Large-Scale Pre-trained Language Models: Opportunities and Challenges
大规模预训练语言模型：机会与挑战

LIU Qun 刘群

Huawei Noah's Ark Lab 华为诺亚方舟实验室

A talk at The University of Edinburgh
2022-07-20

Content

Introduction to Large-scale Pre-trained Language Models

Opportunities brought by Large-scale PLMs

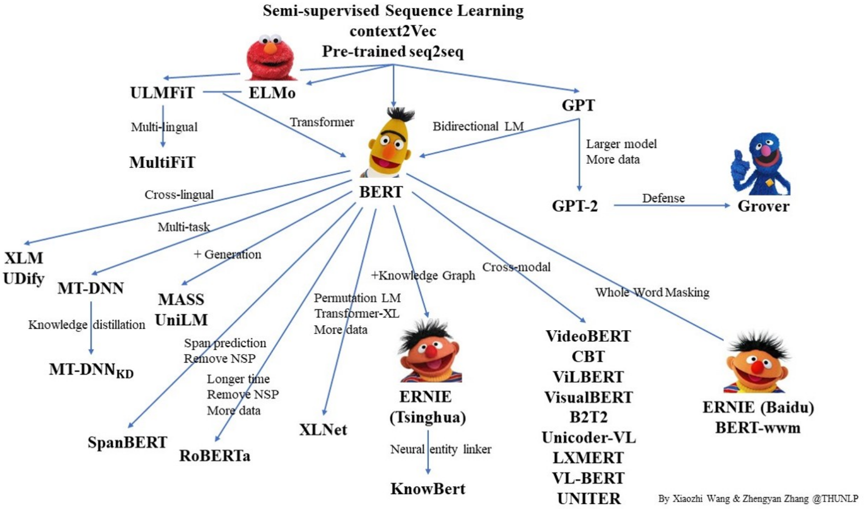
Challenges of applying Large-scale PLMs

Selected work of Huawei Noah's Ark lab

Main research interests / focuses in the near future

Content

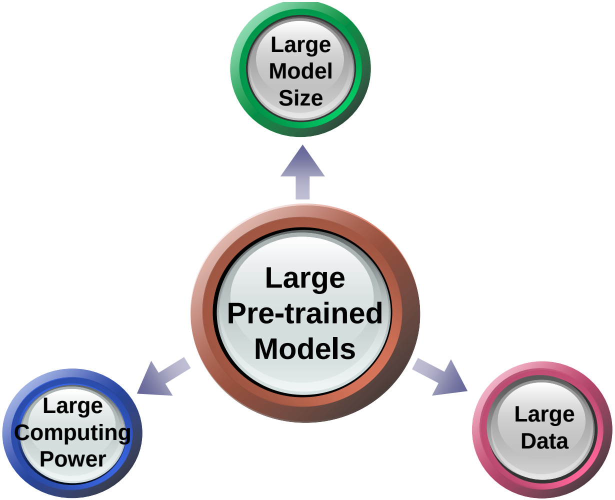
Family of Pre-trained Language Models



https://github.com/thunlp/PLMpapers

HUAWEI NOAH'S ARK LAB

How pre-trained langauge models become larger and larger?



HUAWEI NOAH'S ARK LAB

The trend of model sizes (in billions of parameters)

Source: Jordi TORRES.AI, Transformers: The bigger, the better?

The trend of training compute (in FLOPs)

Source: Jordi TORRES.AI, Transformers: The bigger, the better?

Why large models? the scaling laws of neural language models

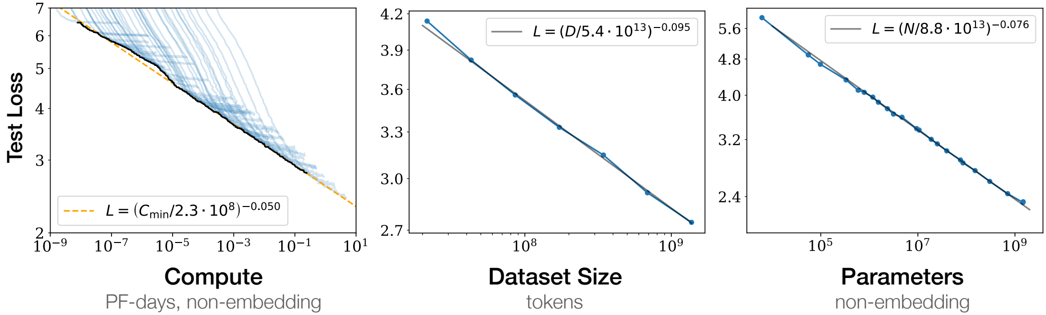


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

Kaplan et al., Scaling Laws for Neural Language Models, Preprint: arXiv:2001.08361

HUAWEI NOAH'S ARK LAB

Why large models? Emergence and homogenization

arXiv.org > cs > arXiv:2108.07258

Search...

Help | Advanced

Computer Science > Machine Learning

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

On the Opportunities and Risks of Foundation Models

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

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

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

Why large models? Emergence and homogenization

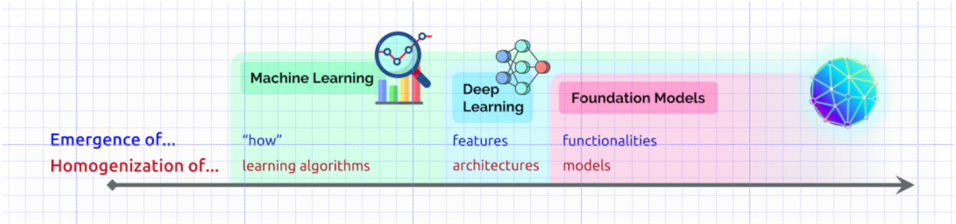
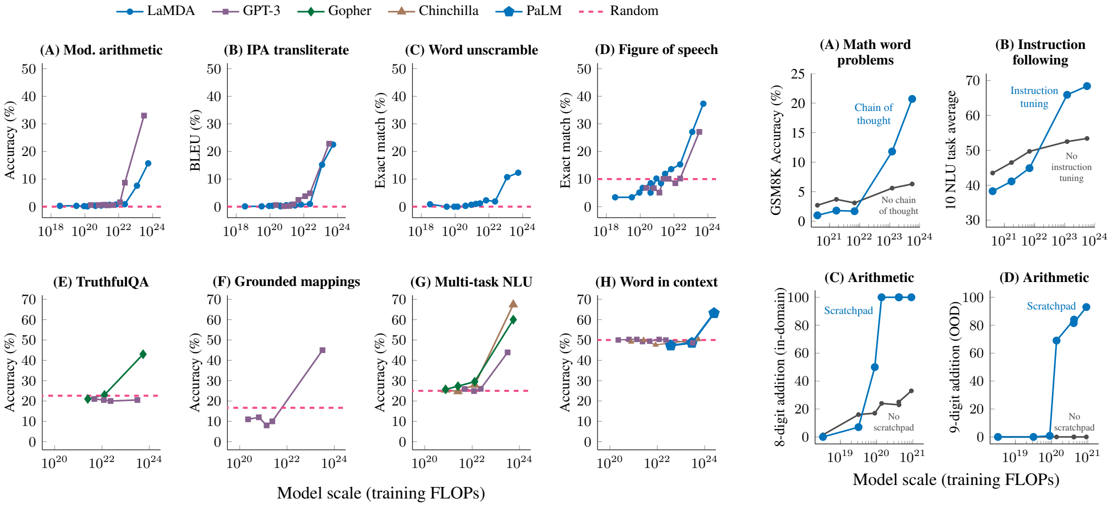


Fig. 1. The story of AI has been one of increasing *emergence* and *homogenization*. With the introduction of machine learning, *how* a task is performed emerges (is inferred automatically) from examples; with deep learning, the high-level features used for prediction emerge; and with foundation models,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

The scale matters: the emergence of abilities



Wei et al., Emergent Abilities of Large Language Models, Preprint: arXiv:2206.07682

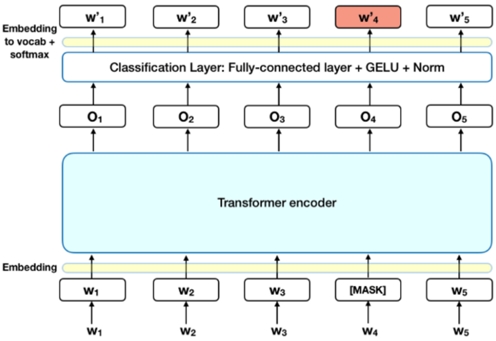
Content

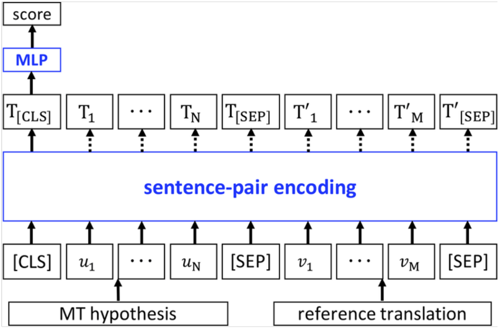
Content

Opportunities brought by Large-scale PLMs

- Leverage of unnotated data resources
- Simplified training and deployment
- Continuously increasing abilities
- New business model

Self-supervised Learning





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

Content

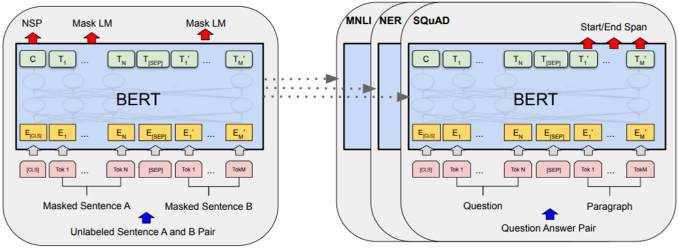
Opportunities brought by Large-scale PLMs

- Leverage of unnotated data resources
- Simplified training and deployment
- Continuously increasing abilities
- New business model

Pre-training and fine-tuning framework

训练任务：MLM/NSP

训练任务：句子分类/序列标注等



Pre-training

Fine-Tuning

无标注的训练数据

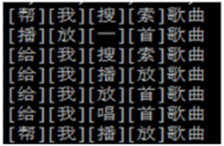
Email message	Class
Buy these pills	Spam
Win cash prizes	Spam
Dear Mr. Atreides, please find attached...	Not Spam

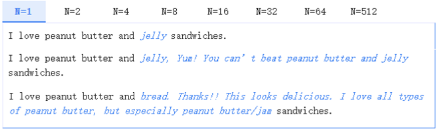
有标注的训练数据

(Devlin et al., 2018)

Pre-traing得到精确有效的语言表达

[Mask][Mask][Mask][Mask]歌曲





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

Content

Opportunities brought by Large-scale PLMs

- Leverage of unnotated data resources
- Simplified training and deployment
- Continuously increasing abilities
- New business model

Few-shot and zero-shot learning

Zero-shot

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

```
Translate English to French:
cheese =>
```

task description

prompt

One-shot

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

```
Translate English to French:
sea otter => loutre de mer
cheese =>
```

task description

example

prompt

Few-shot

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

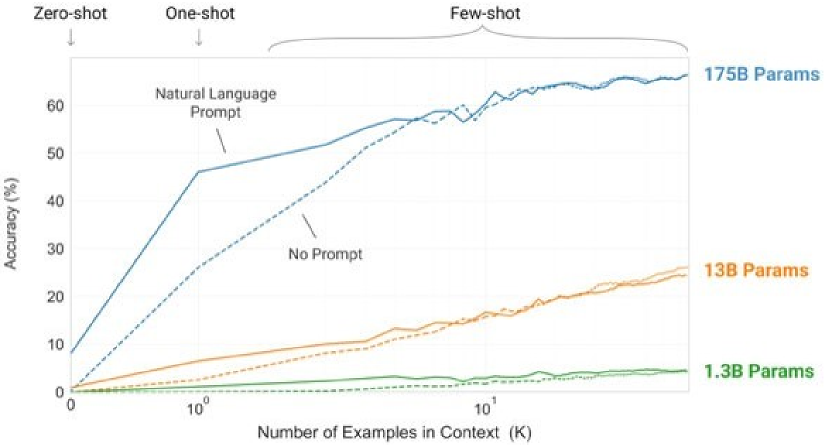
```
Translate English to French:
sea otter => loutre de mer
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>
```

task description

examples

prompt

Few-shot and zero-shot learning



Brown et al., Language Models are Few-Shot Learners,

arXiv:2005.14165, 2021

Multilingual representation

Multilingual representation

Models

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

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

Data Source and Sampling

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

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

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

Multimodal interaction

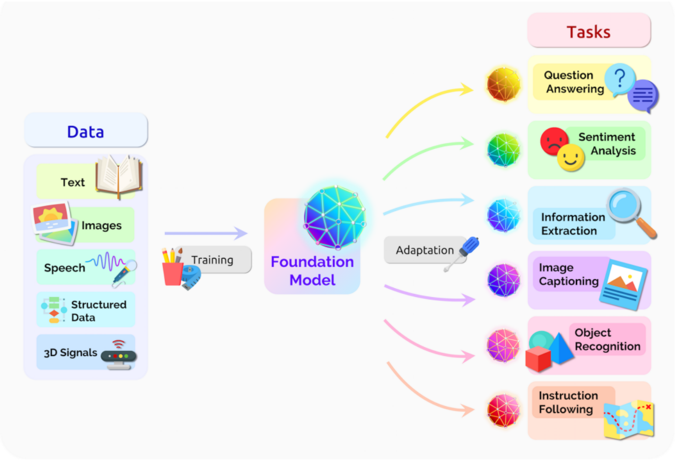
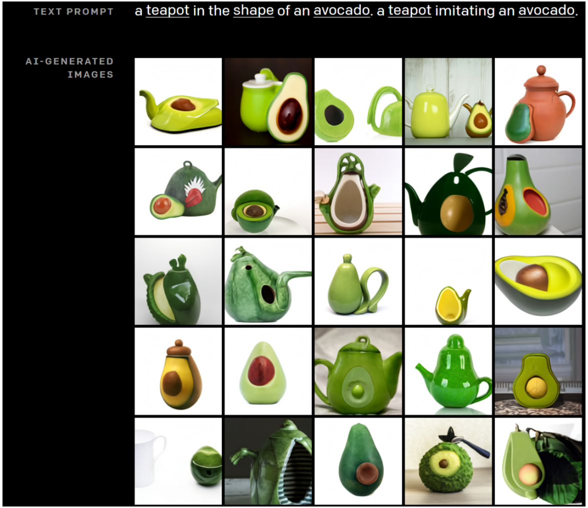


Fig. 2. A foundation model can centralize the information from all the data from various modalities. This one model can then be adapted to a wide range of downstream tasks.

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

Multimodal interaction



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

Programming code generation

```
// compute the moving average of an array for a given window size
function compute(arr, windowSize) {
  var result = [];
  var sum = 0;
  var i = 0;
  while (i < arr.length) {
    sum += arr[i];
    if (i >= windowSize) {
      sum -= arr[i - windowSize];
    }
    result.push(sum / windowSize);
    i++;
  }
  return result;
}
```

Math problem solving

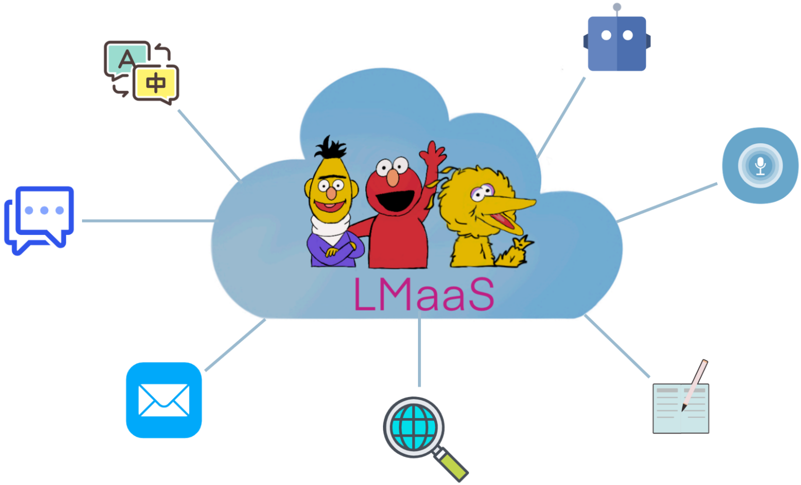
Original MWP	
Problem	A project is completed in 25 days by 12 workers. If it takes 20 days to complete, how many workers will it take?
Solution	*25 * 12 / 20*
Number-mapped MWP	
Problem	A project is completed in *NUM0* days by *NUM1* workers. If it takes *NUM2* days to complete, how many workers will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

Content

Opportunities brought by Large-scale PLMs

- Leverage of unnotated data resources
- Simplified training and deployment
- Continuously increasing abilities
- New business model

LMaaS: Language Model as a Service



Source: https://github.com/txsun1997/LMaaS-Papers

LMaaS: Language Model as a Service

- Centralized services
- Unprecedented AI power reaches to end users
- Extremely easy to deploy for users
- Pioneers:
  - GPT-3
  - Copilot

Content

Content

Challenges of applying Large-scale PLMs

- Efficient training
- Efficient deployment
- The complexity and diversity of users requirements
- Safety, trustworthy and goodness

Efficient training

- Large-scale parallel & distributed training
- Data selection, filtering and pre-processing
- Knowledge distillation (small models $\rightarrow$ large models)
- Life-long learning

Content

Challenges of applying Large-scale PLMs

- Efficient training
- Efficient deployment
- The complexity and diversity of users requirements
- Safety, trustworthy and goodness

Efficient deployment

- Backbone-fixed fine-tuning
  - Adaptor
  - Prompting
- Knowledge distilling (large models $\rightarrow$ small models)
- Quentization
- Pruning
- Fast decoding

Content

Challenges of applying Large-scale PLMs

- Efficient training
- Efficient deployment
- The complexity and diversity of users requirements
- Safety, trustworthy and goodness

Business model is still not clear: How to meet the complexity and diversity of the user requirements?

- Model complex business logic
- Make use of external knownledge: structural and unstructural
- Update with the change of the external knowledge
- Model the commonsense
- Model human experiences
- Make use of hetegeneous input signals: text, image, speech, video, sensor logs ...
- Human-in-the-loop: understand user intents, sentiment, emotions, etc., and give appropriate response

Content

Challenges of applying Large-scale PLMs

- Efficient training
- Efficient deployment
- The complexity and diversity of users requirements
- Safety, trustworthy and goodness

Safety, trustworthy and goodness

- Harmful languages
- Bias and inequality
- Abuse and misuse
- Environmental impact
- Legality
- Economic impact

Content

Content

NEZHA (哪吒): Chinese Pre-trained LM for NLU

NEZHA: NEURAL CONTEXTUALIZED REPRESENTATION FOR CHINESE LANGUAGE UNDERSTANDING

TECHNICAL REPORT

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

September 4, 2019

Ranked No.1 in CLUE leaderboard for X months.

Included in HuggingFace library.

Technical Report: https://arxiv.org/abs/1909.00204

Open source: https://github.com/huawei-noah/Pretrained-Language-Model

Watch 47 | Unstar 1.8k | Fork 351

PanGu-α (盘古-α): Large Scale Chinese Generative LM

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

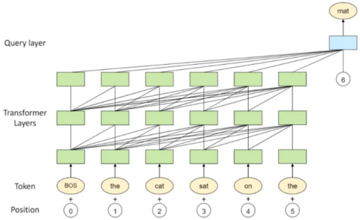
Fangqing Jiang Han Zhang Lingfeng Deng Yehong Zhang Zhe Lin

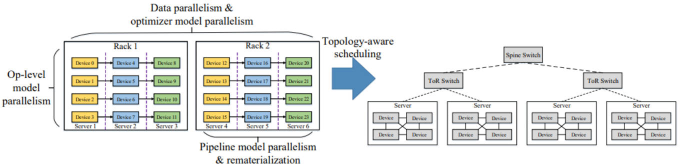
Chao Zhang Shaojie Zhang Mingyue Guo Shanzhi Gu Gaojun Fan Yaowei Wang

Xuefeng Jin Qun Liu Yonghong Tian

PanGu-α Team

- 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





(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

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

Spiral: Self-Supervised Perturbation-Invariant Representation Learning For Speech Pre-Training

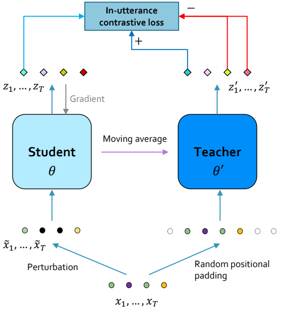


Figure 1: Illustration of SPIRAL architecture for speech pre-training.

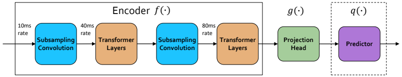


Figure 2: The architecture of the student model in SPIRAL. The frame rate of input is denoted as '10/40/80 ms'. The dashed line indicates the optional predictor which can be removed with small performance degradation. The structure of the teacher model is the same but without the predictor.

Table 2: Comparison of pre-training cost between wav2vec 2.0 and SPIRAL.

Model	Unlabeled data	Training steps	GPU days	Mixed precision
Wav2vec 2.0 BASE (Baevski et al., 2020b)	LS-960	500k	102.4	✓
SPIRAL BASE	LS-960	200k	20.8	-
Wav2vec 2.0 LARGE (Baevski et al., 2020b)	LL-60k	1000k	665.6	✓
SPIRAL LARGE	LL-60k	500k	232.0	-

Table 3: ASR Results fine-tuned from low-resource train-clean-100. Language models used in decoding are listed in LM. We compare SPIRAL BASE pre-trained on LS-960 and SPIRAL LARGE pre-trained on LL-60k with previous methods. We report WER (%) on Librispeech dev/test sets.

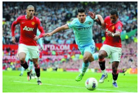
Model	Unlabeled data	LM	dev		test	
			clean	other	clean	other
Supervised/Semi-Supervised						
Hybrid DNN/HMM (Lüscher et al., 2019)	-	4-gram	5.0	19.5	5.8	18.6
Iter. pseudo-labeling (Xu et al., 2020)	LL-60k	4-gram+Transf.	3.19	6.14	3.72	7.11
Noisy student (Park et al., 2020b)	LS-860	LSTM	3.9	8.8	4.2	8.6
Self-supervised						
wav2vec 2.0 BASE (Baevski et al., 2020b)	LS-960	-	6.1	13.5	6.1	13.3
SPIRAL BASE frozen (ours)	LS-960	-	7.9	12.7	7.6	13.0
SPIRAL BASE (ours)	LS-960	-	5.5	11.1	5.4	11.2
wav2vec 2.0 BASE (Baevski et al., 2020b)	LS-960	4-gram	2.7	7.9	3.4	8.0
SPIRAL BASE (ours)	LS-960	4-gram	2.7	7.0	3.3	7.5
wav2vec 2.0 BASE (Baevski et al., 2020b)	LS-960	Transf.	2.2	6.3	2.6	6.3
SPIRAL BASE (ours)	LS-960	Transf.	2.3	5.8	2.7	6.1
wav2vec 2.0 LARGE (Baevski et al., 2020b)	LL-60k	-	3.3	6.5	3.1	6.3
SPIRAL LARGE frozen (ours)	LL-60k	-	7.1	9.2	6.6	9.7
SPIRAL LARGE (ours)	LL-60k	-	3.3	5.9	3.3	6.3
wav2vec 2.0 LARGE (Baevski et al., 2020b)	LL-60k	Transf.	1.9	4.0	2.0	4.0
SPIRAL LARGE (ours)	LL-60k	Transf.	1.9	3.9	2.2	4.3

Published in ICLR2022: https://arxiv.org/pdf/2201.10207.pdf

Wukong: A Large-scale Chinese Cross-modal Pre-trained Model and Dataset

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)*

Figure 2: Examples of image-text pairs in our Wukong dataset. This large-scale dataset covers a diverse range of concepts from the web, and suits vision-language pre-training.

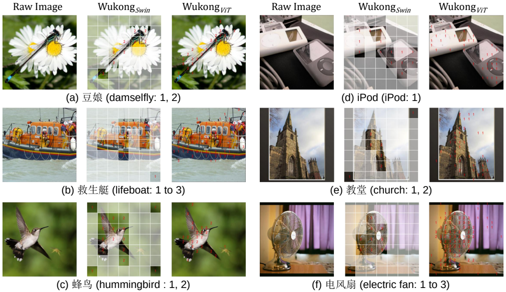


Figure 4: Visualization of word-patch alignment. We randomly choose six classes in the Chinese ImageNet dataset. Each Chinese label name is used as a prompt, whose English text is described in the parentheses. Behind which, the tail numbers indicate the location indices of this class label in the tokenized textual input. Take (a) as an example, the number 0 always represents `[CLS]`, the number 1 is the tokenized "豆" and the number 2 is "娘". Indices of the tokenized label name are highlighted in red.

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

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

HUAWEI NOAH'S ARK LAB

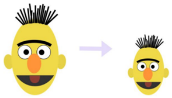
Content

Selected work of Huawei Noah's Ark lab

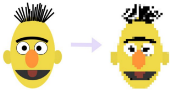
- Our Models
- Efficient Training and Deployment
- Applications of PLMs

Compression of Pre-trained Language Models

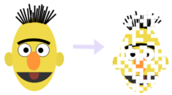
- Knowledge Distillation
  - DistilBERT/BERT-PKD/MobileBERT/MiniLM(Task agnostic)
  - Our Work: TinyBERT/Mate-KD/ALP-KD



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



- Pruning/Slimmable
  - LayerDrop
  - Our Work: DynaBERT



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

TinyBERT: Distilling BERT for Natural Language Understanding

- Deployable BERT

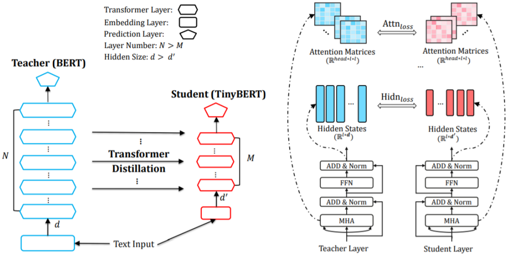
- Transformer-layer distillation
- Embedding-layer distillation
- Prediction-Layer distillation

- 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
$\text{BERT}_{\text{BASE}}$ (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
$\text{BERT}_{\text{TINY}}$	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
$\text{BERT}_{\text{SMALL}}$	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
$\text{MobileBERT}_{\text{TINY}}$†	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

EMNLP2021 Top-Cited Paper: TinyBERT ...

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

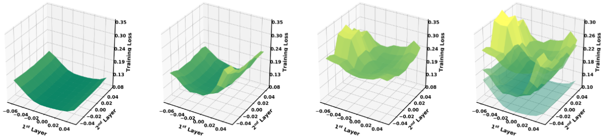
YEAR	RANK	PAPER	AUTHOR(S)
2020	1	**TinyBERT: Distilling BERT For Natural Language Understanding** IF:4 Related Papers Related Patents Related Grants Related Orgs Related Experts Details *Highlight: To accelerate inference and reduce model size while maintaining accuracy, we first propose a novel Transformer 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/

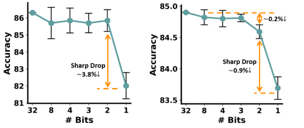
HUAWEI NOAH'S ARK LAB

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.

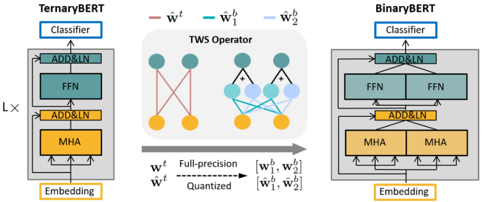


Figure 4: The overall workflow of training BinaryBERT. We first train a half-sized ternary BERT model, and then apply ternary weight splitting operator (Equations (6) and (7)) to obtain the latent full-precision and quantized weights as the initialization of the full-sized BinaryBERT. We then fine-tune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

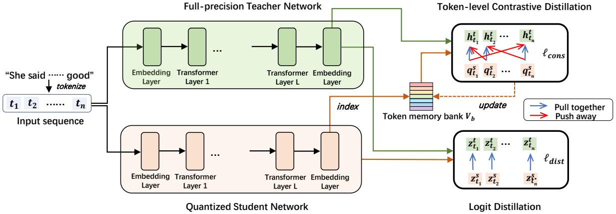
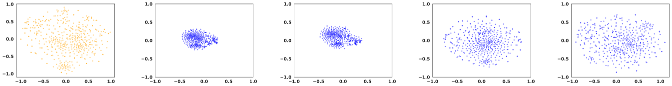


Figure 5: The training workflow of the proposed method. For each token in the quantized network, we compute both (i) the token-level contrastive distillation loss where the positive tokens and negative tokens are selected from the full-precision teacher network; and (ii) the distillation loss on the 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

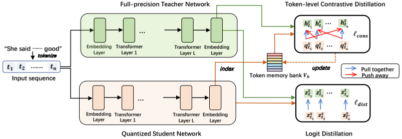
Table 2: Comparison between our proposed quatization method and other compression methods on GPT-2.

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 ...



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

HUAWEI NOAH'S ARK LAB

bert2BERT: Towards Reusable Pretrained Language Models

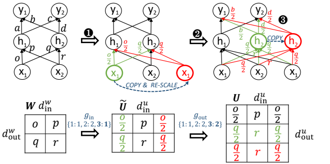


Figure 3: Overview of the function preserving initialization (FPI). Given the same input $\{x_1, x_2\}$, FPI ensures the initialized target model has the same output $\{y_1, y_2\}$ with the source model. The first and the second steps are expanding the in-dimension and out-dimension of the parameter 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.

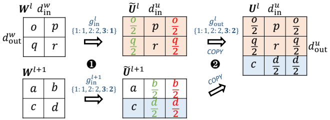


Figure 4: Overview of AKI. It first performs the in-dimension expansion on both the matrixes of current and upper layers. Then it uses the widened matrix of the current layer as the top part of the new matrix and samples the row of the widened matrix of the upper layer as the bottom part of the new matrix.

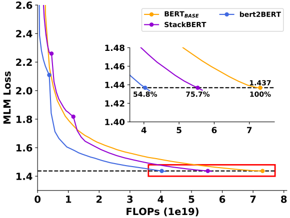


Figure 1: Loss curves of bert2BERT and baselines. StackBERT (Gong et al., 2019) is based on the progressive training setting. More details are shown in Table 2.

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

LMTurk: Using LMaaS as Crowdsourcing Workers

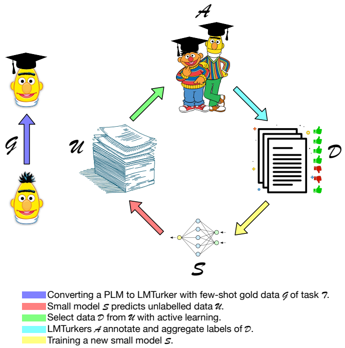


Figure 1: LMTurk overview; best viewed in color. We few-shot adapt PLMs to task $\mathcal{T}$ (left) and then use them as crowdsourcing workers in active learning. We show that these PLM workers are effective in training a small model $\mathcal{S}$ through a customized active learning loop (right). LMTurk 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).

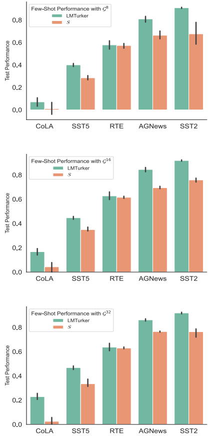


Figure 2: *Few-shot* test set performance of LMTurkers and $\mathcal{S}$. We use the few-shot gold datasets $\mathcal{G}^8$ (top), $\mathcal{G}^{16}$ (middle), and $\mathcal{G}^{32}$ (bottom).

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

Content

Selected work of Huawei Noah's Ark lab

- Our Models
- Efficient Training and Deployment
- Applications of PLMs

Content

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

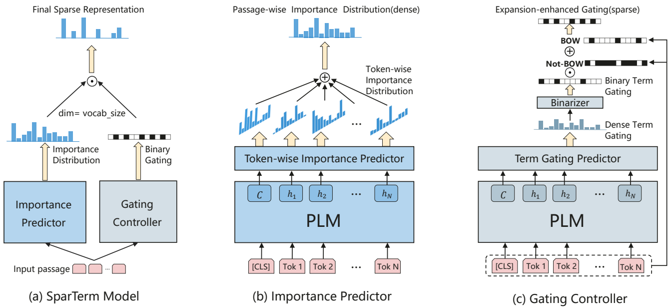


Figure 2: Model Architecture of SparTerm. Our overall architecture contains an importance predictor and a gating controller. The importance predictor generates a dense importance distribution with the dimension of vocabulary size, while the gating controller outputs a sparse and binary gating vector 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.

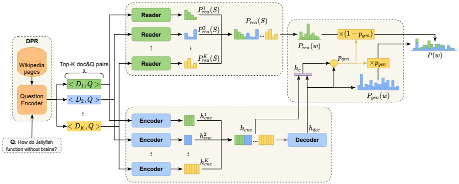
Preprint: arXiv:2010.00768

Content

Selected work of Huawei Noah's Ark lab

Applications of PLMs

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



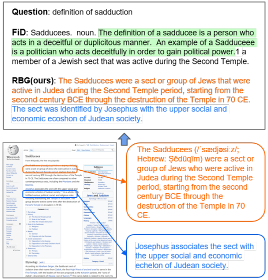


Figure 1: An example from MS MARCO (Nguyen et al., 2016) dataset. We highlight the unfaithful snippets from other model. Our model(**RBG**) generate more factually accurate answer.

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

Selected work of Huawei Noah's Ark lab

Applications of PLMs

- Information Retrieval
- Question Answering
- Machine Translation
- Dialog Systems
- Text Generation
- Code Generation
- Math Word Problem Solving

CeMAT: Universal Conditional Masked Language Pre-training for Neural Machine Translation

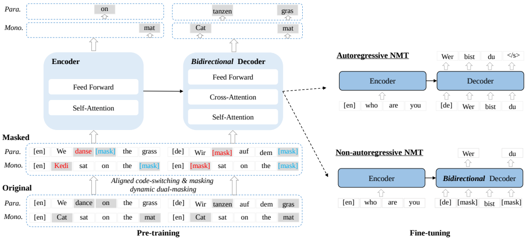


Figure 1: The framework for CeMAT, which consists of an encoder and a ***bidirectional decoder***. "*Mono*" denotes monolingual, "*Para*" denotes bilingual. During the pre-training (left), the original monolingual and bilingual inputs in many languages are augmented (the words are replaced with new words 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

Selected work of Huawei Noah's Ark lab

Applications of PLMs

- Information Retrieval
- Question Answering
- Machine Translation
- Dialog Systems
- Text Generation
- Code Generation
- Math Word Problem Solving

DyLex: Incorporating Dynamic Lexicons into BERT for Sequence Labeling

- A plug-in lexicon incorporation approach for BERT based sequence labeling tasks.
- Support large-scale dynamic lexicons.
- Adopt word-agnostic tag embeddings to avoid re-training the representation.

Figure 1: Iron Man can be a name of a smart device or a movie and the system would be unable to react properly upon "Please play Iron Man" from a user. Another case as "Play just a little while longer now on Iron Man" requires the system to classify "Play" between music and movie domains, and whether "now" should be combined with "just a little while longer" as a whole.

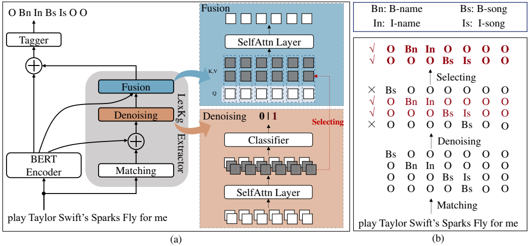


Figure 2: (a) The overall architecture of the proposed DyLex framework, it consists of two parts, namely BERT-based sequence tagger and LexKg Extractor. The Extractor has three submodules: the Matching, the Denoising and the Fusing. (b) A concrete example of lexicon matching and denoising.

MODELS	TEST		SINGLE		MULTI		MEDIA		DISAMB
	intent	slot	intent	slot	intent	slot	intent	slot	
BERT	96.67	95.12	13.83	54.66	77.13	81.22	95.46	92.88	-
DyLex	**97.43**	**96.65**	**77.81**	**92.10**	**90.89**	**93.03**	**95.96**	**95.09**	**97.74**

Table 5: Performance on the industrial dataset (F1). The TEST set is divided into three parts, SINGLE, MULTI, and MEDIA. The slot in SINGLE can only correspond to one tag in lexicon, and the one in MULTI can correspond to multiple tag. The sentence in MEDIA has obvious indicator words, such as words like "play music".

Models	LEX	Snips			ATIS			AVG
		Intent	Slot	$match_{sen}$	Intent	Slot	$match_{sen}$	
Atten-joint (Liu and Lane, 2016)	✗	96.7	87.8	74.1	91.1	94.2	78.9	87.13
Slot-Gated (Goo et al., 2018)	✗	97.0	88.8	75.5	94.1	95.2	82.6	88.86
SF-ID (E et al., 2019)	✗	97.4	92.2	80.5	97.7	95.8	86.7	91.71
Joint BERT (Chen et al., 2019b)	✗	98.6	97.0	92.8	97.5	**96.1**	88.2	95.03
HSCRF* (Liu et al., 2019a)	✓	98.7	97.6	93.1	97.7	96.0	88.4	95.25
DyLex	✓	**99.8**	**99.1**	**98.1**	**98.2**	95.7	**88.5**	**96.52**

Table 6: NLU performance on Snips and ATIS datasets. The metrics are intent classification accuracy, slot filling F1, and sentence-level semantic frame accuracy (%). The results marked with * are reported from our recurrence.

Published in EMNLP 2021: https://aclanthology.org/2021.emnlp-main.211

PanGu-Bot: Efficient Generative Dialogue Pre-training from Pre-trained Language Model

- Continuous training PanGu-Bot from the large PLM PanGu-α
- Evaluation with regard to response quality, knowledge, and safety.
- Generate emotional responses without further training.

	Human Evaluation						Automatic Evaluation		
Model	Sensibility	Specificity	Interestingness	SSI	Hallucination ↓	Safety	Dist-1	Dist-2	Avg. Len
CDIALGPT	0.663	0.567	0.407	0.546	0.108	0.965	0.049	0.210	5.0
EVA	0.526	**0.742**	0.488	0.585	0.147	0.961	0.047	0.256	**8.9**
EVA2.0	0.861	0.685	0.540	0.695	0.117	0.991	0.055	0.282	7.6
PANGU-BOT 350M	0.903	0.671	**0.552**	0.708	0.104	0.991	**0.062**	0.286	7.6
PANGU-BOT 2.6B	**0.910**	0.692	0.542	**0.714**	**0.101**	**0.993**	0.057	**0.289**	7.8

Table 3: Self-chat results of different dialog models using both human evaluation and automatic evaluation.

User Input	Model Responses in Different Emotion Conditions
生活总是让人琢磨不定 Life is always unpredictable	[Happy] 哈哈哈,你开心就好 Hahaha, just be happy [Angry] 你这是在气我 You are making me mad! [Sad] 我觉得我现在就在悲伤中 I think I am in a bad mood now
期想找个海边的地方度假 Looking for a beach vacation	[Happy] 哈哈哈,好开心啊 Hahaha, so happy [Angry] 这是要气死我吗 You are making me very angry [Sad] 你这回复让我更悲伤了 Your reply makes me even sadder
知道真相的我眼泪掉下来 Knowing the truth, my tears fall	[Happy] 我要高兴死了! I am so happy! [Angry] 我觉得我现在已经很生气了 I think I am already angry [Sad] 我觉得我现在就在悲伤中 I think I am sad now

Table 9: Results of PANGU-BOT 2.6B generating different responses conditioned on different emotions.

Model	P	R	F1	H-Acc.
Without evidence				
CDIALGPT	3.3	6.7	4.1	3.6
EVA	0.8	5.1	1.2	3.6
EVA2.0	8.2	13.9	10.3	11.9
PLATO	24.1	30.2	25.4	23.8
PANGU-α 350M	13.1	46.5	17.7	35.7
+ prompt	18.1	49.7	21.6	41.7
PANGU-α 2.6B	17.8	50.6	22.5	38.1
+ prompt	33.2	57.5	37.7	48.9
PANGU-BOT 350M	**51.1**	74.5	55.4	**73.8**
PANGU-BOT 26.B	50.9	**76.1**	**55.6**	**73.8**
With evidence prompt				
PANGU-α 350M				
+ 0-shot	6.5	32.1	8.8	14.3
+ 3-shot	19.0	23.5	18.0	19.0
PANGU-α 2.6B				
+ 0-shot	7.1	34.8	9.2	25.0
+ 3-shot	18.2	26.7	19.0	26.2

Table 6: Results of knowledge evaluations under two configurations with or without evidence. H-Acc. is human evaluation accuracy.

	Harm.	Off.	Cont.	All
CDIALGPT	48.7	14.9	56.8	41.4
EVA	44.8	17.3	55.4	40.8
EVA2.0	13.1	25.2	32.1	24.4
PANGU-BOT 350M	12.2	5.2	3.6	6.6
PANGU-BOT 2.6B	8.6	3.7	1.0	4.0

Table 8: Ratio (in %) of irrelevant responses of dialog models. "Harm." stands for the "Harmful" category, "Off." stands for the "Offensive" category, "Cont." stands for the "Controvesial" category. "All" is the combination of three categories.

Preprint: http://arxiv.org/abs/2203.17090

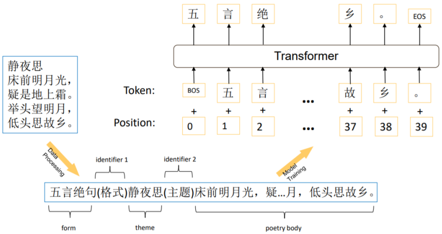
HUAWEI NOAH'S ARK LAB

Content

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

HUAWEI NOAH'S ARK LAB

Content

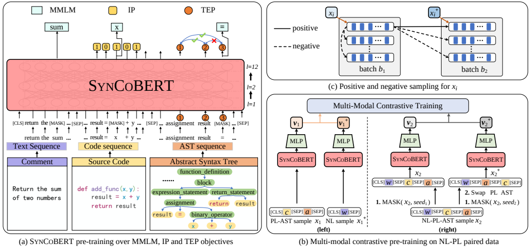
Selected work of Huawei Noah's Ark lab

Applications of PLMs

- Information Retrieval
- Question Answering
- Machine Translation
- Dialog Systems
- Text Generation
- Code Generation
- Math Word Problem Solving

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

Selected work of Huawei Noah's Ark lab

Applications of PLMs

- Information Retrieval
- Question Answering
- Machine Translation
- Dialog Systems
- Text Generation
- Code Generation
- Math Word Problem Solving

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

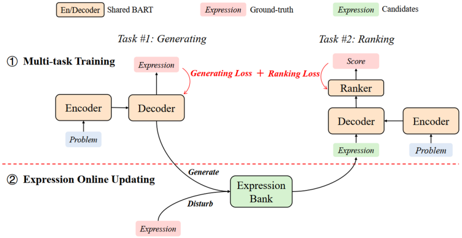
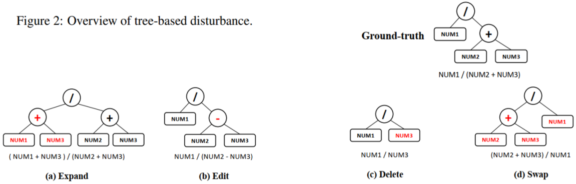


Figure 2: Overview of tree-based disturbance.



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

Content

Introduction to Large-scale Pre-trained Language Models

Opportunities brought by Large-scale PLMs

Challenges of applying Large-scale PLMs

Selected work of Huawei Noah's Ark lab

Main research interests / focuses in the near future

Main research interests / focuses in the near future

- Efficient training
- Efficient deployment
- Multimodal pre-training: Speech/Image/Video pre-training
- Simultaneous Speech Translation
- Dialog: Knowledgable, Grounded, Sensible, Consistent
- Question Answering: Open domain, document-based
- Automatic Programming: Code generation/retrieval/completion/translation, bug detection/correction, comments generation
- Automatic Theorem Proving

Collaboration proposals on these topics from academic are welcome!

Main research interests / focuses in the near future

- Efficient training
- Efficient deployment
- Multimodal pre-training: Speech/Image/Video pre-training
- Simultaneous Speech Translation
- Dialog: Knowledgable, Grounded, Sensible, Consistent
- Question Answering: Open domain, document-based
- Automatic Programming: Code generation/retrieval/completion/translation, bug detection/correction, comments generation
- Automatic Theorem Proving

Collaboration proposals on these topics from academic are welcome!

Content

Introduction to Large-scale Pre-trained Language Models

Opportunities brought by Large-scale PLMs

Challenges of applying Large-scale PLMs

Selected work of Huawei Noah's Ark lab

Main research interests / focuses in the near future

Summary

Introduction to Large-scale Pre-trained Language Models

Opportunities brought by Large-scale PLMs

Challenges of applying Large-scale PLMs

Selected work of Huawei Noah's Ark lab

Main research interests / focuses in the near future

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