基于预训练大模型的多语种机器翻译探讨

刘群 华为诺亚方舟实验室

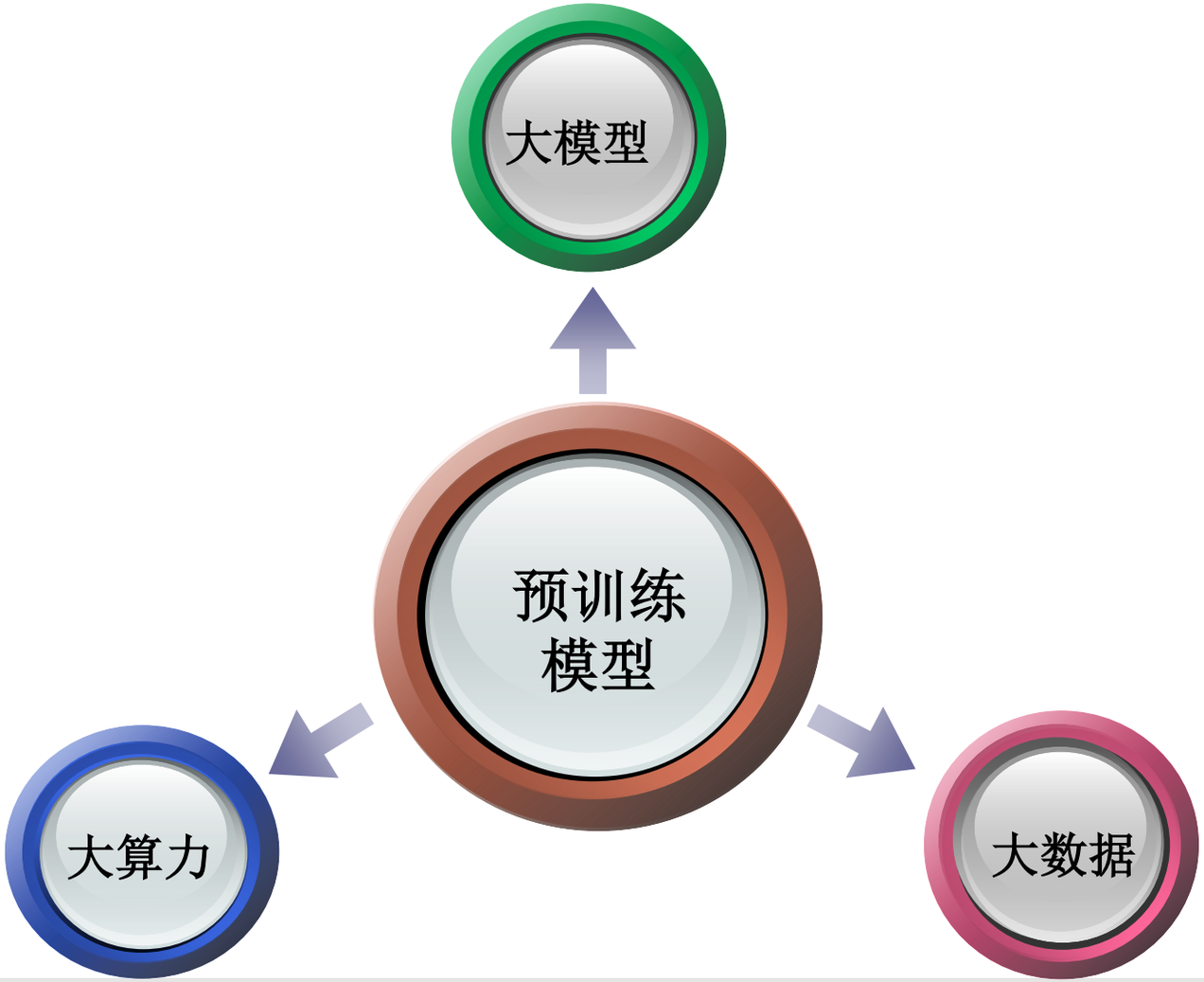
“一带一路”多语言翻译国际大科学计划报告会
2021年6月7日，鹏城实验室

www.huawei.com

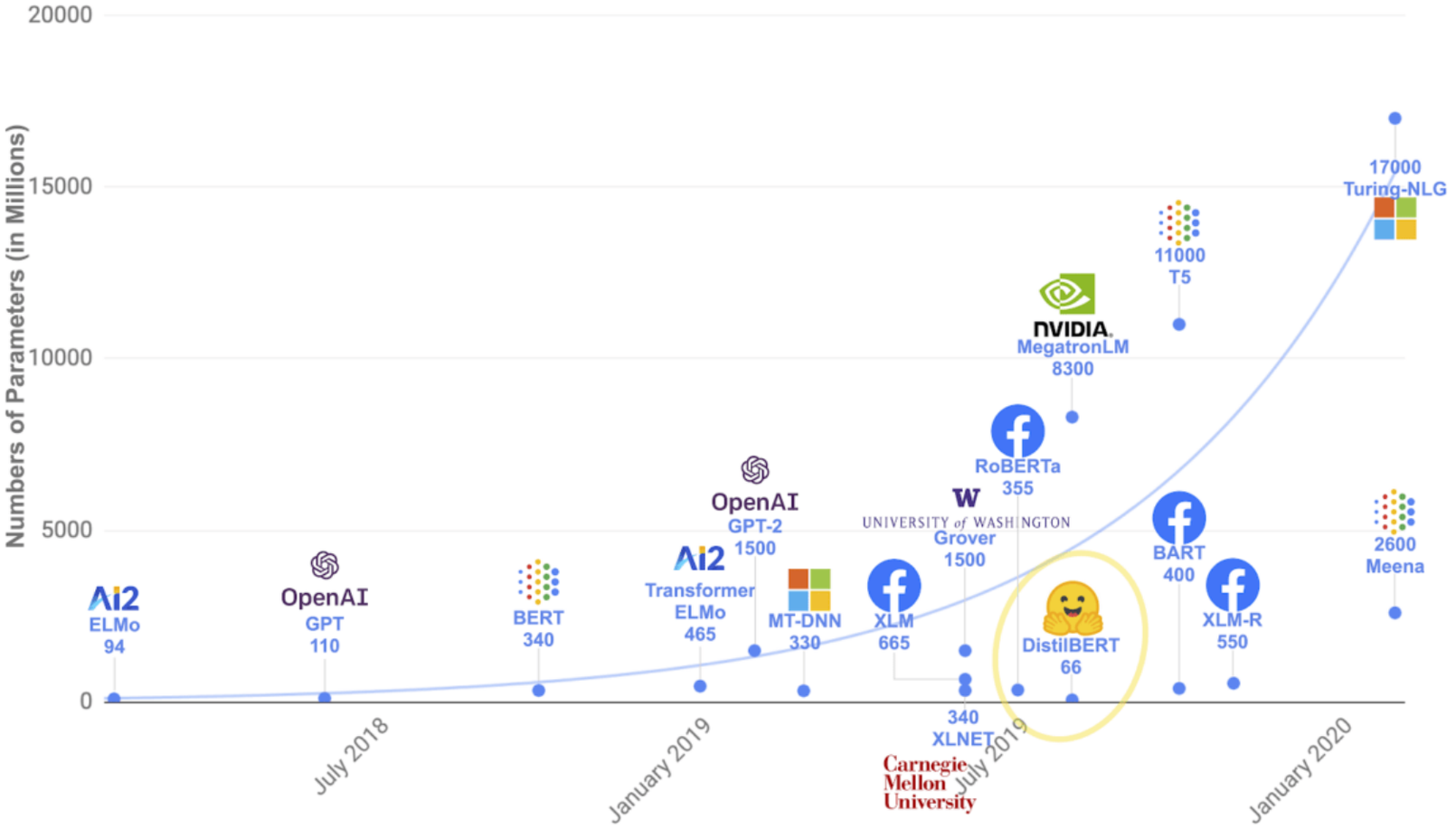
Content

- 预训练大模型的发展趋势
- 多语种机器翻译研究现状
- 基于预训练大模型的多语种机器翻译

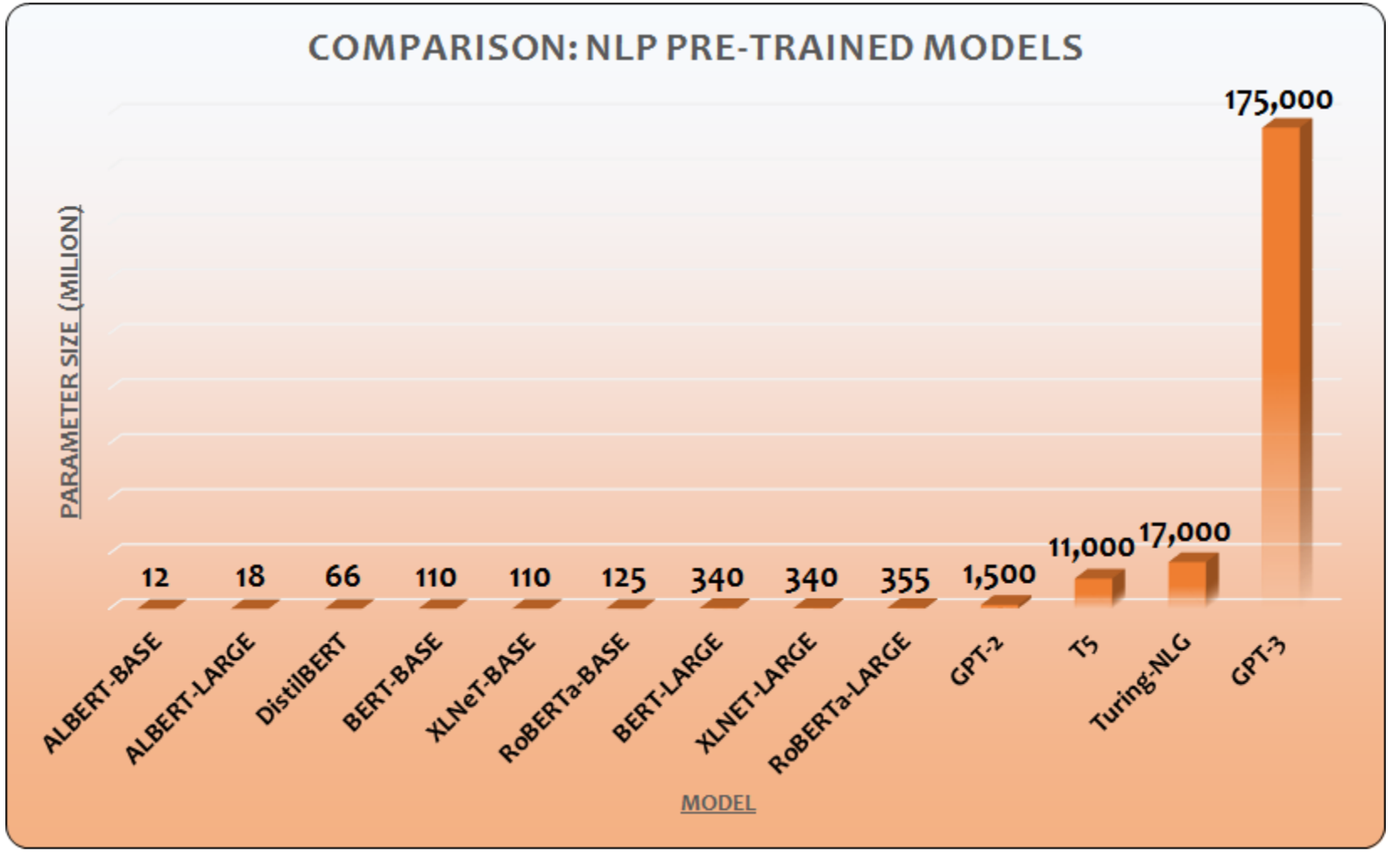
预训练大模型的发展趋势



Pre-trained Models



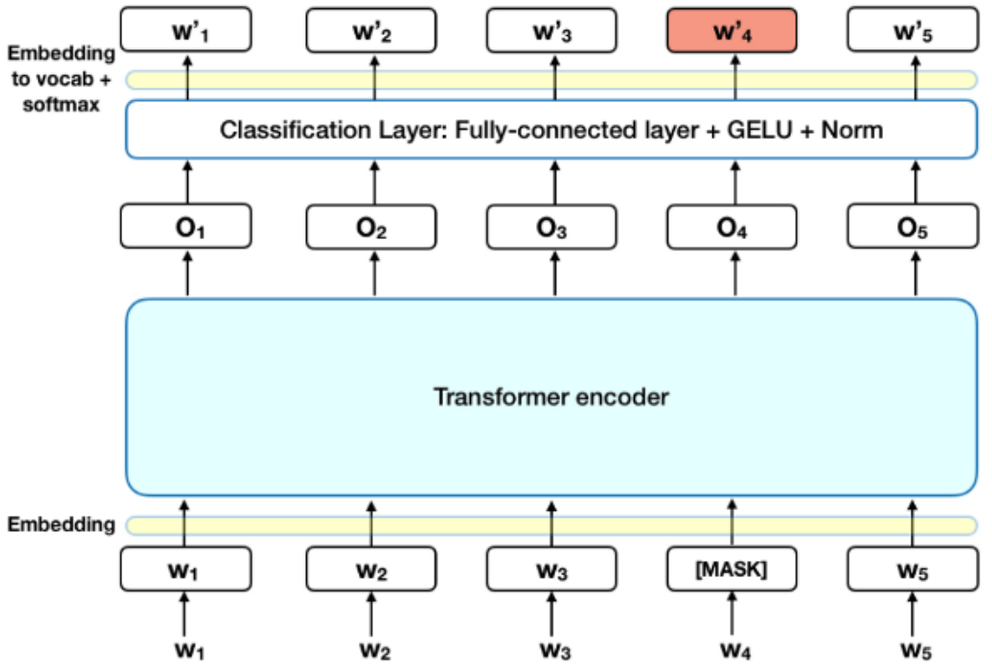
Pre-trained Models

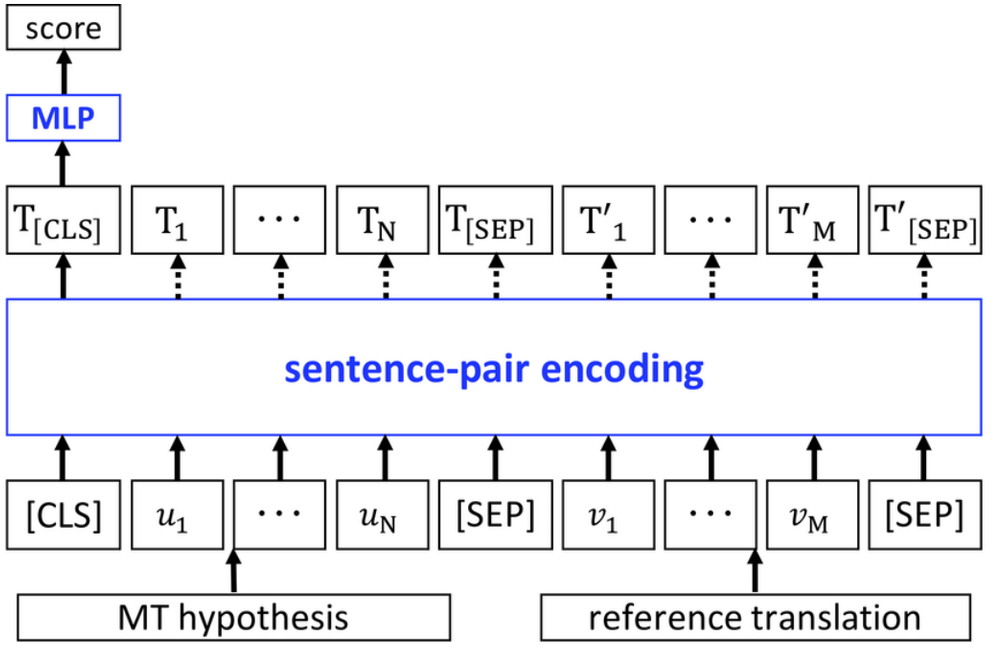


预训练大模型给我们带来了什么？

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

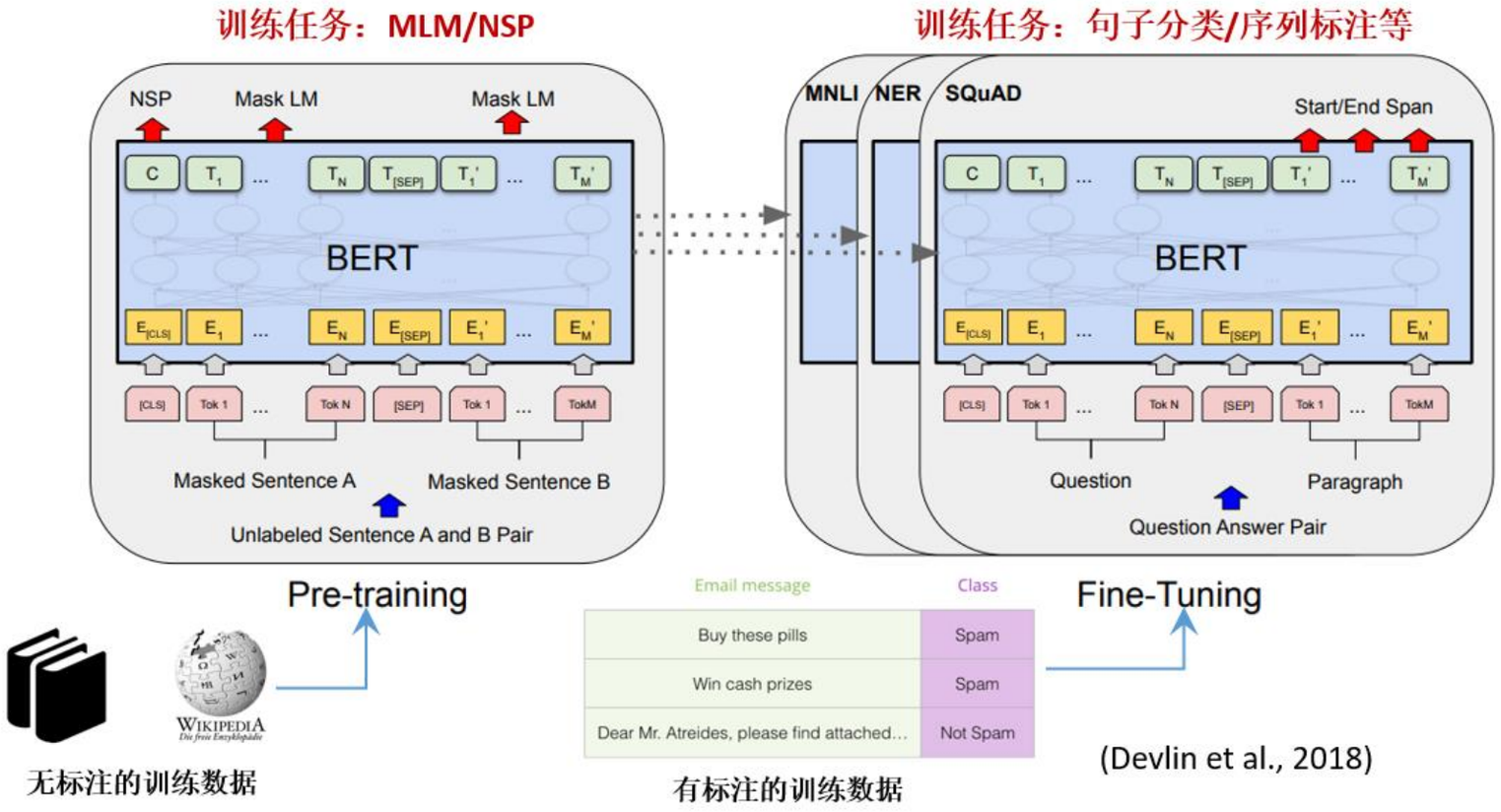
海量无标注或弱标注数据的利用（自监督学习）





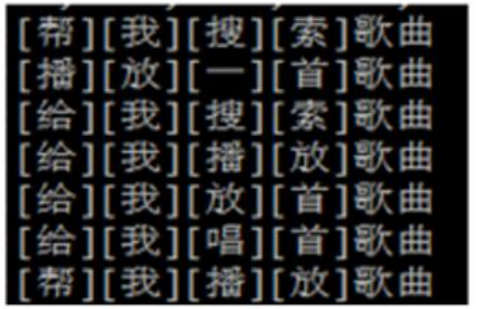
(Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805, 2018)

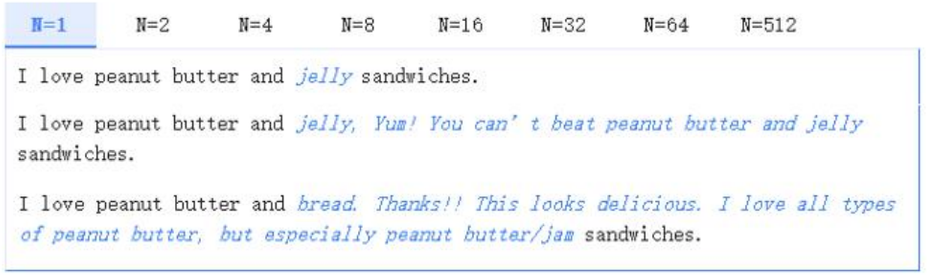
预训练+微调框架：下游任务模型结构的简化 / 性能的普遍提高



Pre-traing得到精确有效的语言表达

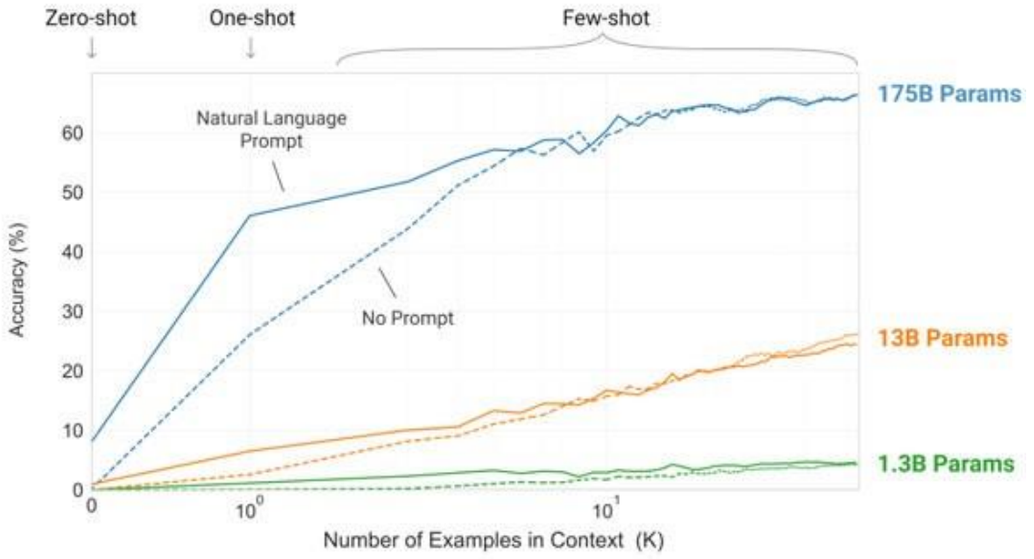
[Mask][Mask][Mask][Mask]歌曲





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

少样本和零样本的学习



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

The three settings we explore for in-context learning

Zero-shot

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

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

One-shot

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

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

Few-shot

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

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

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

Fine-tuning

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

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

多语言表达能力

多语言表达能力

Models

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

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

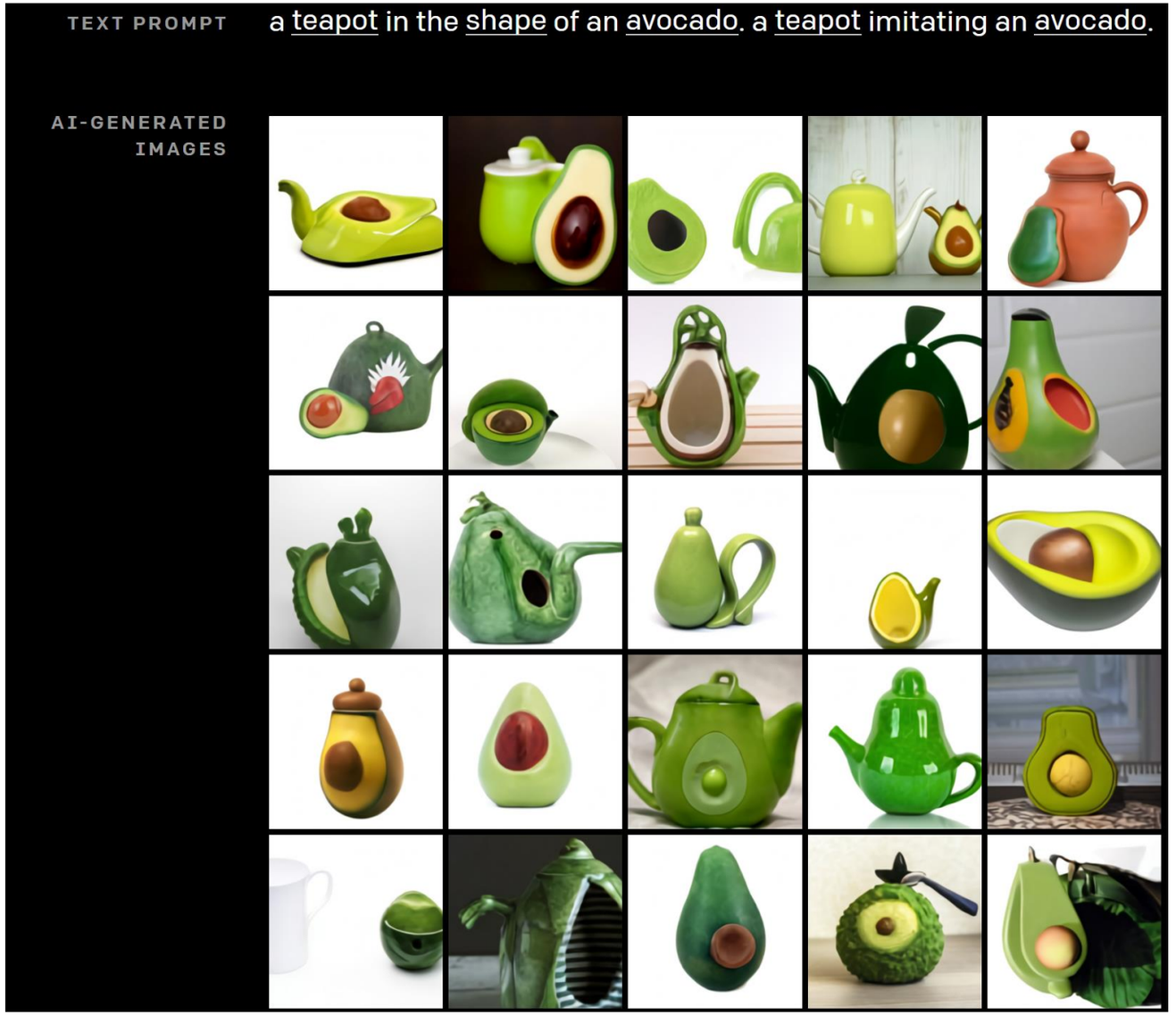
Data Source and Sampling

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

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

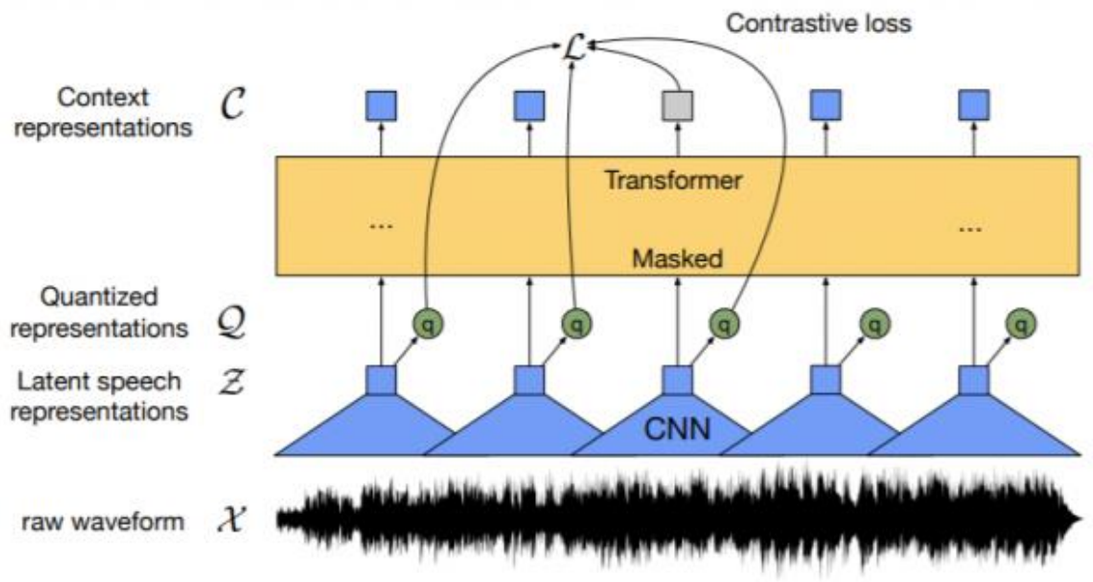
https://github.com/google-research/bert/blob/master/multilingual.md

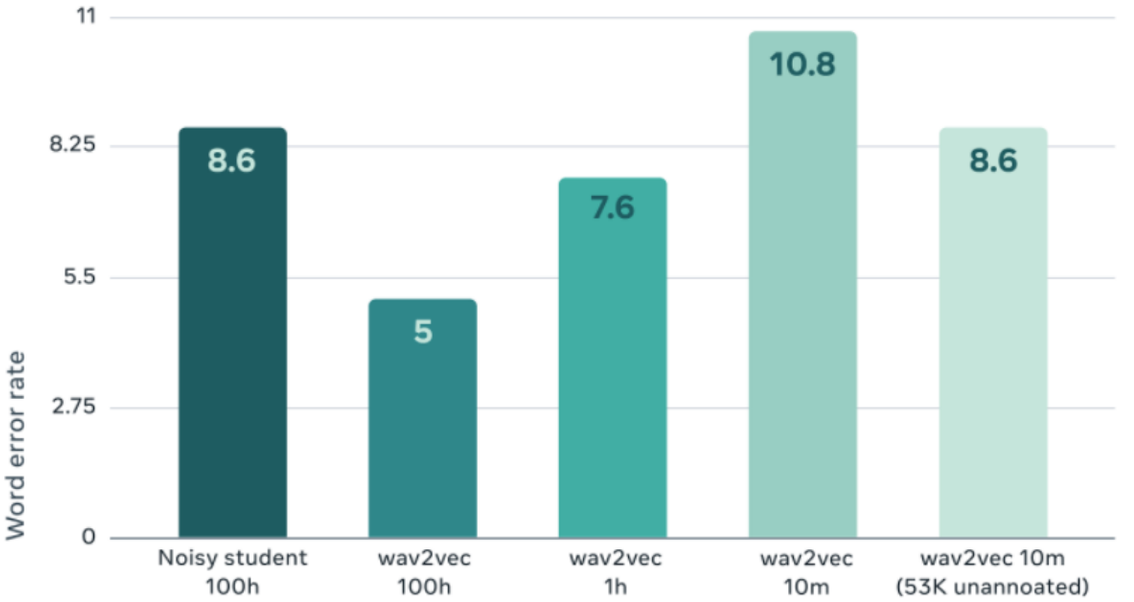
多模态交互



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

多模态交互





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

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

Content

- 预训练大模型的发展趋势
- 多语种机器翻译研究现状
- 基于预训练大模型的多语种机器翻译

Motivation

- **Current MT products usually support translations between up to 100+ languages**
  - Deploy ~2L systems trained on English--centric data, e.g., Chinese <->English
  - English as a pivot for translations between other languages, e.g., Chinese<->English<->German
- **Challenges**
  - The number of deployed systems increases linearly
  - Pivot increases translation latency
  - Low quality on low-source languages
  - Unable to translate zero-resource languages
  - Most data are English centric
- **one-modal-for-all-languages solution for MT and help low-/zero-resource languages?**
  - i.e., multilingual neural machine translation (MNMT)
  - positive transfer between related domains and transferable tasks.

Google’s First Attempt on MNMT

- **The same architecture as bilingual models**
- **Shared model parameters between all languages**
- **Mainly English-centric data with joint subwords**
- **Language-awareness by preprocessing data**

```
How are you? -> ¿Cómo estás?
```



```
<2es> How are you? -> ¿Cómo estás?
```

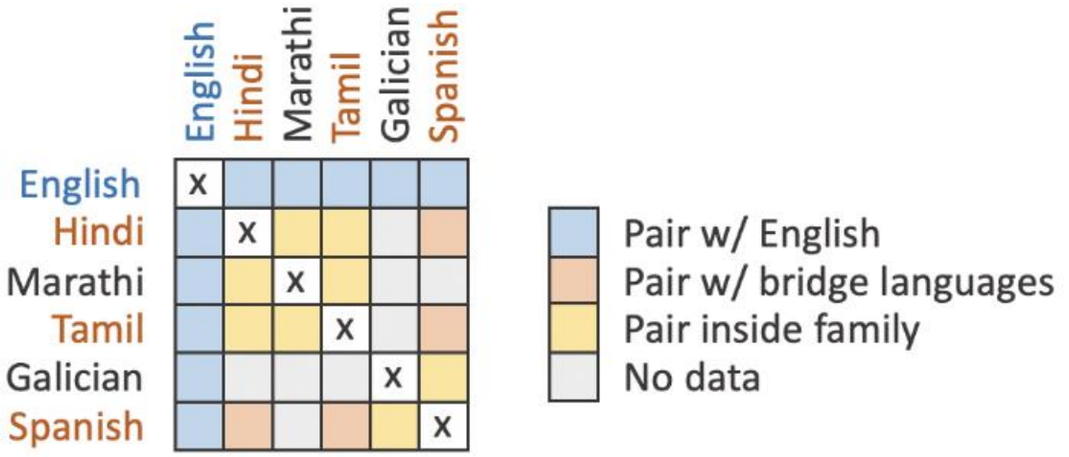
- **Observations**
  - Quality improvement on low-resource languages
  - Direct zero-shot translation is possible
  - Allow mixing languages on the source/target side
  - Visual evidence for interlingua representation

(Johnson et al. 2017. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. TACL.)

Beyond English-Centric MNMT from Facebook

- **Select 100 languages with high coverage on diversity of scripts and resource levels**
- **Mining parallel data with LASER embeddings with considerations of language similarity to avoid mining all directions**
- **Sparse mining include 3 parts:**
  - 14 language groups according to linguistic similarity, geographic and cultural proximity; all languages within a grouping are mined against each other
  - 1-3 bridge languages per group are mined against all other bridge languages
  - English centric

Model	All Avg	Supervised Low	Mid	High
Random 80%	11.9	3.6	16.1	31.5
Random 80% w/ En	16.3	8.9	22.4	36.6
Bridge Language, 80%	**17.2**	**10.4**	**23.2**	**37.4**

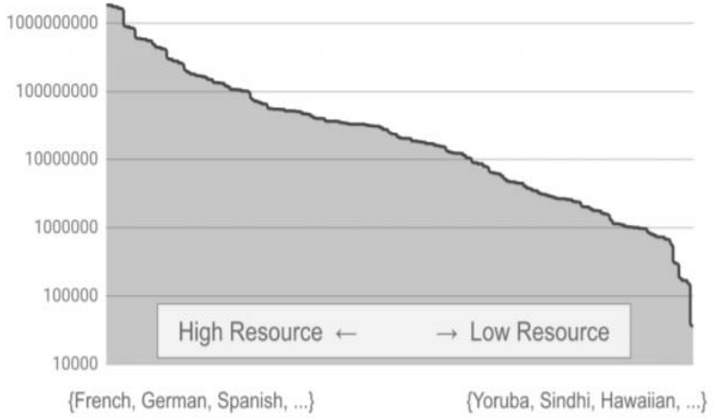


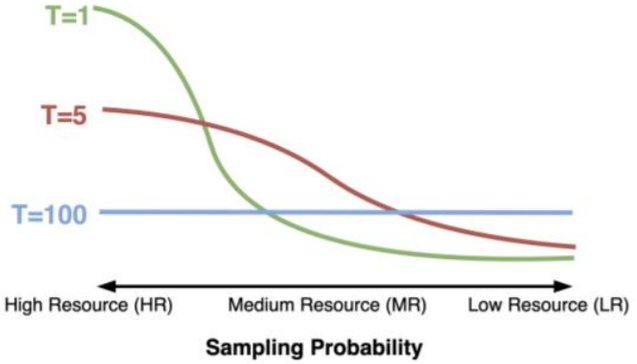
Setting	To English	From English	Non-English
Bilingual baselines	27.9	**24.5**	8.3
English-Centric	31.0	24.2	5.7
English-Centric with Pivot	—	—	10.4
Many-to-Many	**31.2**	24.1	**15.9**

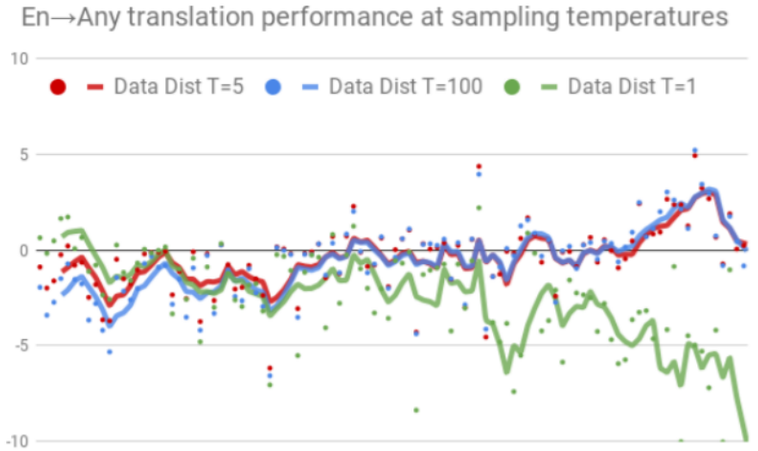
(Fan et al. 2020. Beyond English-Centric Multilingual Machine Translation. arXiv:2010.11125)

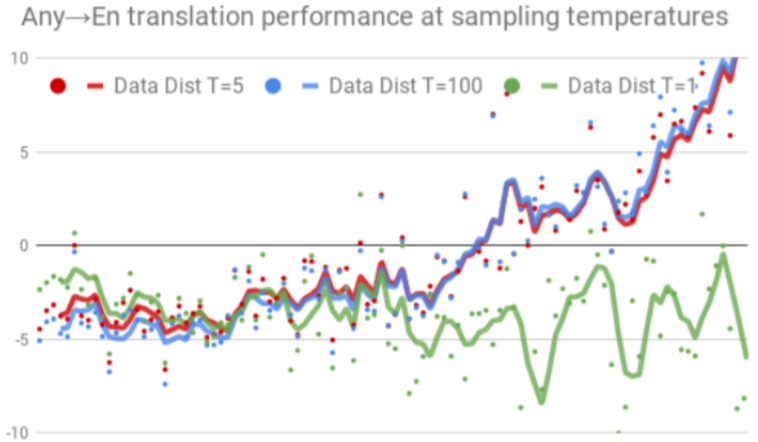
Data Imbalance

- **Data imbalance problem naturally exists across languages**
- **Sample equally (T=100): maximize improvement on low resource and large deterioration on high-resource**
- **Sample truly (T=1): retrain more performance on high-resource with sacrificing on low-resource languages**
- **Appropriate sampling(T=5) can potentially alleviate the problem but be heuristic and still challenging**









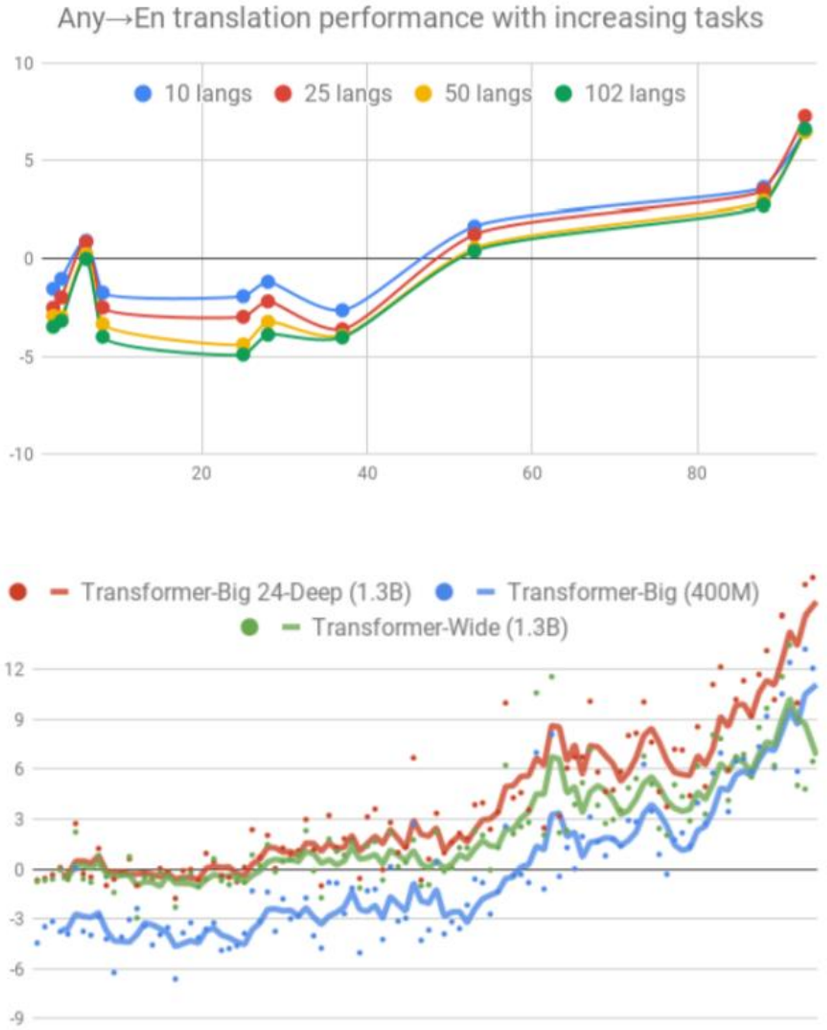
(Arivazhagan et al. 2019. Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges. arXiv:1907.05019.)

Scalability Issue

- **Scalability issue: performance degrades for all language pairs, especially the high and medium resource ones, as the number of tasks grows.**
- **However, the zero-shot performance for most language pairs increases when using more languages**

	$De \rightarrow Fr$	$Be \rightarrow Ru$	$Yi \rightarrow De$	$Fr \rightarrow Zh$	$Hi \rightarrow Fi$	$Ru \rightarrow Fi$
10 langs	11.15	36.28	8.97	15.07	2.98	6.02
102 langs	14.24	50.26	20.00	11.83	8.76	9.06

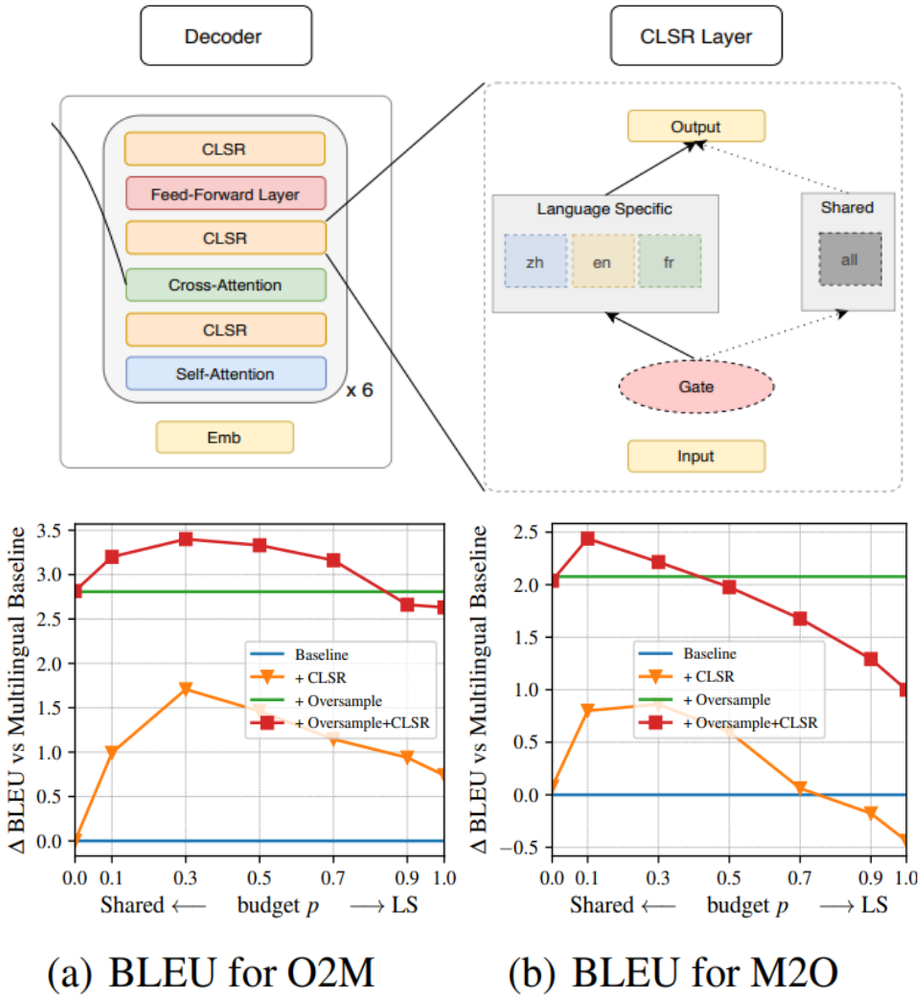
- **Model capacity might be one of the most important factors**
- **However, naively scaling capacity might result in poor performance, e.g., the Transformer-Wide fails to show similar gains in the low-resource setting**



(Arivazhagan et al. 2019. Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges. arXiv:1907.05019.)

Model Capacity and Language Diversity

- **Experiments show sharing encoders among multiple languages is effective and widely used. Keeping decoders separate is important.**
- **Investigate language-specific parameters to improve model capacity and handle language diversity, manually or automatically**
- **MNMT is sensitive to both the amount and the position of language-specific modules**
- **one-to-many translation benefits more than many-to-one translation, especially with data imbalance**
- **Automatic routing suggests 10%-30% language-specific capacity to reach to best performance on top and/or bottom layers**
- **Automatic routing has little to do with linguistic characteristics.**

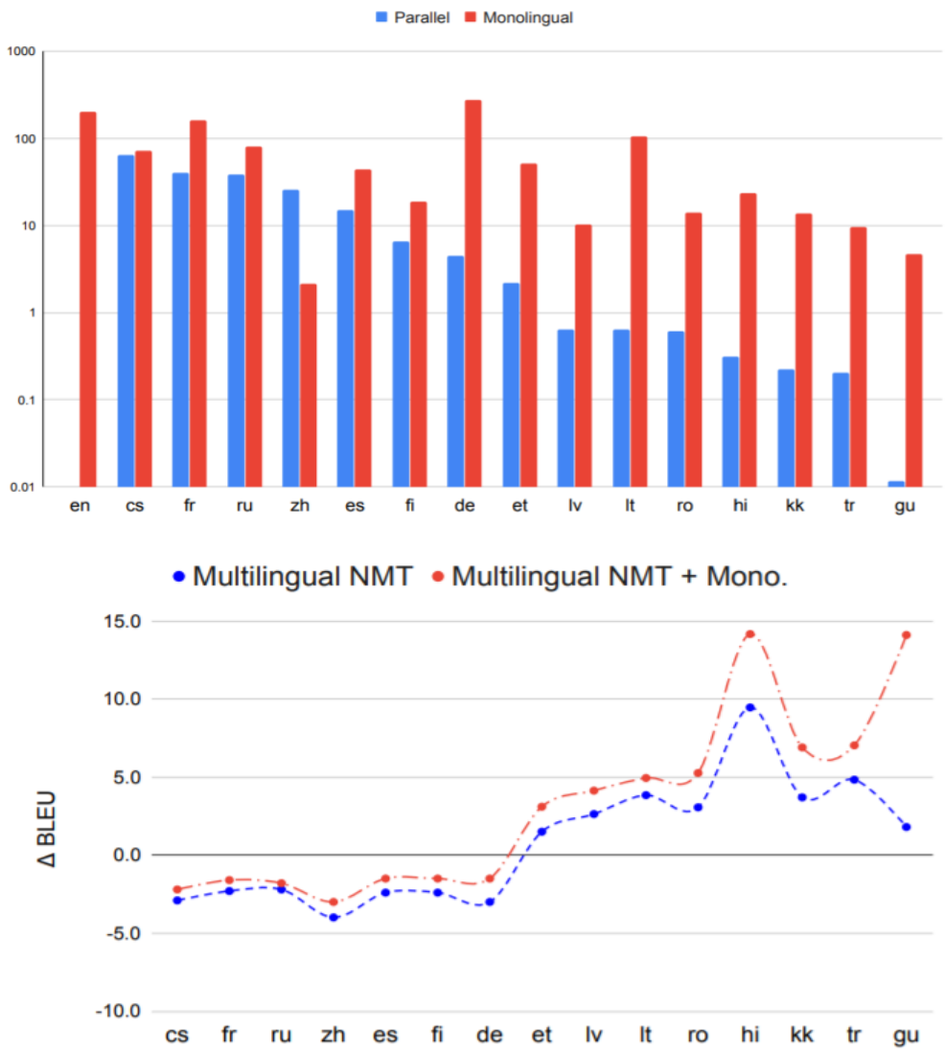


(Zhang et al. 2021. Share Or Not? Learning To Schedule Language-Specific Capacity For Multilingual Translation. ICLR.)

Monolingual data

- **Monolingual data is in much larger scale specially for low-resource languages**
- **proven to be very effective for bilingual models**
- **By back-translation or denoising**
- **Improvement on both high- and low-resource languages**

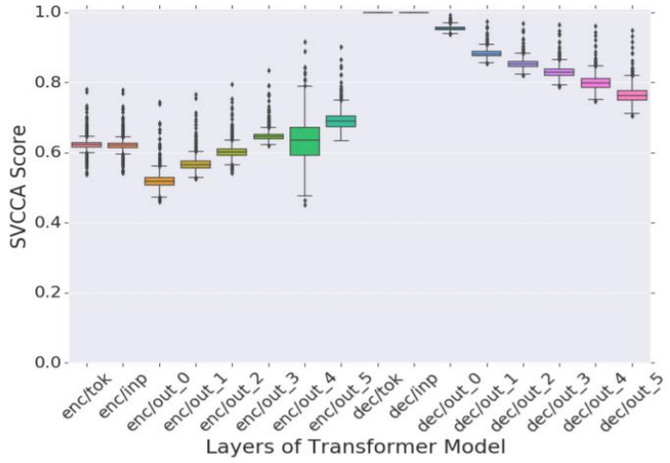
- **Pretraining on mutilingual mono or parallel data, then finetuning with bilingual data is also effective for low-resource languages**
- **However, this deviates from the goal of one-model-for-all-languages**



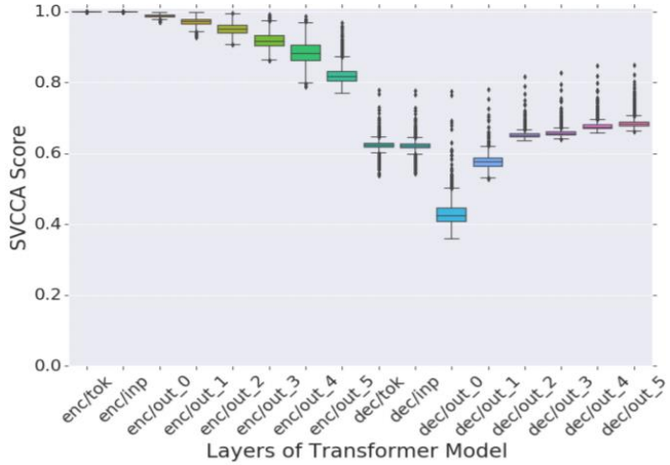
(Siddhant et al., Leveraging Monolingual Data with Self-Supervision for Multilingual Neural Machine Translation, ACL 2020)

One-Model-For-All-Languages is Still Challenging

- **Current studies suggest that the source language representation depends on the target language and vice versa. Representation similarity varies across layers.**
- **Representation Learning**
  - Handling language divergence: languages diverse in many aspects, e.g., domain, scripts, grammar etc.
  - Representation bottleneck: performance degrades when increasing the number of tasks
  - Balancing the sharing of representations between languages
- **Pretrained Models**
  - Pretrain+finetune procedure currently shows significant improvements on low-resource languages
  - How to make use of various kinds of knowledge from the web to help the representation learning
  - A large and stronger universal model is also desired to improve translation quality in general



(a) X-En Language Pairs



(b) En-X Language Pairs

(Kudugunta et al. 2020. Investigating Multilingual NMT Representations at Scale. EMNLP.)

Pre-trained Multilingual NMT on Parallel Data

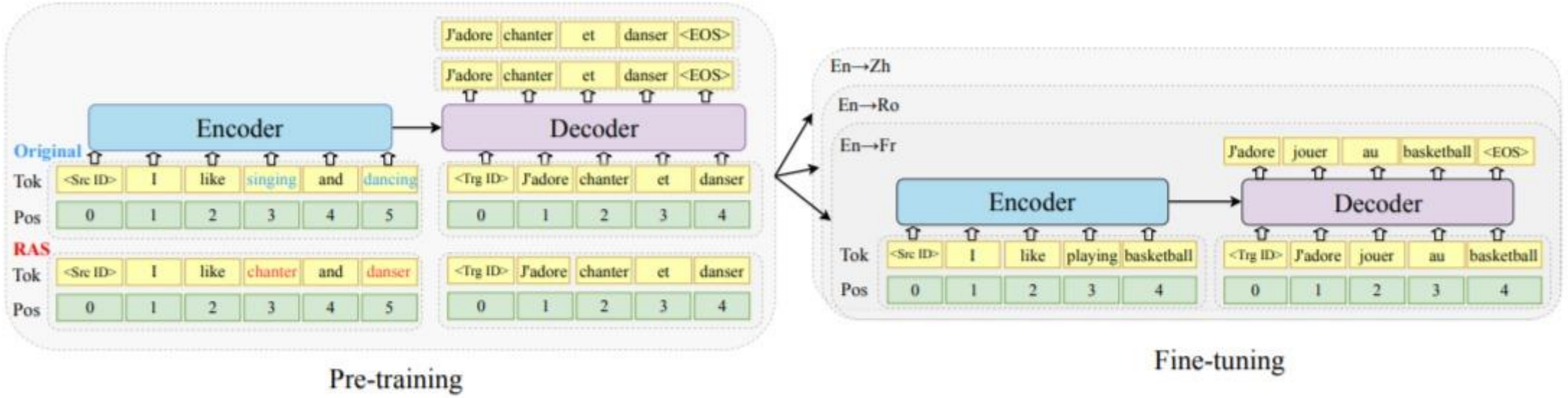
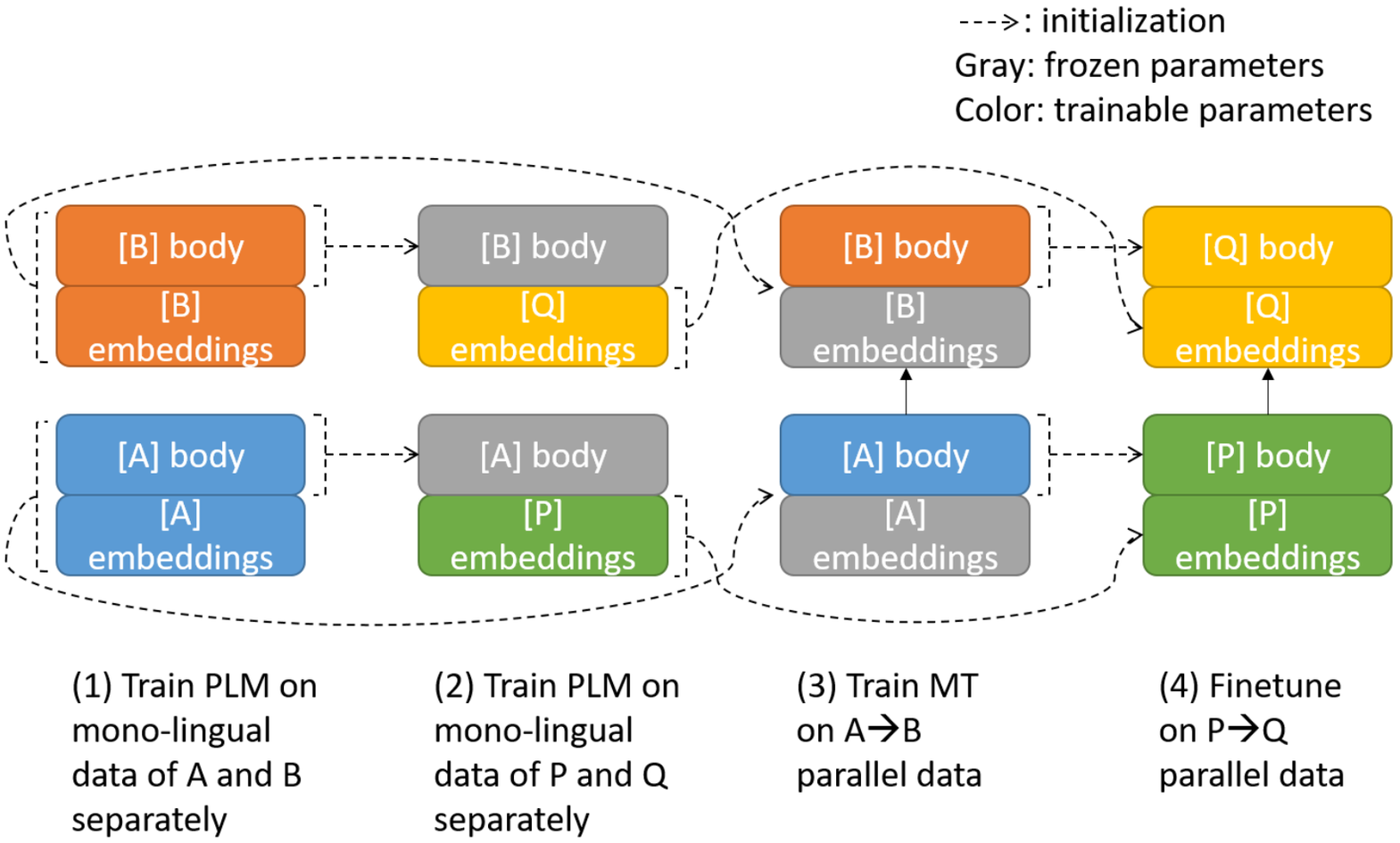


Figure 1: The proposed mRASP method. "Tok" denotes token embedding while "Pos" denotes position embedding. During the pre-training phase, parallel sentence pairs in many languages are trained using translation loss, together with their substituted ones. We randomly substitute words with the same meanings in the source and target sides. During the fine-tuning phase, we further train the model on the downstream language pairs to obtain specialized MT models.

(Lin et al., Pre-training Multilingual Neural Machine Translation by Leveraging Alignment Information, EMNLP 2020)

Dual Transfer for Low-Resource NMT



(Zhang et al., Learning Multilingual Representation via Cross-Lingual Deep Alignments, Findings of ACL 2021)

Content

- 预训练大模型的发展趋势
- 多语种机器翻译研究现状
- 基于预训练大模型的多语种机器翻译

基于预训练大模型的多语种机器翻译的设想

- 在现有大语言模型（盘古-α）基础上持续训练
- 在文本基础上引入多模态自监督训练，特别是语音自监督训练
- 充分利用大数据优势：
  - 多语言平行语料库
  - 多语言单语语料库
  - 语音识别数据
  - 纯语音数据
- 充分利用大算力优势（鹏城云脑、昇腾芯片、MindSpore框架）
- 充分利用国家实验室、企业、高校强强联合的优势

Thank you

www.huawei.com

Copyright©2008 Huawei Technologies Co., Ltd. All Rights Reserved.
The information contained in this document is for reference purpose only, and is subject to change or withdrawal according to specific customer requirements and conditions.