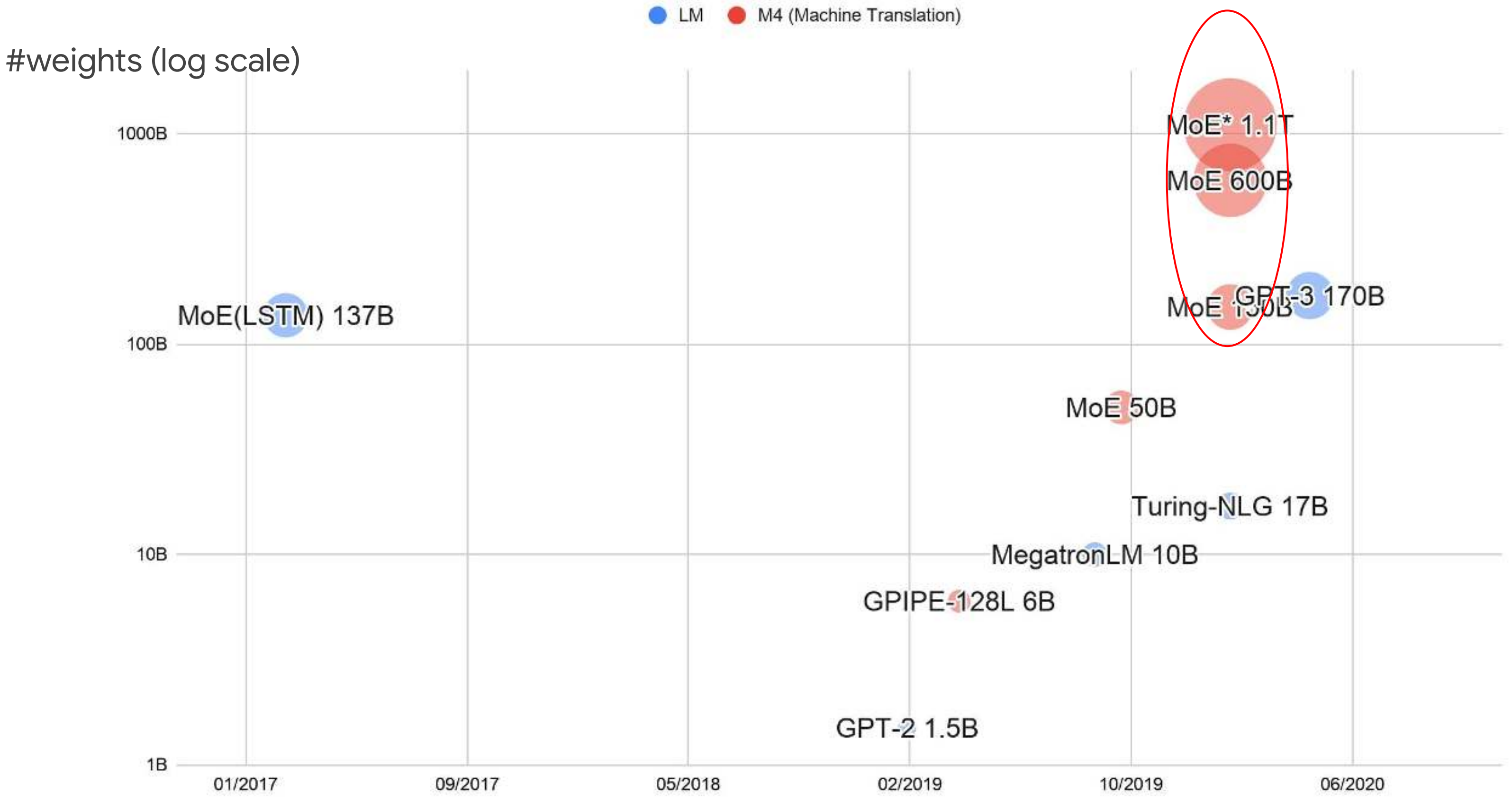
Big models



Google

Google Neural Machine Translation

Our goal

Develop a universal machine translation model (i.e. one model for all languages and domains)

*“Perhaps the way [of translation] is to descend, from each language, down to the common base of human communication -- the real but as yet **undiscovered universal language** -- and then re-emerge by whatever particular route is convenient.”*

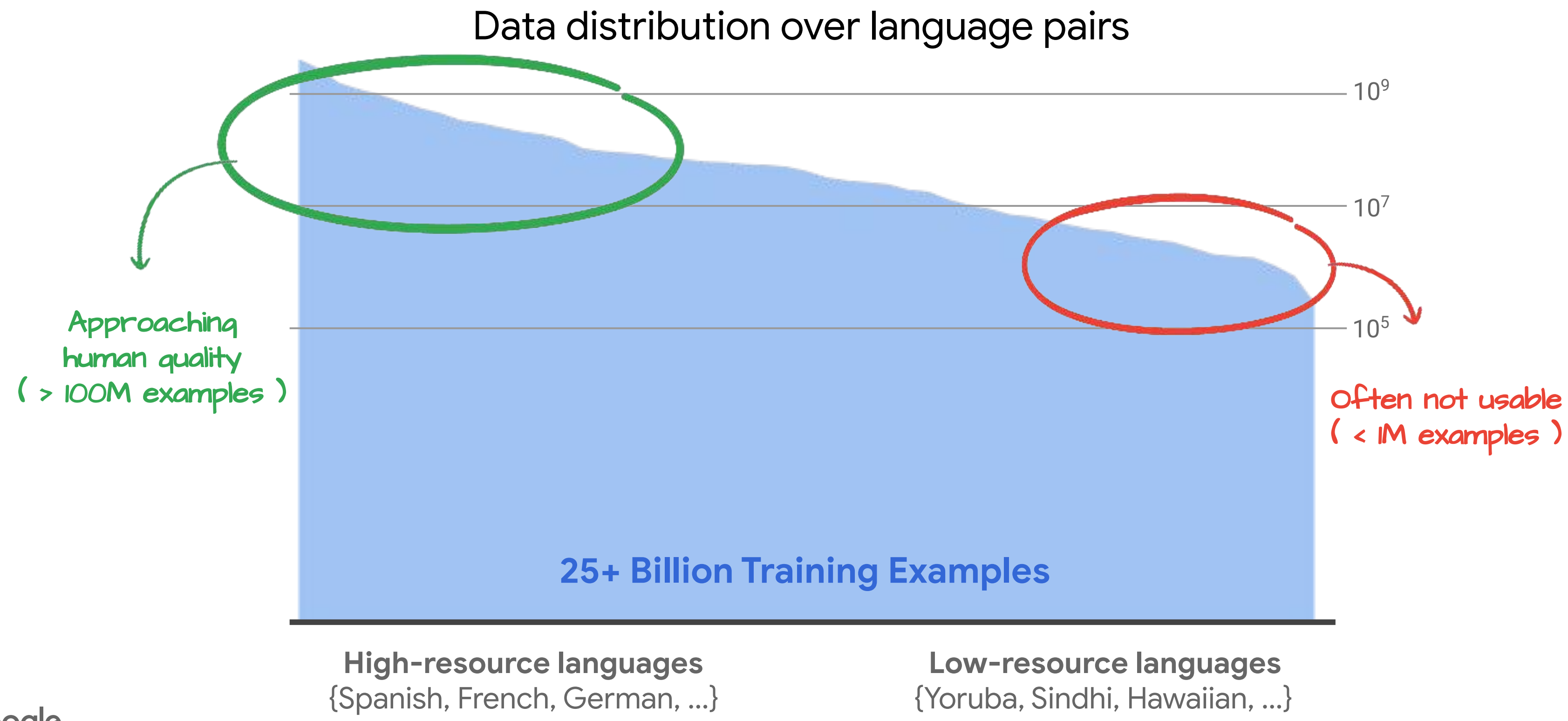
Warren Weaver (1949)

Google AI Blog

The latest news from Google AI

Exploring Massively Multilingual, Massive Neural Machine Translation

Motivation 1:
Improve translation quality for all language pairs

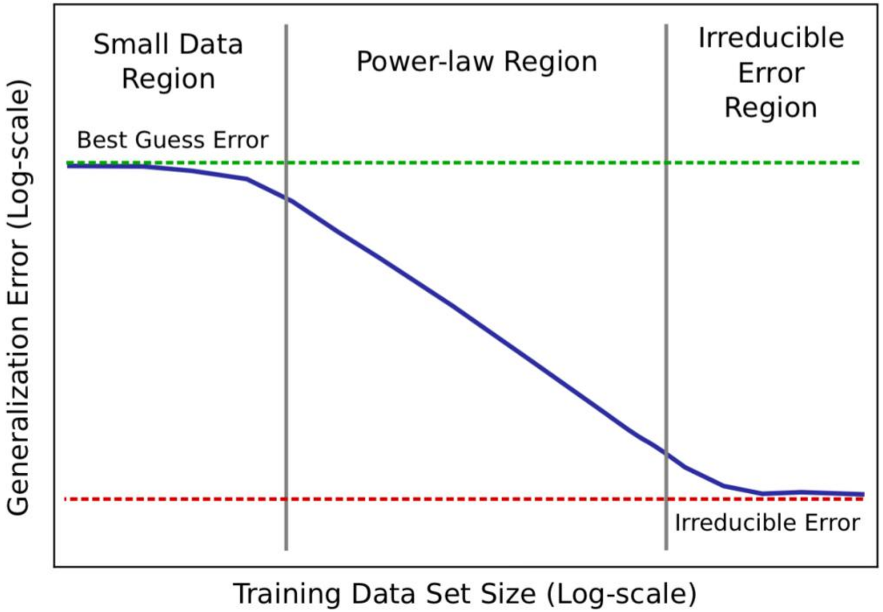


Motivation 2:
Expand language coverage

In the world, there are...	7,000+ Total languages	2,000+ African languages	700+ Native Am. languages[1]
But Translate only supports...	103 Total languages	11 African languages	0 Native Am. languages

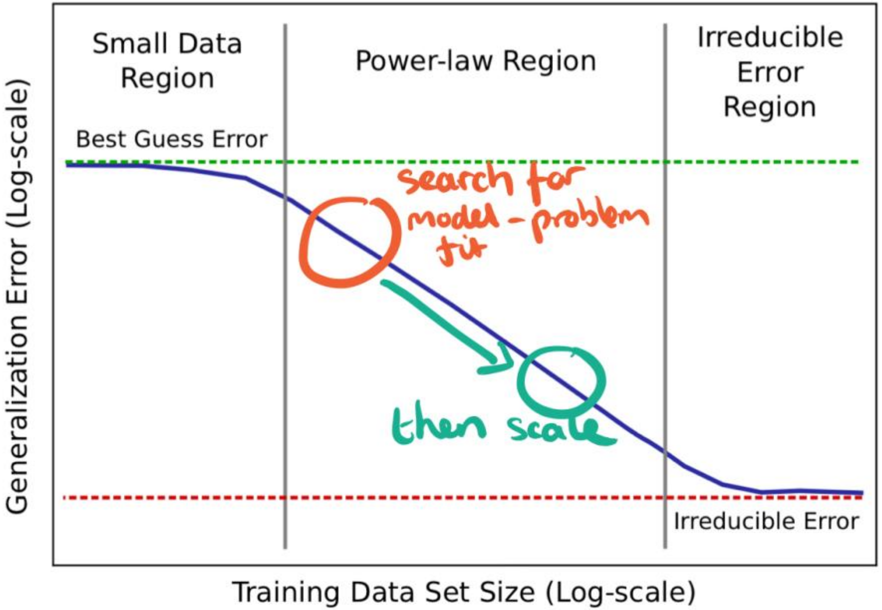
1.. Estimate is 766 Native Am languages globally: 383 in South America (source); 176 in central America (source); 207 in North America (source)

Motivation 3:
Neural network scaling and the new understanding of generalization



Hestness et al. 2018

Motivation 3:
Neural network scaling and the new understanding of generalization



Motivation 4:
This is a compelling test bed for ML research

Massive multilinguality requires advances in :

- Multi-task learning
- Meta-learning
- Continual learning

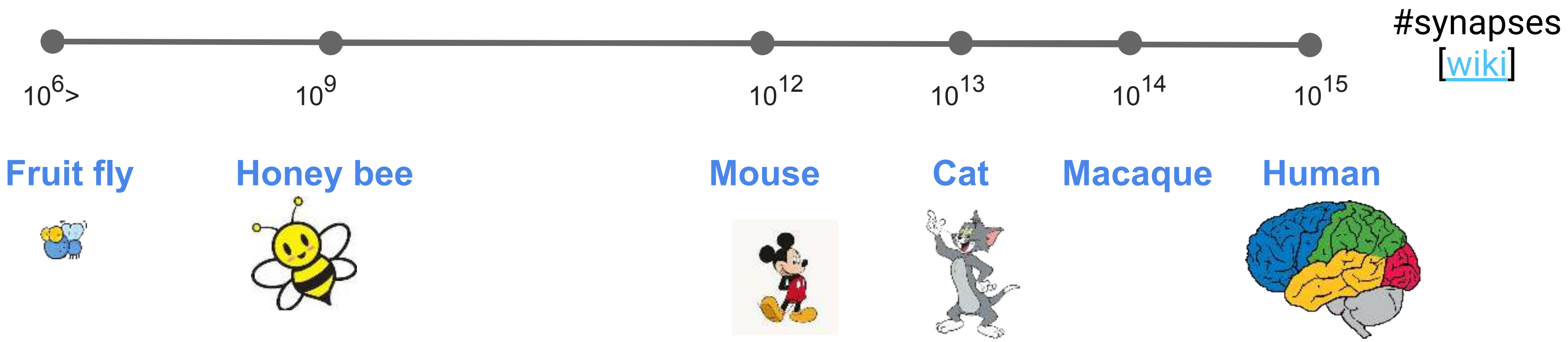
To achieve massive multilinguality, we need massive scale, requires advances in:

- **Model capacity**
- Trainability and optimization
- Efficiency improvements

Progress and Future

Google

Number of Synapses



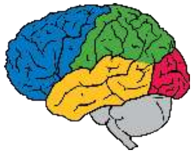
Number of Synapses

NMT with Attention
Resnet50
[25-50M]

10^6>	10^9	10^{12}	10^{13}	10^{14}	10^{15}	#synapses [wiki]
Fruit fly	**Honey bee**	**Mouse**	**Cat**	**Macaque**	**Human**	





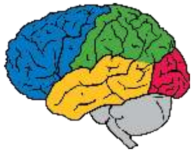


Number of Synapses

GNMT production model (2016) 200M

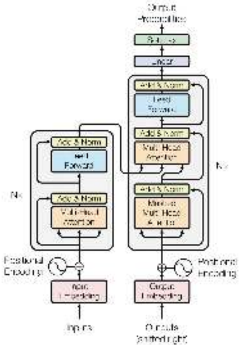
10^6> — 10^9 — 10^{12} — 10^{13} — 10^{14} — 10^{15}

#synapses [wiki]

Fruit fly	Honey bee	Mouse	Cat	Macaque	Human
					

Google AI

Number of Synapses



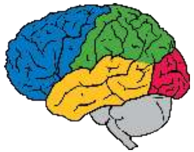
Transformer
[400M]

#synapses
[wiki]

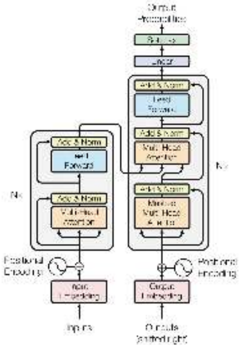
10^6>	10^9	10^12	10^13	10^14	10^15
Fruit fly	Honey bee	Mouse	Cat	Macaque	Human







Number of Synapses



Transformer
[400M]

Facebook ResNeXt101
Open AI GPT2
MSR ZeRO
NVidia Megatron-LM
Google-T5
[1-10B]

M4:
Massively Multilingual Massive Machine Translation
[80B]

10^6> | 10^9 | 10^{12} | 10^{13} | 10^{14} | 10^{15} | #synapses [wiki]

Fruit fly

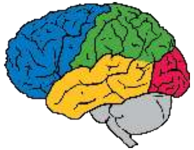
Honey bee

Mouse

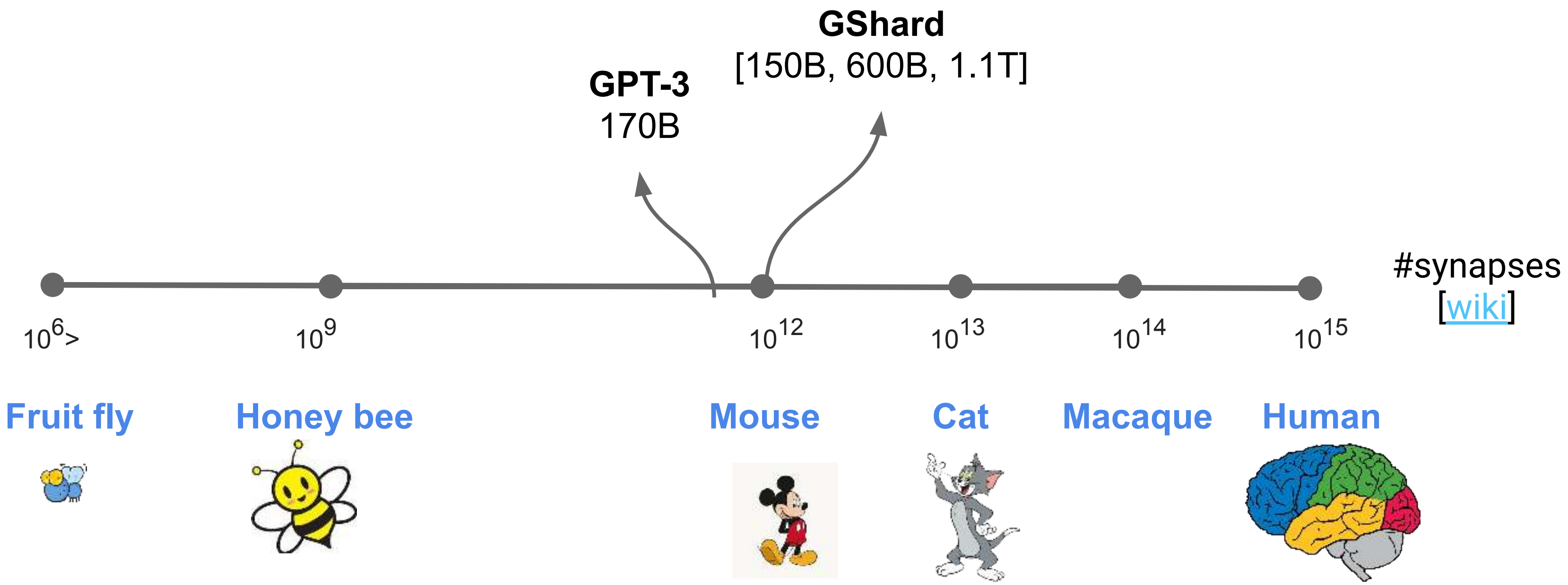
Cat

Macaque

Human

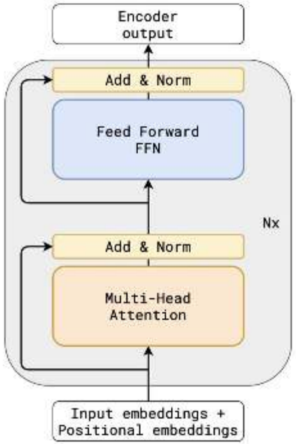


Number of Synapses



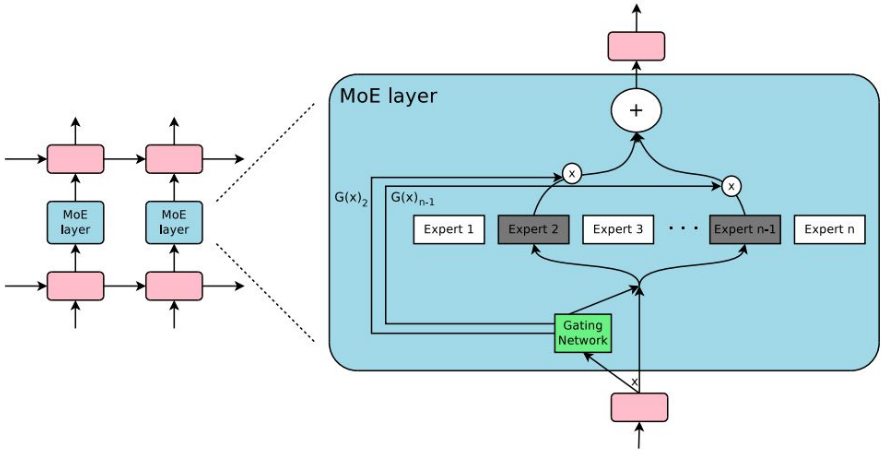
Transformer

Transfomer
Encoder



- Powerful
    - Cores to many SOTA results.
- Simple
    - Easy to express in linear algebra.
    - Reproduced many many times.
- Originally proposed in the [paper]

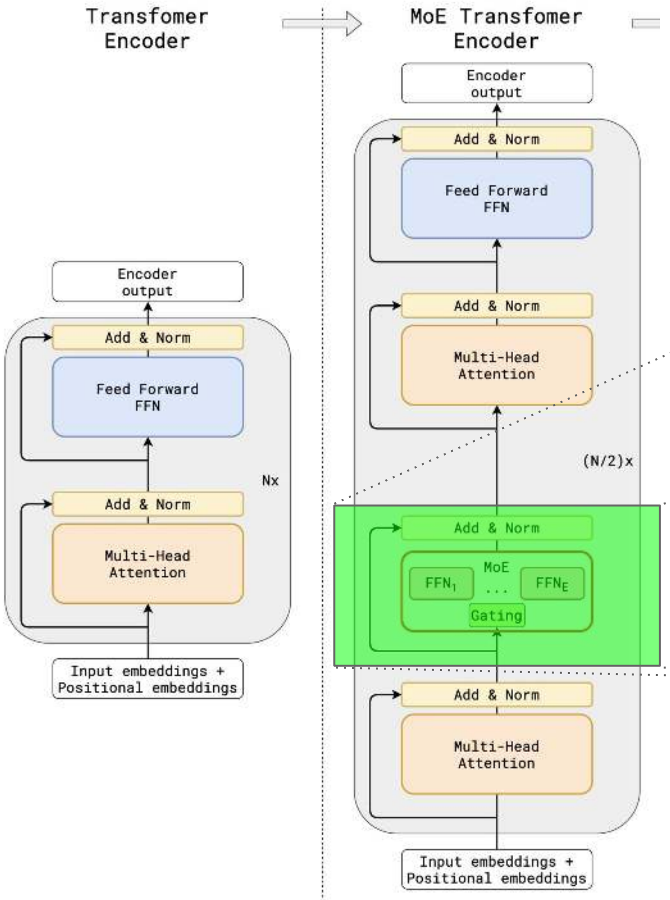
Mixture of Experts (MoE)



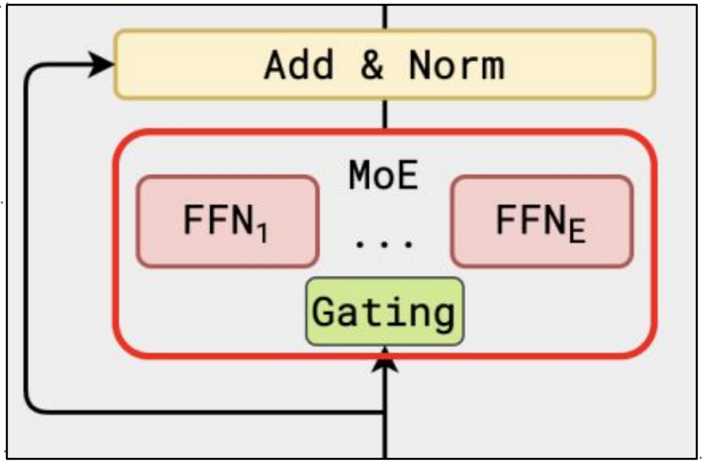
- Sparsely gated
  - Cost-effective inference
- Embarrassingly parallelizable
  - Nice to accelerators
- Originally proposed in this paper

Google

Mixture-of-Experts Transformer



Replace every-other FFN with an MoE layer

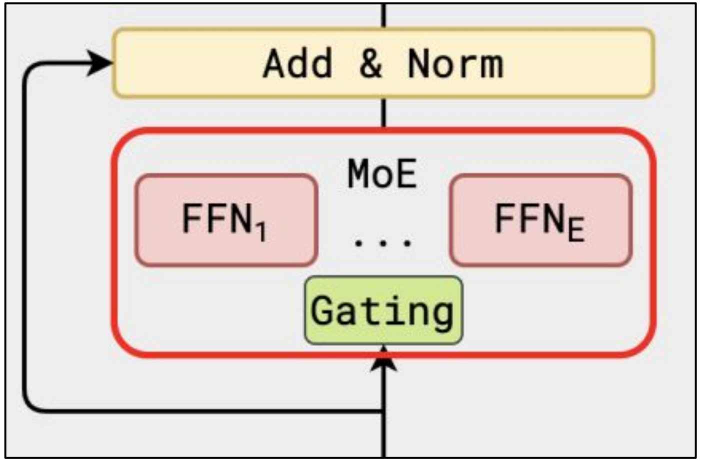


Position-wise Mixture-of-Experts Layer

x_s is the input token

$$\mathcal{G}_{s,E} = \text{GATE}(x_s)$$
$$\text{FFN}_e(x_s) = wo_e \cdot \text{ReLU}(wi_e \cdot x_s)$$
$$y_s = \sum_{e=1}^{E} \mathcal{G}_{s,e} \cdot \text{FFN}_e(x_s)$$

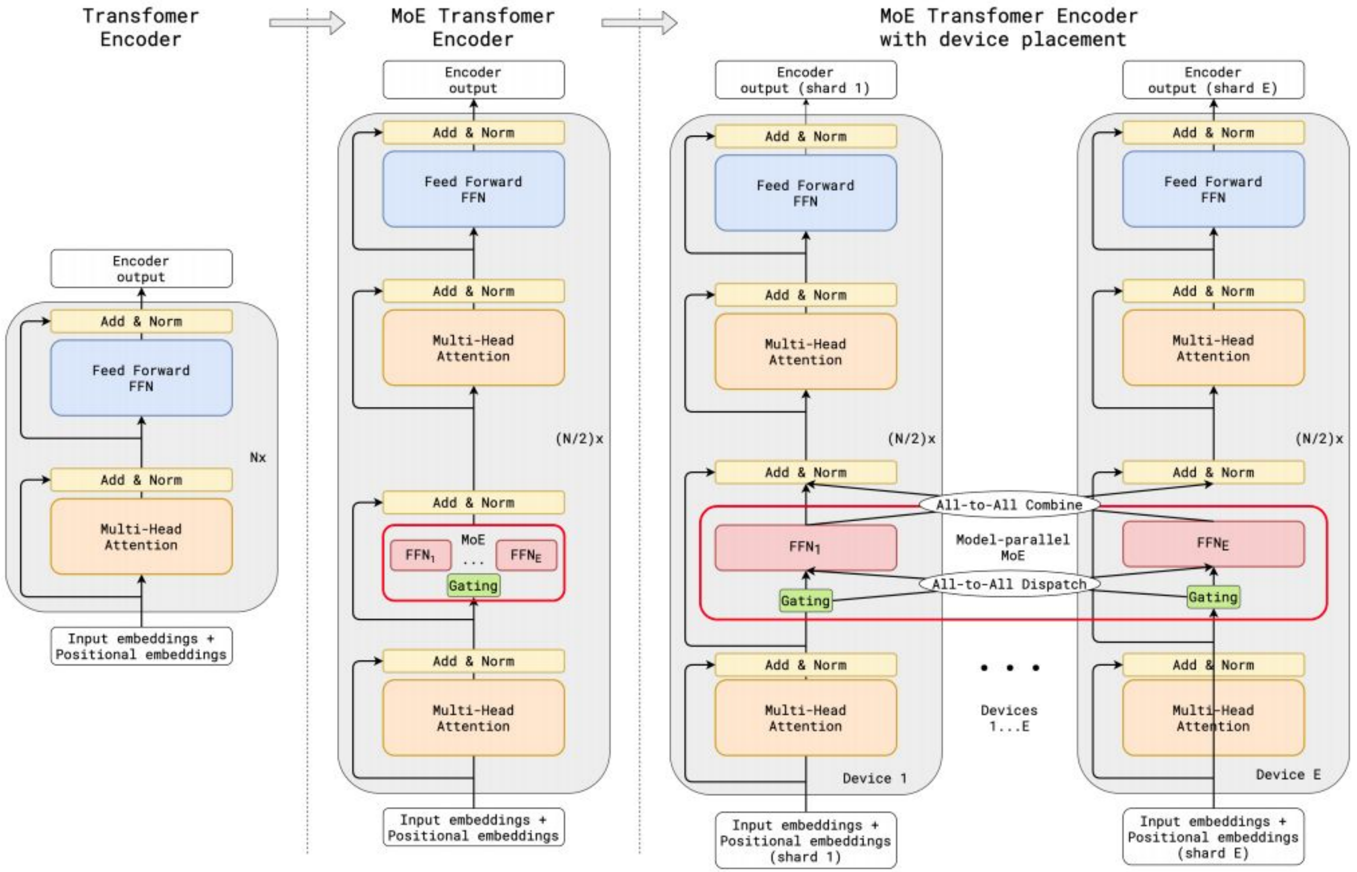
E feed-forward networks $\text{FFN}_1 \ldots \text{FFN}_E$



$\mathcal{G}_{s,E}$ is computed by a gating network.
y_s, is the weighted average

Algorithm details

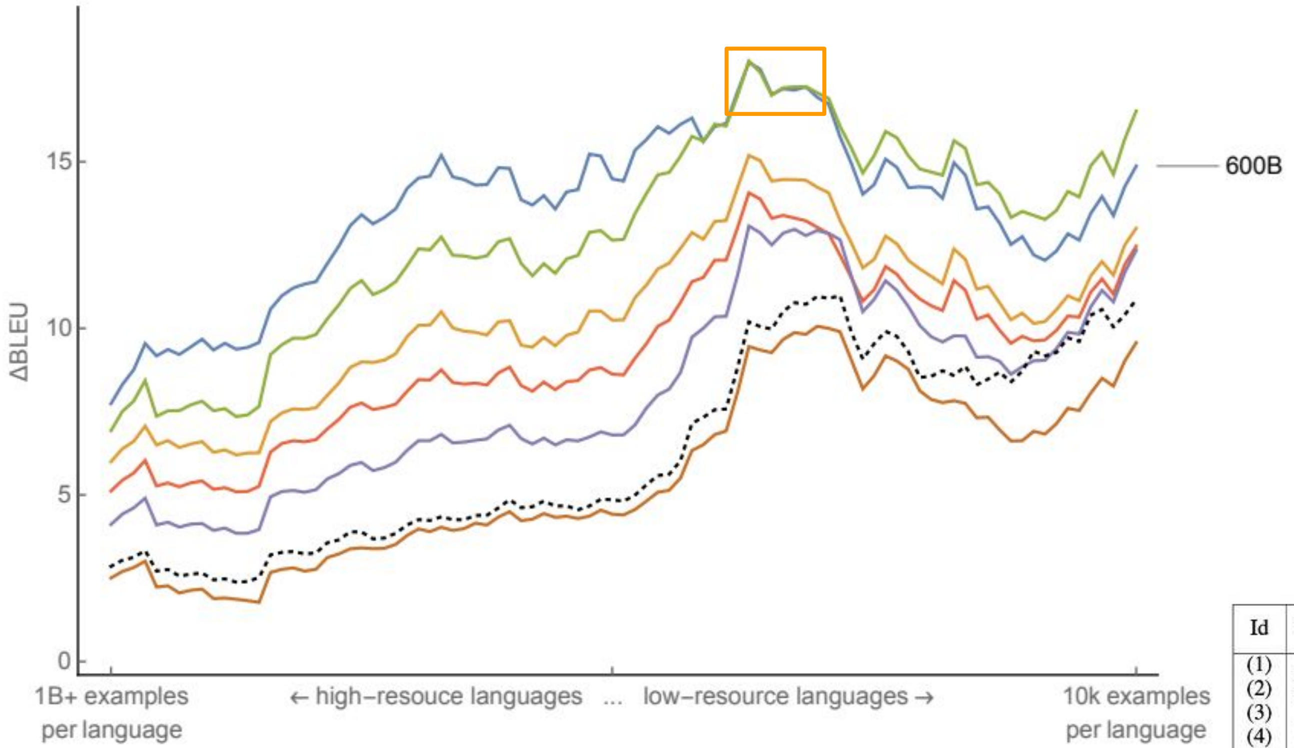
- **Gate function written in linear algebra**
  - Easy to express in a sequential program
- **Experts load balancing during training**
  - Auxiliary loss helps
- **Uniform routing during warming up phase**
- **Random second expert dispatch**
- **Flat beam search for inference**



Transfomer Encoder
MoE Transfomer Encoder
MoE Transfomer Encoder with device placement
Encoder output
Add & Norm
Feed Forward FFN
Multi-Head Attention
Nx
Input embeddings + Positional embeddings
(N/2)x
MoE
FFN_1
...
FFN_E
Gating
Encoder output (shard 1)
Encoder output (shard E)
All-to-All Combine
Model-parallel MoE
All-to-All Dispatch
Devices 1...E
Device 1
Device E
Input embeddings + Positional embeddings (shard 1)
Input embeddings + Positional embeddings (shard E)

M4 ΔBLEU

Google Research



Id	Model	BLEU avg.	ΔBLEU avg.	Weights
(1)	MoE(2048E, 36L)	**44.3**	13.5	600B
(2)	MoE(2048E, 12L)	41.3	10.5	200B
(3)	MoE(512E, 36L)	43.7	12.9	150B
(4)	MoE(512E, 12L)	40.0	9.2	50B
(5)	MoE(128E, 36L)	39.0	8.2	37B
(6)	MoE(128E, 12L)	36.7	5.9	12.5B
*	T(96L)	36.9	6.1	2.3B
*	Baselines	30.8	-	100×0.4B

Quality vs. Cost