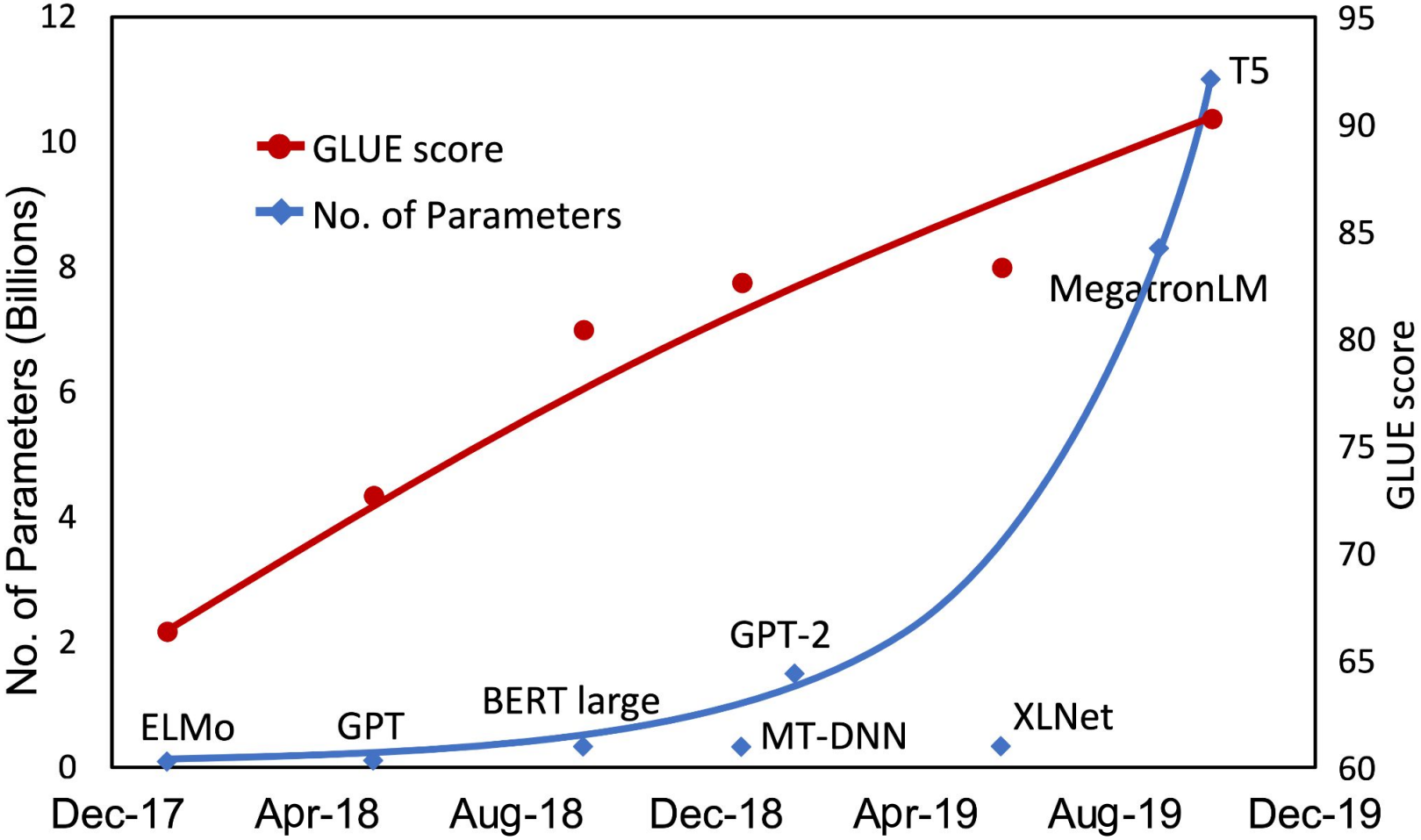
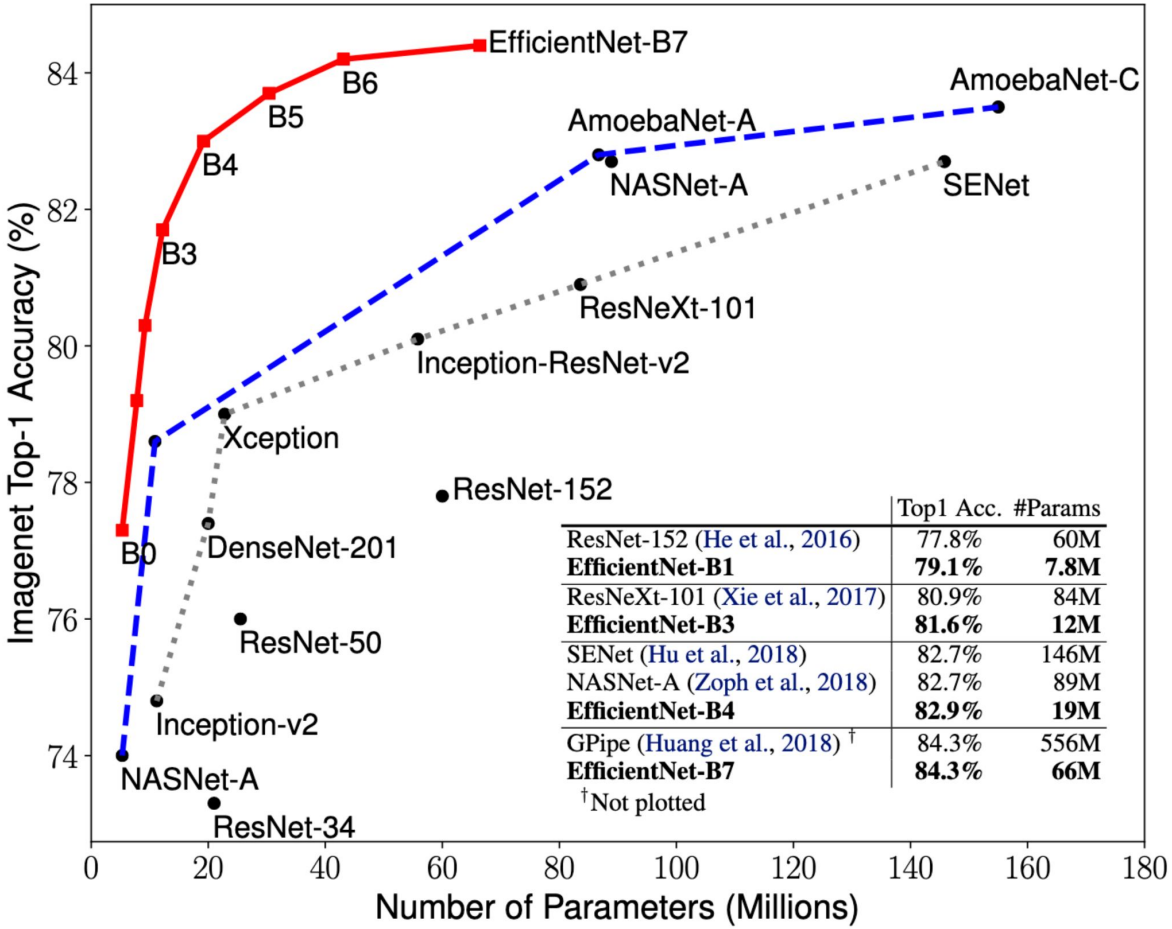
Scaling up models and data

Colin Raffel

CIFAR Deep Learning & Reinforcement Learning Summer School



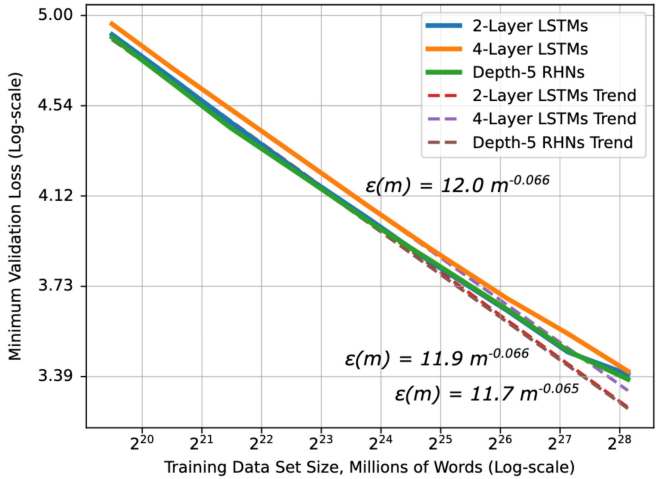
From "Real-Time Social Media Analytics with Deep Transformer Language Models: A Big Data Approach" by Ahmet and Abdullah

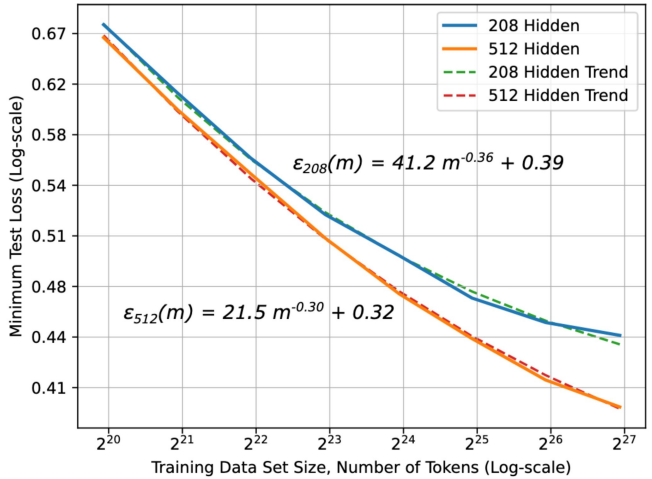


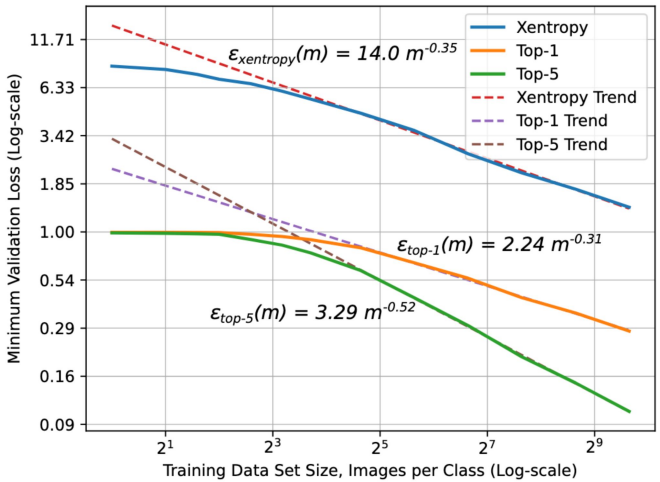
	Top1 Acc.	#Params
ResNet-152 (He et al., 2016)	77.8%	60M
EfficientNet-B1	**79.1%**	**7.8M**
ResNeXt-101 (Xie et al., 2017)	80.9%	84M
EfficientNet-B3	**81.6%**	**12M**
SENet (Hu et al., 2018)	82.7%	146M
NASNet-A (Zoph et al., 2018)	82.7%	89M
EfficientNet-B4	**82.9%**	**19M**
GPipe (Huang et al., 2018) †	84.3%	556M
EfficientNet-B7	**84.3%**	**66M**

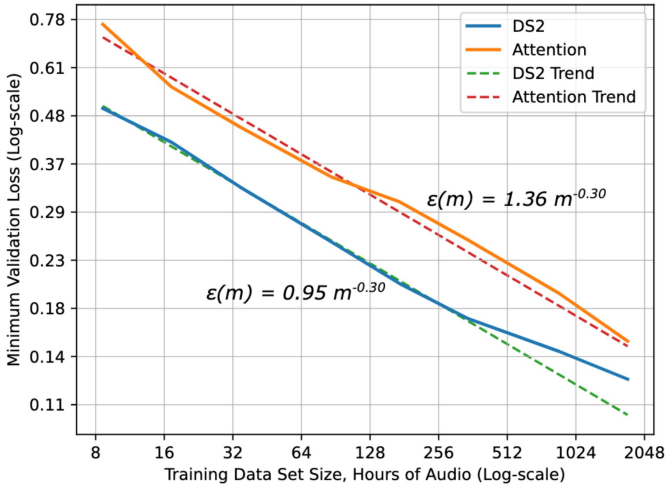
†Not plotted

From "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks" by Tan and Le

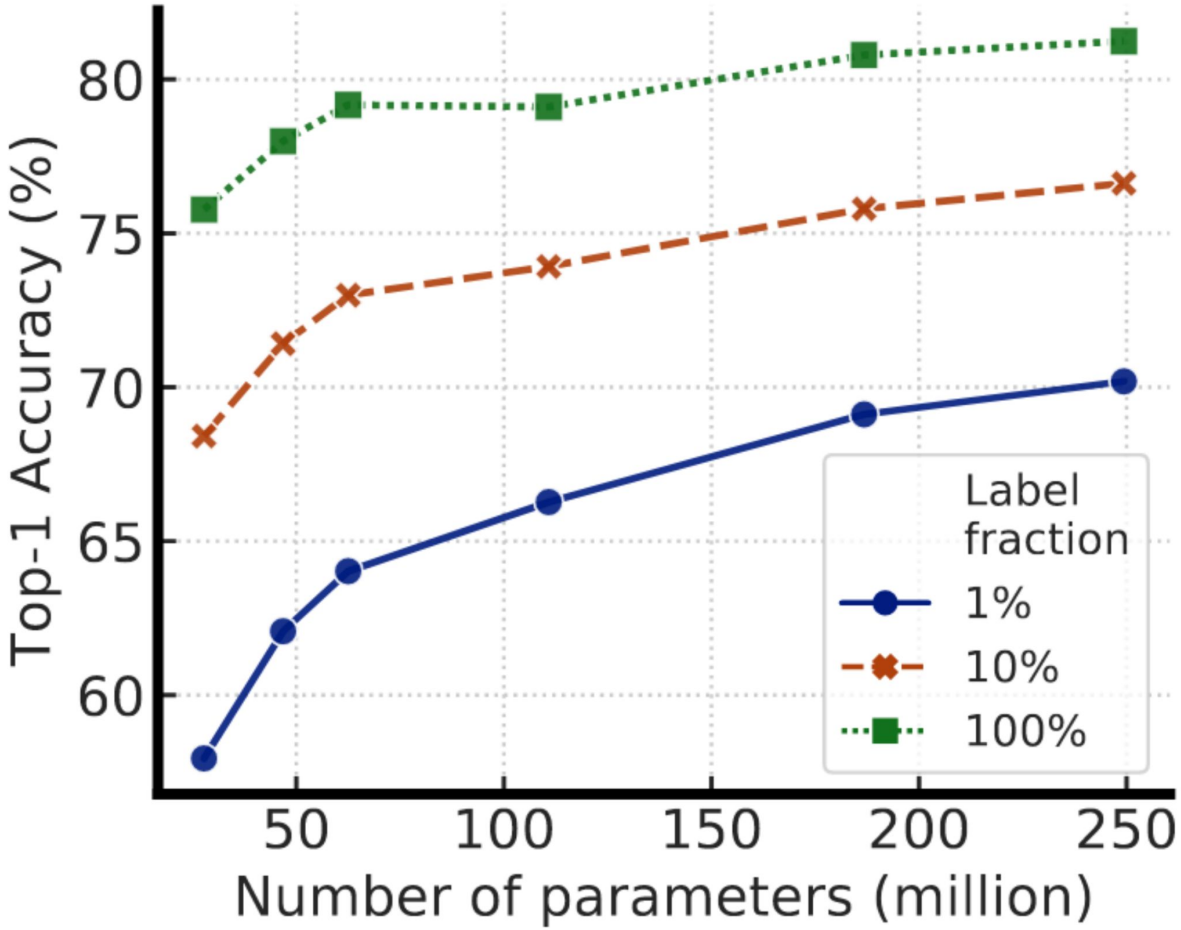




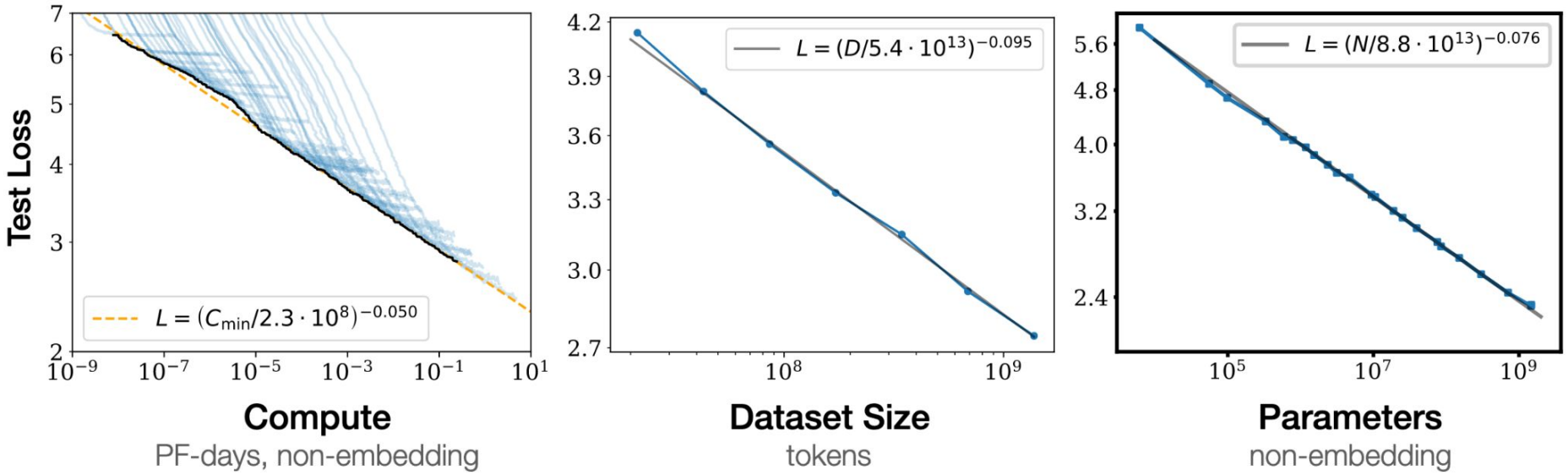




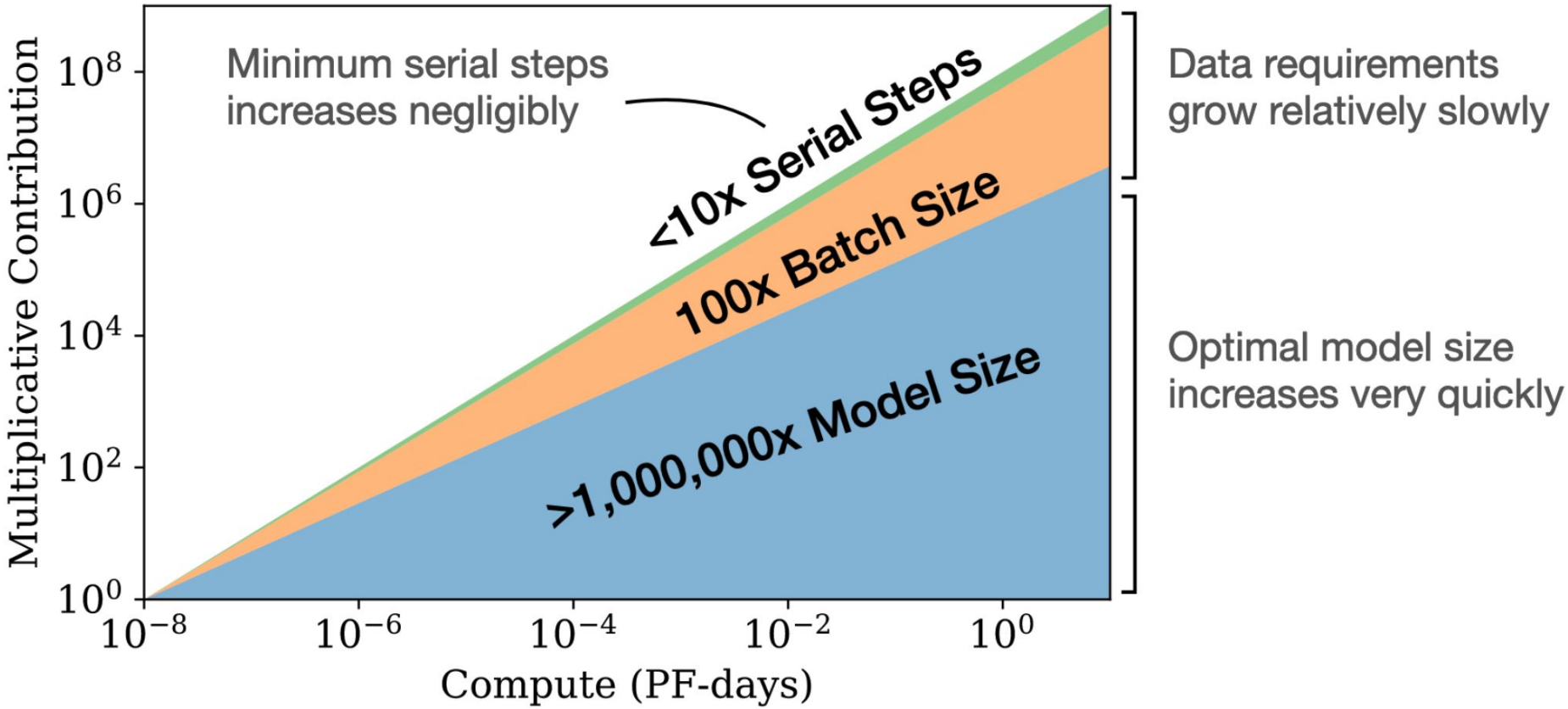
From "Deep Learning Scaling is Predictable, Empirically" by Hestness et al.



From "Big Self-Supervised Models are Strong Semi-Supervised Learners" by Chen et al.



From "Scaling Laws for Neural Language Models" by Kaplan et al.



From "Scaling Laws for Neural Language Models" by Kaplan et al.

$$\hat{y}_i = f_\theta(x_i)$$

$$\underbrace{\ldots W h \ldots}$$

$$\partial\theta = \sum_{i=1}^{N} \nabla_\theta \mathcal{L}(\hat{y}_i, y_i)$$

$$\theta \leftarrow \theta + \texttt{optimizer}(\partial\theta)$$

$$\hat{y}_i = \underbrace{f_\theta(x_i)}_{\dots W h \dots}$$

$$\partial\theta = \sum_{i=1}^{N} \nabla_\theta \mathcal{L}(\hat{y}_i, y_i)$$

$$\theta \leftarrow \theta + \texttt{optimizer}(\partial\theta)$$

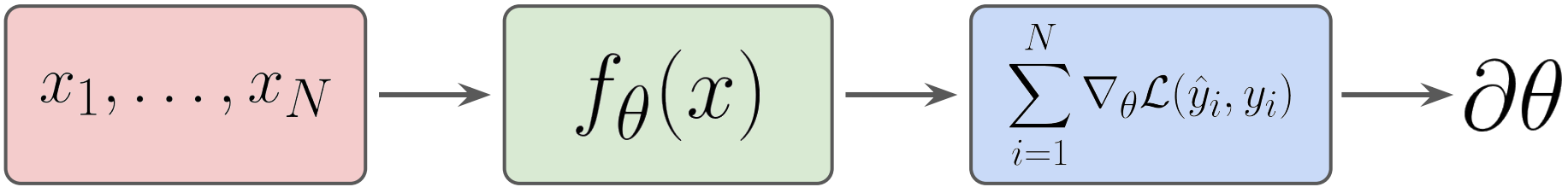
Memory

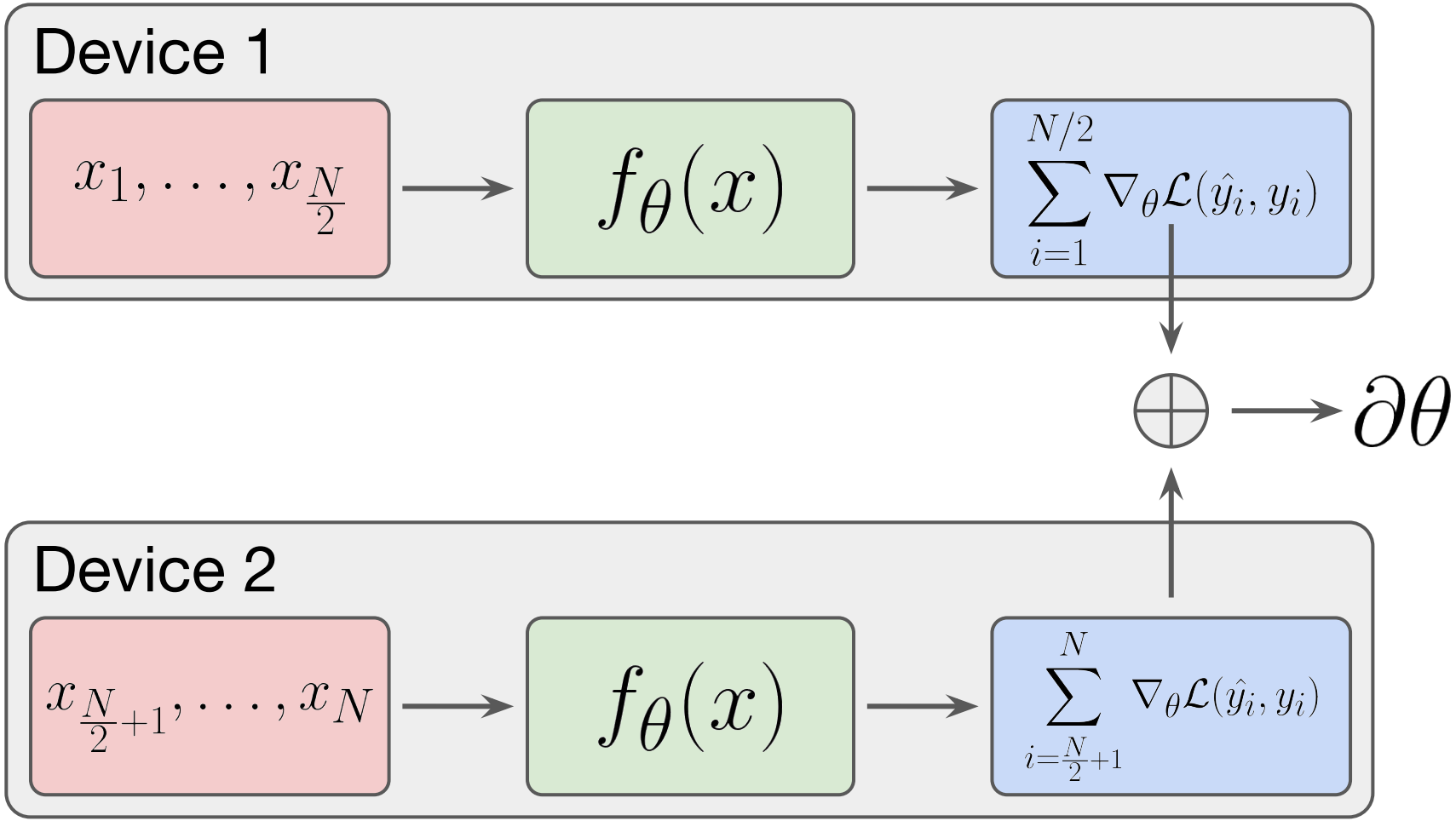
$$\hat{y}_i = \underbrace{f_\theta(x_i)}_{\ldots Wh \ldots}$$

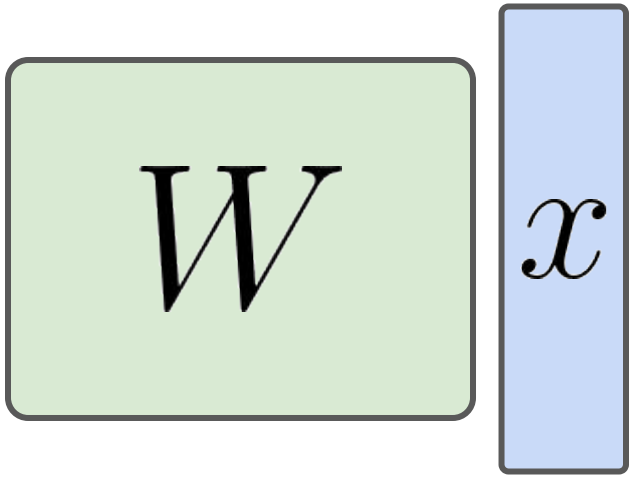
Compute

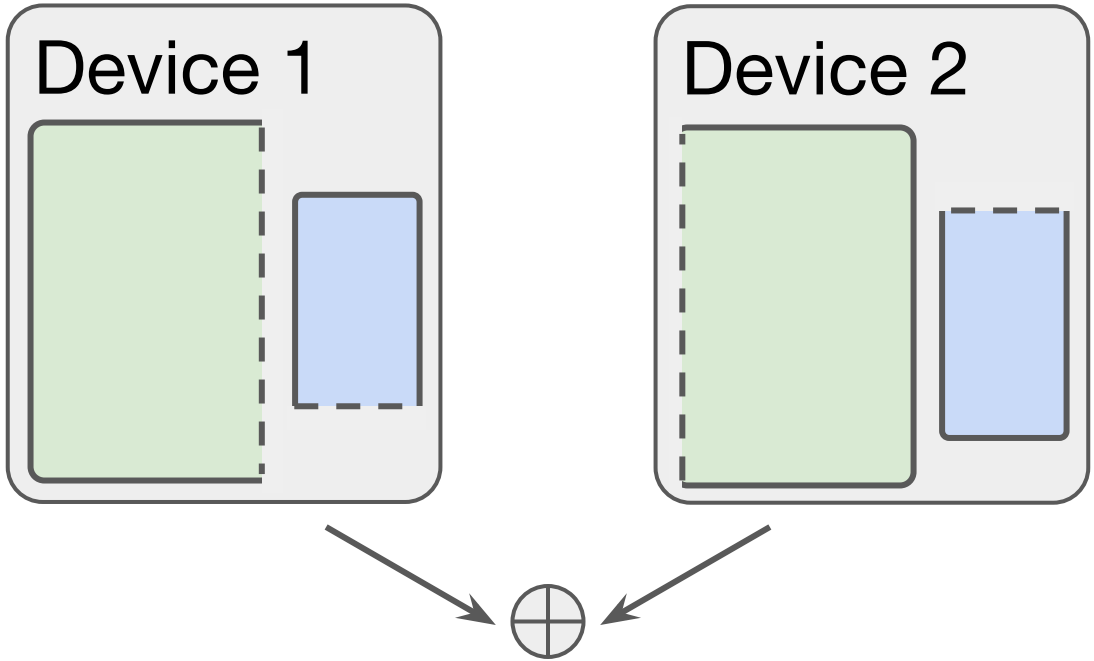
$$\partial\theta = \sum_{i=1}^{N} \nabla_\theta \mathcal{L}(\hat{y}_i, y_i)$$

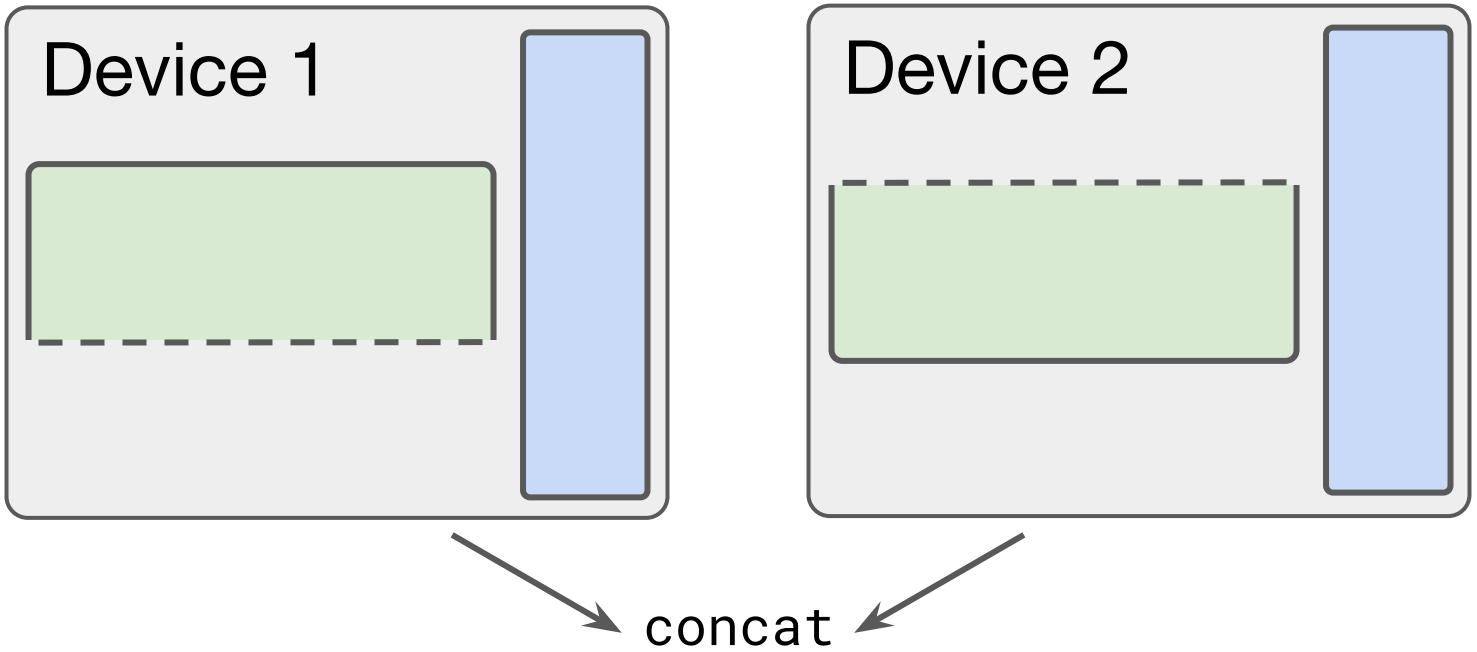
$$\theta \leftarrow \theta + \texttt{optimizer}(\partial\theta)$$

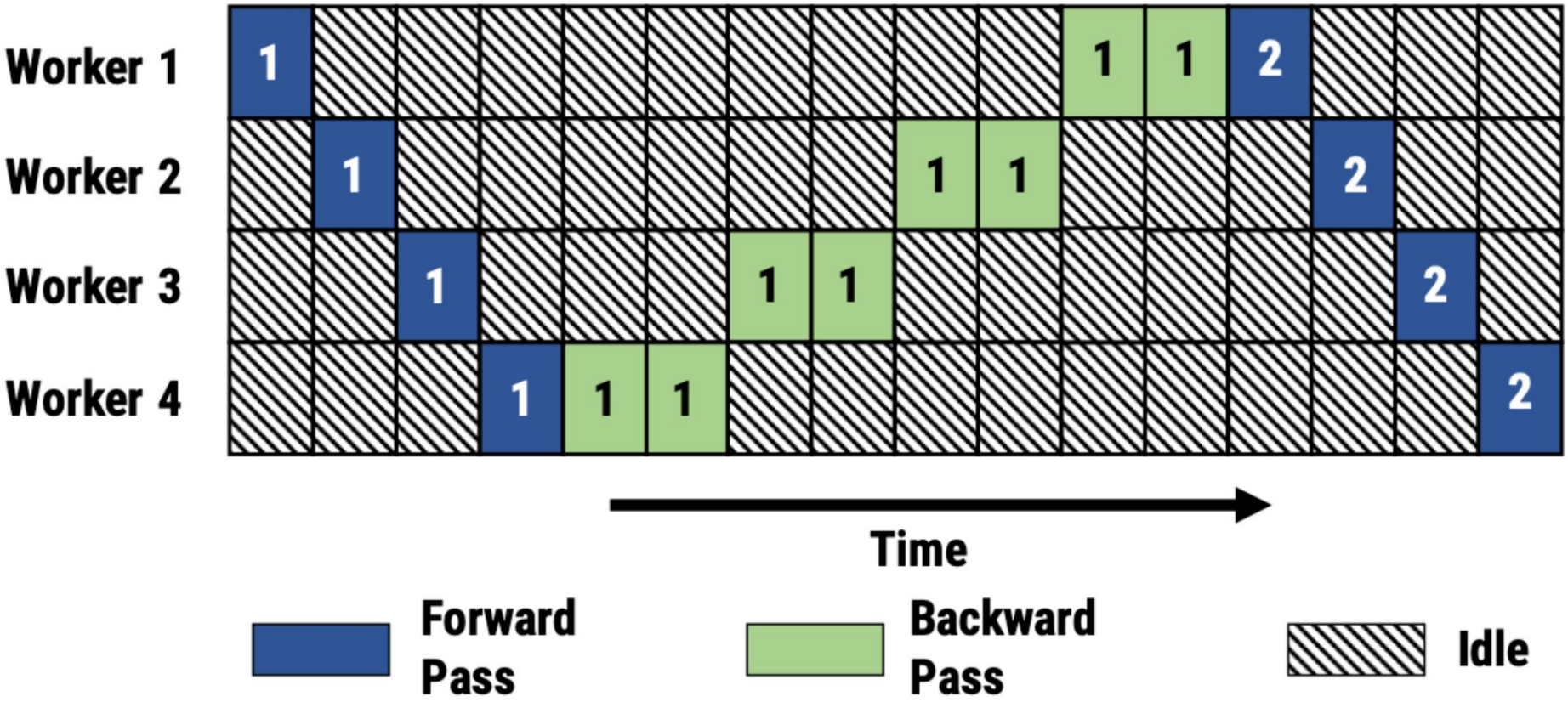

$x_1, \ldots, x_N$
$f_\theta(x)$
$\sum_{i=1}^{N} \nabla_\theta \mathcal{L}(\hat{y}_i, y_i)$
$\partial\theta$


Device 1
$x_1, \ldots, x_{\frac{N}{2}}$
$f_\theta(x)$
$\sum_{i=1}^{N/2} \nabla_\theta \mathcal{L}(\hat{y}_i, y_i)$
$\partial\theta$
Device 2
$x_{\frac{N}{2}+1}, \ldots, x_N$
$f_\theta(x)$
$\sum_{i=\frac{N}{2}+1}^{N} \nabla_\theta \mathcal{L}(\hat{y}_i, y_i)$

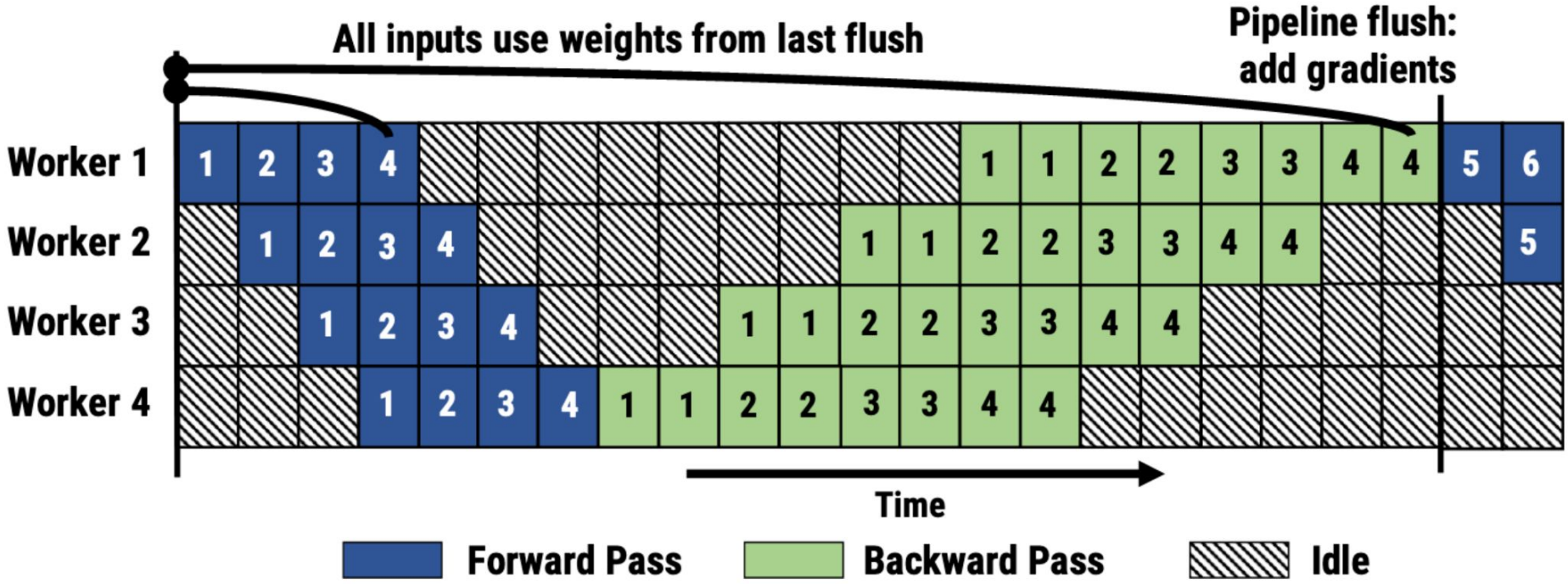

W
x


Device 1
Device 2


Device 1
Device 2
concat

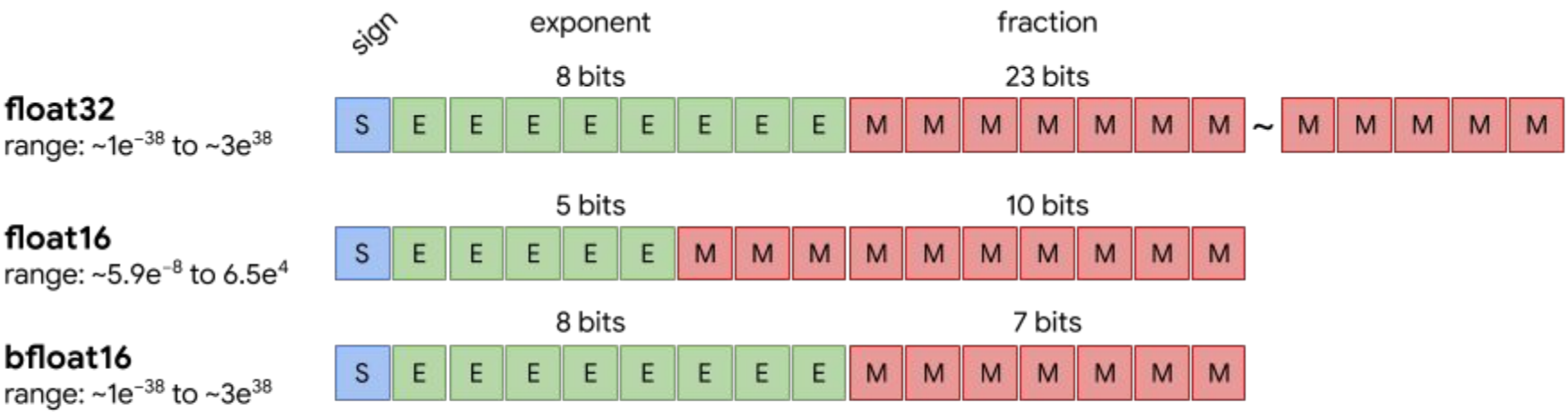


From "PipeDream: Generalized Pipeline Parallelism for DNN Training" by Narayanan et al.

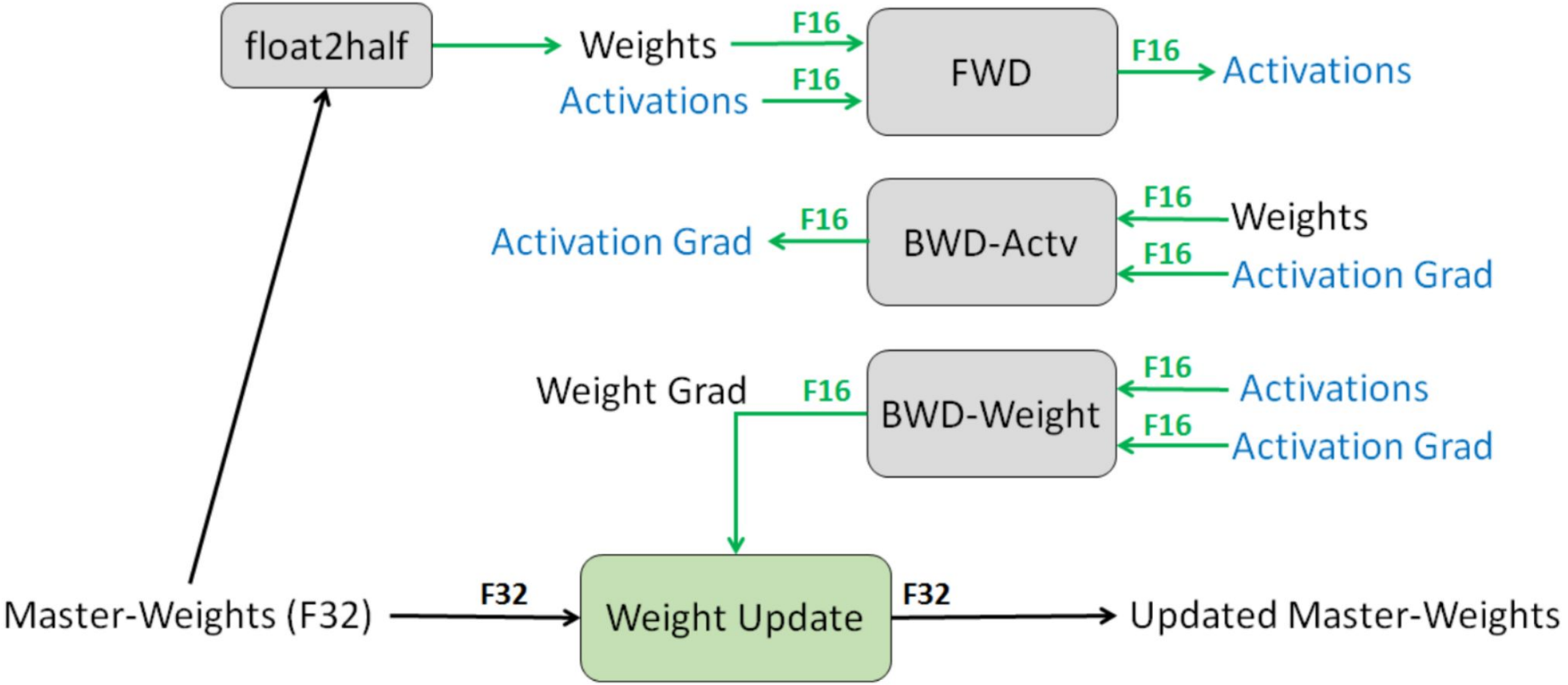


From "PipeDream: Generalized Pipeline Parallelism for DNN Training" by Narayanan et al.

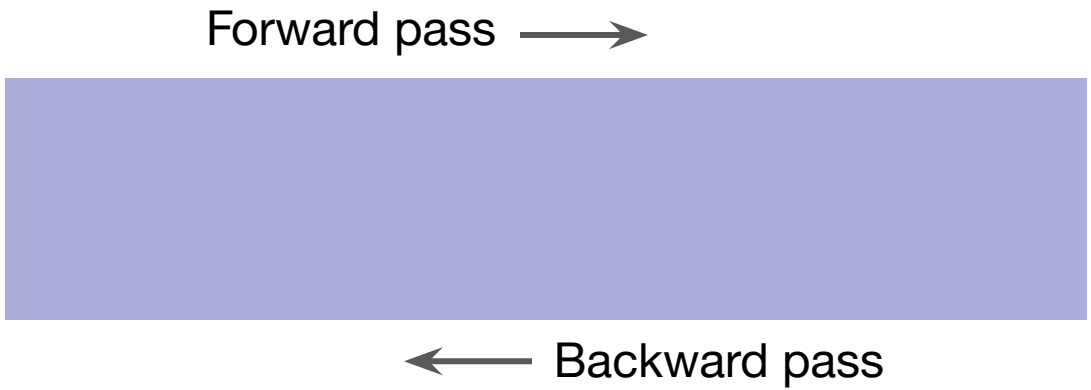
From "Efficient Large-Scale Language Model Training on GPU Clusters" by Narayanan et al.



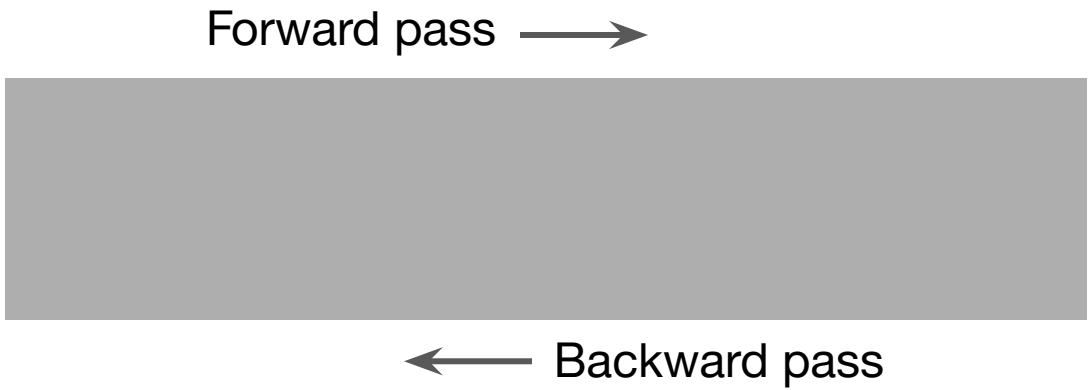
From https://cloud.google.com/tpu/docs/bfloat16



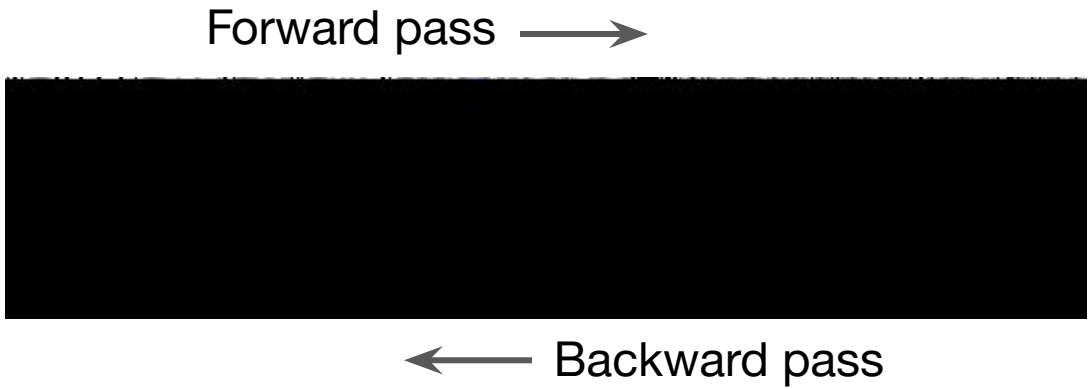
From "Mixed Precision Training" by Micikevicius et al.



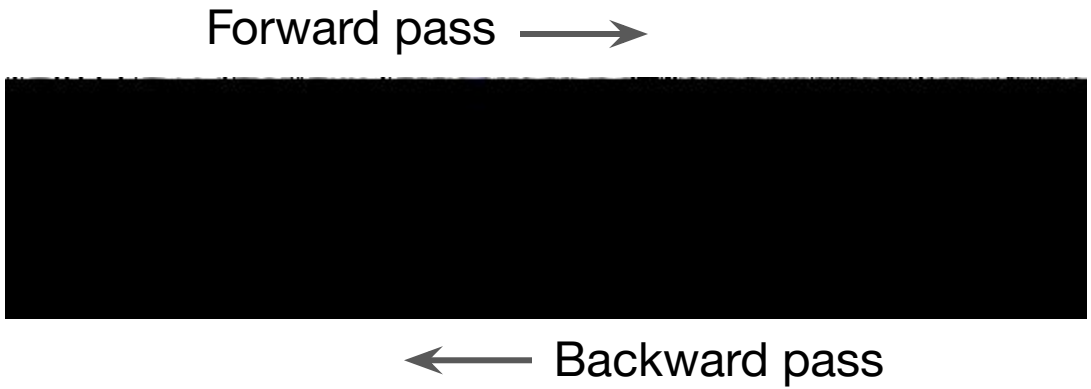
From https://medium.com/tensorflow/fitting-larger-networks-into-memory-583e3c758ff9



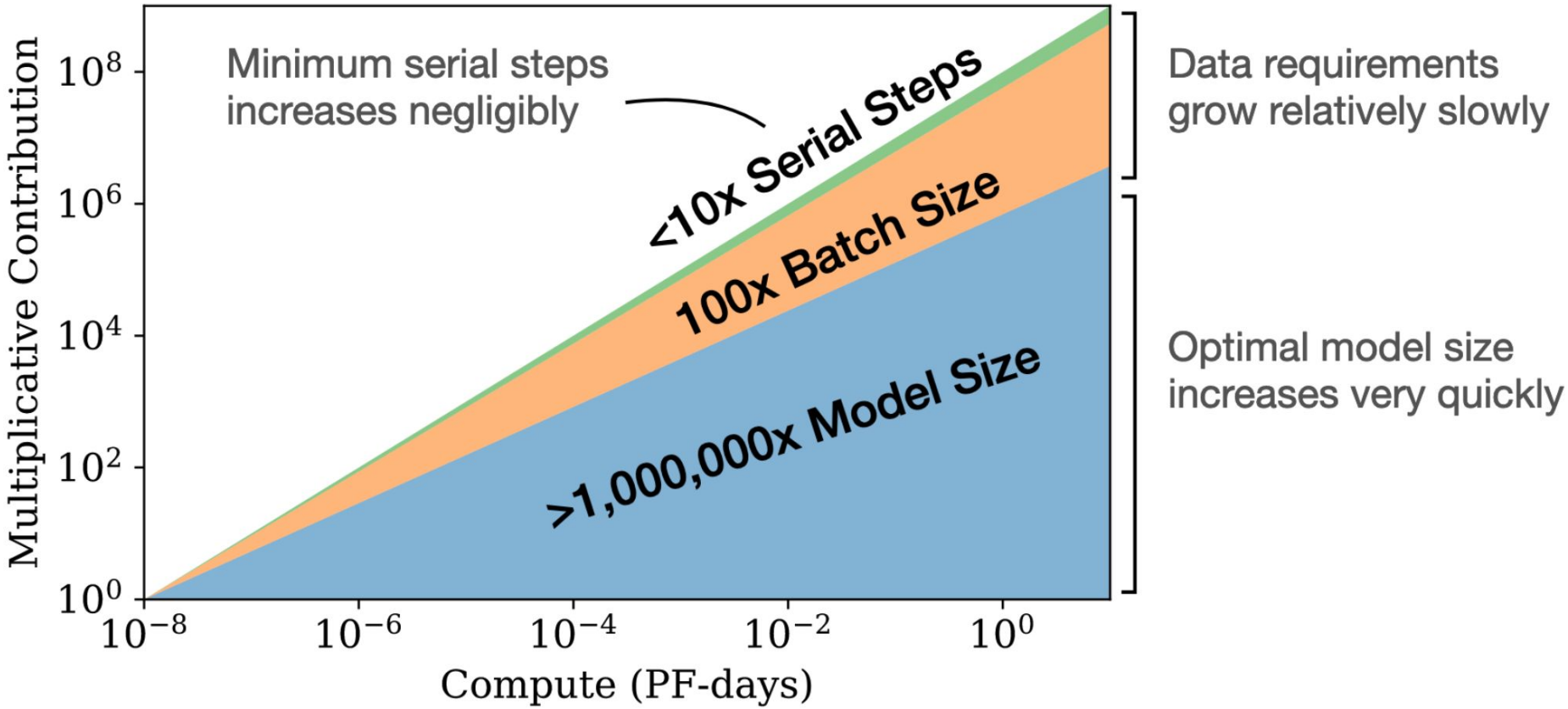
From https://medium.com/tensorflow/fitting-larger-networks-into-memory-583e3c758ff9



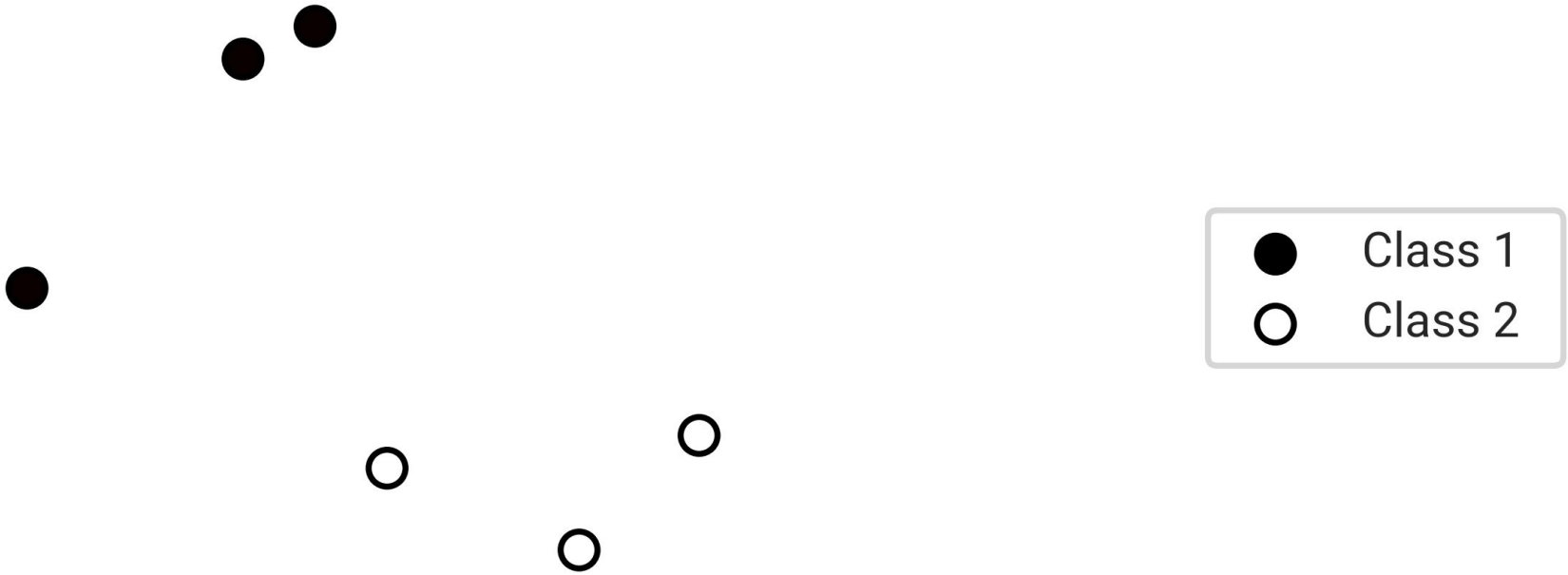
From https://medium.com/tensorflow/fitting-larger-networks-into-memory-583e3c758ff9

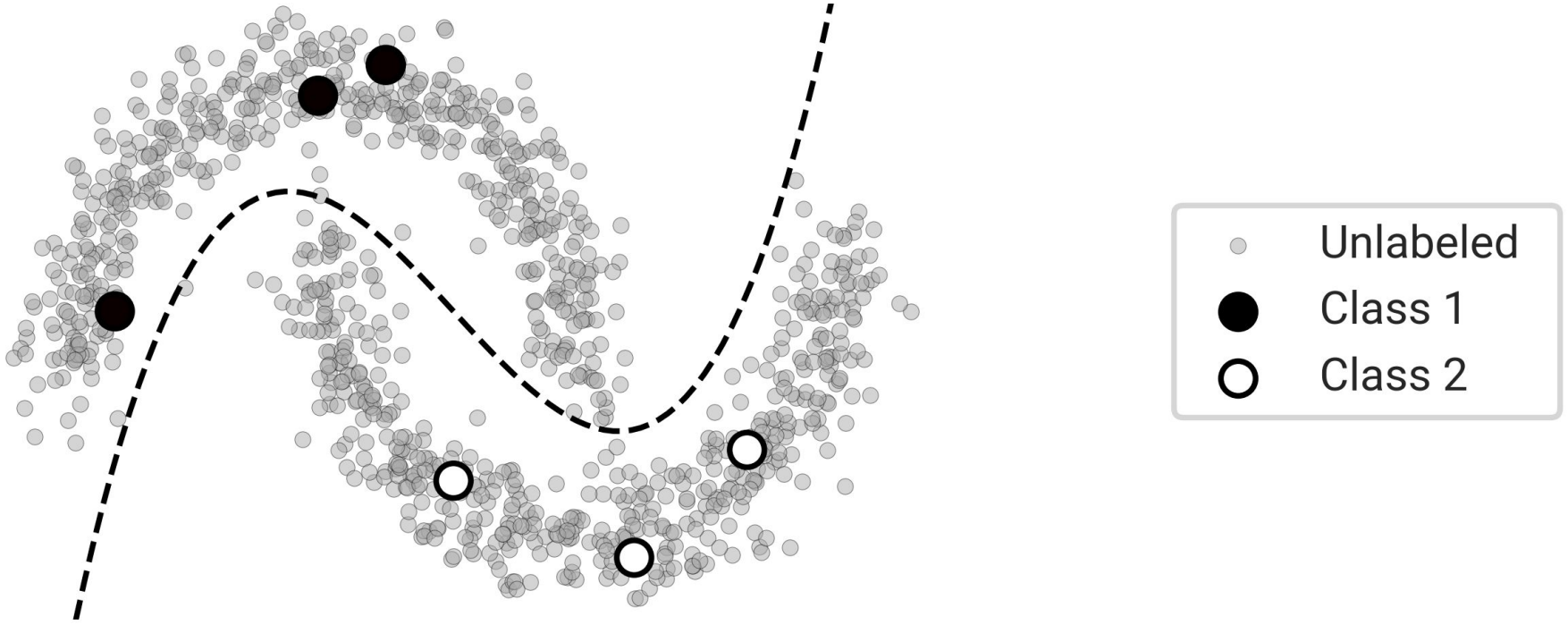


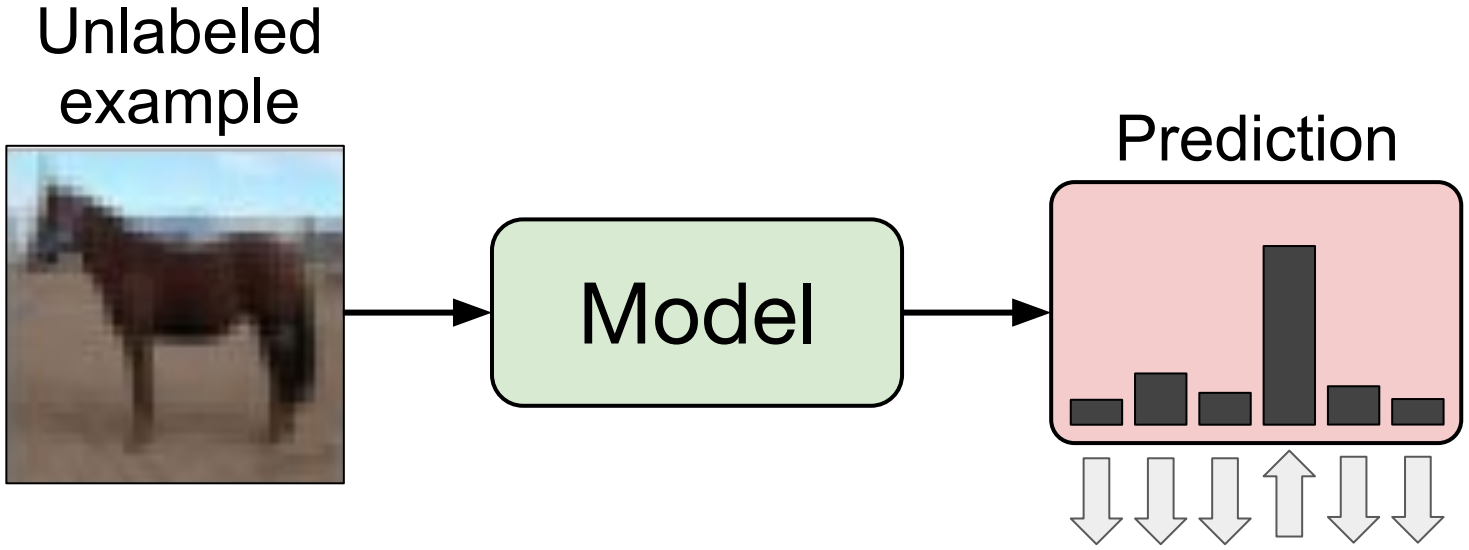
From https://medium.com/tensorflow/fitting-larger-networks-into-memory-583e3c758ff9



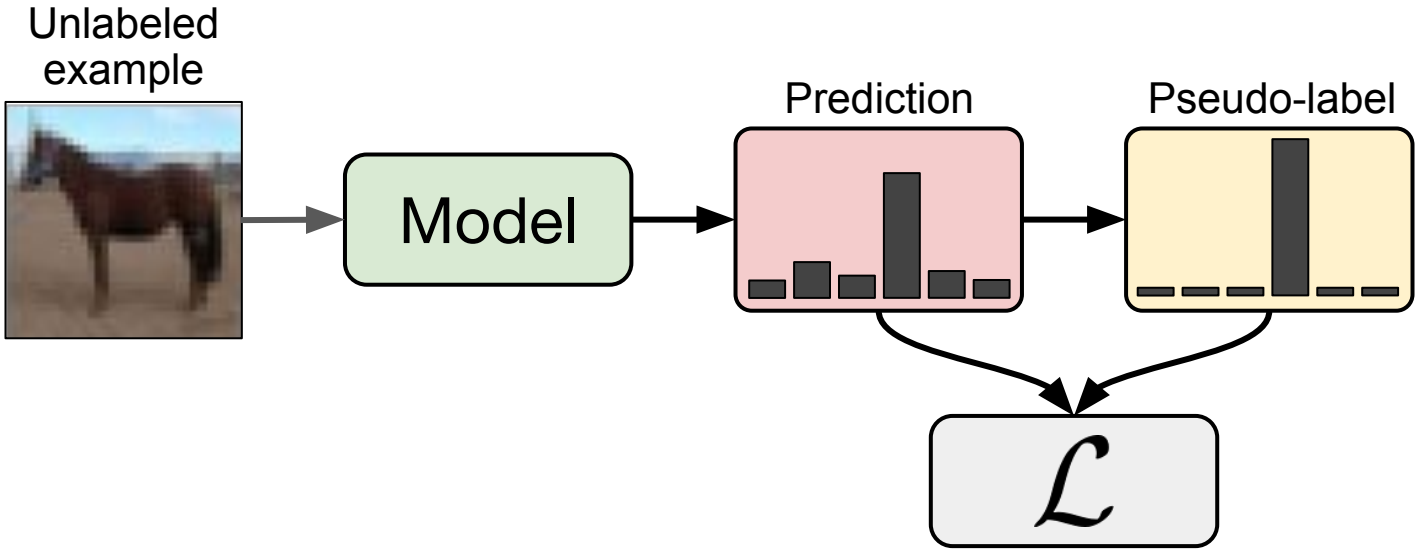
From "Scaling Laws for Neural Language Models" by Kaplan et al.


Class 1
Class 2

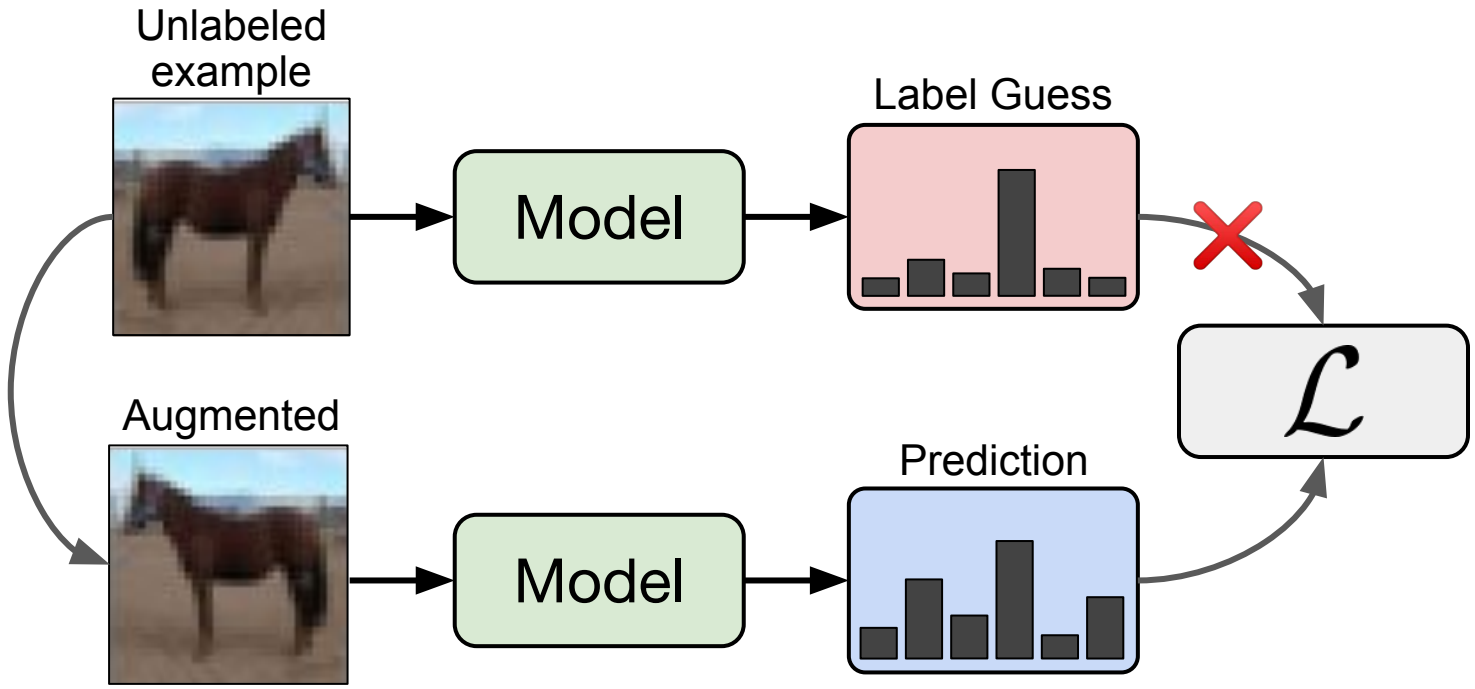

Unlabeled
Class 1
Class 2



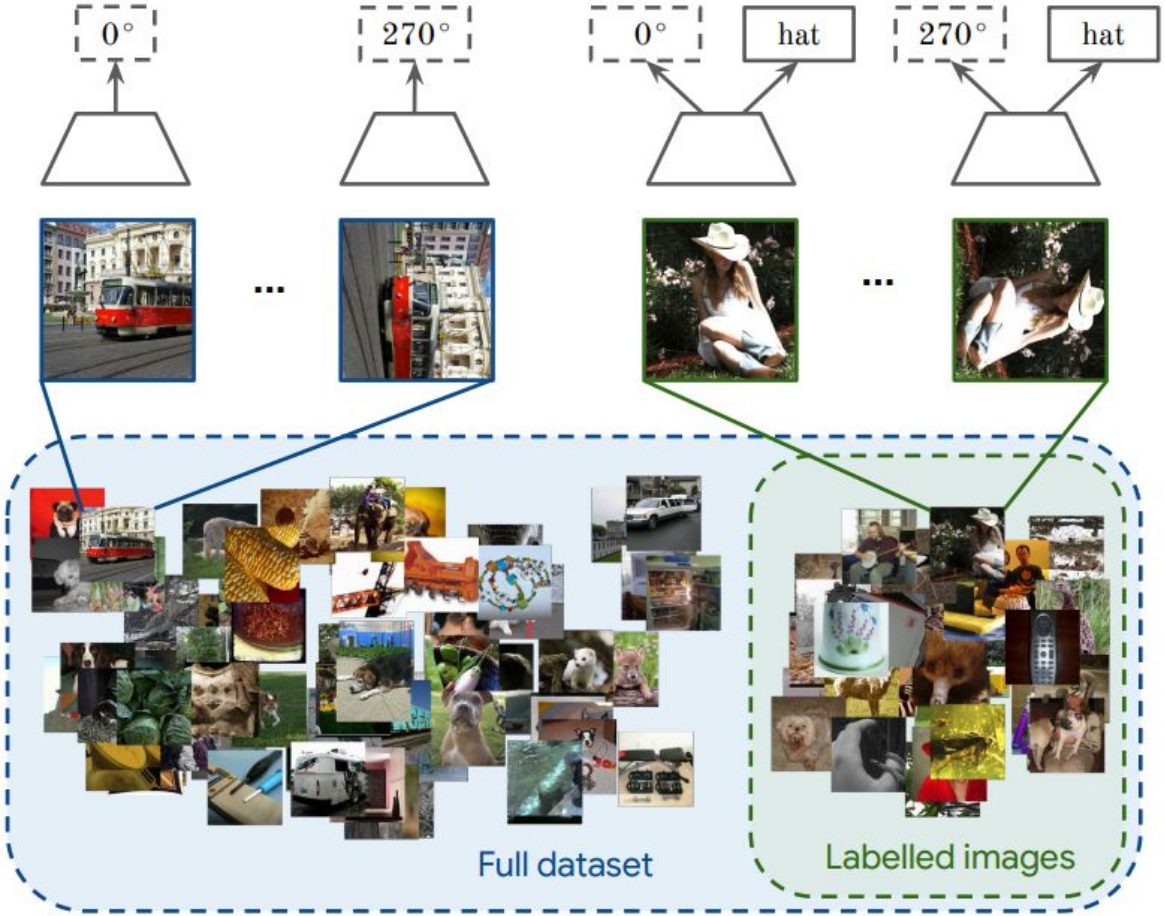
$$\mathcal{L} = -\sum_{y} p_\theta(y|x) \log p_\theta(y|x)$$



$$\mathcal{L} = -\arg\max_{y}[p_\theta(y|x)] \log p_\theta(y|x)$$

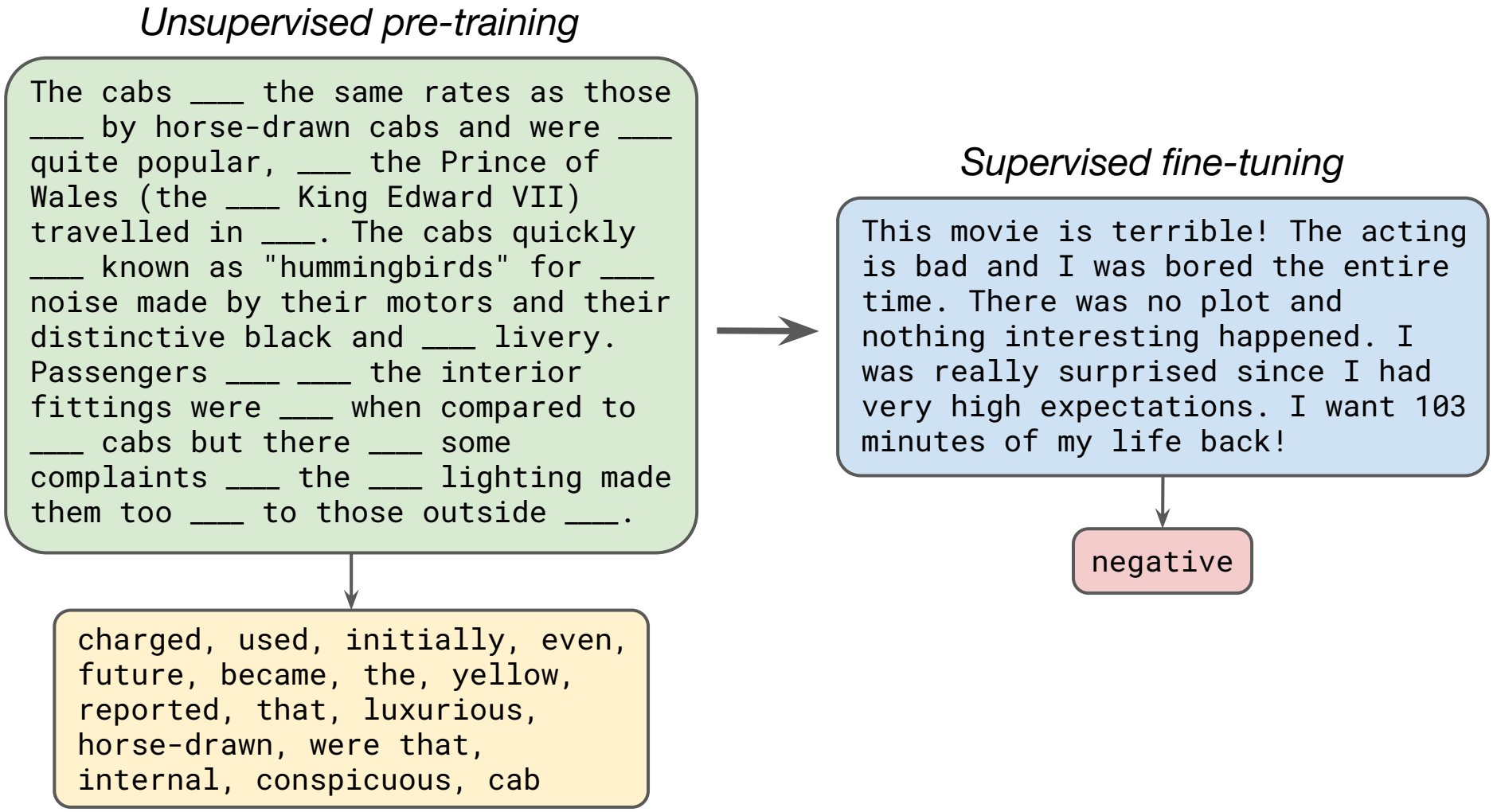


$$\mathcal{L} = \text{sg}[p_\theta(y|x)] \log p_\theta(y|x')$$



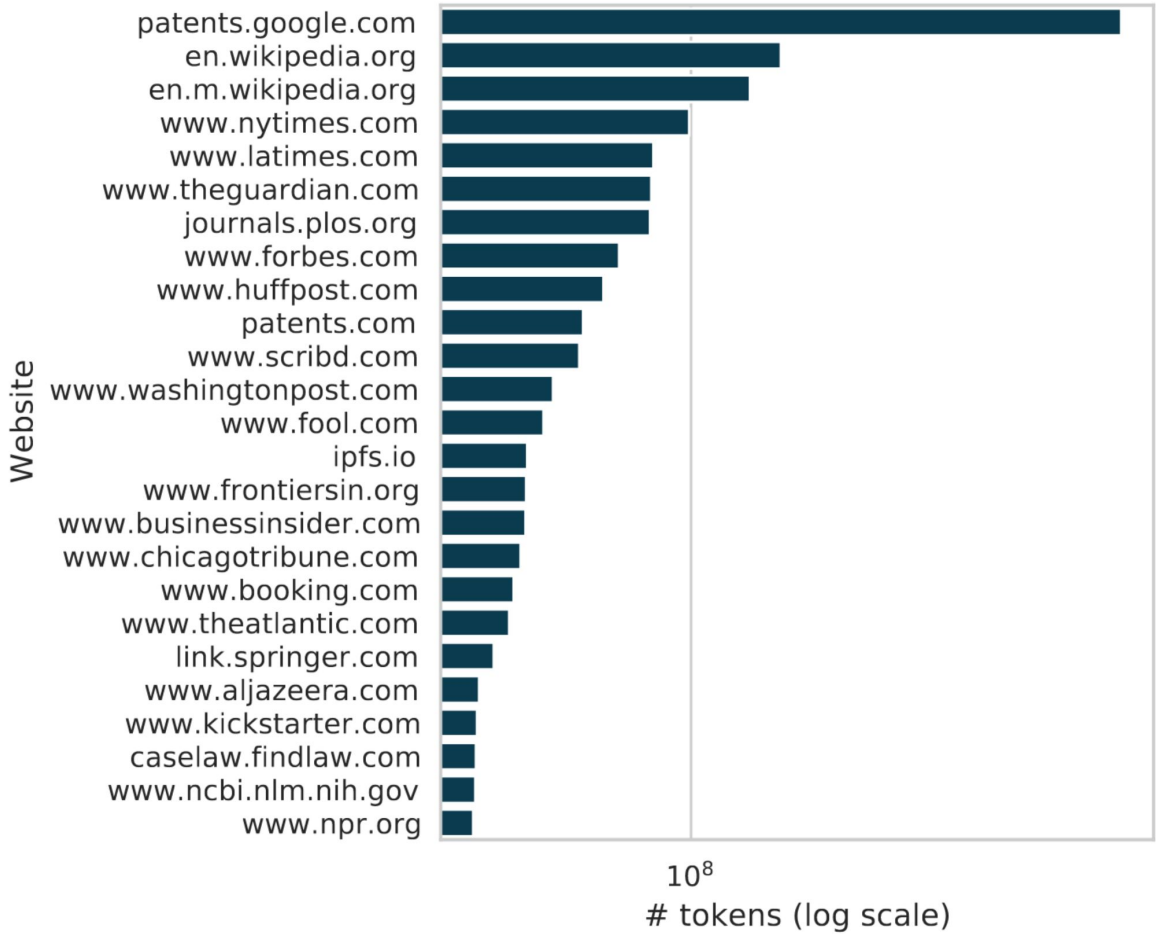
From "S⁴L: Self-Supervised Semi-Supervised Learning" by Zhai et al.

From https://ai.googleblog.com/2020/04/advancing-self-supervised-and-semi.html


Unsupervised pre-training
The cabs ___ the same rates as those ___ by horse-drawn cabs and were ___ quite popular, ___ the Prince of Wales (the ___ King Edward VII) travelled in ___. The cabs quickly ___ known as "hummingbirds" for ___ noise made by their motors and their distinctive black and ___ livery. Passengers ___ ___ the interior fittings were ___ when compared to ___ cabs but there ___ some complaints ___ the ___ lighting made them too ___ to those outside ___.
charged, used, initially, even, future, became, the, yellow, reported, that, luxurious, horse-drawn, were that, internal, conspicuous, cab
Supervised fine-tuning
This movie is terrible! The acting is bad and I was bored the entire time. There was no plot and nothing interesting happened. I was really surprised since I had very high expectations. I want 103 minutes of my life back!
negative

The size of data available on the web has enabled deep learning models to achieve high accuracy on specific benchmarks in NLP and computer vision applications. However, in both application areas, the training data has been shown to have problematic characteristics resulting in models that encode stereotypical and derogatory associations along gender, race, ethnicity, and disability status.

From “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜” by Bender et al.



From "Documenting the English Colossal Clean Crawled Corpus" by Dodge et al.

… accuracy improvements depend on the availability of exceptionally large computational resources that necessitate similarly substantial energy consumption. As a result these models are costly to train and develop, both financially, due to the cost of hardware and electricity or cloud compute time, and environmentally, due to the carbon footprint required to fuel modern tensor processing hardware.

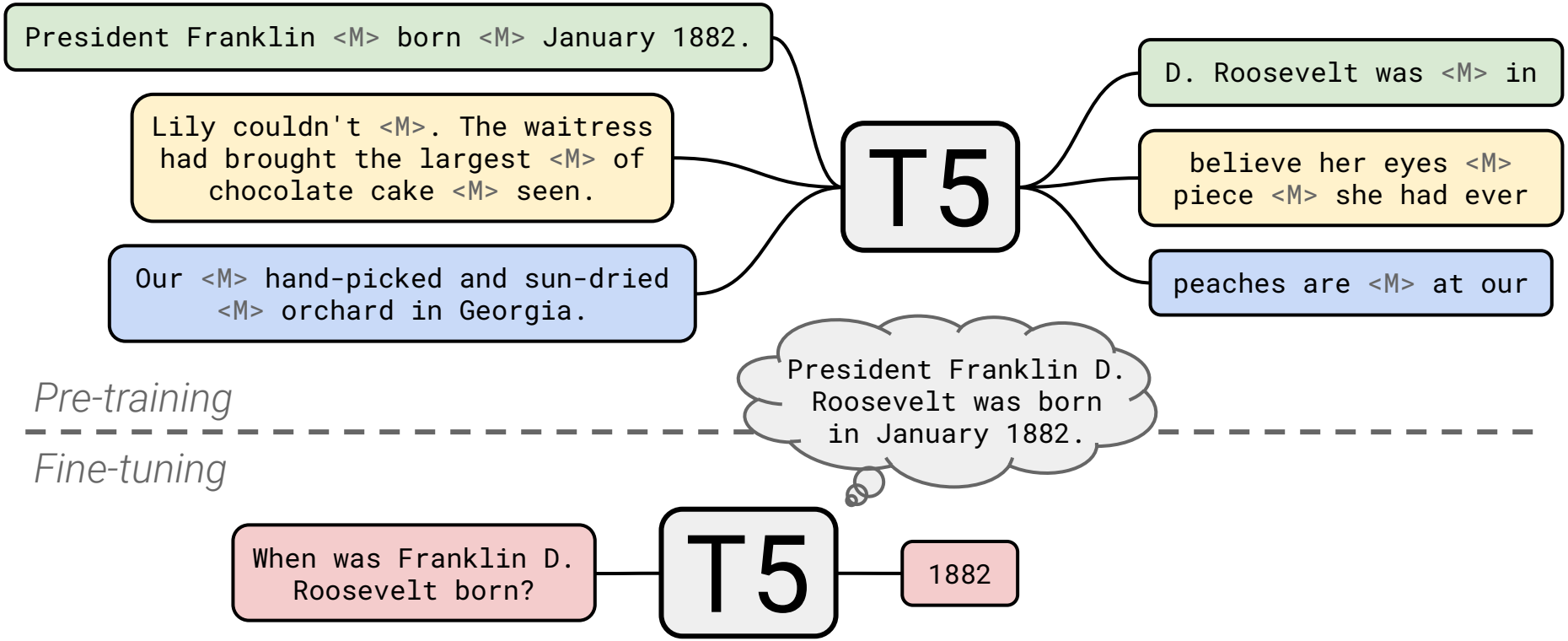
From “Energy and Policy Considerations for Deep Learning in NLP” by Strubell et al.

Model	*Evolved Trans-former NAS*	*T5*	*Meena*	*Gshard -600B*	*Switch Trans-former*	*GPT-3*
Number of Parameters (B)	0.064 per model	11	2.6	619	1500	175
Percent of model activated on every token	100%	100%	100%	0.25%	0.10%	100%
Developer	Google					OpenAI
Datacenter of original experiment	Google Georgia	Google Taiwan	Google Georgia	Google North Carolina	Google Georgia	Microsoft
When model ran	Dec 2018	Sep 2019	Dec 2019	Apr 2020	Oct 2020	2020
Datacenter Gross CO_2e/KWh (kg/KWh when it was run)	0.431	0.545	0.415	0.201	0.403	0.429
Datacenter Net CO2e/KWh (kg/KWh when it was run)	0.431	0.545	0.415	0.177	0.330	0.429
Datacenter PUE (when it was run)	1.10	1.12	1.09	1.09	1.10	1.10
Processor	TPU v2	TPU v3				V100
Chip Thermal Design Power (TDP in Watts)	280	450				300
Measured System Average Power per Accelerator, including memory, network interface, fans, host CPU (W)	208	310	289	288	245	330
Measured Performance (TFLOPS/s)[12]	24.8	45.6	42.3	48.0	34.4	24.6
Number of Chips	200	512	1024	1024	1024	10,000
Training time (days)	6.8	20	30	3.1	27	14.8
Total Computation (floating point operations)	2.91E+21	4.05E+22	1.12E+23	1.33E+22	8.22E+22	3.14E+23
Energy Consumption (MWh)	7.5	85.7	232	24.1	179	1,287
% of Google 2019 total energy consumption (12.2 TWh = 12,200,000 MWh) [Goo20]	0.00006%	0.00070%	0.00190%	0.00020%	0.00147%	0.01055%
Gross tCO_2e for Model Training	3.2	46.7	96.4	4.8	72.2	552.1
Net tCO_2e for Model Training	3.2	46.7	96.4	4.3	59.1	552.1

From "Carbon Emissions and Large Neural Network Training" by Patterson et al.

GPT-3 175B model required 3.14E23 FLOPS of computing for training. Even at theoretical 28 TFLOPS for V100 and lowest 3 year reserved cloud pricing we could find, this will take 355 GPU-years and cost $4.6M for a single training run.

From https://lambdalabs.com/blog/demystifying-gpt-3/



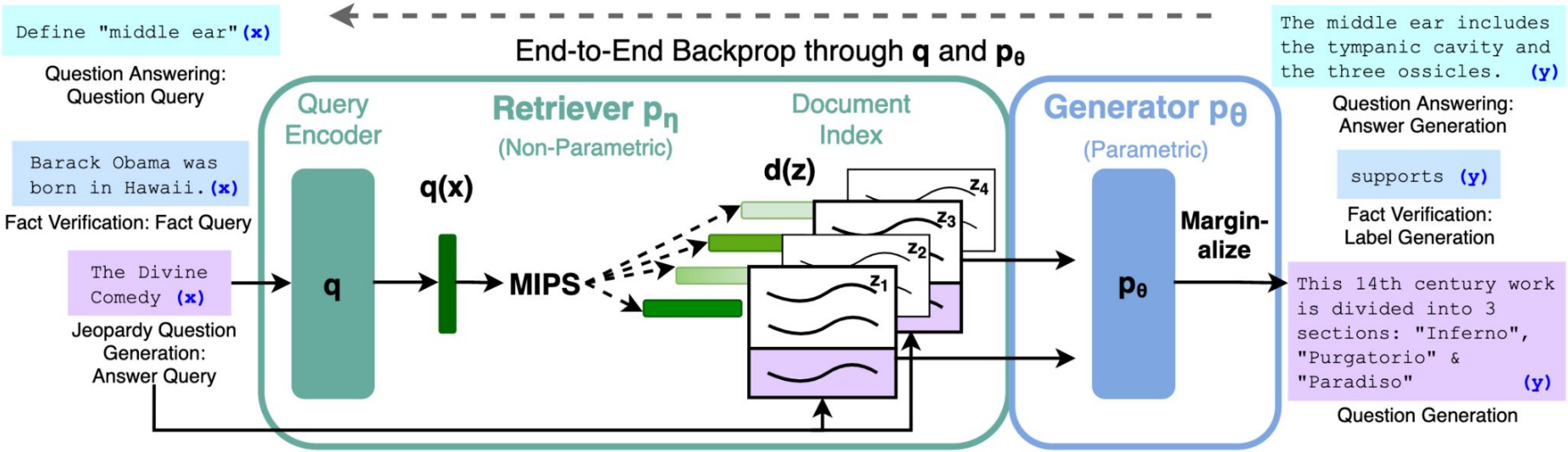
From "How Much Knowledge Can You Pack Into the Parameters of a Language Model?" by Roberts et al.

Index of /wikidatawiki/entities/

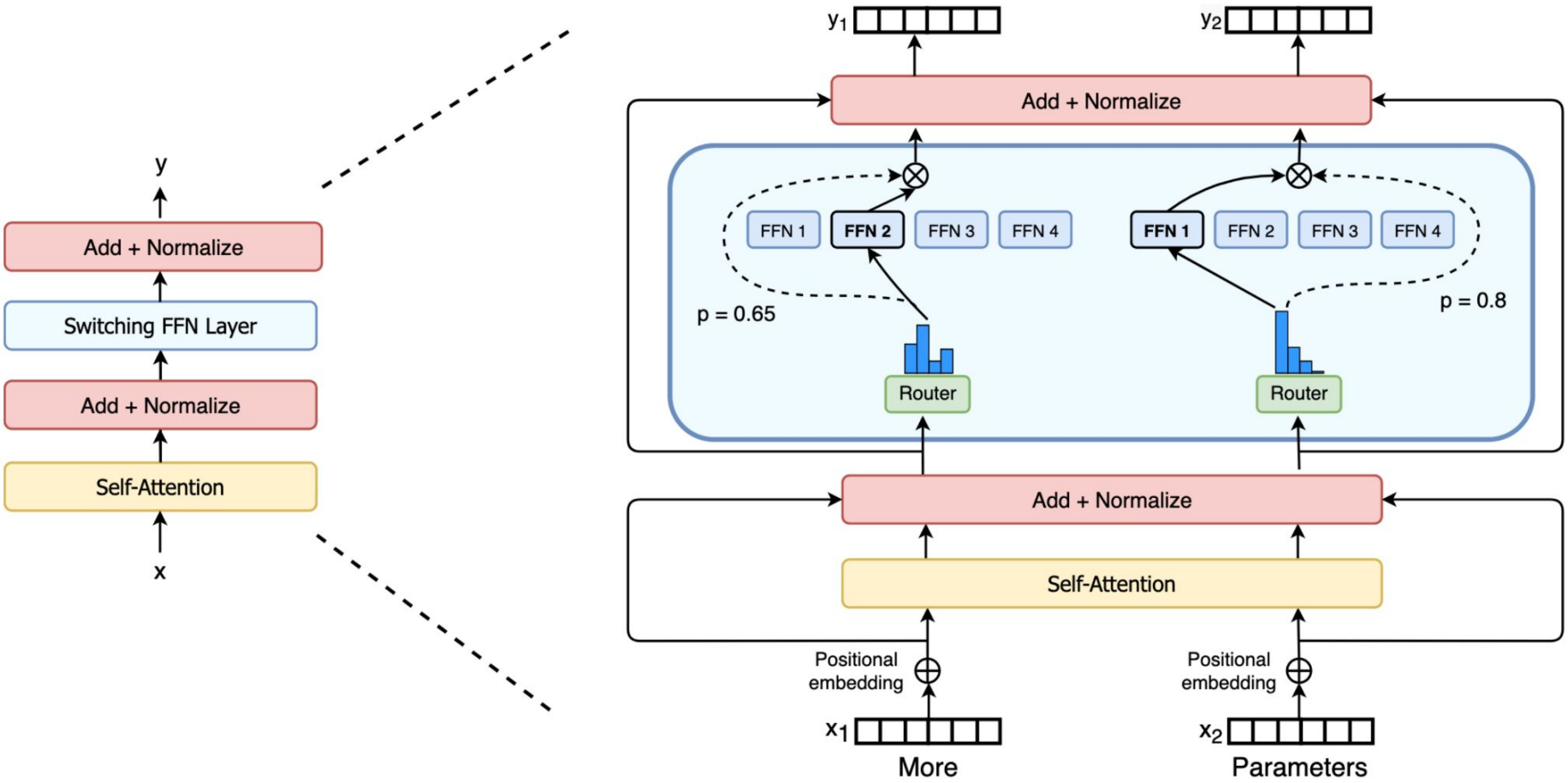
latest-all.json.bz2	27-Jul-2021 11:32	68418560489
latest-all.json.gz	27-Jul-2021 04:58	102963487951

~70GB compressed = 13B float32 parameters

	NQ	WQ	TQA
Open-domain SoTA	41.5	42.4	57.9
T5.1.1-Base	25.7	28.2	24.2
T5.1.1-Large	27.3	29.5	28.5
T5.1.1-XL	29.5	32.4	36.0
T5.1.1-XXL	32.8	35.6	42.9
T5.1.1-XXL + SSM	35.2	42.8	51.9



From "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks" by Lewis et al.



From "Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity" by Fedus et al.

The Summer of Language Models 21 (BigScience)

🐠 Overview

[↑]

The "Summer of Language Models 21 🌸" (in short "BigScience") is a **one-year long research workshop on very large language models** as used and studied in the field of Natural Language Processing and more generally Artifical Intelligence research.

The workshop is

- conducted from May 2021 to May 2022
- with several **live sessions/events** spread over the year (first session on April 28th, second session planned for end of July). Find all events here.
- with **collaborative tasks** aimed at creating, sharing and evaluating a **very large multilingual dataset** and a **very large language model** as tools for research

From https://bigscience.huggingface.co/

Thanks.

craffel@gmail.com