Trillion Parameter Language Model...” arxiv:2303.10845. 2023-03-19.

Pangu-Σ : A multi-domian one-trillion-parameters sparse language model

- One trillion parameters sparse LM
- Fully based on Huawei technology stack
- Long time stable training based on 512 Ascend D910 Card + MindSpore
- Extremely simple expert routing strategy: Ramdomly Routing Experts (RRE)
- Pluggable multi-domain multi-task life-long learning, with lossless expert-tailor
- Enabling industrial deployment on single server (with 8 Ascend cards)
- SotA on zero-shot and fine-tuned performance on Chinese downstream tasks, including QA, dialog and translation

Ren et al. "PanGu-Σ : Towards Trillion Parameter Language Model..." arxiv:2303.10845. 2023-03-19.

Pangu- Σ : Architecture



- **Mixed Dense/Sparse Architecture**

  While M=0 or Expert=1
  Degraded to dense architecture
  Equivalent to Pangu-Alpha

  While N=0
  Fully sparse architecture

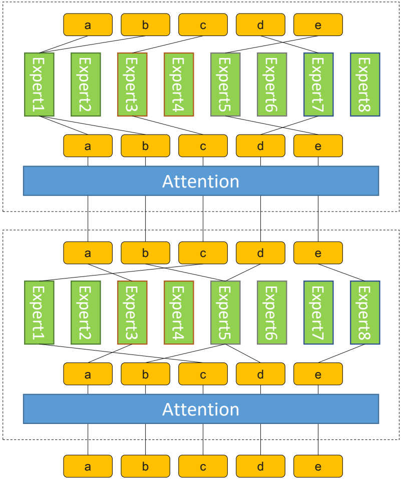
- **Higher Layers: Grouped RRE Repre.**
- **Lower Layers: Universal Dense Repre.**
- **Mixted Data/Task Training**

HUAWEI NOAH'S ARK LAB

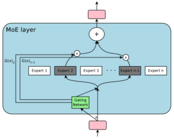
Pangu- Σ : Randomly-Routed Experts (RRE)

[a:1, b:1,c:3, d:7, e:5]

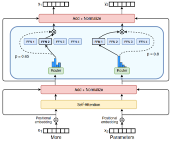
[a:3, b:5,c:1, d:5, e:8]



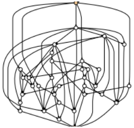
Mixture of Expert
Google 2017



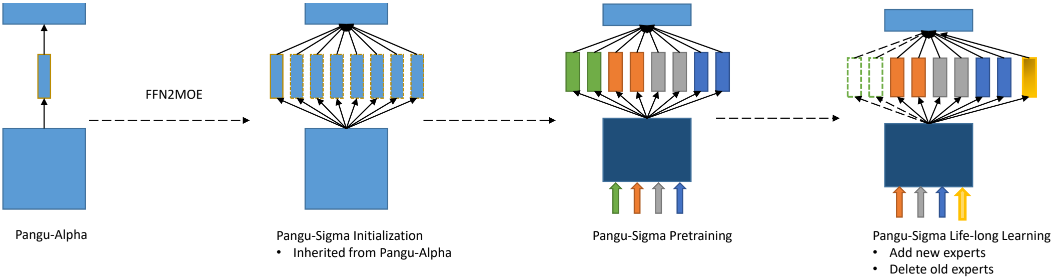
Switch Transformer
Google 2021



Randomly Connected NN
Meta 2019



Pangu- Σ : Multi-task Life-long Learning



FFN2MOE

- ✓ Inherited knowledge from Pangu-Alpha
- ✓ Speed-up convergence

Two-layer RRE

- ✓ Task-expert fine-grained Control
- ✓ Expert Workload Balance
- ✓ Grouped All-to-All Communication

Expert Editing

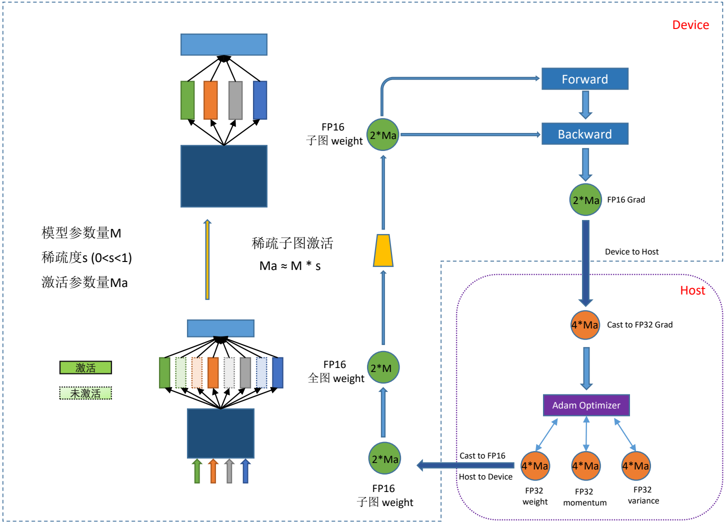
- ✓ Expert add/delete/edit
- ✓ Single Domain Fast deployment

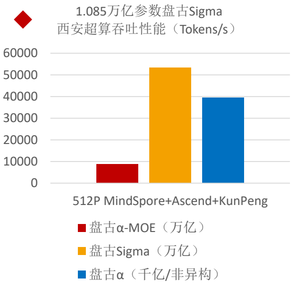
Domain Expanding

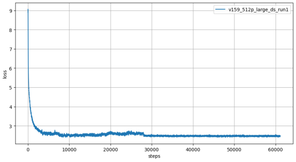
- ✓ Monolingual to multilingual
- ✓ Single domain to multi-domain

HUAWEI NOAH'S ARK LAB

Pangu-Σ : High Performance Heterogeneous Training





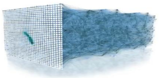


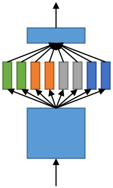
1.085万亿参数盘古Sigma收敛曲线
(持续稳定训练中)

HUAWEI NOAH'S ARK LAB

Pangu- Σ : Industrial Deployment

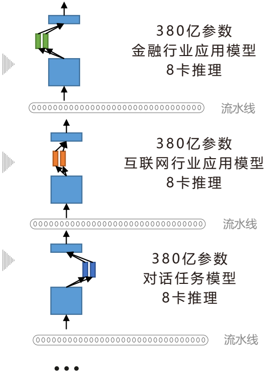
盘古Sigma
512卡
Ascend+MindSpore
高性能训练





万亿参数基础模型

盘古Sigma
30倍
无损剪枝



380亿参数
金融行业应用模型
8卡推理

380亿参数
互联网行业应用模型
8卡推理

380亿参数
对话任务模型
8卡推理

AI工业化开发

模型泛化　极大节省训练投入
流水线　工具集成，训练更快

效率提升
10-100倍
更佳模型性能

万亿参数盘古Sigma预训练 → 行业/领域子模型部署

迭代

Pangu- Σ : Training Configuration

Data

Data size：304.12B Tokens（1 TB）：

- Chinese：75.47B Tokens
  - ✓ General domain
  - ✓ Financial, Medical, Law, ...
- English：75.9B Tokens
  - ✓ General domain
- English-Chinese：77.51B Tokens
  - ✓ General domain
  - ✓ News
- Code：75.24B Tokens
  - ✓ Python
  - ✓ Java

Computing

Platform：Huawei AI Stack

- Hardware：
  - ✓ 西安超算：512卡Ascend910
- Software：
  - ✓ MindSpore v1.6
  - ✓ CANN c80
- Training time：
  - ✓ 100 days

Model

#Parameters：1.085Trillion

- Model Arch.：Decoder+RRE
- Model Specification:
  - ✓ Layers：40
  - ✓ RRE Layers：8
  - ✓ Hidden：5120
  - ✓ FFN：20480
  - ✓ Expert Per Layer：640
- Training Specification：
  - ✓ Ascend+KunPeng Heterogeneous Training
  - ✓ 8-way model parallel
  - ✓ 64-way expert parallel
  - ✓ 64way data parallel

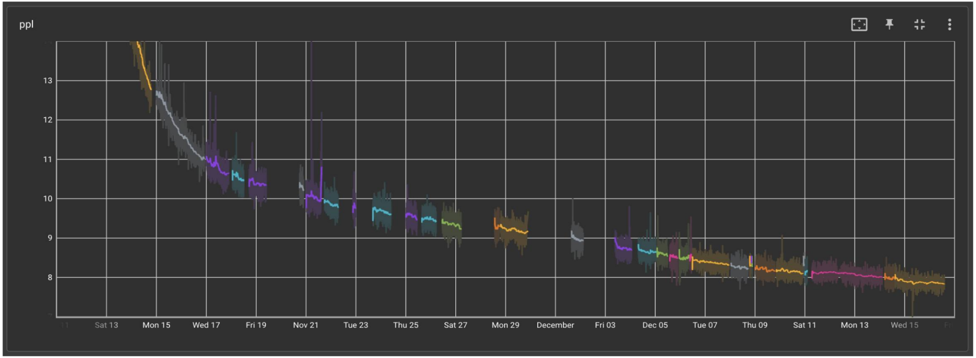
HUAWEI NOAH'S ARK LAB

OPT-175B: Longest stable training duration: 2.8 days

Infrastructure Stability

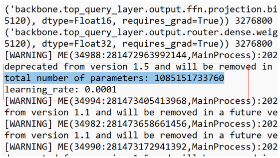
992 80GB A100 GPUs + PyTorch Megatron

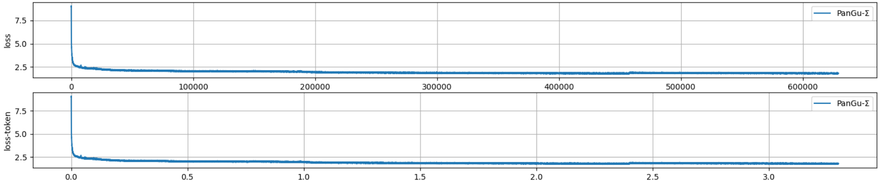
We managed to hit our top three record long runs of the experiment these past two weeks, lasting 1.5, 2.8, and 2 days each! If we were to look at only the runs that have contributed to pushing training further and plot training perplexity against wall clock time, we get the following:



https://github.com/facebookresearch/metaseq/blob/main/projects/OPT/chronicles/56_percent_update.md

Pangu- Σ : Stable training for 25 days until manual termination

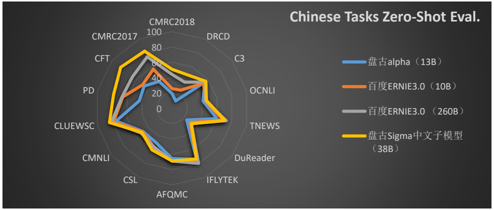


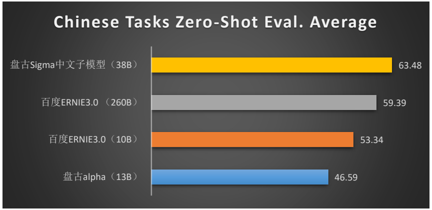


HUAWEI NOAH'S ARK LAB

Pangu- Σ : Performance on NLP downstream tasks

- Outperform Baidu Ernie 3.0 on zero-shot evaluation on 10 Chinese downstream tasks





Model	CMRC (em/f1)	DRCD (em/f1)	C3 (acc)	OCNLI (acc)	Tnews (acc)	Dureader (rouge-1)	Iflytek (acc)	Afqmc (acc)	CSL (acc)	Cmnli (acc)	Cluewsc (acc)	Pd (acc)	Cft (acc)	Cmrc2017 (acc)	Average
Pangu alpha 13B	1.46/ 19.28	0.66/ 10.55	54.47	41.53	60.26	24.46	73.80	65.76	49.30	48.44	75.00	43.86	46.60	38.90	46.59
Baidu ERNIE3.0 10B	7.61/ 25.61	10.58/ 26.29	52.62	44.31	68.40	29.79	75.34	68.99	55.63	49.41	78.38	66.07	49.30	56.66	53.34
Baidu ERNIE3.0 260B	16.62/ 44.2	21.08/ 37.83	54.85	44.61	72.60	32.13	79.84	68.99	55.80	51.70	81.08	67.06	66.14	74.63	59.39
Pangu Sigma38B Chinese Sub-model	20.88/ 50.83	26.82/ 48.09	56.49	46.03	70.82	32.29	73.57	68.84	60.33	50.53	83.88	78.47	85.87	82.73	63.48

Pangu- Σ : Downstream Tasks - Dialog

Human Metrics

Models	Adequacy	具体性	趣味性	事实性错误	安全性	综合
CDialGPT	0.597	0.680	0.143	0.070	0.970	0.664
EVA	0.507	0.743	0.210	0.080	0.953	0.667
EVA2.0	0.677	0.783	0.287	0.070	0.980	0.731
Pangu-BOT	0.810	0.803	0.293	0.050	0.987	0.769
Pangu-Sigma	0.830	0.857	0.340	0.040	0.993	0.796

Automatic Metrics

Model	Distinct 1	Distinct 2	BLEU-2	BLEU-3	Semantic-coherence
CDialGPT	0.035	0.172	0.1405	0.088	0.344
EVA	0.067	0.313	0.169	0.103	0.393
EVA2.0	0.073	0.341	0.169	0.104	0.458
Pangu-BOT	0.089	0.337	0.171	0.106	0.459
Pangu-Sigma	0.109	0.369	0.177	0.110	0.502

Dialog Examples

```
---------------------------------------------------------------------------
Context:   你好 ！
 你好 。
 你 最近 知道 我们 中国男篮 进 了 世界杯 吗 ？
 不怎么 关注 中国男篮 ，不是 很 喜欢 。
 还是 可以 看看 的 嘛 ，我们 自己 的 队伍 能够 打 进 世界杯 还是 有 很大 进步 了 。
 算了吧 ，我们 中国 的 球类运动 除了 乒乓球 男女 一样 的 厉害 以外 ，其他 球类运动 男队 真的 不如 女队 。
 嗨 ，肯定 有好有坏 嘛 ，你 看 这 打篮球 的 姚明 、 易建联 还是 很 厉害 啊 ，都 在 NBA 效力 过 的 。
Response:  姚明 也 是 打篮球 的 啊 。
---------------------------------------------------------------------------
Context:   你好 呀 ！
 嗨 ，你好 。
 你 平时 喜欢 篮球 吗 ？你 有 看 最近 的 广东 中国男篮 世界杯 吗 ？
 看 了 有 关注 一点 ，这次 中国男篮 有点 不 在 状态 啊 ，挺 可惜 的 。
 是 的 ，这次 中国男篮 遭 失利 ，姚明 也 把 全部 责任 都 推到 自己 身上 。
 我 觉得 这 不是 姚明 的 主要 责任 ，至少 这 不 应该 姚明 一个 人去 承担 。
 是 的 ，这次 的 中国男篮 确实 不 在 状态 ，广东队 应该 还要 加强 训练 加强 球技 啊 。
Response:  对 啊 ，杜 大爷 言重 了 ，篮球 是 外线投篮 ，外线投篮 需要 2 个 大个子 的 外线投篮 。 朱芳雨 就 有 身高 啊 ，还有 这个 实力 。
---------------------------------------------------------------------------
Context:   嗨 i, 你好 呀 ！
 嗨 ，你好 。
 你 在 看 关于 网球 的 比赛 吗 ？
 对 呀 ，我 在 看 WTA 的 女单 1/4 决赛 。
 是 吗 ，你 很 喜欢 网球 这项 运动 吗 ？
 我 呀 ，我 还 挺 喜欢 的 ，我 平时 就 比较 喜欢 看 网球 比赛 。 你 呢 ？
 我 不是 很 喜欢 ，我 觉得 打网球 太 费力气 了 。
Response:  打网球 确实 是 挺 耗费 体力 的 ，尤其 是 初学者 。
---------------------------------------------------------------------------
```


HUAWEI NOAH'S ARK LAB

Pangu- Σ : Downstream Tasks - Question and Answering

QA metrics

Model	Precision	Recall	F1	Human Eval.
CDialGPT	0.033	0.067	0.041	0.036
EVA	0.008	0.051	0.012	0.036
EVA2.0	0.082	0.139	0.103	0.119
Pangu-BOT	0.509	0.761	0.556	0.738
Pangu-Sigma	0.574	0.822	0.621	0.774

QA Examples

Context: 最大 的 金字塔 是 哪 一个 ？
Response: 目前 ， 胡夫 金字塔 是 世界 上 最大 的 金字塔 。

Context: 世界 上 海拔 最高 的 高原 是 什么 高原 ？
Response: 青藏高原 是 世界 上 海拔 最高 的 高原 。

Context: 世界 上 最大 的 沙漠 是 什么 ？
Response: 撒哈 啦

Context: 世界 面积 最大 的 大洲 是 ？
Response: 北美洲 。

Context: 世界 最长 的 河 是 ？
Response: 尼罗河 ！

Pangu- Σ : Downstream Tasks - Translation

WMT20 Chinese-English

Models	BLEU
MT5-XXL-13B	24
CPM-2-11B	26.2
ERNIE3.0-10B	26.8
Pangu-Sigma-38B Chinese-English Low Resources	30.99
Pangu-Sigma-38B Chinese-English High Resources	36.63

WMT17 Chinese-English

Model	BLEU
CEMAT	22.82
Pangu-Sigma-38B Chinese-English	23.52

WMT17 English-Chinese

Model	BLEU
CEMAT	35.82
Pangu-Sigma-38B Chinese-English	38.82

Chinese-English Translation Examples

Source: [摘要] 浙江 小镇 的 潜力 被 逐步 唤醒 。
Predict: [Abstract] The potential of small towns in Zhejiang is gradually awakened.

Source: 预计 12 月 固定资产 投资 累计 增速 为 7.2%, 与 上 月 持平 , 低于 去年 。
Predict: The cumulative rate of fixed-asset investment is expected to be 7.2% in December, unchanged from the previous month, and lower than last year.

Source: " 漫步 绍兴 安昌 古镇 , 随处可见箍桶 、 竹编 、 打铁 、 纳鞋 、 挑 花边 、 纺 棉花 等 传统 手工 技艺 , 充分 展现 了 安昌 古镇 的 非 物质 文化遗产 , 四方 游客 陶醉 在 观赏 体验 绍兴 民俗风情 、 领略 绍兴
Predict: Walking around the ancient town of Anchang, visitors can see the traditional handicrafts of barrel-making, bamboo weaving, iron-making, shoe-making, lace-making and cotton-spinning. Thes

Source: 现在 我们 所有 的 信息 来源 沟通 渠道 都 被 切断 。
Predict: Now all of our sources of information are cut off.

Content

Large Language Models (LLMs): Background

Pangu Models

LLM Research in Huawei Noah's Ark Lab

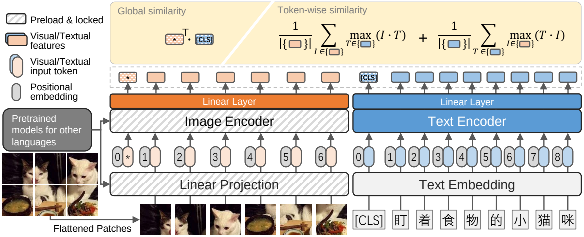
Future Work

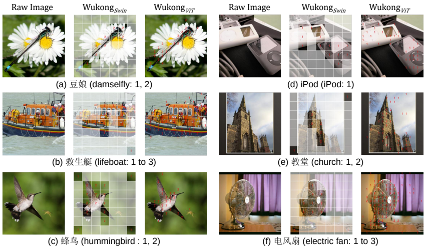
Content

LLM Research in Huawei Noah's Ark Lab

- Multi-modal Language Models
- Efficient Training and Deployment
- Arabic Language Models
- Information Retrieval
- Question Answering
- Machine Translation
- Poem Generation
- Code Generation
- Math Word Problem Solving

Wukong FILIP：
Fine-grained Interactive Language-Image Pre-Training





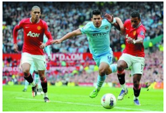
Technical report: https://arxiv.org/abs/2111.07783, https://arxiv.org/abs/2202.06767.pdf

Wukong Dataset:

A 100 Million Large-scale Chinese Cross-modal Pre-training Benchmark

狗子示意来访人员要想进去，先过来扫码，狗子还特意下来用嘴巴对着 (*The dog signaled to the visitors to scan the code first before entrance, and the dog also deliberately came down and pointed his mouth at it.*)

你好，我们是社区工作人员，是来做接种疫苗排查工作的 (*Hello, we are community workers and are here to do vaccination screening.*)



13-14赛季 英超第5轮 曼城 vs 曼联 13. 09. 22 (*13-14 Premier League Round 5 Manchester City vs Manchester United 13.09.22*)

中国骄傲中国女排成功抵达东京不到6天就将在赛场上再展风采 (*China pride, the Chinese women's volleyball team, will show its style on the field in less than 6 days right after its arrival in Tokyo*)

简欧三居室酒柜装修效果图 (*Renderings of the decoration of the wine cabinet in the three bedrooms of Europe*)

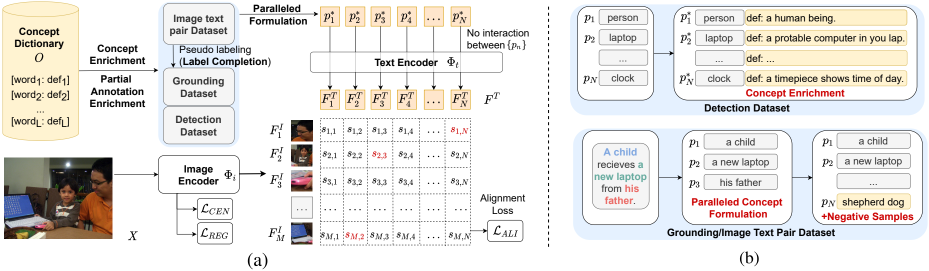
【互邦工厂旗舰店】上海互邦轮椅钢管轻便手动折叠轮椅 (*【Hubang factory flagship store】 Shanghai Hubang wheelchair steel pipe lightweight manual folding wheelchair*)

Dataset release: https://wukong-dataset.github.io/wukong-dataset/

Wukong DetCLIP：

Dictionary-Enriched Visual-Concept Paralleled Pre-training for Open-world Detection

- 外部知识库引入：引入wordnet来提供类别之间的先验关系
- 自动目标类别生成：通过融合开集检测和captioning任务来直接生成预测目标的类别，无需人工指定。
- 细粒度文本对齐预训练，百万级高分辨率数据大规模多机多卡并行训练。
- 在LVIS数据集上的检测精度已超过GLIP模型14.4% mAP，获得ECCV2022开集检测竞赛冠军。

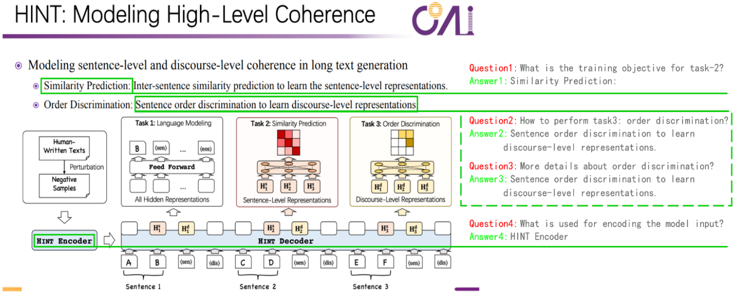


Paper: https://arxiv.org/abs/2209.09407

Wukong Reader：

Multi-modal Pre-training for Fine-grained Visual Document Understanding

- 构建了文本行对比学习、掩码区域建模和文本行方格匹配等多种预训练目标，综合文本、视觉表征和空间布局信息进行细粒度建模，学习统一的文档表示
- 在千万级文档数据（涵盖表单，宣传单，简历，科研论文等）上进行了无监督预训练，在下游文档信息抽取、分类等多种下游任务超越业界SOTA
- 具备强大的多任务和领域迁移能力，支持扫描文档、PDF、幻灯片、海报、网页截图等不同领域的文档理解与开放域信息抽取。



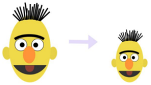
Paper: https://arxiv.org/abs/2212.09621

Content

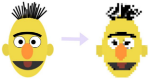
LLM Research in Huawei Noah's Ark Lab

Compression of Pre-trained Language Models

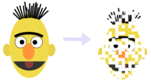
- Knowledge Distillation
  - DistilBERT/BERT-PKD/MobileBERT/MiniLM(Task agnostic)
  - Our Work: TinyBERT/Mate-KD/ALP-KD



- Quantization
  - Q-BERT/Q8BERT
  - Our Work: TernaryBERT/BinaryBERT
  - Our Work: QuantGPT/QuantBART (ACL2022 Outstanding Paper Award)



- Pruning/Slimmable
  - LayerDrop
  - Our Work: DynaBERT



- Model archetecture search
  - Our Work: AutoTinyBERT
- Automatic feature generation:
  - Our Work: GhostBERT

TinyBERT: Distilling BERT for Natural Language Understanding

- Deployable BERT

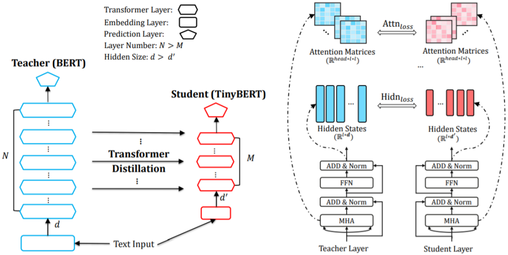
- Transformer-layer distillation
- Embedding-layer distillation
- Prediction-Layer distillation

- Two-stage learning: general (pre-training) distillation and the task-specific distillation

- 7.5x smaller and 9.4x faster on inference

- Ranked 1st at CLUE

- Accelerated on Bolt, on-device inference cost 6ms on ARM A76 CPU



System	#Params	#FLOPS	Speedup	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	Avg
$BERT_{BASE}$ (Teacher)	109M	22.5B	1.0x	83.9/83.4	71.1	90.9	93.4	52.8	85.2	87.5	67.0	79.5
$BERT_{TINY}$	14.5M	1.2B	9.4x	75.4/74.9	66.5	84.8	87.6	19.5	77.1	83.2	62.6	70.2
$BERT_{SMALL}$	29.2M	3.4B	5.7x	77.6/77.0	68.1	86.4	89.7	27.8	77.0	83.4	61.8	72.1
$BERT_4$-PKD	52.2M	7.6B	3.0x	79.9/79.3	70.2	85.1	89.4	24.8	79.8	82.6	62.3	72.6
$DistilBERT_4$	52.2M	7.6B	3.0x	78.9/78.0	68.5	85.2	91.4	32.8	76.1	82.4	54.1	71.9
$MobileBERT_{TINY}$†	15.1M	3.1B	-	81.5/81.6	68.9	**89.5**	91.7	**46.7**	80.1	**87.9**	65.1	**77.0**
$TinyBERT_4$ (ours)	14.5M	1.2B	9.4x	**82.5/81.8**	**71.3**	87.7	**92.6**	44.1	**80.4**	86.4	**66.6**	**77.0**
$BERT_6$-PKD	67.0M	11.3B	2.0x	81.5/81.0	70.7	89.0	92.0	-	-	85.0	65.5	-
$DistilBERT_6$	67.0M	11.3B	2.0x	82.6/81.3	70.1	88.9	92.5	49.0	81.3	86.9	58.4	76.8
$TinyBERT_6$ (ours)	67.0M	11.3B	2.0x	**84.6/83.2**	**71.6**	**90.4**	**93.1**	**51.1**	**83.7**	**87.3**	**70.0**	**79.4**

Published in EMNLP 2020: https://aclanthology.org/2020.findings-emnlp.372.pdf

HUAWEI NOAH'S ARK LAB

EMNLP2021 Top-Cited Paper: TinyBERT ...

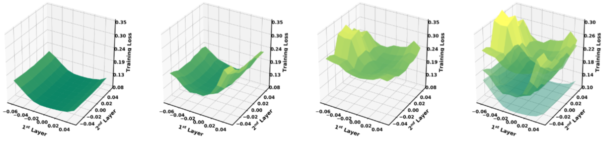
TABLE 1: Most Influential EMNLP Papers (2021-02)

YEAR	RANK	PAPER	AUTHOR(S)
2020	1	**TinyBERT: Distilling BERT For Natural Language Understanding** **IF:4** Related Papers Related Patents Related Grants Related Orgs Related Experts Details *Highlight: To accelerate inference and reduce model size while maintaining accuracy, we first propose a novel Transformer distillation method that is specially designed for knowledge distillation (KD) of the Transformer-based models.*	XIAOQI JIAO et. al.

"Paper Digest Team analyze all papers published on EMNLP in the past years, and presents the 10 most influential papers for each year."

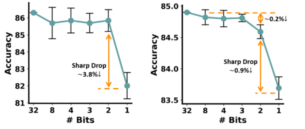
https://www.paperdigest.org/2021/02/most-influential-emnlp-papers/

BinaryBERT: Pushing the Limit of BERT Quantization



(a) Full-precision Model. (b) Ternary Model. (c) Binary Model. (d) All Together.

Figure 2: Loss landscapes visualization of the full-precision, ternary and binary models on MRPC. For (a), (b) and (c), we perturb the (latent) full-precision weights of the value layer in the 1st and 2nd Transformer layers, and compute their corresponding training loss. (d) shows the gap among the three surfaces by stacking them together.



(a) MRPC. (b) MNLI-m.

Figure 1: Performance of quantized BERT with varying weight bit-widths and 8-bit activation. We report the mean results with standard deviations from 10 seeds on MRPC and 3 seeds on MNLI-m, respectively.

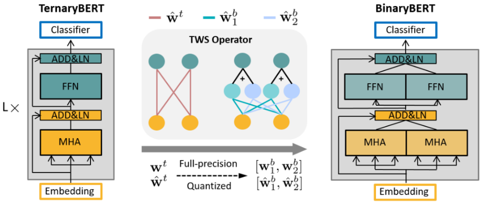


Figure 4: The overall workflow of training BinaryBERT. We first train a half-sized ternary BERT model, and then apply ternary weight splitting operator (Equations (6) and (7)) to obtain the latent full-precision and quantized weights as the initialization of the full-sized BinaryBERT. We then fine-tune BinaryBERT for further refinement.

Method	#Bits (W-E-A)	Size (MB)	Ratio (↓)	SQuAD v1.1	MNLI -m
BERT-base	*full-prec.*	418	1.0	80.8/88.5	84.6
DistilBERT	*full-prec.*	250	1.7	79.1/86.9	81.6
LayerDrop-6L	*full-prec.*	328	1.3	-	82.9
LayerDrop-3L	*full-prec.*	224	1.9	-	78.6
TinyBERT-6L	*full-prec.*	55	7.6	79.7/87.5	82.8
ALBERT-E128	*full-prec.*	45	9.3	82.3/89.3	81.6
ALBERT-E768	*full-prec.*	120	3.5	81.5/88.6	82.0
Quant-Noise	PQ	38	11.0	-	83.6
Q-BERT	2/4-8-8	53	7.9	79.9/87.5	83.5
Q-BERT	2/3-8-8	46	9.1	79.3/87.0	81.8
Q-BERT	2-8-8	28	15.0	69.7/79.6	76.6
GOBO	3-4-32	43	9.7	-	83.7
GOBO	2-2-32	28	15.0	-	71.0
TernaryBERT	2-2-8	28	15.0	79.9/87.4	83.5
BinaryBERT	**1-1-8**	**17**	**24.6**	**80.8/88.3**	**84.2**
BinaryBERT	**1-1-4**	**17**	**24.6**	**79.3/87.2**	**83.9**

Table 4: Comparison with other state-of-the-art methods on development set of SQuAD v1.1 and MNLI-m.

Published in ACL-IJCNLP2021: https://arxiv.org/pdf/2012.15701.pdf

QuantGPT and QuantBART

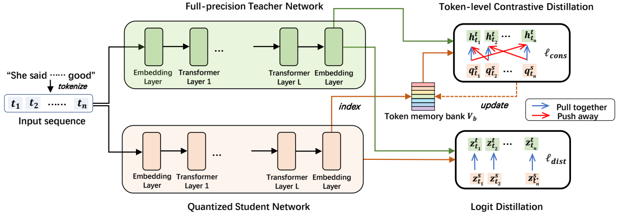
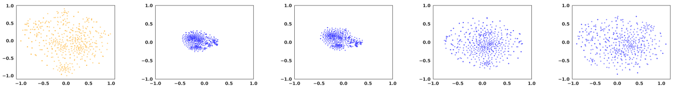


Figure 5: The training workflow of the proposed method. For each token in the quantized network, we compute both (i) the token-level contrastive distillation loss where the positive tokens and negative tokens are selected from the full-precision teacher network; and (ii) the distillation loss on the logits. The embedding layer and all weights in the Transformer layers are quantized with the proposed module-dependent dynamic scaling.



(a) Full-precision. (b) PACT. (c) LSQ. (d) LAQ. (e) Ours.

Figure 2: T-SNE visualization of the most frequent 500 word embeddings, of the full-precision and different 2-bit quantized models trained on PTB dataset. Embeddings of different methods show different degrees of homogeneity.

Method	Size (MB)(↓)	WikiText2 PPL(↓)	PTB PPL(↓)	WikiText103 PPL(↓)
full-prec.	474.9 (1.0x)	14.4	14.6	13.9
KnGPT2	332.0 (1.4x)	-	-	20.5
DistilGPT2	329.6 (1.4x)	-	-	21.1
LightPAFF	268.0 (1.8x)	18.8	22.8	16.4
Ours(8-8-8)	121.4 (3.9x)	**15.3**	**14.9**	**14.6**
Ours(4-4-8)	62.4 (7.6x)	15.6	15.0	15.3
Ours(2-2-8)	33.0 (**14.4x**)	17.3	16.1	17.0

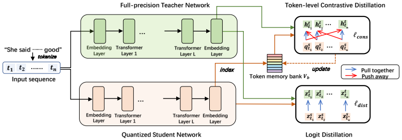
Table 2: Comparison between our proposed quatization method and other compression methods on GPT-2.

Method	#Bits (W-E-A)	Size (MB)(↓)	XSum		
Metric			R1 (↑)	R2 (↑)	RL (↑)
-	*full-prec.*	532.0	40.75	18.10	33.05
PACT	8-8-8	138.1	39.16	16.60	31.60
LSQ	8-8-8	138.1	39.09	16.72	31.56
LAQ	8-8-8	138.1	39.10	16.74	31.65
QuantBART	8-8-8	138.1	**40.25**	**17.78**	**32.70**
PACT	4-4-8	72.4	32.68	11.52	26.03
LSQ	4-4-8	72.4	38.94	16.48	31.46
LAQ	4-4-8	72.4	39.03	16.68	31.63
QuantBART	4-4-8	72.4	**40.24**	**17.71**	**32.69**
PACT	2-2-8	39.6	7.76	1.30	6.96
LSQ	2-2-8	39.6	37.09	14.88	29.76
LAQ	2-2-8	39.6	37.48	15.27	30.13
QuantBART	2-2-8	39.6	**39.15**	**16.72**	**31.72**

Table 3: Results of abstractive summarization on the test set of the XSum dataset, with quantized BART.

Published in ACL2022: http://arxiv.org/abs/2203.10705

ACL2022 Outstanding Paper Award: Compression of ...



ACL 2022

Association for Computational Linguistics

Outstanding Paper Award

Presented to

Chaofan Tao, Lu Hou, Wei Zhang, Lifeng Shang,
Xin Jiang, Qun Liu, Ping Luo, Ngai Wong

"Compression of Generative Pre-trained Language Models via Quantization"

25 May 2022

Bernardo Magnini, General Chair

Smaranda Muresan, Program Co-chair

Preslav Nakov, Program Co-chair

Aline Villavicencio, Program Co-chair

https://aclanthology.org/2022.acl-long.331/

HUAWEI NOAH'S ARK LAB

bert2BERT: Towards Reusable Pretrained Language Models

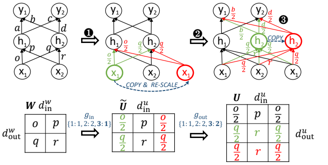


Figure 3: Overview of the function preserving initialization (FPI). Given the same input $\{x_1, x_2\}$, FPI ensures the initialized target model has the same output $\{y_1, y_2\}$ with the source model. The first and the second steps are expanding the in-dimension and out-dimension of the parameter matrix according to mapping functions g_{in} and g_{out} respectively. After we expand the matrix $\boldsymbol{W}$ into $\boldsymbol{U}$, we use the in-dimension expansion on the upper parameter matrix again to ensure the output $\{y_1, y_2\}$ same as the original one. From the view of neurons, FPI copies the corresponding input and output neurons to expand the neural network.

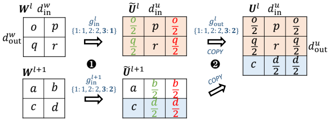


Figure 4: Overview of AKI. It first performs the in-dimension expansion on both the matrixes of current and upper layers. Then it uses the widened matrix of the current layer as the top part of the new matrix and samples the row of the widened matrix of the upper layer as the bottom part of the new matrix.

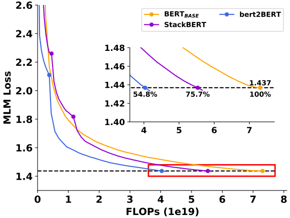
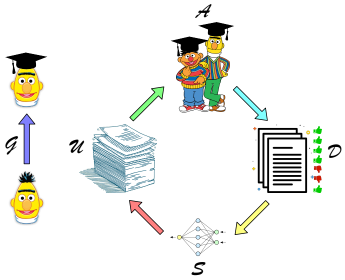


Figure 1: Loss curves of bert2BERT and baselines. StackBERT (Gong et al., 2019) is based on the progressive training setting. More details are shown in Table 2.

Published in ACL2022: https://aclanthology.org/2022.acl-long.151

LMTurk: Using LMaaS as Crowdsourcing Workers



Converting a PLM to LMTurker with few-shot gold data $\mathcal{G}$ of task $\mathcal{T}$.
Small model $\mathcal{S}$ predicts unlabelled data $\mathcal{U}$.
Select data $\mathcal{D}$ from $\mathcal{U}$ with active learning.
LMTurkers $\mathcal{A}$ annotate and aggregate labels of $\mathcal{D}$.
Training a new small model $\mathcal{S}$.

Figure 1: LMTurk overview; best viewed in color. We few-shot adapt PLMs to task $\mathcal{T}$ (left) and then use them as crowdsourcing workers in active learning. We show that these PLM workers are effective in training a small model $\mathcal{S}$ through a customized active learning loop (right). LMTurk is a novel way to take advantage of large-scale PLMs: It creates models small enough to be deployed in resource-limited real-world settings.

	Schick and Schütze (2021a,b)	Gao et al. (2021)	Ours
SST2	n/a	93.0±0.6	93.08±0.62
SST5	n/a	49.5±1.7	46.70±0.93
RTE	69.8	71.1±5.3	70.88±1.70
AGN.	86.3±0.0	n/a	87.71±0.07
CoLA	n/a	21.8±15.9	19.71±1.89

Table 1: LMTurkers achieve comparable few-shot performance with the literature. We refer to *PET* results in Schick and Schütze (2021a,b) and results of *Prompt-based FT (auto) + demonstrations* in Gao et al. (2021).

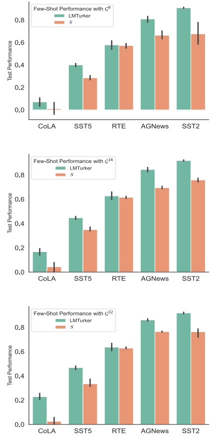


Figure 2: *Few-shot* test set performance of LMTurkers and $\mathcal{S}$. We use the few-shot gold datasets $\mathcal{G}^8$ (top), $\mathcal{G}^{16}$ (middle), and $\mathcal{G}^{32}$ (bottom).

Published in NAACL2022 Findings: https://aclanthology.org/2022.findings-naacl.511

Content

LLM Research in Huawei Noah's Ark Lab

JABER and SABER: Junior and Senior Arabic BERt

Model	Arabic-BERT	AraBERT	CAMeLBERT	ARBERT	MARBERT	JABER	SABER
#Params (w/o emb)	110M (85M)	135M (85M)	108M (85M)	163M (85M)	163M (85M)	135M (85M)	369M (307M)
Vocab Size	32k	64k	30k	100k	100k	64k	64k
Tokenizer	WordPiece	WordPiece	WordPiece	WordPiece	WordPiece	BBPE	BBPE
Normalization	✘	✓	✓	✘	✘	✓	✓
Data Filtering	✘	✘	✘	✘	✘	✓	✓
Textual Data Size	95GB	27GB	167GB	61GB	128GB	115GB	115GB
Duplication Factor	3	10	10	-	-	3	3
Training epochs	27	27	2	42	36	15	5

Table 1: Configuration comparisons of various publicly available Arabic BERT models and ours (JABER and SABER). AraBERT and MARBERT didn't provide their data duplication factor.

	Arabic-BERT	AraBERT	CAMeLBERT	ARBERT	MARBERT	JABER	SABER
MQ2Q*	73.3±0.6	73.5±0.5	68.9±1.1	74.7±0.1	69.1±0.9	75.1±0.3	**77.7±0.4**
MDD	61.9±0.2	61.1±0.3	62.9±0.1	62.5±0.2	63.2±0.3	65.7±0.3	**67.7±0.1**
SVREG	83.6±0.8	82.3±0.9	86.7±0.1	83.5±0.6	88.0±0.4	87.4±0.7	**89.3±0.3**
SEC	42.4±0.4	42.2±0.6	45.4±0.5	43.9±0.6	47.6±0.9	46.8±0.8	**49.0±0.5**
FID	83.9±0.6	85.2±0.2	84.9±0.6	85.3±0.3	84.7±0.4	84.8±0.3	**86.1±0.3**
OOLD	88.8±0.5	89.7±0.4	91.3±0.4	90.5±0.5	91.8±0.3	92.2±0.5	**93.4±0.4**
XNLI	66.0±0.6	67.2±0.4	55.7±1.2	70.8±0.5	63.3±0.7	72.4±0.7	**75.9±0.3**
OHSD	79.3±1.0	79.9±1.8	81.1±0.7	81.9±2.0	83.8±1.4	85.0±1.6	**88.9±0.3**
Avg.	72.4±0.6	72.6±0.6	72.1±0.6	74.1±0.6	73.9±0.7	76.2±0.7	**78.5±0.3**

Table 4: DEV performances and standard deviations over 5 runs on the ALUE benchmark. Bold entries describe the best results among all models, while underlined entries show best results among BERT-base models. * indicates that the results are on our own MQ2Q dev set.

ALUE

Arabic Language Understanding Evaluation

ALUE Paper Cite Code Tasks Leaderboard FAQ Diagnostics Submit Login

Leaderboard

Rank	Name	Model	Details	Score	MQ2Q	MDD	SVREG	SEC	FID	OOLD	XNLI	OHSD	DIAG
1	Huawei Noah's Ark Lab MTL	SABER		77.3	93.3	66.5	79.2	38.8	86.5	93.4	76.3	84.1	26.2
2	Huawei Noah's Ark Lab MTL	JABER		73.7	93.1	64.1	70.9	31.7	85.3	91.4	73.4	79.6	24.4
3	ALUE Baseline	ARABIC-BERT		67.1	85.7	59.7	55.1	25.1	82.2	89.5	61.0	78.7	19.6
4	ALUE Baseline	BERT Multi-lingual Cased		61.0	83.2	61.3	33.9	14.0	81.6	80.3	63.1	70.5	19.0
5	ALUE Baseline	BERT Multi-lingual Uncased		58.6	75.8	58.0	32.0	13.8	81.0	79.8	57.9	70.6	15.1

Preprint: https://arxiv.org/pdf/2112.04329v3.pdf

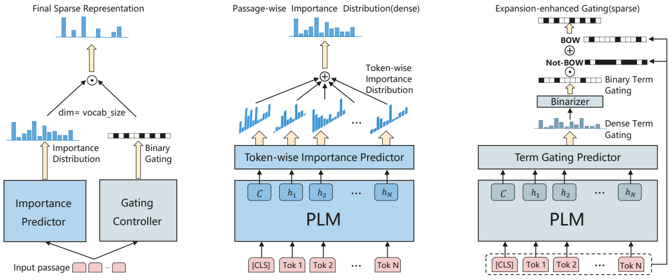
ALUE Leaderboard https://www.alue.org/leaderboard

HUAWEI NOAH'S ARK LAB

Content

LLM Research in Huawei Noah's Ark Lab

SparTerm:
Learning Term-based Sparse Representation for Fast Text Retrieval



(a) SparTerm Model (b) Importance Predictor (c) Gating Controller

Figure 2: Model Architecture of SparTerm. Our overall architecture contains an importance predictor and a gating controller. The importance predictor generates a dense importance distribution with the dimension of vocabulary size, while the gating controller outputs a sparse and binary gating vector to control term activation for the final representation. These two modules cooperatively ensure the sparsity and flexibility of the final representation.

Query	Can hives be a sign of pregnancy?	
Type	Term frequency	SparTerm
Literal term Weights	hives are caused by allergic reactions . the dryness and stretching of your skin along with other changes can make you more susceptible to experiencing hives during pregnancy . hives can be caused by an allergic reaction to almost anything . some common causes of hives during pregnancy are noted below : medicine	hives are caused by allergic reactions . the dryness and stretching of your skin along with other changes can make you more susceptible to experiencing hives during pregnancy . hives can be caused by an allergic reaction to almost anything . some common causes of hives during pregnancy are noted below : medicine
Term expansion		symptoms:1.0, women:0.99, rash:0.98, feel:0.99, causing:0.97, body:0.96, affect:0.96, baby:0.94, pregnant:0.93, **sign:0.91**, …

Figure 1: The comparison between BoW and SparTerm representation. The depth of the color represents the term weights, deeper is higher. Compared with BoW, SparTerm is able to figure out the semantically important terms and expand some terms not appearing in the passage but very semantically relevant, even the terms in the target query such as "sign".

Model	MRR@10
BM25+*PassageRetrievalMax*	23.6
HDCT+*PassageRetrievalMax*	26.1
BM25	24.5
HDCT(sum)	28.0
HDCT(decay)	28.7
SparTerm(literal-only)+*PassageRetrievalMax*	28.5
SparTerm(expansion-enhanced)+*PassageRetrievalMax*	**29.0**

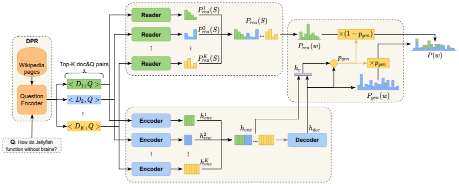
Table 3: Performance of baselines and our models on dev set of MSMARCO document ranking dataset. All use the max score of passages in the document as the document score at the query time.

Preprint: arXiv:2010.00768

Content

LLM Research in Huawei Noah's Ark Lab

Read before Generate!
Faithful Long Form Question Answering with Machine Reading



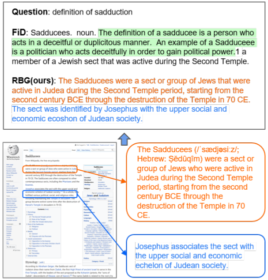


Figure 1: An example from MS MARCO (Nguyen et al., 2016) dataset. We highlight the unfaithful snippets from other model. Our model(**RBG**) generate more factually accurate answer.

Models	Eli5		MS MARCO	
	ROUGE-L	F1	ROUGE-L	F1
T5(base)	21.02	18.36	21.19	20.03
BART(large)	22.69	22.19	23.26	25.6
DPR+BART	17.41	17.88	23.01	25.13
RAG	16.11	17.24	-	-
FiD	25.70	28.55	24.64	27.08
RBG(ours)	**26.46**	**29.04**	**24.72**	**27.52**

Table 1: Performance comparison between our RBG method and the baselines on the KILT-ELI5 (Petroni et al., 2021) and MS MARCO (Nguyen et al., 2016) evaluation sets.

Model	Retrieval		Generation		
	PRr.	R@5	F1	R-L	KRL
RBG(ours)	10.83	27.25	**24.53**	**27.13**	**2.62**
DPR_kilt_wiki	14.83	27.69	16.45	15.91	2.46
c-REALM[1]	10.67	24.56	23.19	22.88	2.36
DPR+BART	10.67	26.92	17.41	17.88	1.90
RAG	11.00	22.92	14.05	14.51	1.69
BART-large	0.00	0.00	20.55	19.23	0.00
T5-base	0.00	0.00	19.08	16.10	0.00

Published in ACL2022 Findings: https://aclanthology.org/2022.findings-acl.61

Content

LLM Research in Huawei Noah's Ark Lab

CeMAT:

Universal Conditional Masked Language Pre-training for Neural Machine Translation

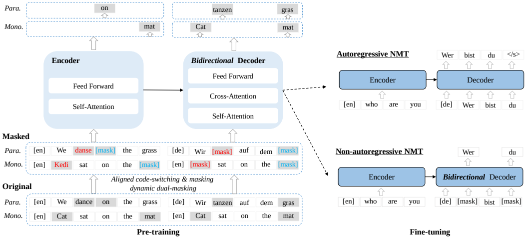


Figure 1: The framework for CeMAT, which consists of an encoder and a ***bidirectional decoder***. "*Mono*" denotes monolingual, "*Para*" denotes bilingual. During the pre-training (left), the original monolingual and bilingual inputs in many languages are augmented (the words are replaced with new words with same semantics or "[mask]", please see Figure 2 for more details) and fed into the model. Finally, we predict all the "[mask]" words on the source side and target side respectively. For fine-tuning (right), CeMAT provides unified initial parameter sets for AT and NAT.

Autoregressive MMT results:

Lang-Pairs	En-Kk		En-Tr		En-Et		En-Fi		En-Lv		En-Cs	En-De	En-Fr	Avg
Source	WMT19		WMT17		WMT18		WMT17		WMT17		WMT19	WMT19	WMT14	
Size	91k(low)		207k(low)		1.94M(medium)		2.66M(medium)		4.5M(medium)		11M(high)	38M(extr-high)	41M(extr-high)	
Direction	→	←	→	←	→	←	→	←	→	←	→	→	→	
Direct	0.2	0.8	9.5	12.2	17.9	22.6	20.2	21.8	12.9	15.6	16.5	30.9	41.4	17.1
mBART	2.5	7.4	17.8	22.5	21.4	27.8	22.4	28.5	15.9	19.3	18.0	30.5	41.0	21.2
mRASP	8.3	12.3	20.0	23.4	20.9	26.8	24.0	28.0	21.6	**24.4**	19.9	35.2	**44.3**	23.8
CeMAT	**8.8**	**12.9**	**23.9**	**23.6**	**22.2**	**28.5**	**25.4**	**28.7**	**22.0**	24.3	**21.5**	**39.2**	43.7	25.0
Δ	+8.6	+12.1	+14.4	+11.4	+4.3	+5.9	+5.2	+6.9	+9.1	+8.7	+5.0	+8.3	+2.3	+7.9

Table 2: Comprehensive comparison with mRASP and mBART. Best results are highlighted in **bold**. CeMAT outperforms them on AT for all language pairs but two directions. Even for extremely high-resource scenarios(denoted as "extr-high"), we observe gains of up to +8.3 BLEU on En→De language pair.

Non-autoregressive MMT results:

Source	IWSLT14		WMT16		WMT14		Avg
Lang-Pairs	En→De	De→En	En→Ro	Ro→En	En→De	De→En	
Transformer (Vaswani et al., 2017)	23.9	32.8	34.1	34.5	28.0	32.7	31.0
Mask-Predict (Ghazvininejad et al., 2019)	22.0	28.4	31.5	31.7	26.1	29.0	28.1
mRASP (Lin et al., 2020)	23.9	30.3	32.2	32.1	26.7	29.8	29.2
CeMAT (Ours)	**26.7**	**33.7**	**33.3**	**33.0**	**27.2**	**29.9**	30.6

Table 5: Comprehensive comparison with two strong baselines. "mRASP" denotes using mRASP to initialize Mask-Predict, "CeMAT (Ours)" denotes using our CeMAT to initialize. We obtain consistent and significant improvements on all language pairs, outperforming AT on IWSLT14 tasks. Best non-autoregressive results are highlighted in **bold**.

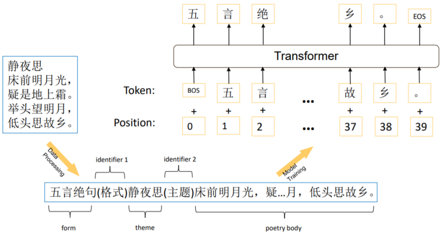
Published in ACL2022: https://aclanthology.org/2022.acl-long.442

Content

LLM Research in Huawei Noah's Ark Lab

GPT-based Classical Chinese Poetry Generation

- Pre-trained GPT model on Chinese news corpus, then fine-tuned with 250,000 Chinese poetries and couplets
- No human crafted rules or features
- Generate well-formed and high-quality poetries given the title, with good diversity
- Online demo on Huawei Cloud, gaining great popular on Chinese social media



五绝(Wujue)·秋思
暮燕翻惊户， 飞鸿却唤人。 西风卷梧叶， 触落一庭秋。
七绝(Qijue)·秋思
年华冉冉飞无翼， 风物萧萧滞故乡。 万里重云正愁绝， 洞庭湖外见清霜。

Preprint: https://arxiv.org/abs/1907.00151

Content

LLM Research in Huawei Noah's Ark Lab

Pangu-Coder: a Function Level Code Generation Model

MODEL	SIZE	n_{CNTX}	n_{VOCAB}	DATA (GB)	TRAIN TOKENS	HUMANEVAL (%) PASS@1	PASS@10	PASS@100
GPT-NEO [10]	125 M	2,048	50 K	825	300 B	0.75	1.88	2.97
CODEX [16]	300 M	4,096	50 K	729	400 B	13.17	20.37	**36.27**
ALPHACODE [47]	302 M	2,304	8 K	715	-	11.60	18.80	31.80
CODEGEN MULTI [51]	350 M	2,048	50 K	1,595	250 B	6.67	10.61	16.84
CODEGEN MONO [51]	350 M	2,048	50 K	1,812	325 B	12.76	23.11	35.19
PANGU-CODER	317 M	1,024	42 K	147	211 B	**17.07**	**24.05**	34.55
CODEX	679 M	4,096	50 K	729	400 B	16.22	25.70	40.95
ALPHACODE	685 M	2,304	8 K	715	-	14.20	24.40	38.80
ALPHACODE	1.1 B	2,304	8 K	715	-	17.10	28.20	45.30
GPT-NEO	1.3 B	2,048	50 K	825	380 B	4.79	7.47	16.30
CODEX	2.5 B	4,096	50 K	729	400 B	21.36	35.42	**59.50**
PANGU-CODER	2.6 B	1,024	42 K	147	387 B	**23.78**	35.36	51.24
CODEGEN MULTI	2.7 B	2,048	50 K	1,595	500 B	14.51	24.67	38.56
CODEGEN MONO	2.7 B	2,048	50 K	1,812	650 B	23.70	**36.64**	57.01
GPT-NEO	2.7 B	2,048	50 K	825	420 B	6.41	11.27	21.37
GPT-J [67]	6 B	2,048	50 K	825	402 B	11.62	15.74	27.74
CODEGEN MULTI	6.1 B	2,048	50 K	1,595	1 T	18.20	28.70	44.90
CODEGEN MONO	6.1 B	2,048	50 K	1,812	1.3 T	**26.13**	**42.29**	**65.82**
INCODER [27]	6.7 B	2,048	27.6 K	216	52 B	15.20	27.80	47.00

Table 4: Pass@k rates on the HumanEval dataset, among various models. Sizes are reported in thousands (K), millions (M), billions (B) and trillions (T).[9]

MODEL	# LAYERS (L)	HIDDEN SIZE (d)	FFN size (d_{ff})	# HEADS (N_h)	CONTEXT SIZE (n_{CNTX})	VOCAB (n_{VOCAB})
PANGU-CODER 317 M	24	1,024	4,096	16	1,024	41,865
PANGU-CODER 2.6 B	32	2,560	10,240	32	1,024	41,865

Table 2: PANGU-CODER model sizes and configurations.

- Autoregressive LM architecture (317M/2.6B) herited from Pangu-alpha
- Two-stage training schema, with different training data formating
  - Stage-1: 188B tokens
  - Stage-2: 42B tokens
- Outperform Codex/AlphaCode models with similar sizes on Pass@1 metric on HumanEval dataset.

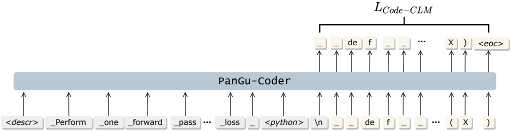
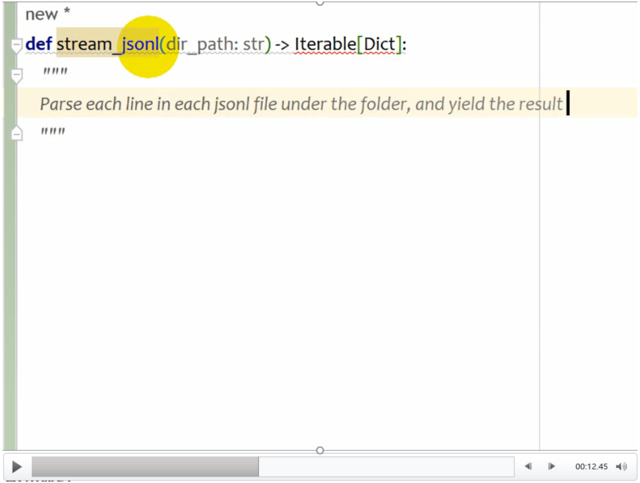


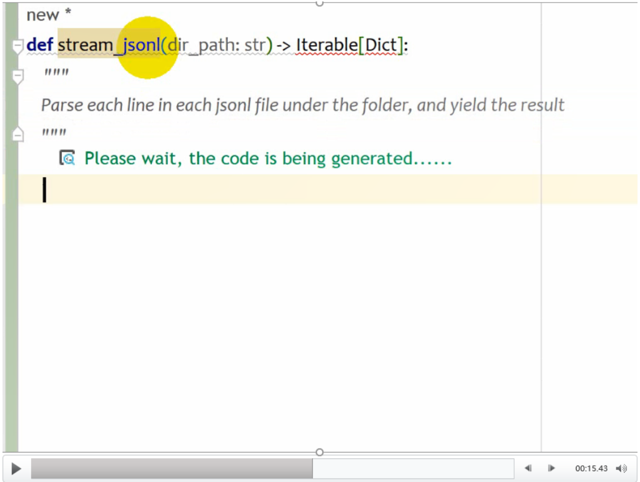
Figure 4: CODE-CLM: Causal Language Modeling over code-only tokens.

Technical Report: https://arxiv.org/abs/2207.11280

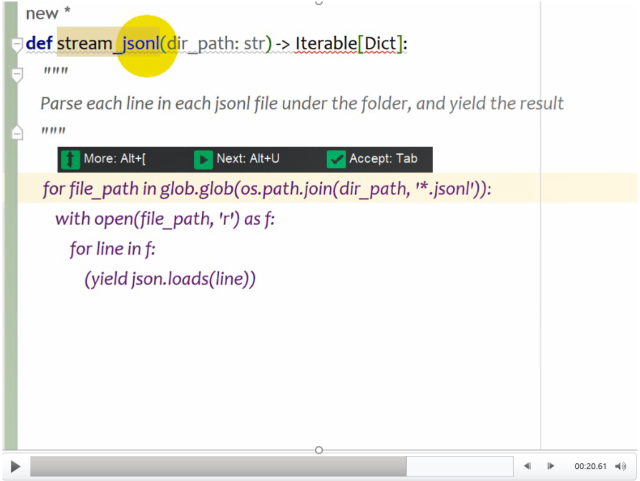
Pangu-Coder: a Function Level Code Generation Model



Pangu-Coder: a Function Level Code Generation Model

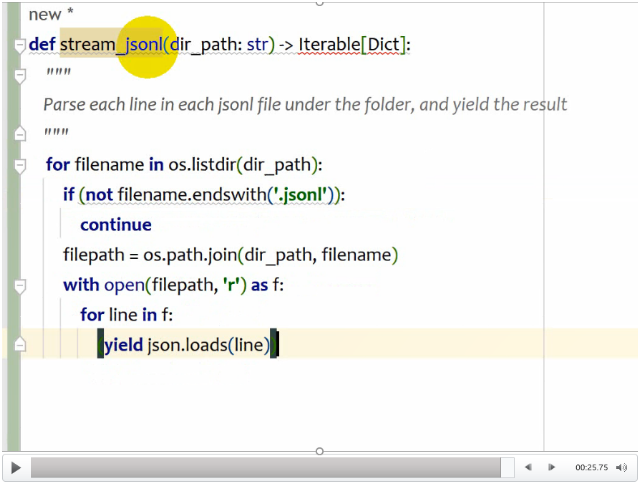


Pangu-Coder: a Function Level Code Generation Model



HUAWEI NOAH'S ARK LAB

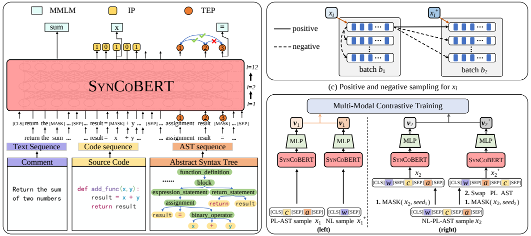
Pangu-Coder: a Function Level Code Generation Model



HUAWEI NOAH'S ARK LAB

SynCoBERT:
Syntax-Guided Multi-Modal Contrastive Pre-Training for Code Representation

- Novel pre-training objectives originating from the symbolic and syntactic properties of source code:
  - Identifier Prediction (IP)
  - AST Edge Prediction (TEP)
- A multi-modal contrastive learning strategy to maximize the mutual information among different modalities.
- Extensive experiments on four downstream tasks: code search, clone detection, code defect detection and code translation.



(c) Positive and negative sampling for x_i

(a) SYNCOBERT pre-training over MMLM, IP and TEP objectives

(b) Multi-modal contrastive pre-training on NL-PL paired data

Figure 3: Different scenes of SYNCOBERT pre-training. (a) SYNCOBERT takes source code paired with comment and the corresponding AST as the input, and is pre-trained with MMLM, IP, TEP objectives. (b) Positive sampling for NL-PL paired data, (left) NL *vs* PL-AST, (right) NL-PL-AST *vs* NL-AST-PL. (c) An illustration about positive and negative pairs, including *in-batch* and *cross-batch* negative sampling.

Table 1: Results on the natural language code search task evaluating with MRR, using the AdvTest and CodeSearch datasets.

Model	AdvTest	CodeSearch						
	Python	Ruby	Javascript	Go	Python	Java	PHP	Average
NBow	-	16.2	15.7	33.0	16.1	17.1	15.2	18.9
CNN	-	27.6	22.4	68.0	24.2	26.3	26.0	32.4
BiRNN	-	21.3	19.3	68.8	29.0	30.4	33.8	33.8
Transformer	-	27.5	28.7	72.3	39.8	40.4	42.6	41.9
RoBERTa	18.3	58.7	51.7	85.0	58.7	59.9	56.0	61.7
RoBERTa (code)	-	62.8	56.2	85.9	61.0	62.0	57.9	64.3
CodeBERT	27.2	67.9	62.0	88.2	67.2	67.6	62.8	69.3
GraphCodeBERT	35.2	70.3	64.4	89.7	69.2	69.1	64.9	71.3
SYNCOBERT	**38.1**	**72.2**	**67.7**	**91.3**	**72.4**	**72.3**	**67.8**	**74.0**

Table 4: Results on the code translation task with BLEU, Accuracy and CodeBLEU score, using the CodeTrans dataset.

Methods	C#→Java			Java→C#		
	BLEU	Exact Match	CodeBLEU	BLEU	Exact Match	CodeBLEU
Naive copy	18.69	0.0	-	18.54	0.0	-
PBSMT	40.06	16.1	43.48	43.53	12.50	42.71
Transformer	50.47	37.90	61.59	55.84	33.00	63.74
RoBERTa (code)	71.99	57.90	80.18	77.46	56.10	83.07
CodeBERT	72.14	58.80	79.41	79.92	59.00	**85.10**
GraphCodeBERT	72.64	58.80	-	80.58	59.40	-
SYNCOBERT	**76.52**	**61.30**	**82.22**	**80.75**	**60.40**	84.85

Preprint: http://arxiv.org/abs/2108.04556

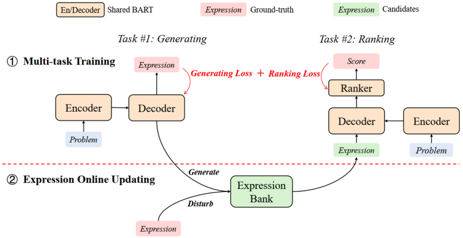
Content

LLM Research in Huawei Noah's Ark Lab

Generate and Rank: A Multi-task Framework for Math Word Problems

Original MWP	
Problem	A project is completed in 25 days by 12 workers. If it takes 20 days to complete, how many workers will it take?
Solution	*25 * 12 / 20*
Number-mapped MWP	
Problem	A project is completed in *NUM0* days by *NUM1* workers. If it takes *NUM2* days to complete, how many workers will it take?
Solution	*NUM0 * NUM1 / NUM2*

- Generator: Finetune BART on MWP seq2seq task
- Ranker: Sequence pair classification task
  - Feed problem into encoder and expression into decoder
- Joint training: Share encoder and decoder



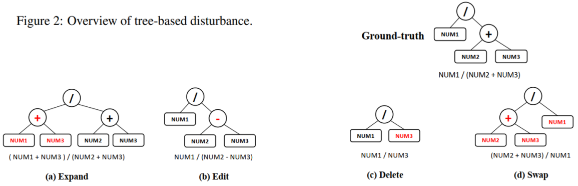


Figure 2: Overview of tree-based disturbance.

Published in Findings of EMNLP 2021: https://aclanthology.org/2021.findings-emnlp.195.pdf

Content

Large Language Models (LLMs): Background

Pangu Models

LLM Research in Huawei Noah's Ark Lab

Future Work

Future Work

-

Content

Large Language Models (LLMs): Background

Pangu Models

LLM Research in Huawei Noah's Ark Lab

Future Work

Summary

Large Language Models (LLMs): Background

Pangu Models

LLM Research in Huawei Noah's Ark Lab

Future 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.

HUAWEI