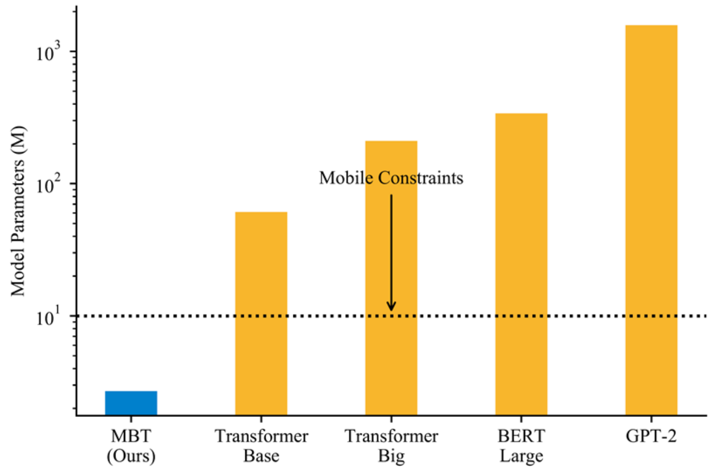
Efficient Transformer for Mobile Application

Zhanghao Wu, Zhijian Liu, Ji Lin, Yujun Lin, Song Han

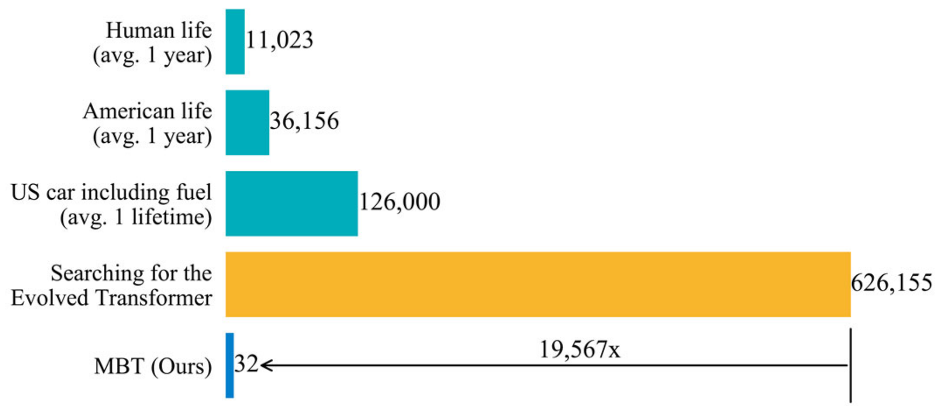
Massachusetts Institute of Technology

77 Massachusetts Avenue, 38-344
Cambridge, MA, 02139
https://hanlab.mit.edu

Modern NLP is EXPENSIVE



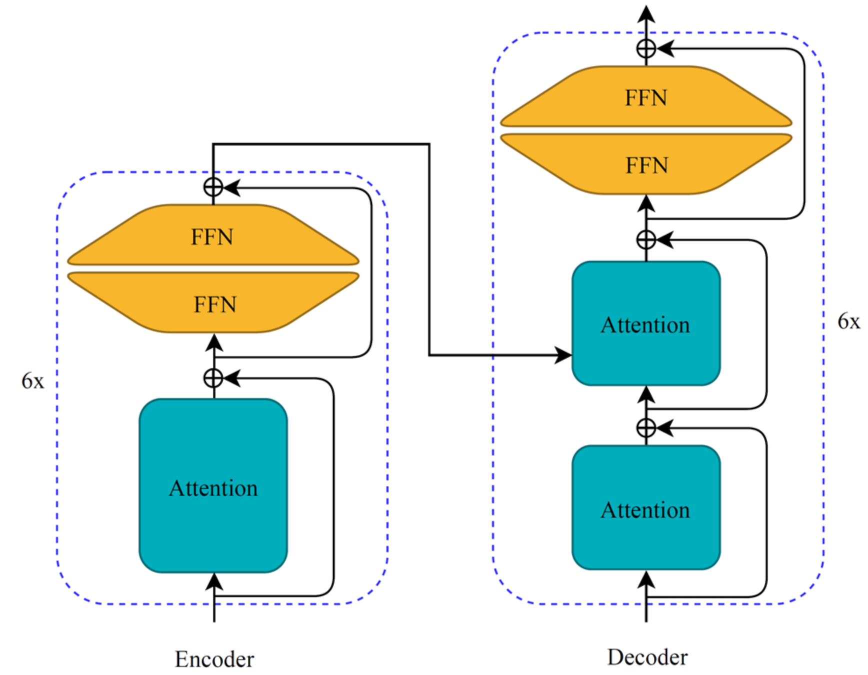
(a) Model sizes of modern NLP models



(b) The design cost measured in pounds of CO_2 emission

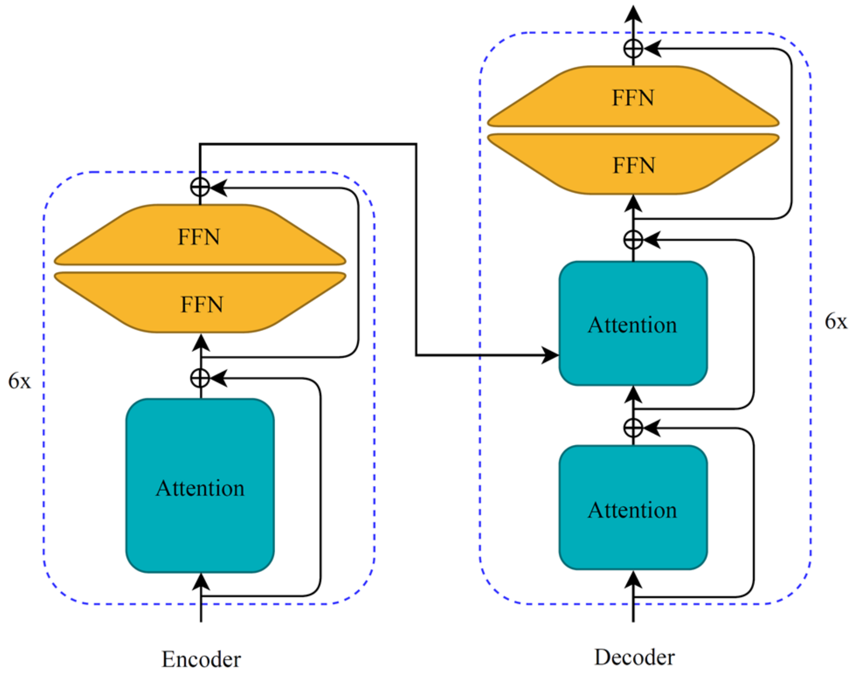
- NLP models are huge — much larger than mobile settings (a);
- Neural Architecture Search is a choice for finding an efficient model, but the massive searching cost raises much concerns (b).

Transformer Framework

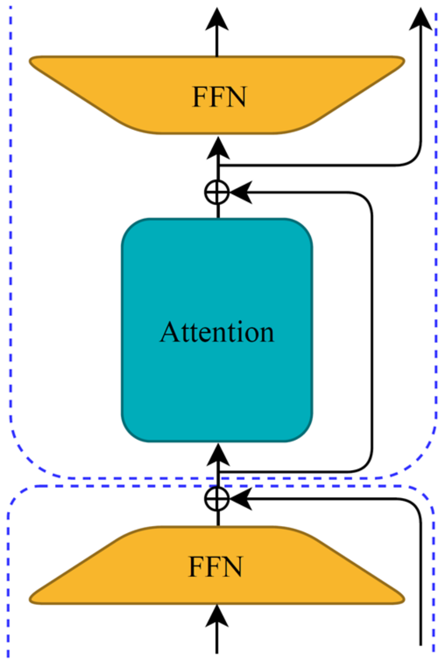


Basic transformer architecture for translation

Transformer Framework

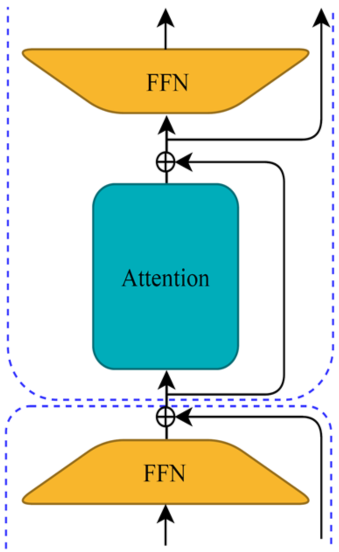


Basic transformer architecture for translation

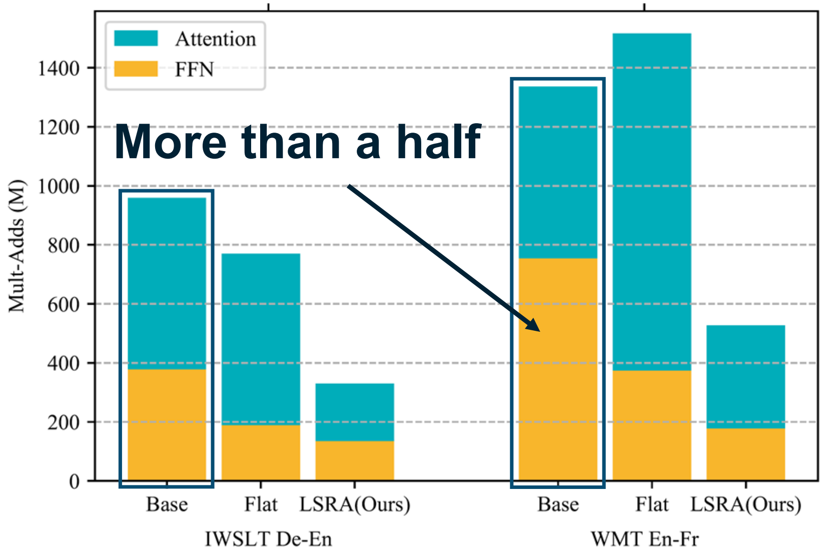


A different view for transformer block

Is Bottleneck Effective for 1-D Attention?



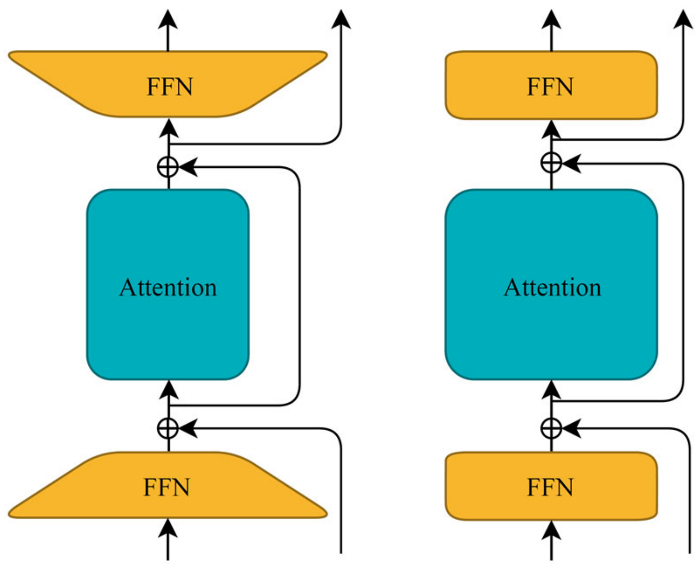
A different view for transformer block



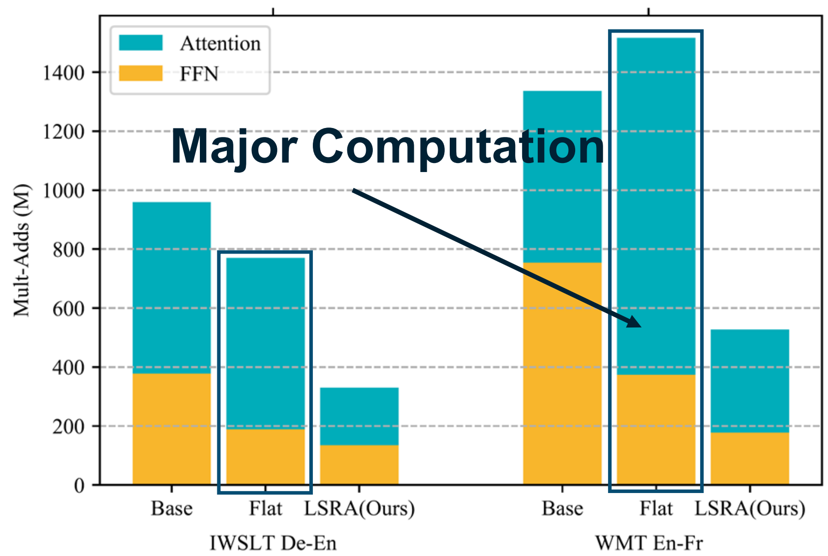
Mult-Adds breakdown for attention and FFN

- Original bottleneck design cannot significantly reduce the computation, also harms the capacity of attention layer due to smaller dimension.

Is Bottleneck Effective for 1-D Attention?



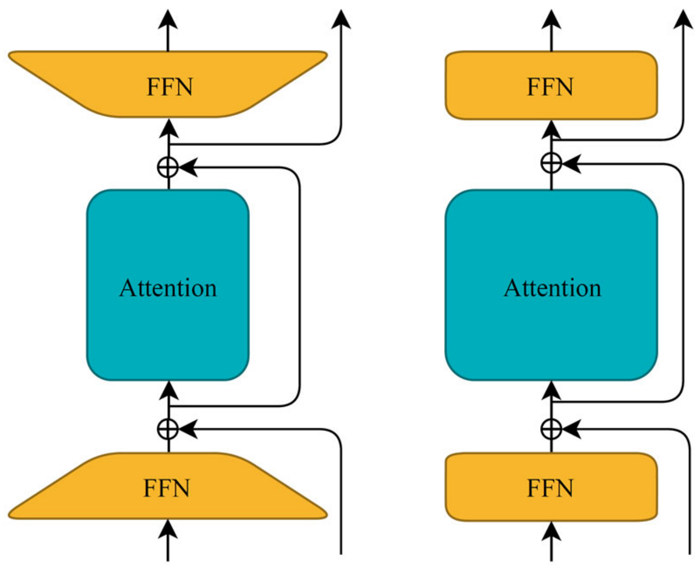
Vanilla and flattened transformer block



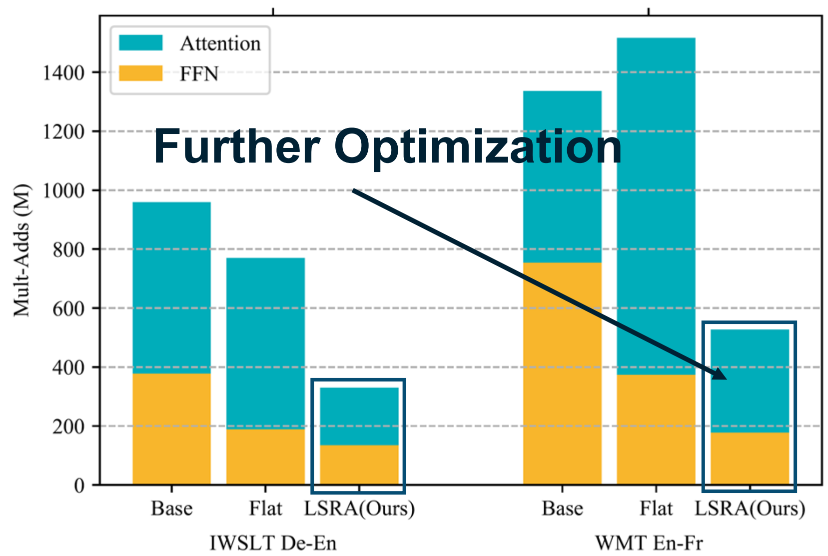
Mult-Adds breakdown for attention and FFN

- Original bottleneck design cannot significantly reduce the computation, also harms the capacity of attention layer due to smaller dimension.

Is Bottleneck Effective for 1-D Attention?



Vanilla and flattened transformer block



Mult-Adds breakdown for attention and FFN

- Original bottleneck design cannot significantly reduce the computation, also harms the capacity of attention layer due to smaller dimension.
- Flatten the transformer leaves larger space for further optimization.

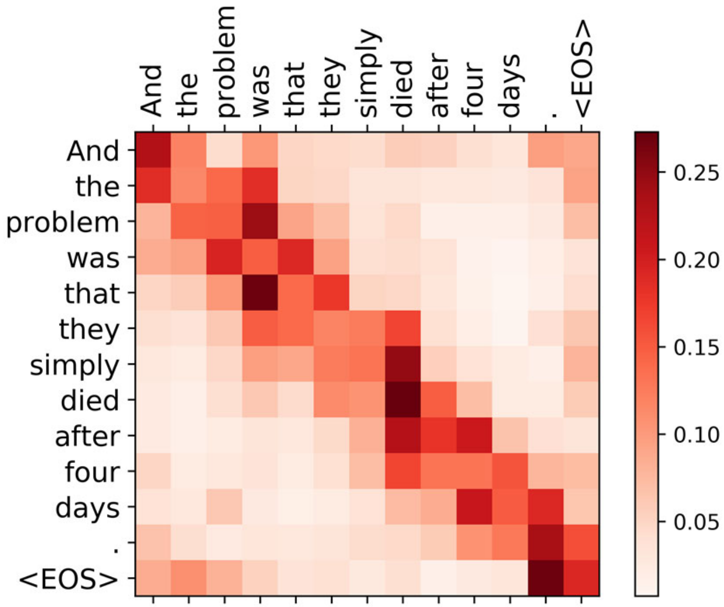
Flatten the Transformer

	IWSLT De-En			WMT En-De				WMT En-Fr
	Embedding	Mult-Adds	BLEU	Embedding	Mult-Adds	Attention	BLEU	BLEU
Vaswani et al. (2017)	512-1024	959M	34.4	512-2048	1.3G	44%	27.3	38.1
So et al. (2019)	–	–	–	512-2048	1.3G	44%	27.7	40.0
Our Reimplementation	512-1024	959M	34.5	512-2048	1.3G	44%	27.7	39.9
Transformer (*Flat*)	512-512	460M	**34.5**	720-720	1.5G	75%	**27.8**	**41.0**

- With the 'flat' version of transformer, the attention part now takes up the major computation.
- 'Flat' transformer can achieve comparable BLEU with the original transformer (slightly increase the computation, when necessary).

What does Attention Learn?

- Strong pattern for the attention: sparse points, vertical lines and diagonal groups
- The former two as "global" information and the latter one as "local" correlation.

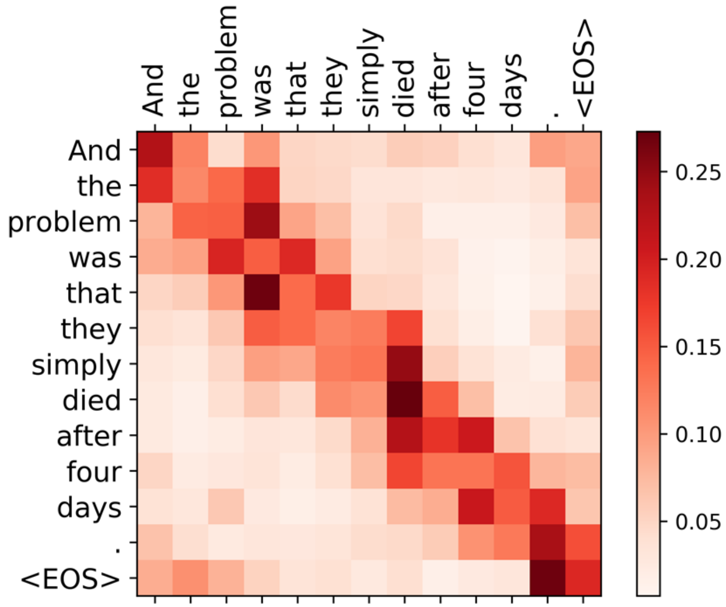


Self Attention

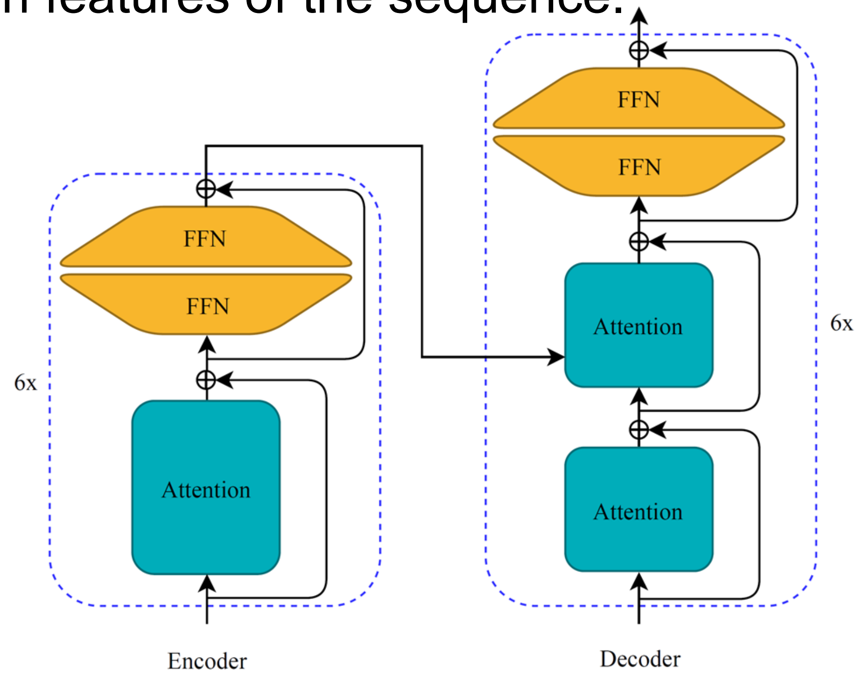
Zhanghao Wu, Zhijian Liu, Ji Lin, Yujun Lin, Song Han, Efficient Transformer for Mobile Applications

What does Attention Learn?

- Original self attention in the transformer captures both local and global information, since the FFN does not learn features of the sequence.



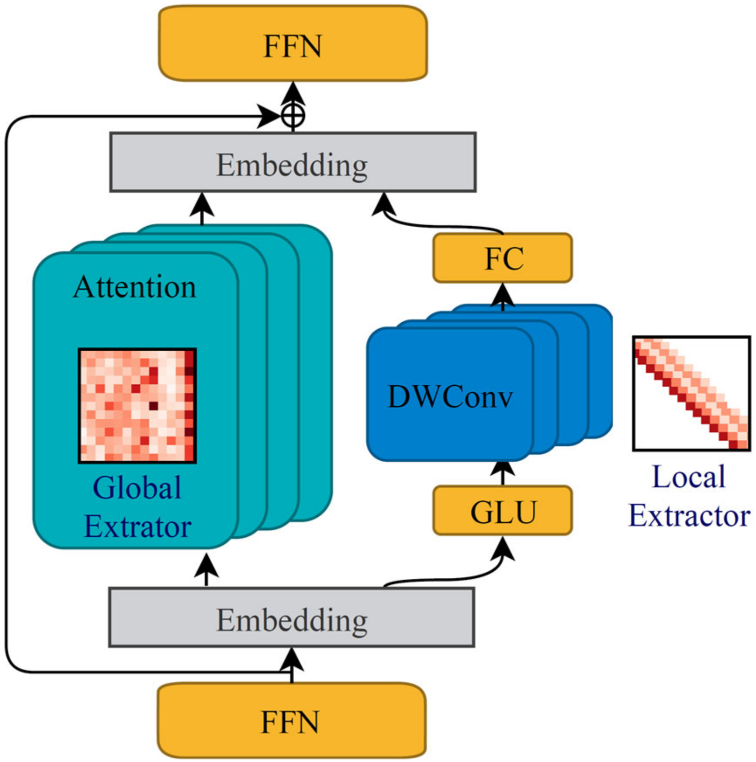
Self Attention



Basic transformer architecture for translation

Long-Short Range Attention (LSRA)

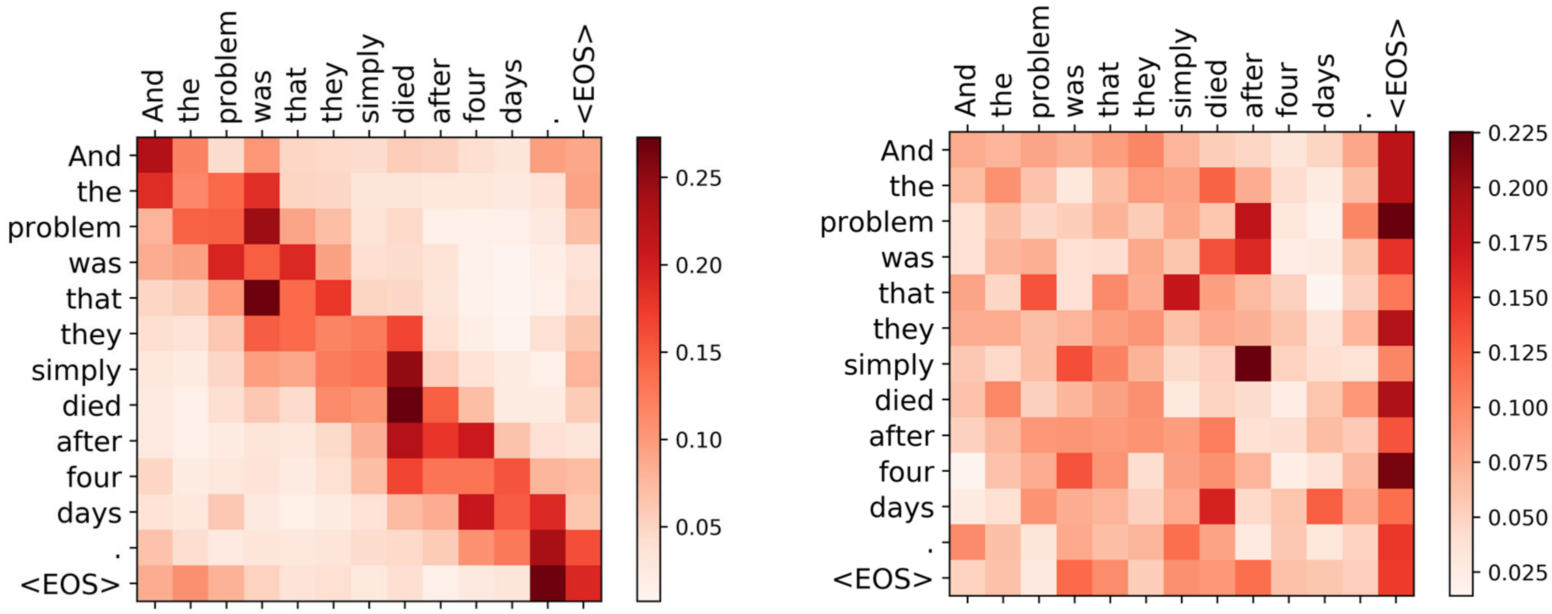
- Motivation: original attention modules must extract all the features with the same architecture, requiring a large capacity.
- Specialization works great in hardware design when the resources are limited.
- LSRA follows a specialized two-branch design: left for global context, right for local information



Mobile Transformer Block (LSRA)

Mobile Transformer with LSRA

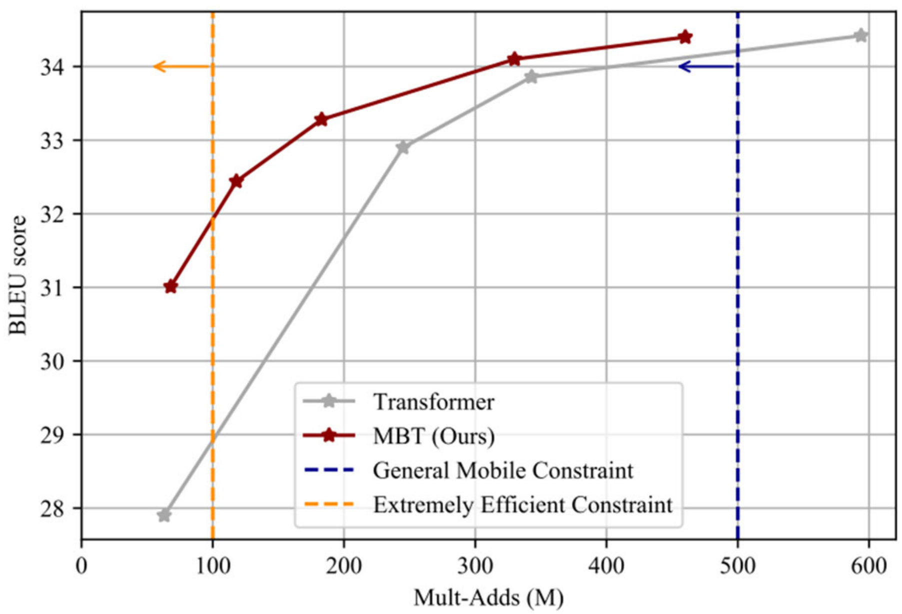
- Original self attention in original transformer needs to capture both local and global information (a).
- With LSRA, the attention branch only captures global contexts (b).



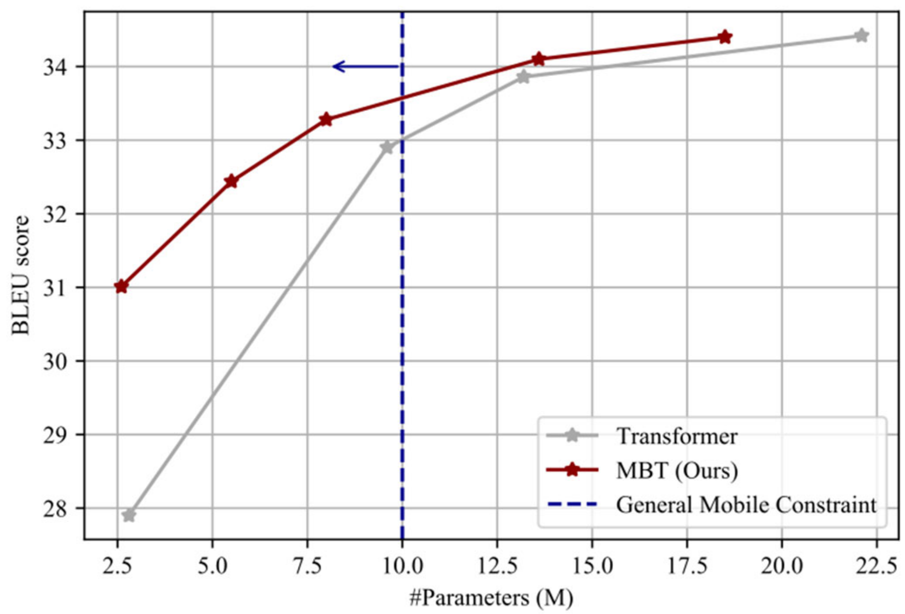
(a) Self Attention **(b) LSRA**

Mobile Transformer

- Our mobile transformer (MBT) with LSRA outperforms the basic transformer.
- On IWSLT'14 De-En dataset with better trade-off for both Mult-Adds and the number of parameters.



(a) IWSLT'14 De-En BLEU vs. Mult-Adds



(b) IWSLT'14 De-En BLEU vs. #Parameters

Mobile Transformer

- Our mobile transformer (MBT) also outperforms the basic transformer on both WMT'14 En-De and WMT'14 En-Fr dataset on mobile settings.
- Specialization is more effective with tighter resource constraints.

	#Parameters	Mult-Adds	WMT'14 En-De BLEU	WMT'14 En-De ΔBLEU	WMT'14 En-Fr BLEU	WMT'14 En-Fr ΔBLEU
Transformer (Vaswani et al., 2017)	2.8M	87M	21.3	–	33.6	–
Mobile Transformer (Ours)	2.9M	90M	**22.5**	**+1.2**	**34.9**	**+1.3**
Transformer (Vaswani et al., 2017)	11.1M	338M	25.1	–	37.6	–
Mobile Transformer (Ours)	11.7M	360M	**25.8**	**+0.7**	**38.7**	**+1.1**
Transformer (Vaswani et al., 2017)	17.3M	527M	26.1	–	38.4	–
Mobile Transformer (Ours)	17.3M	527M	**26.5**	**+0.4**	**39.6**	**+1.2**

Mobile Transformer

- Even compared to Neural Architecture Search-based Evolved Transformer (ET) [1], MBT offers 0.5 and 0.4 more BLEU score under the 100M and 400M mobile settings.
- It saves the design cost by 20000× in CO2 emission and the 250 GPU years of searching.

	#Parameters	Mult-Adds	BLEU	GPU Hours	CO_2e (lbs)	Cloud Computation Cost
Transformer	2.8M	87M	21.3	8×12	26	\$68 - \$227
ET (So et al., 2019)	3.0M	94M	22.0	8×274K	626K	\$1.6M - \$5.5M
Mobile Transformer (Ours)	2.9M	90M	**22.5**	8×14	32	\$83 - \$278
Transformer	11.1M	338M	25.1	8× 16	36	\$93.9 - \$315
ET (So et al., 2019)	11.8M	364M	25.4	8× 274K	626K	\$1.6M - \$5.5M
Mobile Transformer (Ours)	11.7M	360M	**25.8**	8 × 19	43	\$112 - \$376

[1] So, David, Quoc Le, and Chen Liang. "The Evolved Transformer." ICML, 2019.

Summary

- We analyze the computation bottleneck structure and argue that the bottleneck design is not optimal for 1-D attention.
- We propose a novel specialized multi-branch feature extractor, Long-Short Range Attention (LSRA) and a Mobile Transformer (MBT) based on LSRA.
- It alerts us to rethink the practicality of AutoML in terms of design cost.
- The efficient natural language processing designed for mobile settings is vital for the deployment of language related applications, such as machine translation on the edge devices.

Thank you!

Hardware, **A**I and **N**eural-nets

hanlab@mit.edu