华为机器翻译架构和模型加速

魏代猛
华为2012 / 机器翻译实验室

个人简介

- 华为高级技术专家，机器翻译算法负责人，产品落地华为云、HMS、华为手机等
- 北京大学硕士，研究方向：机器翻译、同传翻译、语义理解等
- 带领团队参加 WMT20/21/22 news、biomedical、efficiency等赛道多项 第一，IWSLT 22 多项第一， WAT20比赛多项第一
- 在AAAI，ACL，EMNLP，ICASSP等发表论文30+

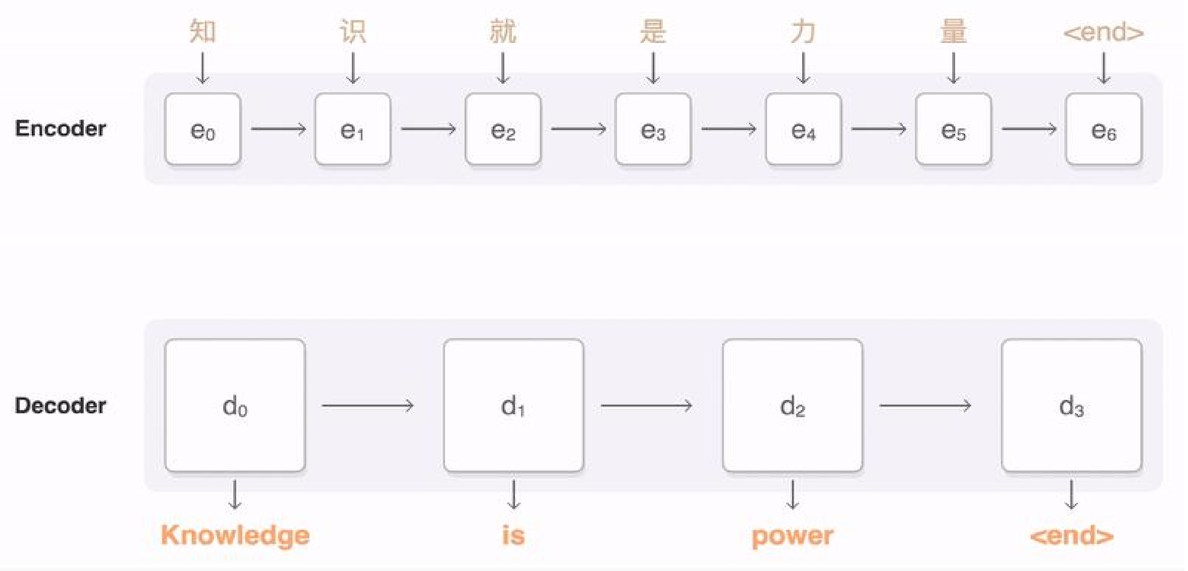
大纲

- 机器翻译简介
- 模型推理问题
- 端测推理加速
- 华为机器翻译
- 总结

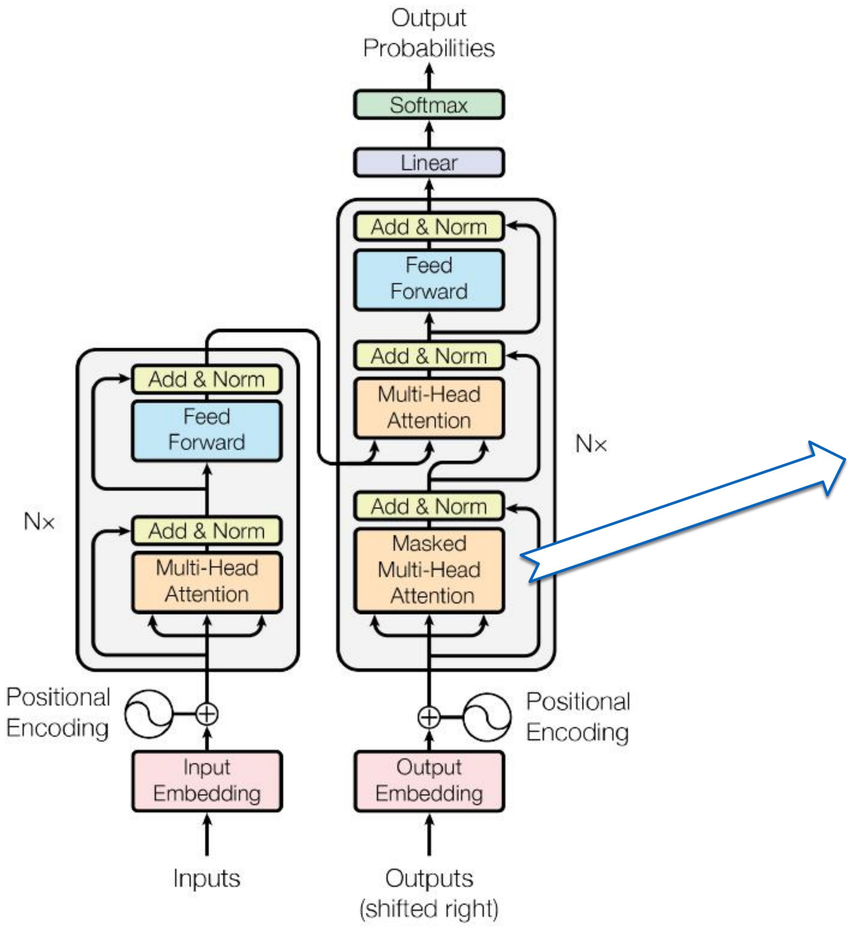
机器翻译简介

1、主流的机器翻译模型包含Encoder和Decoder两部分，Encoder将原文整个序列编码成一个多维向量，Decoder将原文序列的向量解码成译文。

2、Attention模型记录原文和译文的词对齐关系，指导机器翻译在解码译文某个词时，应该更关注与原文的哪一个部分，以提升长句翻译质量。



机器翻译简介



Encoder：6
Decoder：6
Hidden size：1024
参数量：2亿
大小：800M

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin. “Attention is All you Need” Neural Information Processing Systems (2017).

模型推理问题

Transformer模型在GPU，CPU，ARM运行典型值

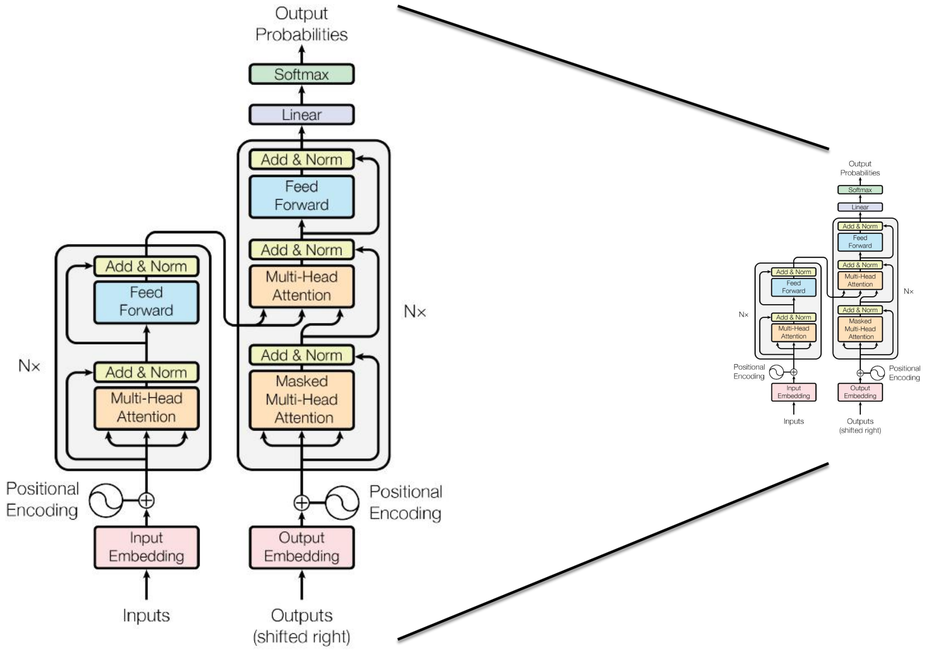
	GPU/T4	CPU/Intel	ARM
耗时	45 ms/token	150 ms/token	-

模型太大，计算量太大

端侧最具挑战

模型推理问题

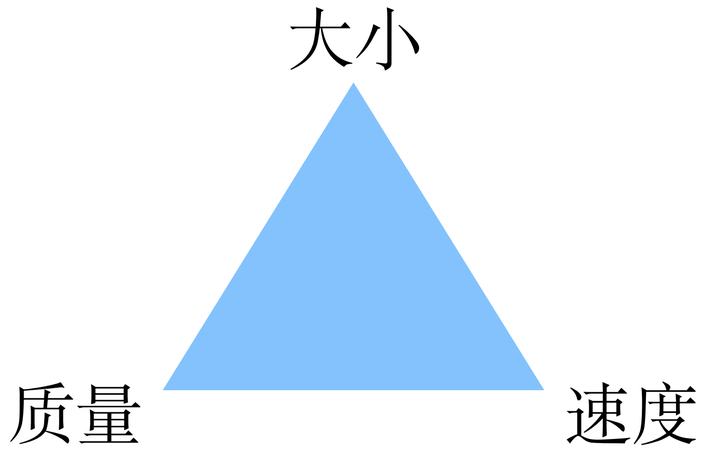
让模型变小



存储 ↓

计算量 ↓

质量 ↓



大纲

- 机器翻译简介
- 模型推理问题
- 端测推理加速
- 华为机器翻译
- 总结

端侧推理加速

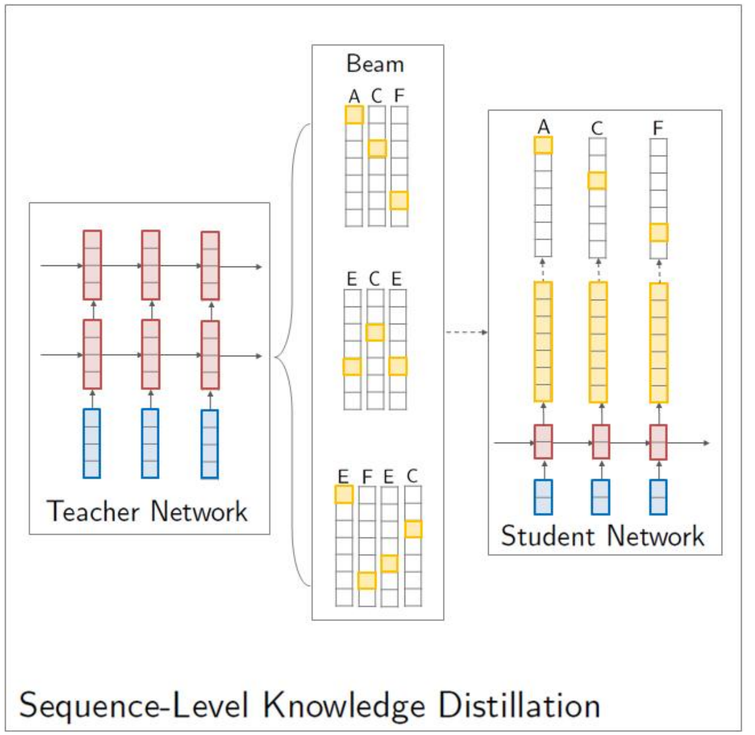
模型存储小

计算量小

小模型 → 质量变差

小模型 → 高质量?

知识蒸馏



Geoffrey E. Hinton, Oriol Vinyals and Jeffrey Dean. “Distilling the Knowledge in a Neural Network” *arXiv: Machine Learning* (2015): n. pag.
Yoon Kim and Alexander M. Rush. “Sequence-Level Knowledge Distillation” *Empirical Methods in Natural Language Processing* (2016).
Markus Freitag, Yaser Al-Onaizan and Baskaran Sankaran. “Ensemble Distillation for Neural Machine Translation” *arXiv: Computation and Language* (2017): n. pag.

端侧推理加速

知识蒸馏

Model	Emb.	FFN	Head	Depth	Params(M)	Size(MB)	wmt19	wmt20
Teacher*4	1024	4096	16	25/6	514	2000	46.71	39.70
Base.12	512	2048	8	12/1	53	210	44.65	38.02

96%

小模型 & 高质量

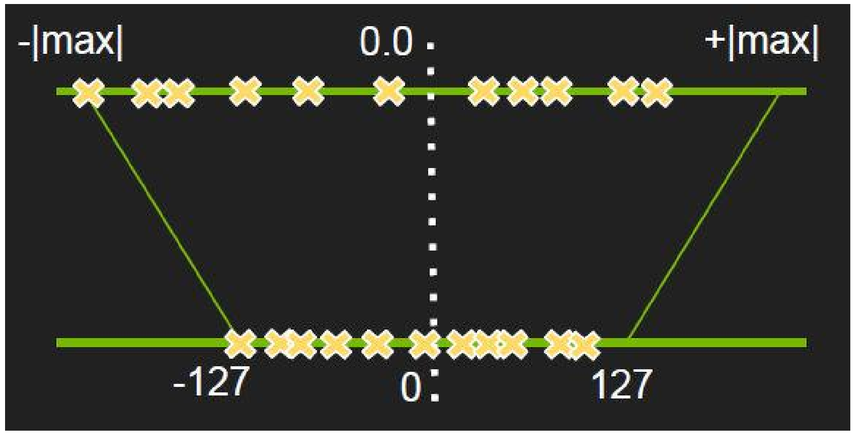
Tiny Bert

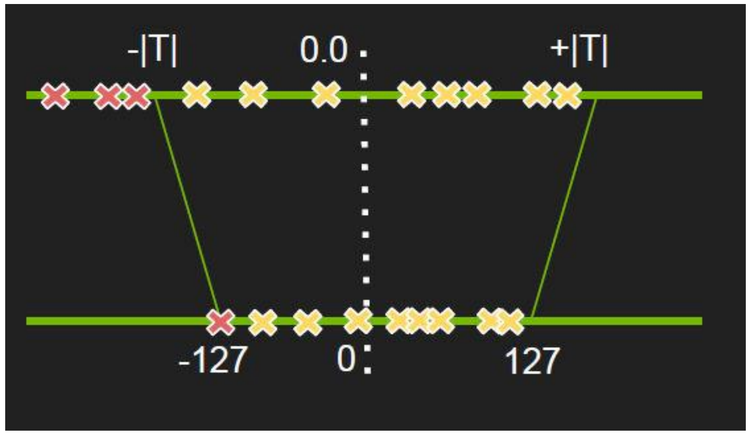
System	#Params	#FLOPs	Speedup	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	Avg
$BERT_{BASE}$ (Teacher)	109M	22.5B	1.0x	83.9/83.4	71.1	90.9	93.4	52.8	85.2	87.5	67.0	79.5
$BERT_{TINY}$	14.5M	1.2B	9.4x	75.4/74.9	66.5	84.8	87.6	19.5	77.1	83.2	62.6	70.2
$BERT_{SMALL}$	29.2M	3.4B	5.7x	77.6/77.0	68.1	86.4	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**

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. TinyBERT: Distilling BERT for Natural Language Understanding. EMNLP 2020

端侧推理加速

小模型：更小空间，更快的推理　　------模型压缩，低精度推理



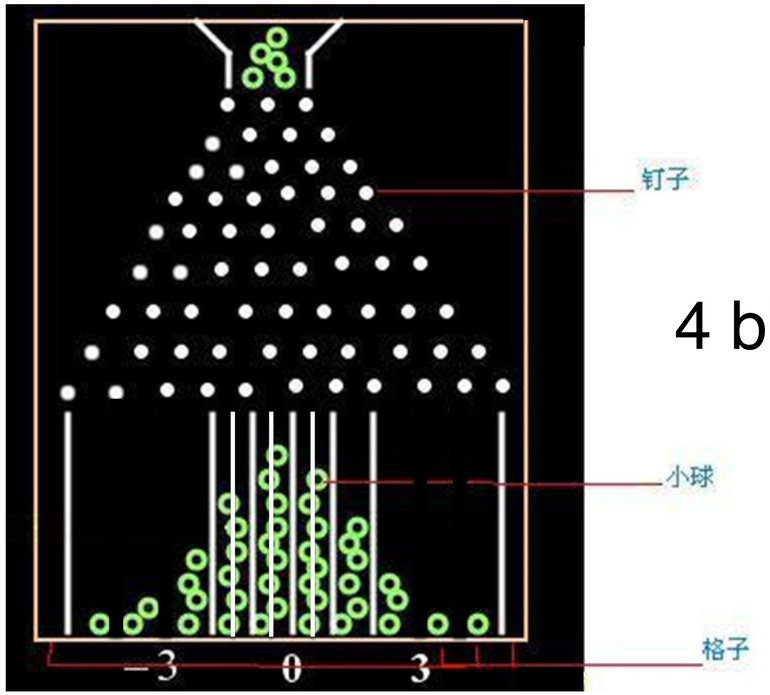


增加量化层，FP32->Int8->FP32，　E2E训练

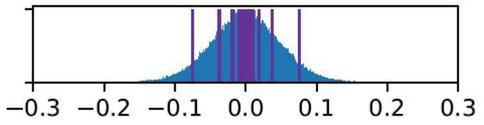
Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam and Dmitry Kalenichenko. “Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference” Computer Vision and Pattern Recognition (2018).

端侧推理加速

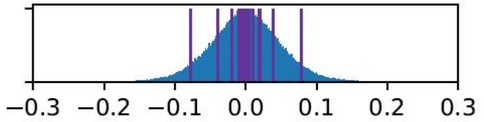
小模型：更小空间，更快的推理　　------模型压缩，低精度推理



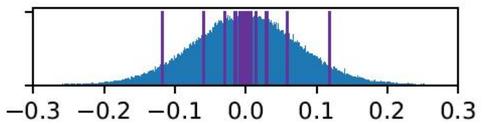
4 bit→log 4bit



(a) First layer encoder self attention weight histogram

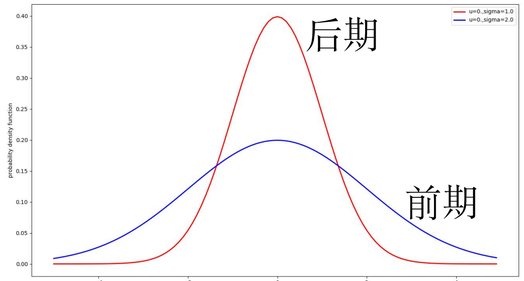


(b) First layer decoder self attention weight histogram



(c) First layer decoder context attention weight

Log→0值问题



Alham Fikri Aji and Kenneth Heafield. 2020. Compressing Neural Machine Translation Models with 4-bit Precision. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 35–42, Online. Association for Computational Linguistics.

端侧推理加速

小模型：更小空间，更快的推理　　------模型压缩，低精度推理

模型	BLEU(WMT14)	Parameter Size
32 bit	26.5	260M
8 bit	26.4(-0.1)	66M
4 bit	24.3(-2.2)	35M
Log 4 bit(后期)	25.1(-1.4)	35M
Log 4 bit(前期)	26.2(-0.3)	35M

直接4bit 影响大，log 4 bit前期介入量化训练很关键

端侧推理加速

小模型：更小空间，更快的推理　　------模型压缩，低精度推理

Int8推理

模型中计算量最大的是矩阵运算(GEMM)

Int8推理：用整型运算代替浮点型运算提速

处理好量化和反量化是提速的关键

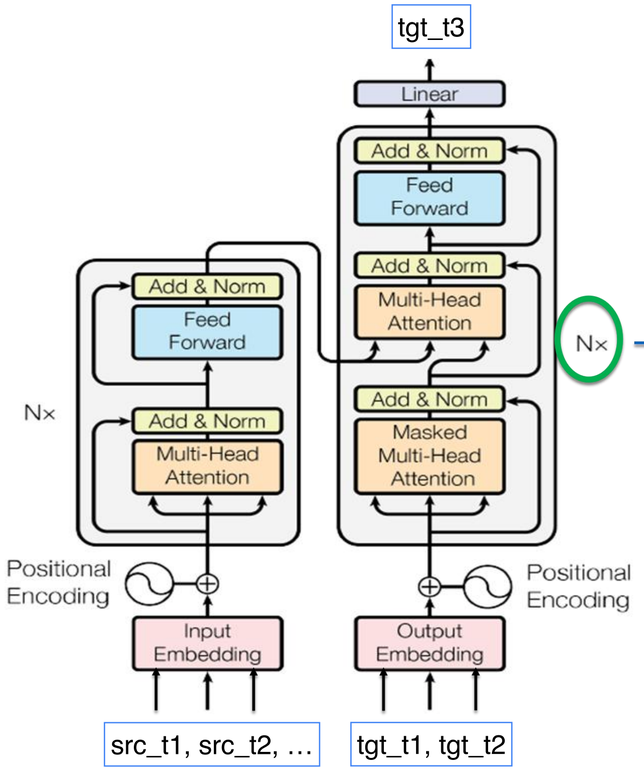
Operator	FP32	INT8
FC	4.806	1.53
Transpose	0.4158	0.228
Eltwise	0.198	0.13
MatMul	0.1134	0.068
Softmax	0.063	0.040
LayerNorm	0.063	0.036
Reshape	0.027	0.013
Activation	0.0216	0.011
Embedding	0.018	0.010
Slice	0.009	0.003
Multiply	0.0054	0.001
Total	**5.94**	**2.1**

华为Noah高性能推理实验室
https://github.com/huawei-noah/bolt

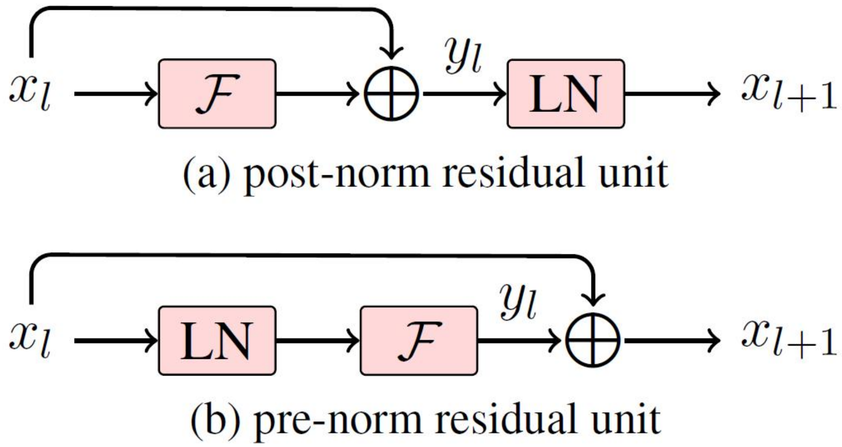
Daya Shanker Khudia, Jianyu Huang, Protonu Basu, Summer Deng, Haixin Liu, Jongsoo Park and Mikhail Smelyanskiy. “FBGEMM: Enabling High-Performance Low-Precision Deep Learning Inference..” arXiv: Learning (2021): n. pag.

端侧推理加速

小模型：更强的能力 ------结构优化，参数共享，多语言模型



能否变小?

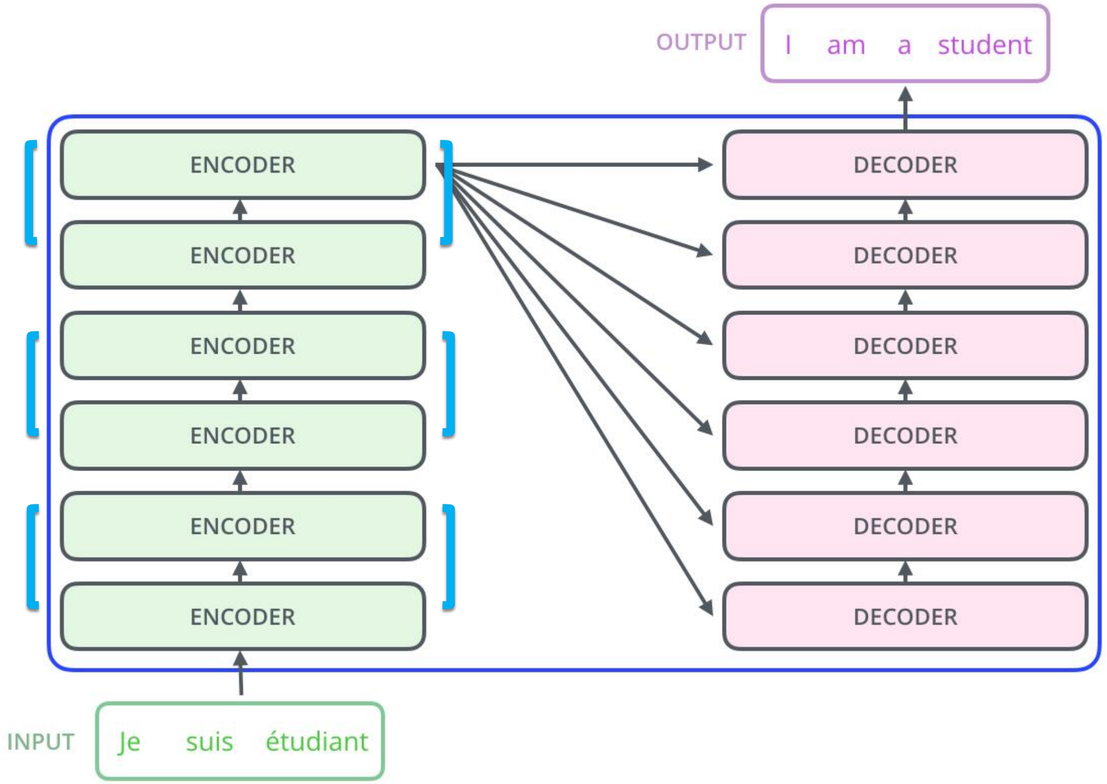


Encoder 25-40层，decoder3层

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F. Wong and Lidia S. Chao. "Learning Deep Transformer Models for Machine Translation.." Meeting of the Association for Computational Linguistics (2019).

端侧推理加速

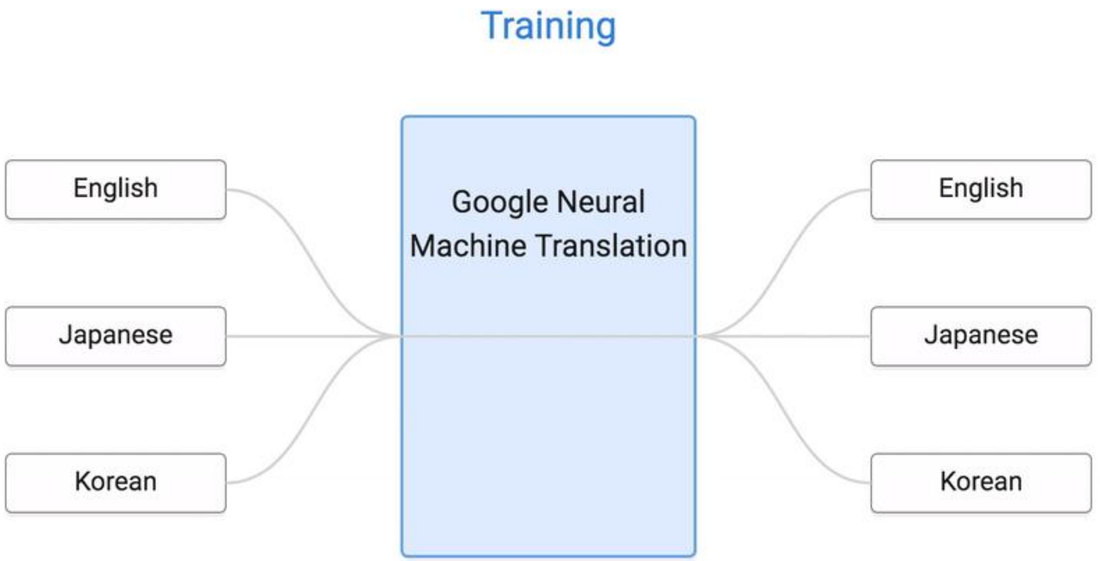
小模型：更强的能力 ------结构优化，参数共享，多语言模型



相邻层共享最好

端侧推理加速

小模型：更强的能力 ------结构优化，参数共享，多语言模型



10个语种放一个模型
30个语种模型容量不够

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg S. Corrado, Macduff Hughes and Jeffrey Dean. “Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation” Transactions of the Association for Computational Linguistics 5 (2017): 339-351.
Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen and Yonghui Wu. “Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges” arXiv: Computation and Language (2019): n. pag.

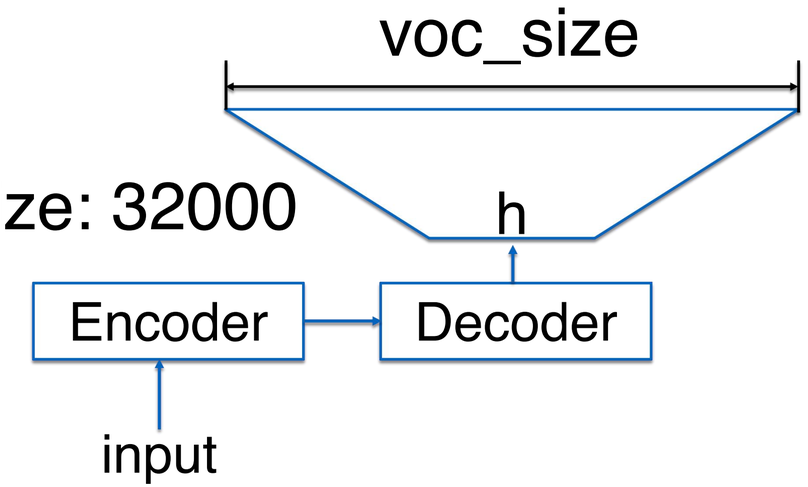
小结

策略	质量	速度	大小
知识蒸馏	√	√	-
量化推理	-	√	√
模型结构	√	√	-
参数共享	√	-	√
多语言	√	-	√
ShortList	专注于减少计算量		
Decoder结构			

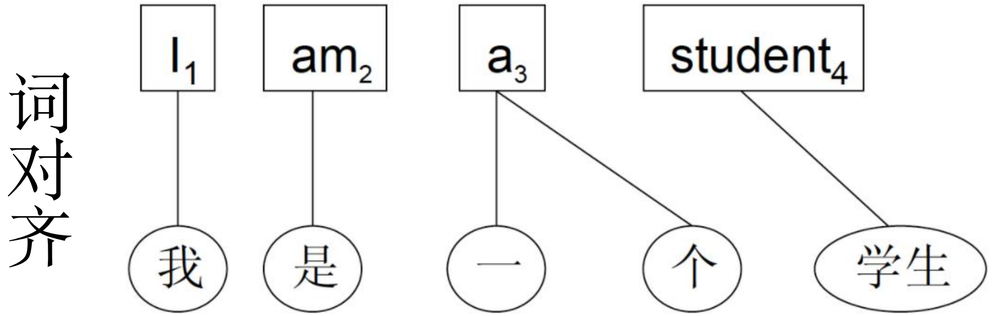
端侧推理加速

ShortList优化

h: 512
voc_size: 32000



10 x FFN



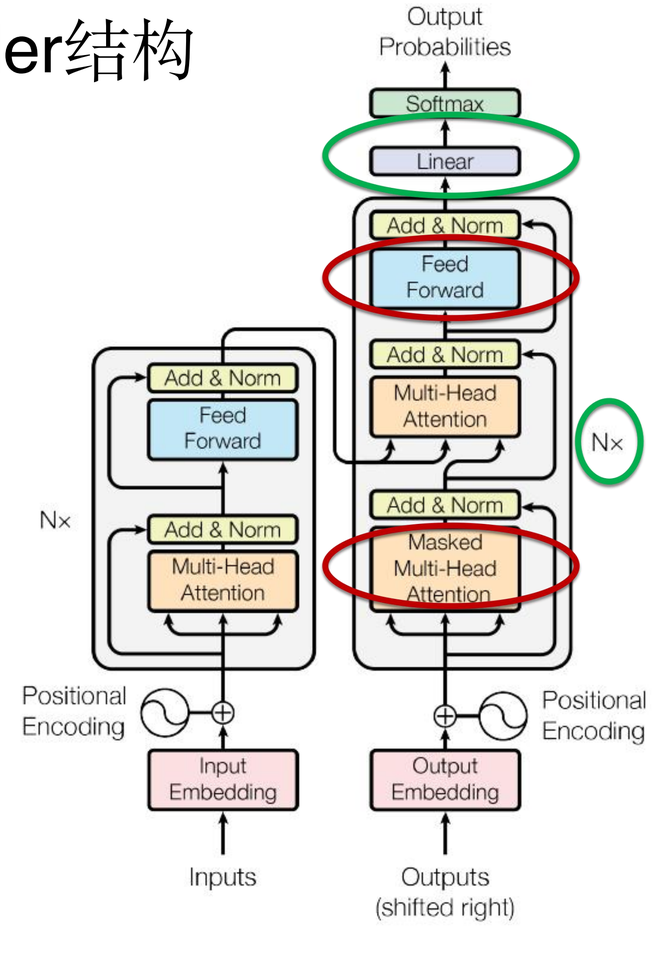
每个词：100候选
16词→300候选(去重)

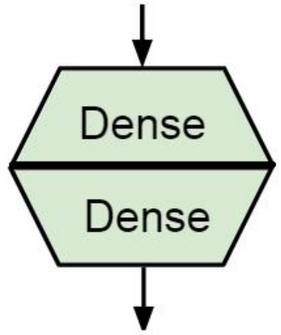
512x32000 → 512x300

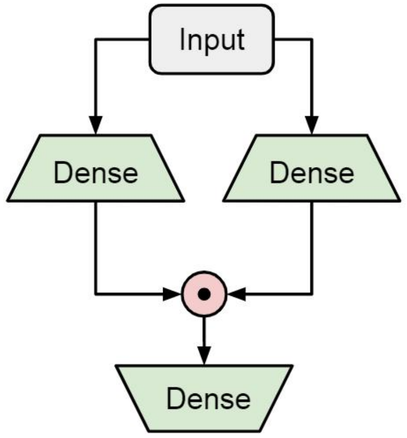
词对齐可以用Fastalign，每个词75候选

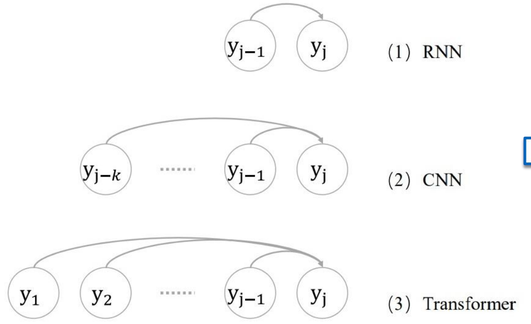
端侧推理加速

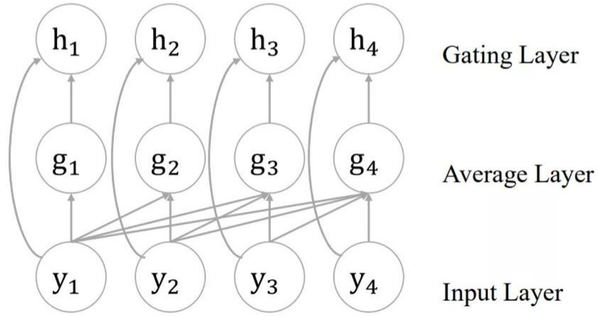
Decoder结构











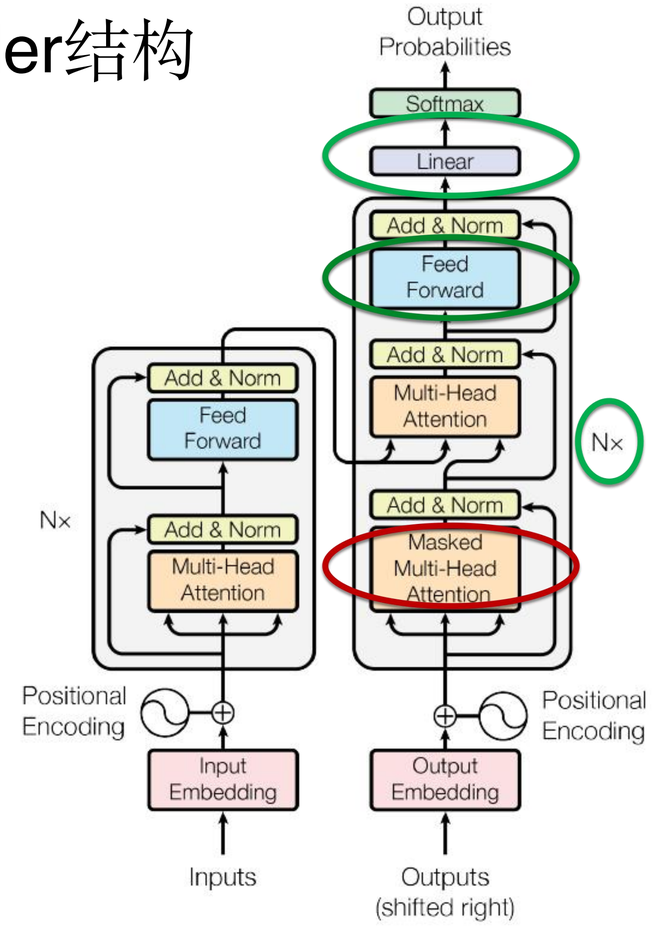
$$SA: O(o_l^2 d) \rightarrow O(o_l$$

Yann N. Dauphin, Angela Fan, Michael Auli and David Grangier. “Language modeling with gated convolutional networks” International Conference on Machine Learning (2017).

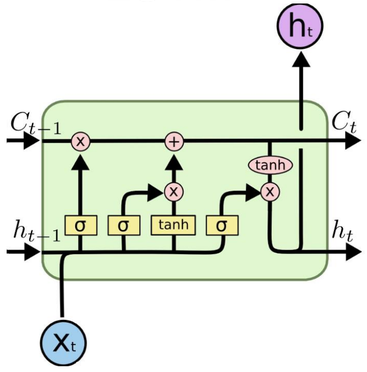
Biao Zhang, Deyi Xiong and Jinsong Su. “Accelerating Neural Transformer via an Average Attention Network” Meeting of the Association for Computational Linguistics (2018).

端侧推理加速

Decoder结构

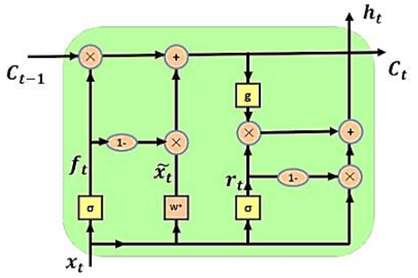


LSTM



$$
\begin{aligned}
f_t &= \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \\
i_t &= \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \\
o_t &= \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \\
c_t &= f_t \circ c_{t-1} + i_t \circ \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \\
h_t &= o_t \circ \sigma_h(c_t)
\end{aligned}
$$

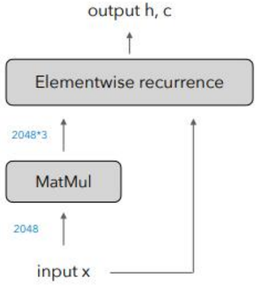
SRU

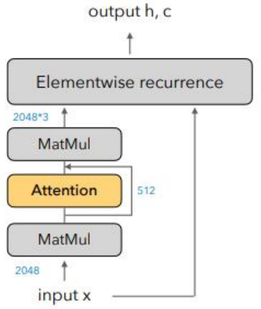


$$
\begin{aligned}
\mathbf{f}_t &= \sigma\left(\mathbf{W}_f \mathbf{x}_t + \mathbf{v}_f \odot \mathbf{c}_{t-1} + \mathbf{b}_f\right) \\
\mathbf{c}_t &= \mathbf{f}_t \odot \mathbf{c}_{t-1} + (1 - \mathbf{f}_t) \odot (\mathbf{W}\mathbf{x}_t) \\
\mathbf{r}_t &= \sigma\left(\mathbf{W}_r \mathbf{x}_t + \mathbf{v}_r \odot \mathbf{c}_{t-1} + \mathbf{b}_r\right) \\
\mathbf{h}_t &= \mathbf{r}_t \odot \mathbf{c}_t + (1 - \mathbf{r}_t) \odot \mathbf{x}_t
\end{aligned}
$$

$$
\mathbf{U}^\top = \begin{pmatrix} \mathbf{W} \\ \mathbf{W}_f \\ \mathbf{W}_r \end{pmatrix} [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_L]
$$

SRU++



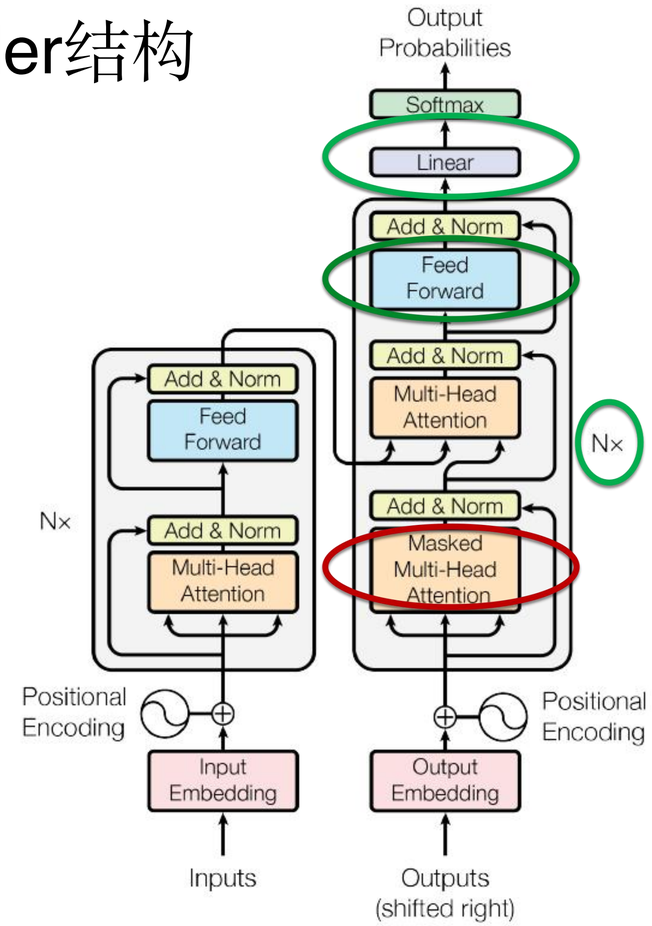


Tao Lei, Yu Zhang, Sida I. Wang, Hui Dai, Yoav Artzi, Simple Recurrent Units for Highly Parallelizable Recurrence, EMNLP 2017

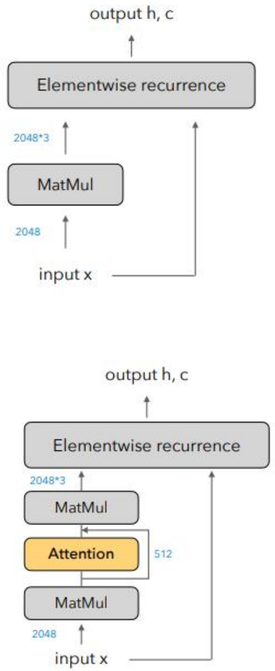
Tao Lei. 2021. When Attention Meets Fast Recurrence: Training Language Models with Reduced Compute. Association for Computational Linguistics 2021.

端侧推理加速

Decoder结构



SRU++



AASRU

$$f[t] = \sigma(U[t,0]) + V \odot C[t-1] + b)$$
$$r[t] = \sigma(U[t,1]) + V^{'} \odot C[t-1] + b^{'})$$
$$c[t] = f[t] \odot C[t-1] + (1 - f[t]) \odot C[t,2]$$
$$h[t] = r[t] \odot C[t] + (1 - r[t]) \odot x[t]$$

$$Q = W^q X^T$$
$$V = W^v X^T$$
$$A^T = AVERAGE(V^T)$$
$$U^T = W^o layernorm(Q + A)$$

Yann N. Dauphin, Angela Fan, Michael Auli and David Grangier. “Language modeling with gated convolutional networks” International Conference on Machine Learning (2017).
Hengchao Shang, Ting Hu, Daimeng Wei, HW-TSC’s Submission for the WMT22 Efficiency Task. In Proceedings of the Seventh Conference on Machine Translation (WMT), pages 677–681, ACL 2022

小结

策略	质量	速度	大小
知识蒸馏	√	√	-
量化推理	-	√	√
模型结构	√	√	-
参数共享	√	-	√
多语言	√	-	√
shortlist	-	√	-
Decoder结构	-	√	-

小结

WMT22 Efficiency Task

Model	Precious	Size	WPS	BLEU
Teacher	FP32	2000	-	36.29
Base.12	FP32	212	237	35.30
	INT8	53	815	35.20
+retrain	INT8	53	815	35.19
Small.9	FP32	112	468	34.40
	INT8	28	1129	34.29
Tiny.12	FP32	68	759	33.62
	INT8	17	1693	33.39
Tiny.6	FP32	52	996	32.27
	INT8	13	2001	31.92
+int4	INT8	6.5	1989	26.10

Table 2: Optimization results. The test set is WMT 2020 News test. The unit of size is MB. WPS refers to the source side. The test environment is Intel(R) Xeon(R) Gold 6278C CPU @ 2.60GH. We submit four models: Base.12, Small.9, Tiny.12 and Tiny.6 and the final Tiny.6+int4

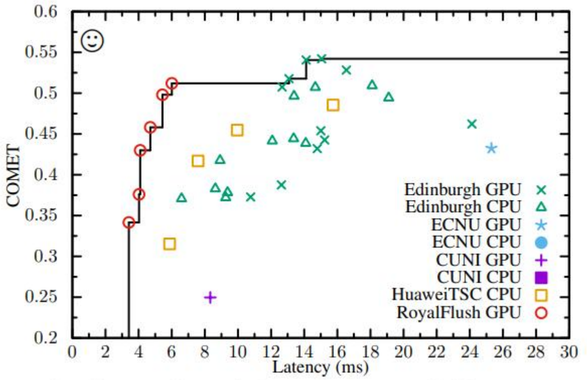
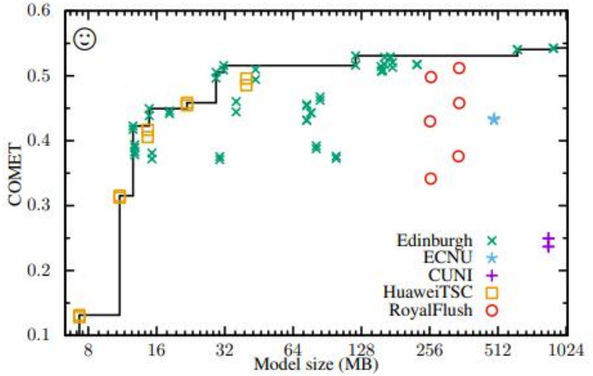
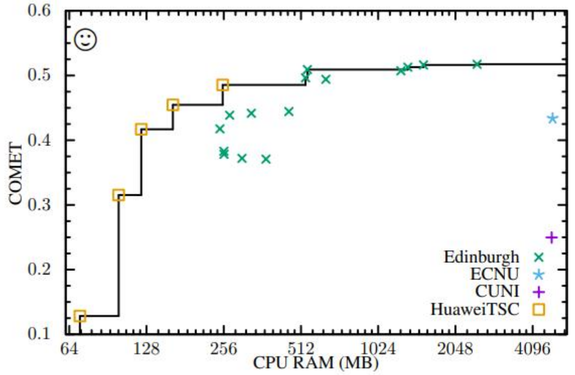


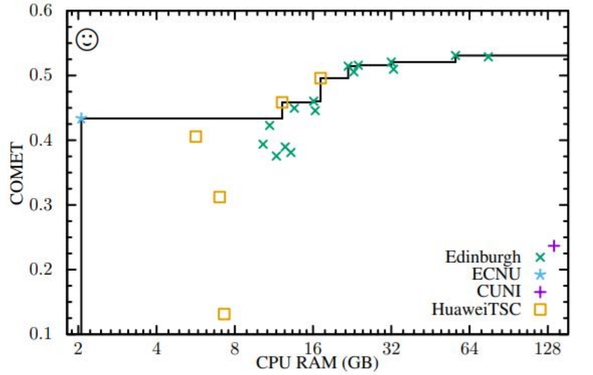
Figure 3: Measured latency for CPU and GPU systems with COMET scores.



(a) Compressed model size.



(c) 1 Core CPU memory consumption with latency.



(d) 36 Core CPU memory consumption with batching.

Hengchao Shang, Ting Hu, Daimeng Wei, Zongyao Li, Xianzhi Yu, Jianfei Feng, Ting Zhu, Lizhi Lei, Shimin Tao, Hao Yang, Ying Qin, Jinlong Yang, Zhiqiang Rao, and Zhengzhe Yu. 2022. HW-TSC's Submission for the WMT22 Efficiency Task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 677–681, ACL 2022

Kenneth Heafield, Biao Zhang, Graeme Nail, Jelmer Van Der Linde, and Nikolay Bogoychev. 2022. Findings of the WMT 2022 Shared Task on Efficient Translation.
In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 100–108, Abu Dhabi, United Arab Emirates (Hybrid). ACL 2022.

大纲

- 机器翻译简介
- 模型推理问题
- 端测推理加速
- 华为机器翻译
- 总结

华为机器翻译


需求
内网
中译外：外籍员工信息不对称
外译中：中方员工查阅外文信息和文献
代码注释翻译
海外舆情分析
外网
2C：海外游购娱，帮助中国人走向世界
2B：中国企业走出去
2D：全球开发者和生态共同成长
接入层
HUAWEI Translate
Web
Client
App
API
WeLink
3MS
iSales+
服务工具市场 Service Tool Market
Linstar Innovation Platform
DBOX
IDEA DBASE
iKnow
AILA
OMRP
CloudDragon
IntelliJ IDEA
累计32万+注册用户，160+内网调用应用，年翻译字符数xxxx亿+，稳定增长中
华为翻译 API over 华为云/HMS
机器翻译 MT
机器翻译(Machine Translation)致力于为企业和个人提供不同语种间标语言文本的自动翻译
商用服务调用费用低至¥35/百万字符
立即使用
帮助文档
服务咨询
浏览器
应用市场
视频
蓝版WeLink
快译
HiAI
支撑华为云和HMS生态商业成功，年翻译字符xxxx亿+，快速增长中
中间层
引擎
通用翻译
领域翻译
低资源翻译
Optimus
算法
RNNSearch
Transformer
Knowledge Distillation
Multilingual
服务
MT
ASR
OCR
AR/VR
基础层
AI框架
MindSpore
TensorFlow/Torch/MindSpore
AI算力
NVIDIA/昇腾
AI数据
BIG DATA
xxx亿+多语言语料库
ICT专业术语库/记忆库/知识库
多模态语料

华为机器翻译


应用层
文本分类
文本生成
实体识别
信息抽取
句法分析
依赖分析
情感分析
工具层
Optimus Seq
模型推理
Optimus Adapter
模型转换
Optimus Constraint kit
受限解码
Optimus Compression
模型压缩
预训练模型
Optimus Domain 领域模型
Optimus ICT
ICT领域模型
Optimus ChinaYoung
网络流行语模型
Optimus Law
法律模型
More
更多
Optimus Task 任务模型
Optimus Sim相识度
Optimus Gen 生成模型
Optimus QE 评估模型
Optimus General 通用模型
Optimus Word 词向量
Optimus Sentence句模型
Optimus KG 知识模型
Optimus Tiny
轻量级模型
核心框架
MindSpore /Optimus Core 核心框架硬件适配

总结

1. 以业务为轴心，按场景优化
2. 立足深度学习不断变化的大背景
3. 结合自身优势，不断迭代

策略	GPU	CPU	ARM
知识蒸馏	-	-	√
量化推理	√	√	√
模型结构	√	√	√
参数共享	-	-	√
多语言	√	√	√
shortlist	-	√	√
Decoder结构	-	-	√

想一想，我该如何把这些技术应用在工作实践中？

THANKS

【议题反馈】华为机器翻译模型训练推理加速实践

扫描二维码 提交议题反馈