Lite Transformer with Long-Short Range Attention

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

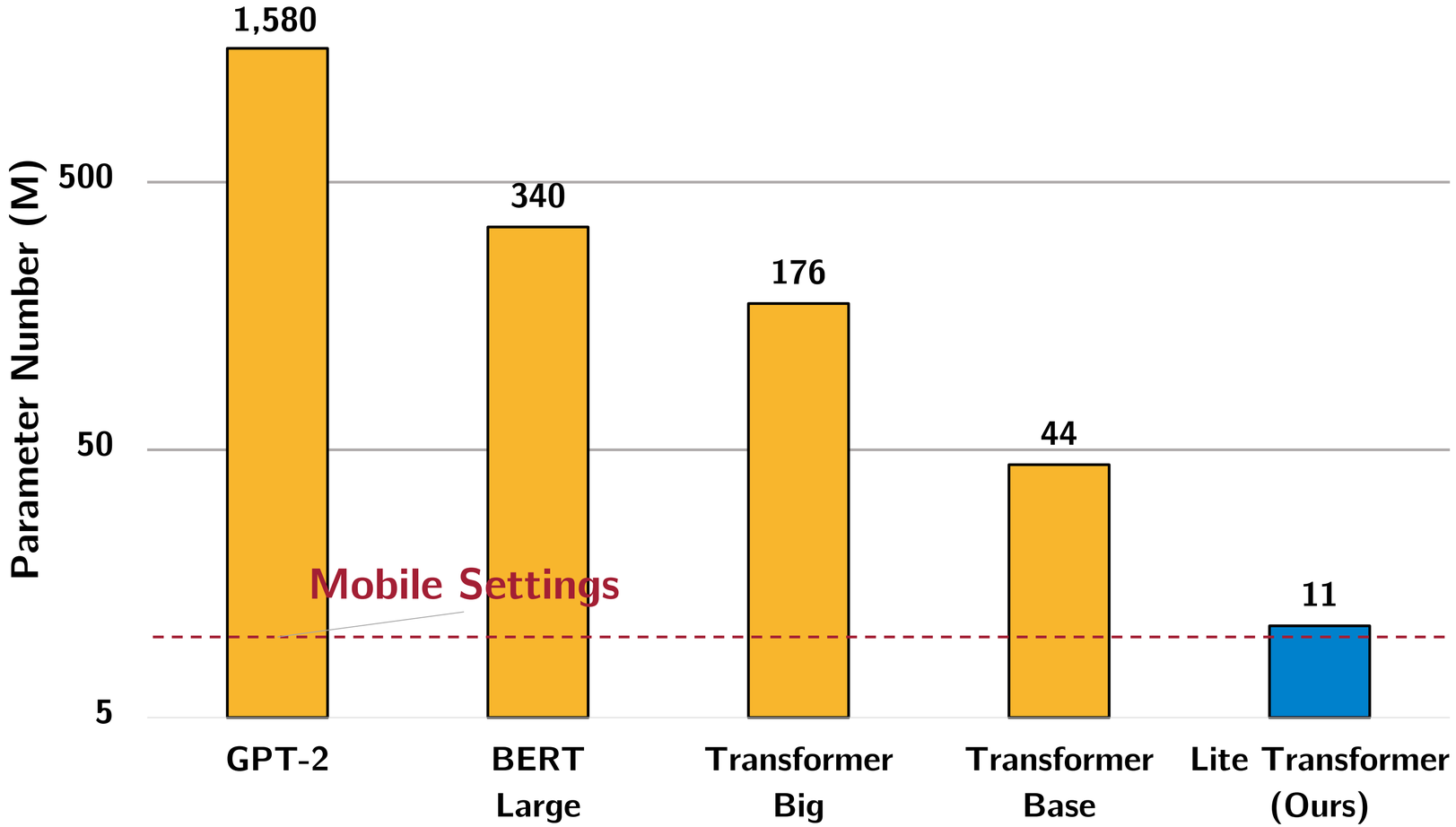


Figure: Parameter numbers of modern NLP models

- NLP models are **huge** — much larger than mobile settings

AutoML is EXPENSIVE

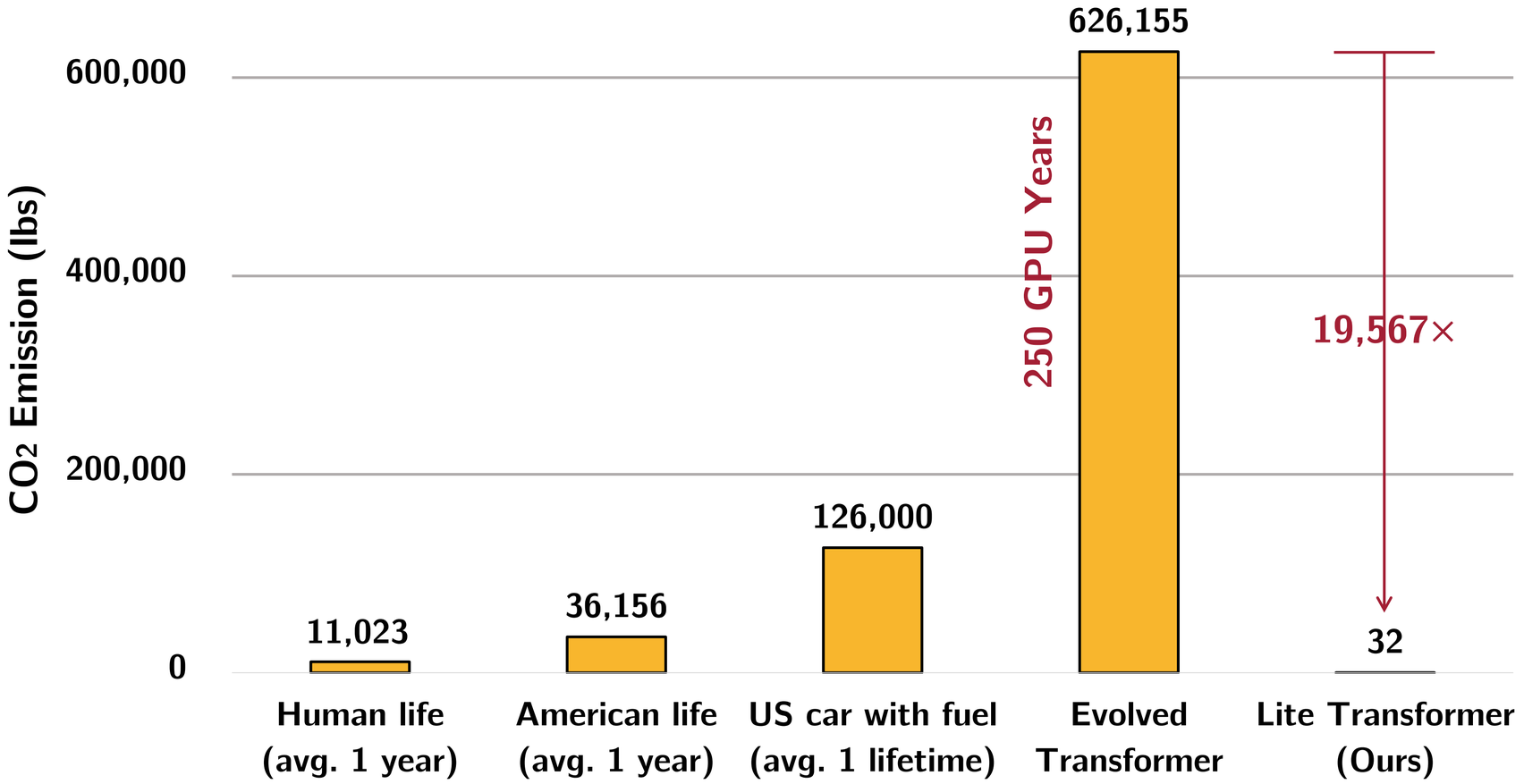
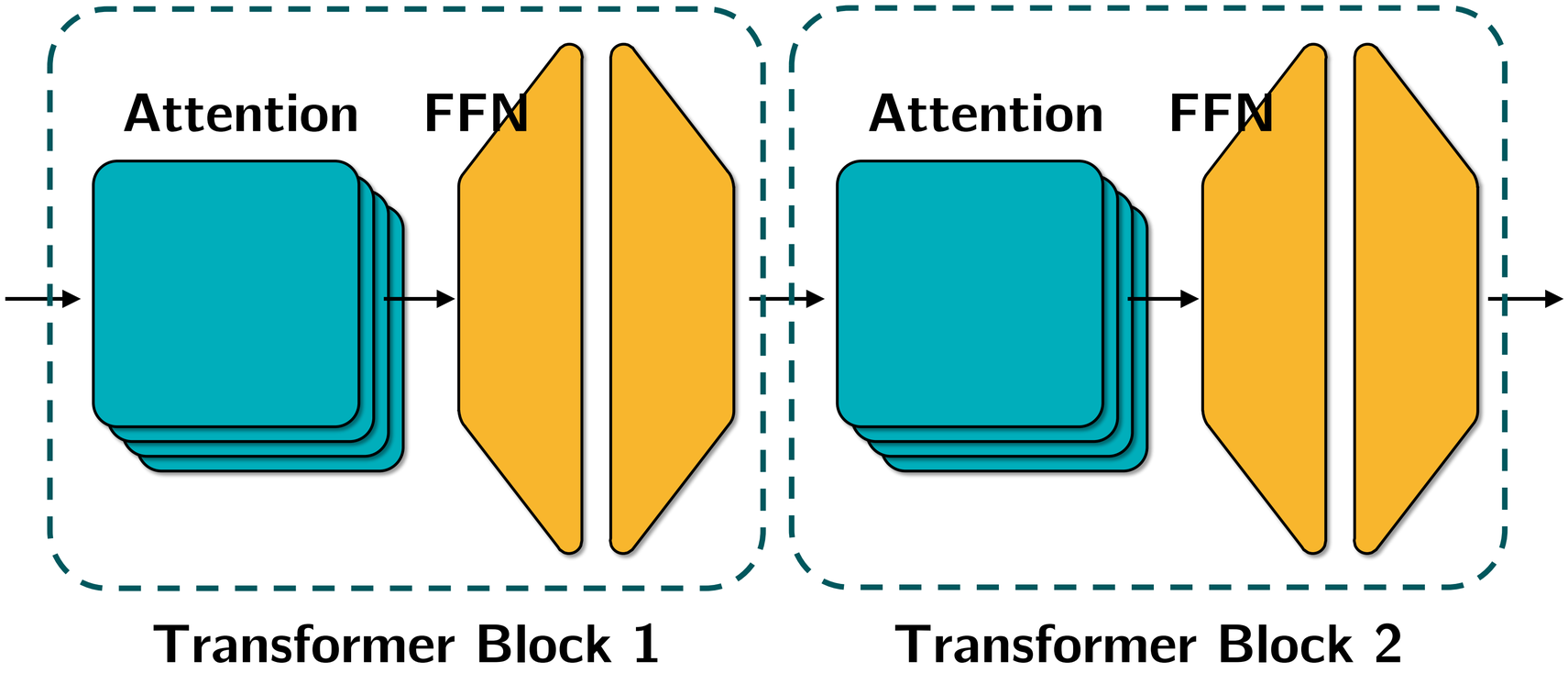


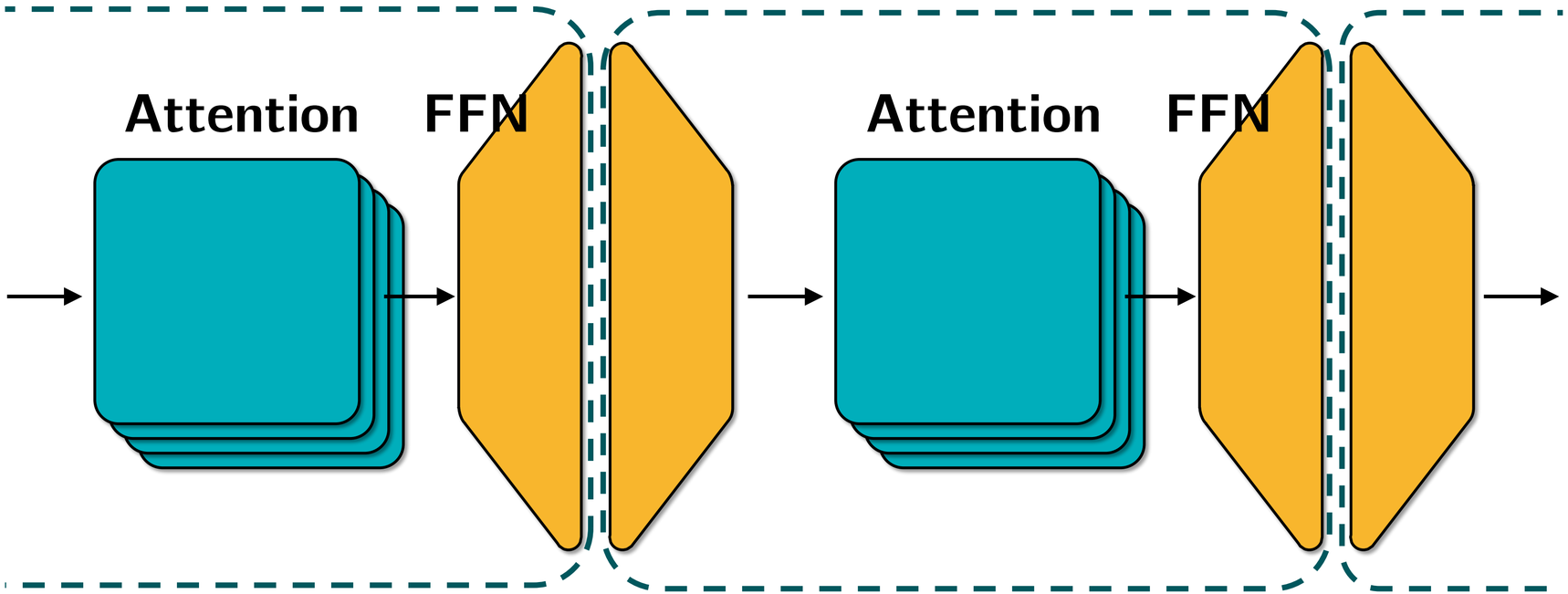
Figure: The design cost measured in CO_2 emission (lbs)

- Auto-ML's **huge searching cost** raises environmental concerns on CO_2.

Transformer Framework

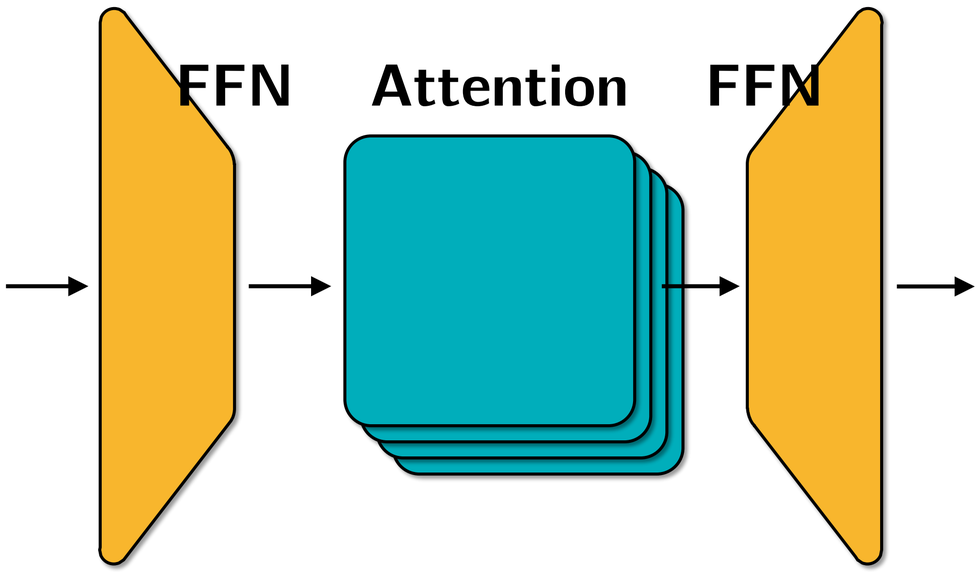


A Different View

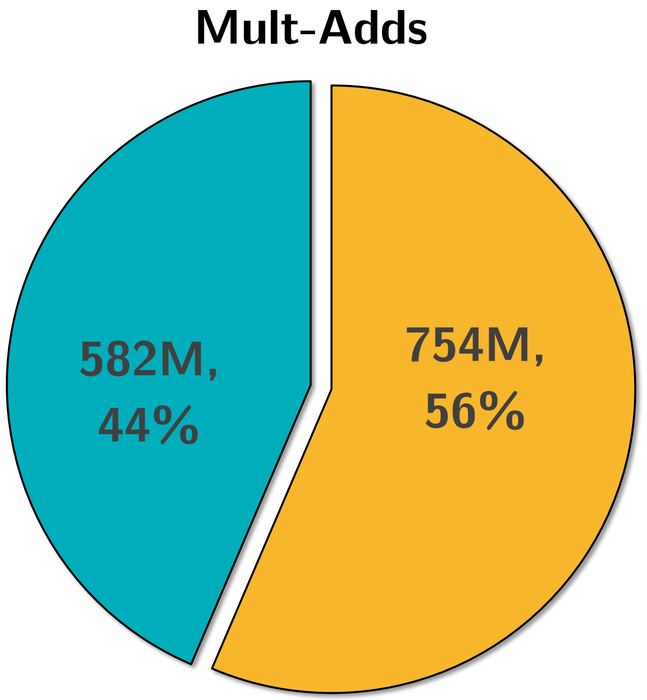


A Different View of Transformer Block

Is Bottleneck Effective for 1-D Attention?



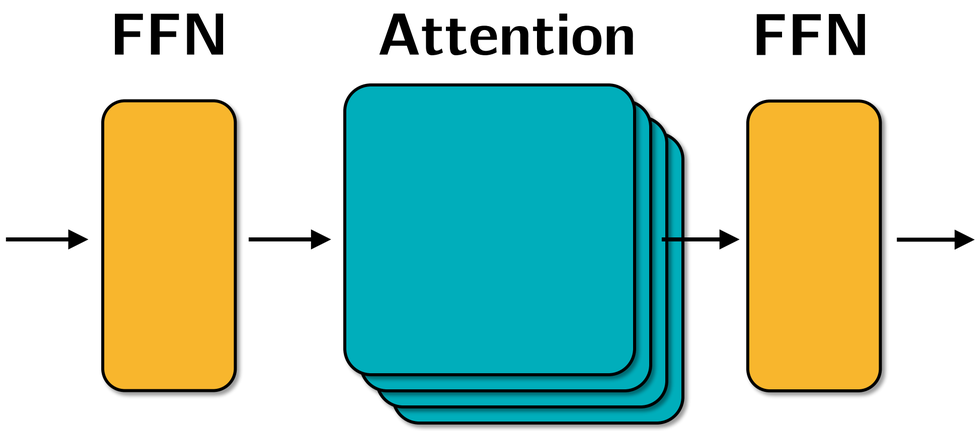
Base Transformer Block



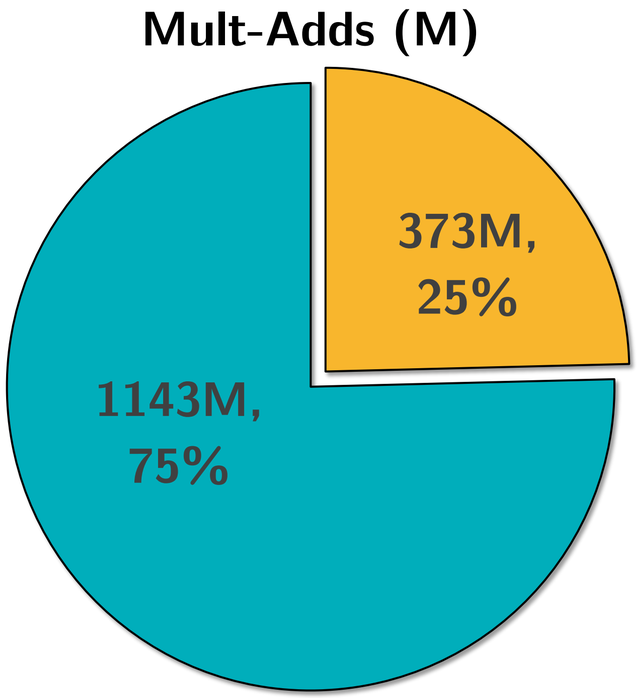
Computation Proportion

- The FFN takes **more than a half** of the computation.

Flattened Transformer



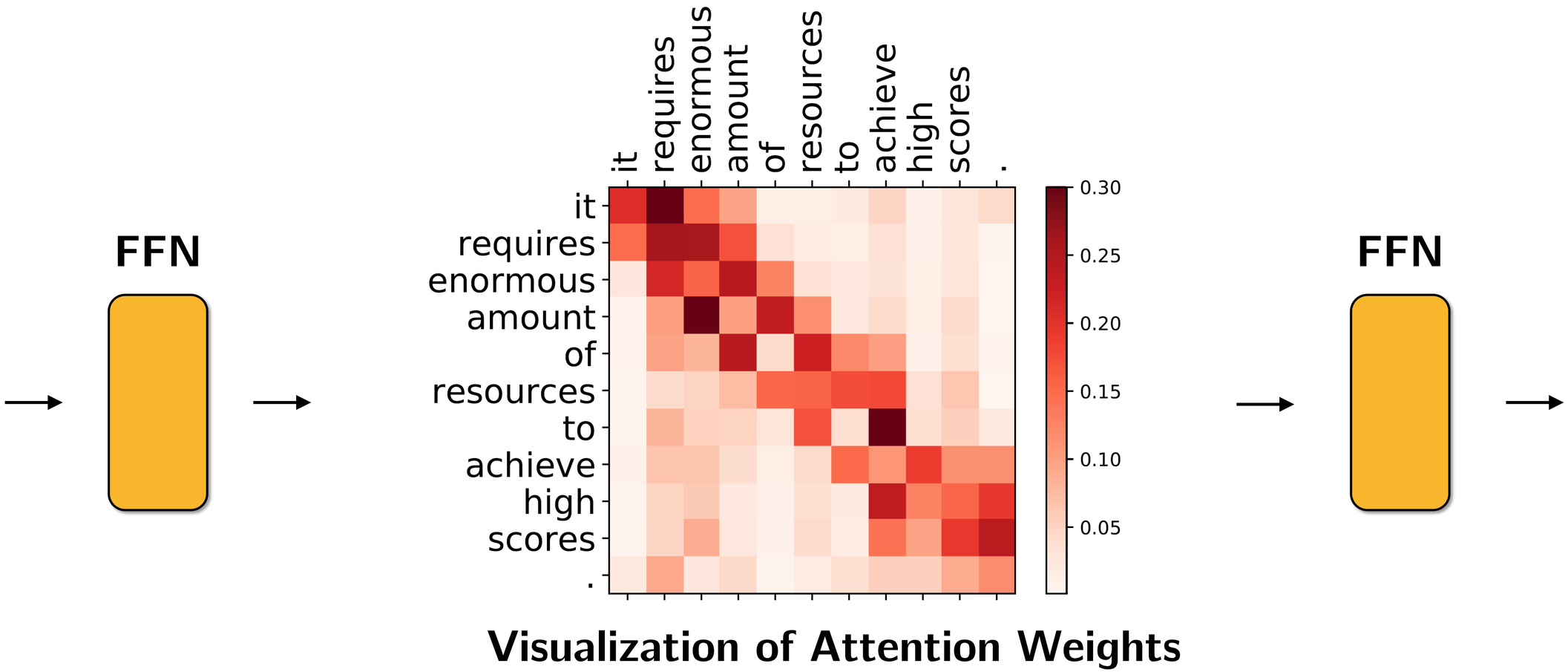
Flattened Transformer Block



Computation Proportion

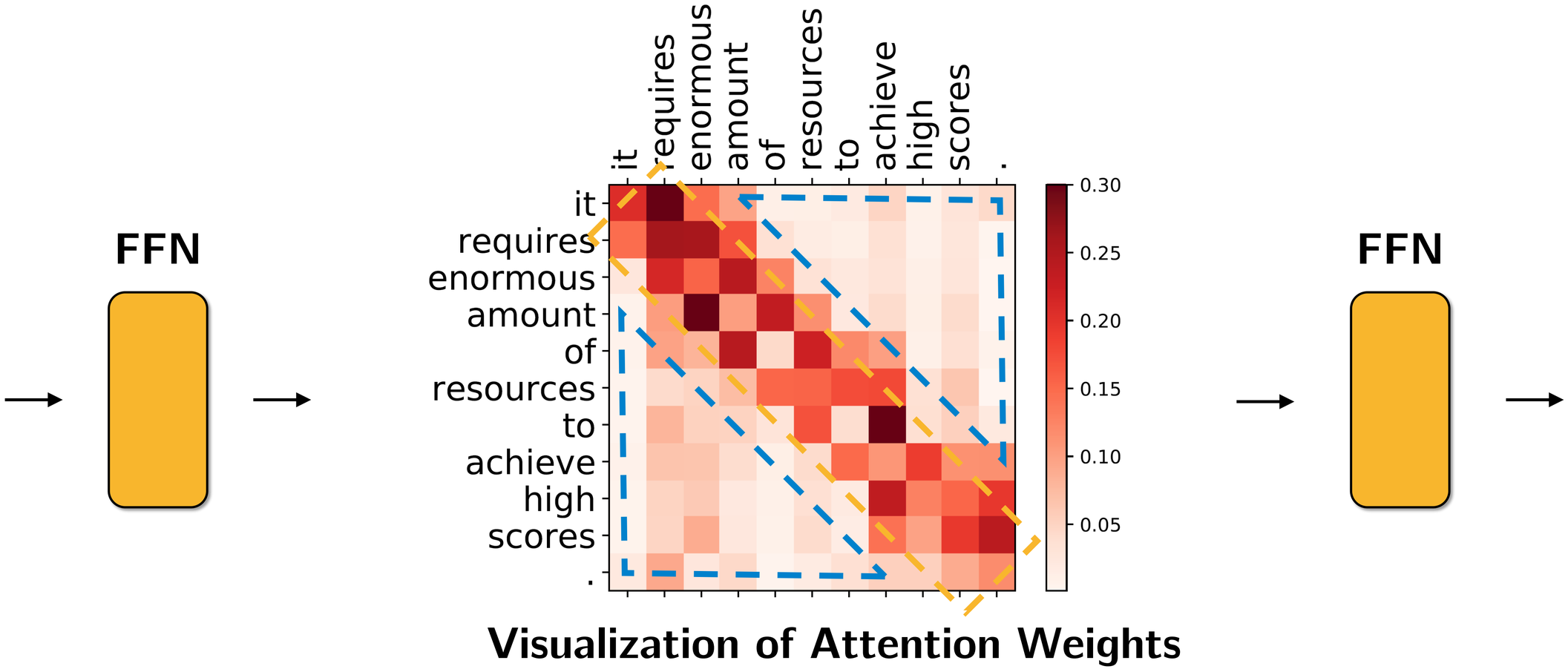
- Attentions take **major computation**, leaving larger space for optimization.

What does Attention Learn?



Visualization of Attention Weights

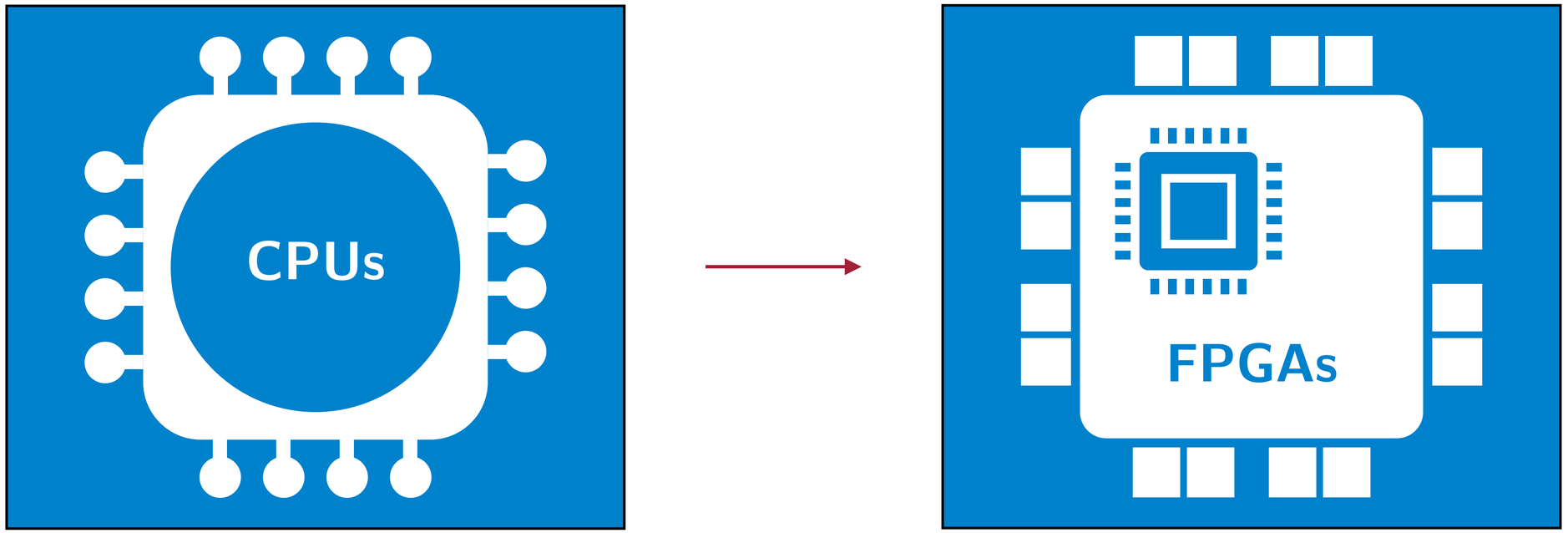
What does Attention Learn?



Visualization of Attention Weights

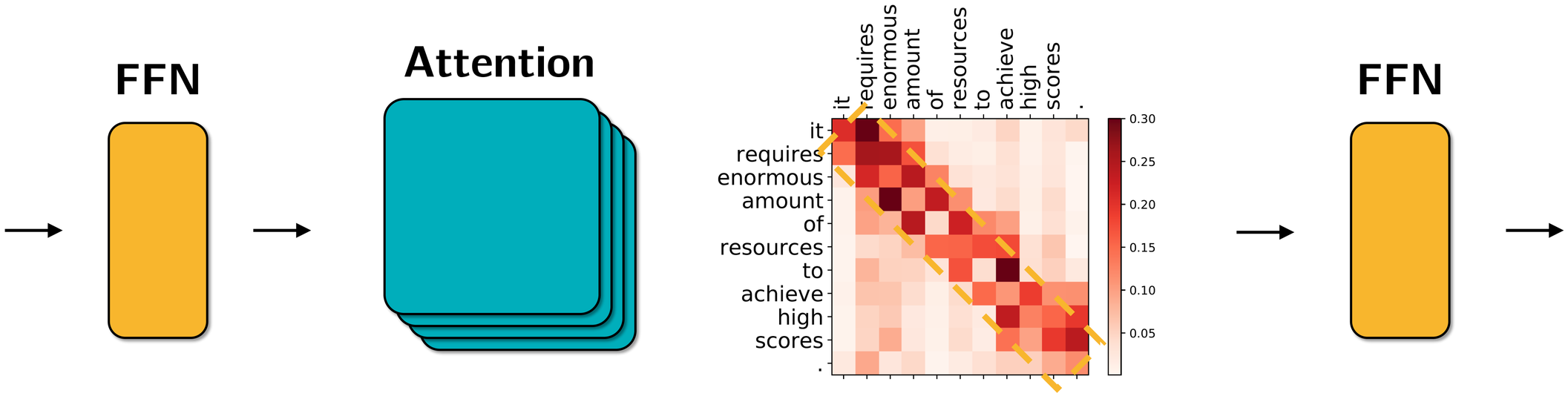
- Attention captures both **sparse global** context and **diagonal local** information.

Motivated by Hardware Design

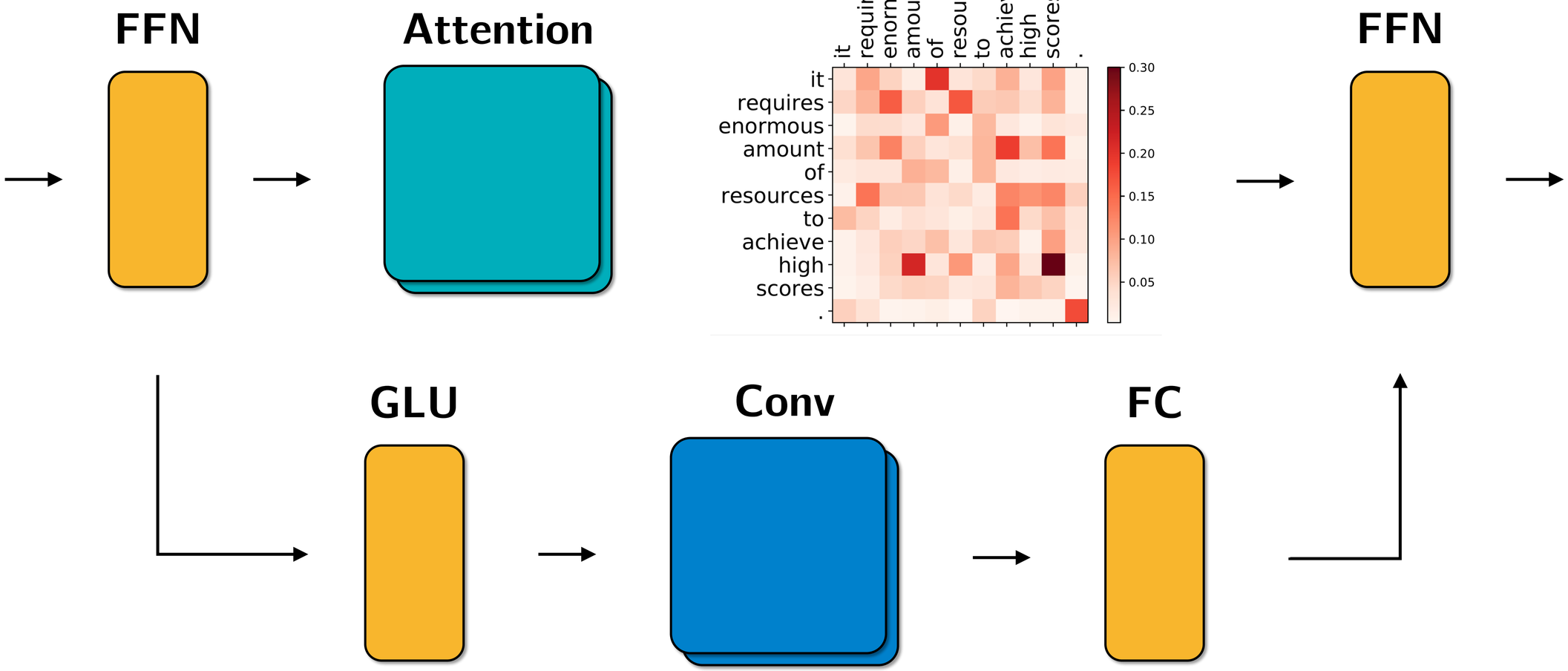


- **Specialization** is the key in efficient hardware design (e.g. FPGA accelerators)

Base Transformer is Redundant

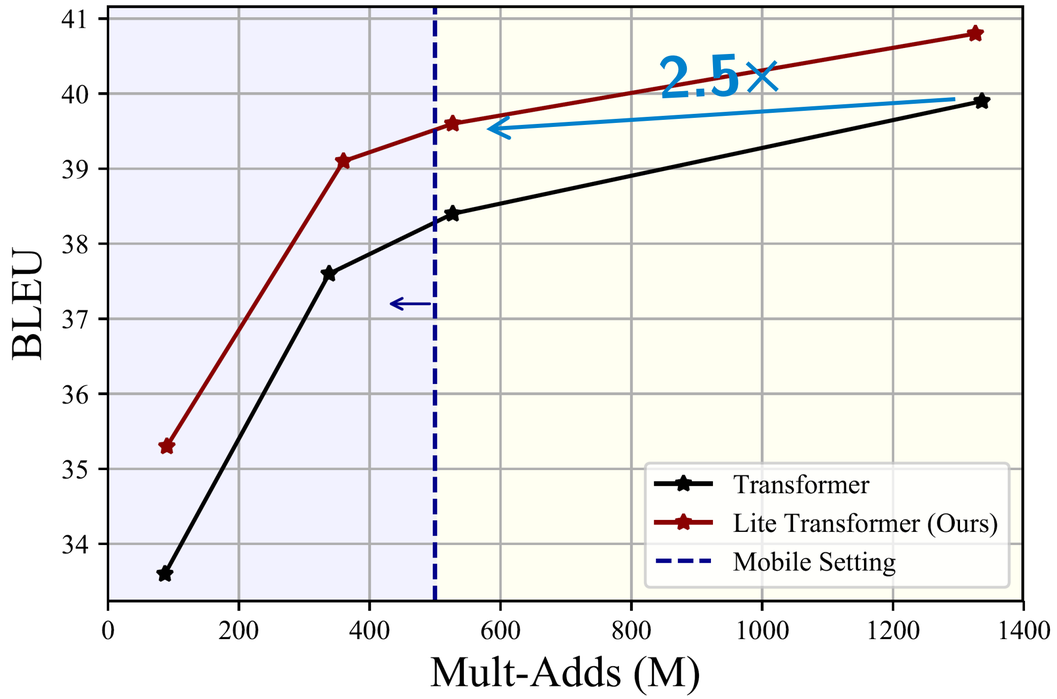


Long-Short Range Attention (LSRA)

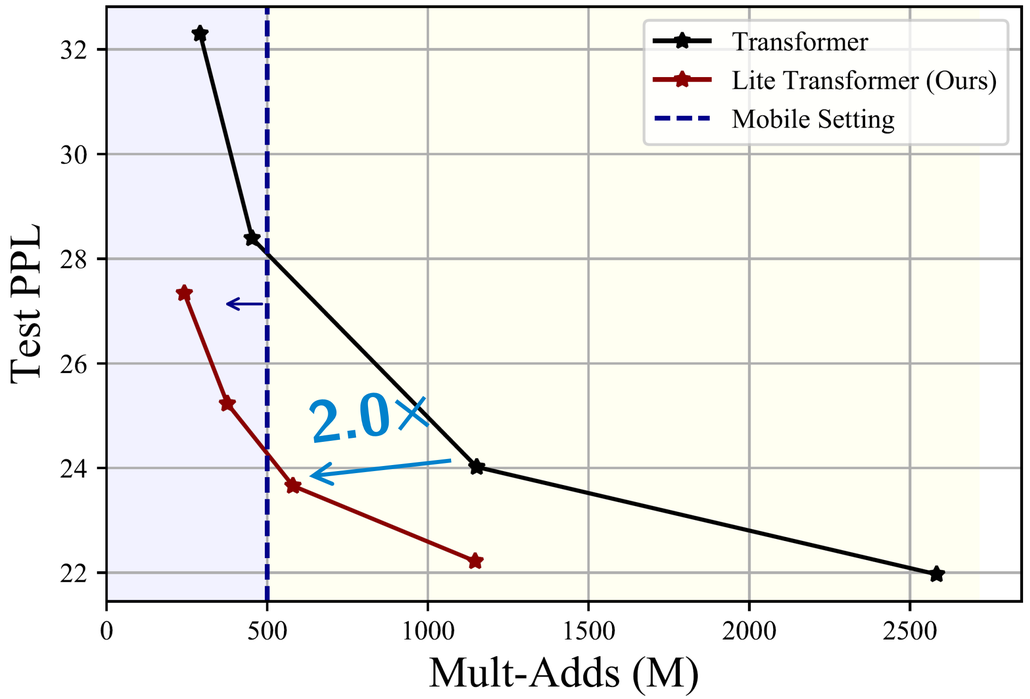


Lite Transformer

- Our Lite Transformer performs well on **machine translation (a)**, **abstractive summarization**, and **language modeling (b)**.



(a) WMT'14 En-Fr



(b) WIKITEXT-103

Lite Transformer vs AutoML

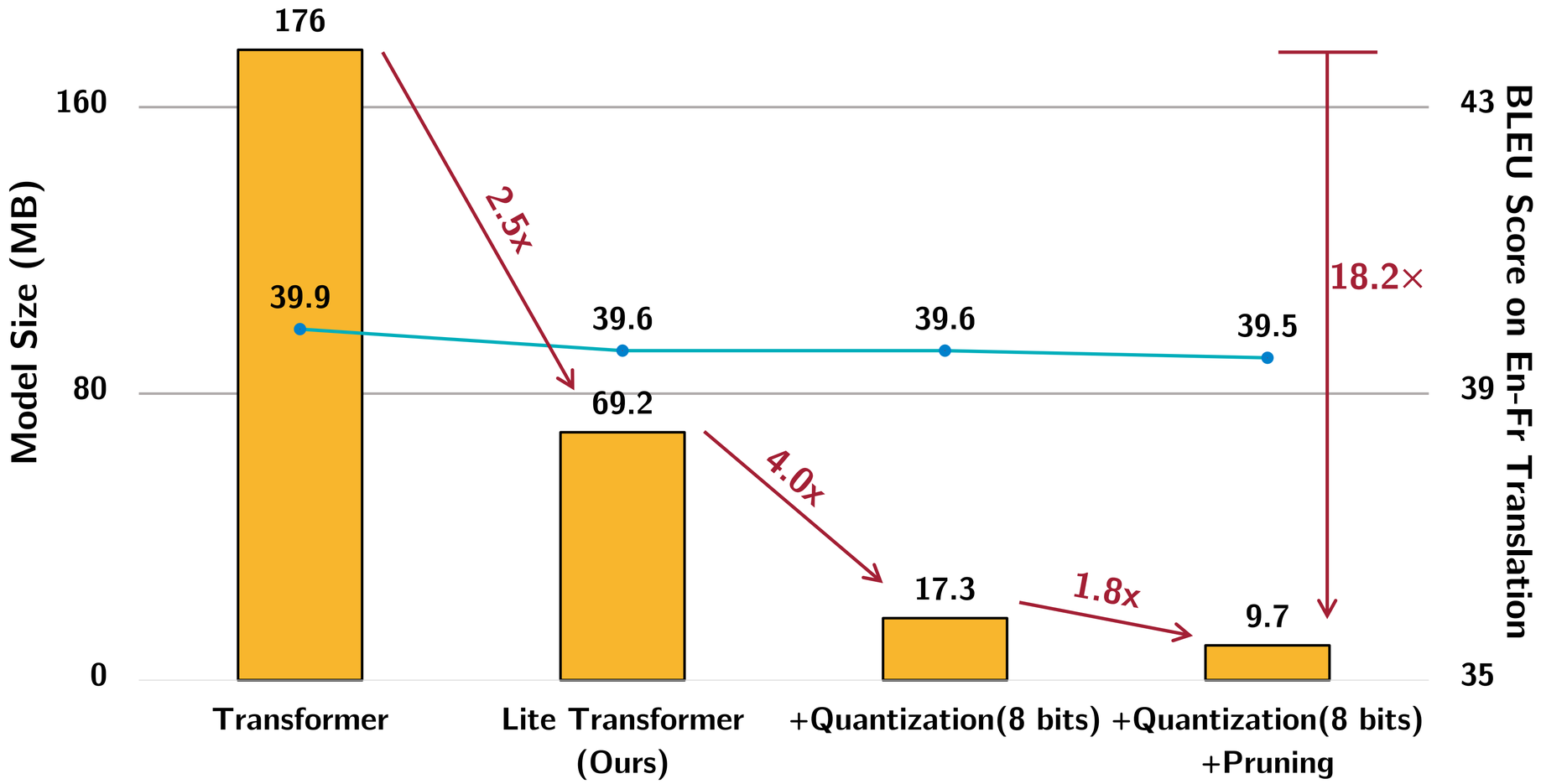
	#Params	#Mult-Adds	BLEU	GPU Hours	CO_2 Emission (lbs)	Cloud Cost ($)
Transformer	2.8M	87M	21.3	1.0×10^2	2.6×10^1	$\$2.3\times 10^2$
Evolved Transformer [AutoML]	3.0M	94M	22.0	2.2×10^6	6.3×10^5	$\$5.5\times 10^6$
Lite Transformer (Ours) [LSRA]	2.9M	90M	**22.5 (+0.5)**	$\mathbf{1.1\times 10^2}$	$\mathbf{3.2\times 10^1}$	$\mathbf{\$2.8\times 10^2}$

20000× Reduction

Better Performance

Further Compress Lite Transformer by 18.2x

- Our Lite Transformer is **orthogonal** to general model compression techniques.



Lite Transformer with Long-Short Range Attention