STAT 453: Introduction to Deep Learning and Generative Models

Sebastian Raschka
http://stat.wisc.edu/~sraschka

Deep Learning & AI News #5

Interesting Things Related to Deep Learning

Feb 27th, 2021

arXiv.org > cs > arXiv:2102.08503

Search...

Help | Advanced S

Computer Science > Machine Learning

[Submitted on 16 Feb 2021]

Federated Evaluation and Tuning for On-Device Personalization: System Design & Applications

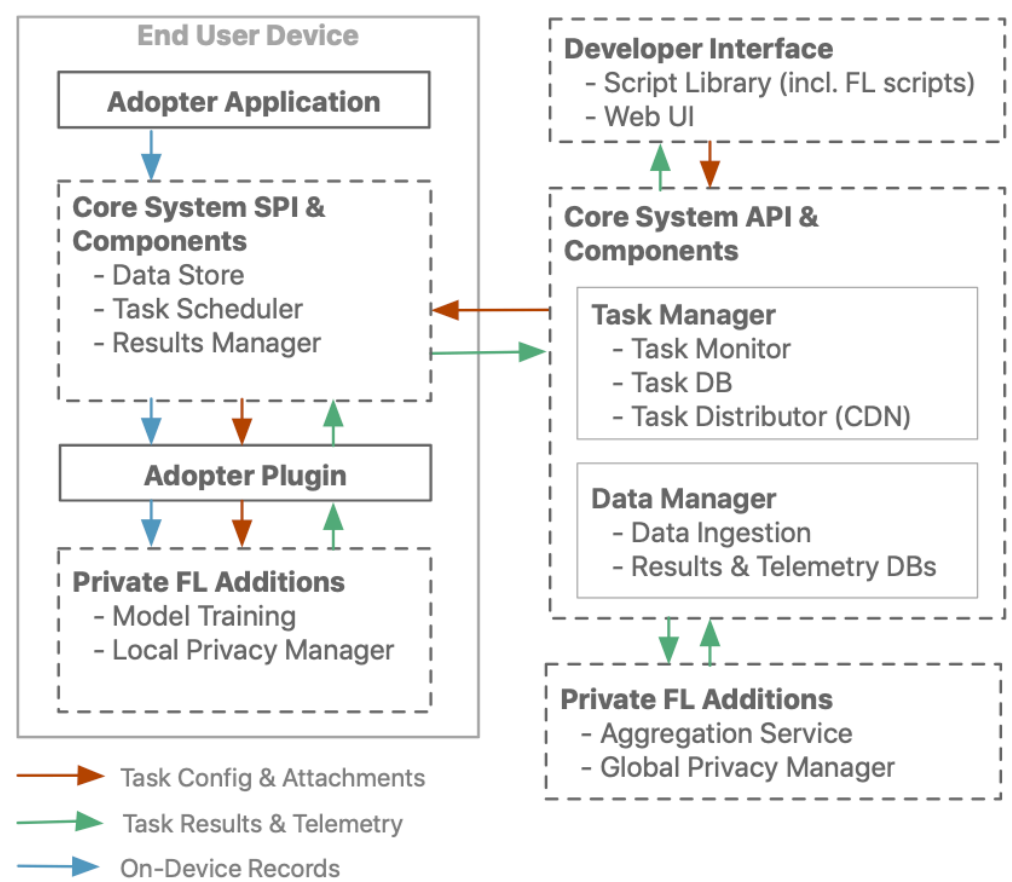
Matthias Paulik, Matt Seigel, Henry Mason, Dominic Telaar, Joris Kluivers, Rogier van Dalen, Chi Wai Lau, Luke Carlson, Filip Granqvist, Chris Vandevelde, Sudeep Agarwal, Julien Freudiger, Andrew Byde, Abhishek Bhowmick, Gaurav Kapoor, Si Beaumont, Áine Cahill, Dominic Hughes, Omid Javidbakht, Fei Dong, Rehan Rishi, Stanley Hung

https://arxiv.org/abs/2102.08503

https://syncedreview.com/2021/02/19/apple-reveals-design-of-its-on-device-ml-system-for-federated-evaluation-and-tuning/

Apple's On-Device ML System for Federated Evaluation and Tuning

- Other companies: use federated learning to tune a global neural network
- Apple:
  - Use global parameters but train local model
  - User data remains inaccessible to server-side



[Submitted on 16 Feb 2021]

Federated Evaluation and Tuning for On-Device Personalization: System Design & Applications

Matthias Paulik, Matt Seigel, Henry Mason, Dominic Telaar, Joris Kluivers, Rogier van Dalen, Chi Wai Lau, Luke Carlson, Filip Granqvist, Chris Vandevelde, Sudeep Agarwal, Julien Freudiger, Andrew Byde, Abhishek Bhowmick, Gaurav Kapoor, Si Beaumont, Áine Cahill, Dominic Hughes, Omid Javidbakht, Fei Dong, Rehan Rishi, Stanley Hung

https://arxiv.org/abs/2102.08503

https://syncedreview.com/2021/02/19/apple-reveals-design-of-its-on-device-ml-system-for-federated-evaluation-and-tuning/

Apple's On-Device ML System for Federated Evaluation and Tuning

Table 1: Federated tuning for news personalization.

	FT Run 1	FT Run 2
Iterations	6	42
Parameters	6	11
Pos. label	tapped	$>= n$ sec in article
Neg. label	not tapped	not tapped
Pred. loss	-86%	-23%

Table 2: Live A/B experimentation results.

	Delta[%]	
	Run 1	Run 2
daily article views	+1.98	+1.87
daily time spent	n/a	+0.90

The optimized parameters from FT run 1 resulted in a 1.98% increase in daily article views, but no statistically significant difference in daily time spent within the application[5]. The optimized parameters from FT run 2 resulted in a 1.87% increase in the daily article views, and a 0.90% increase in the daily time spent within the application.

Microsoft Microsoft 365 Azure Office 365 Dynamics 365 Power Platform Windows 10

Open Source Blog

Create privacy-preserving synthetic data for machine learning with SmartNoise

February 18, 2021 Share

https://cloudblogs.microsoft.com/opensource/2021/02/18/create-privacy-preserving-synthetic-data-for-machine-learning-with-smartnoise/

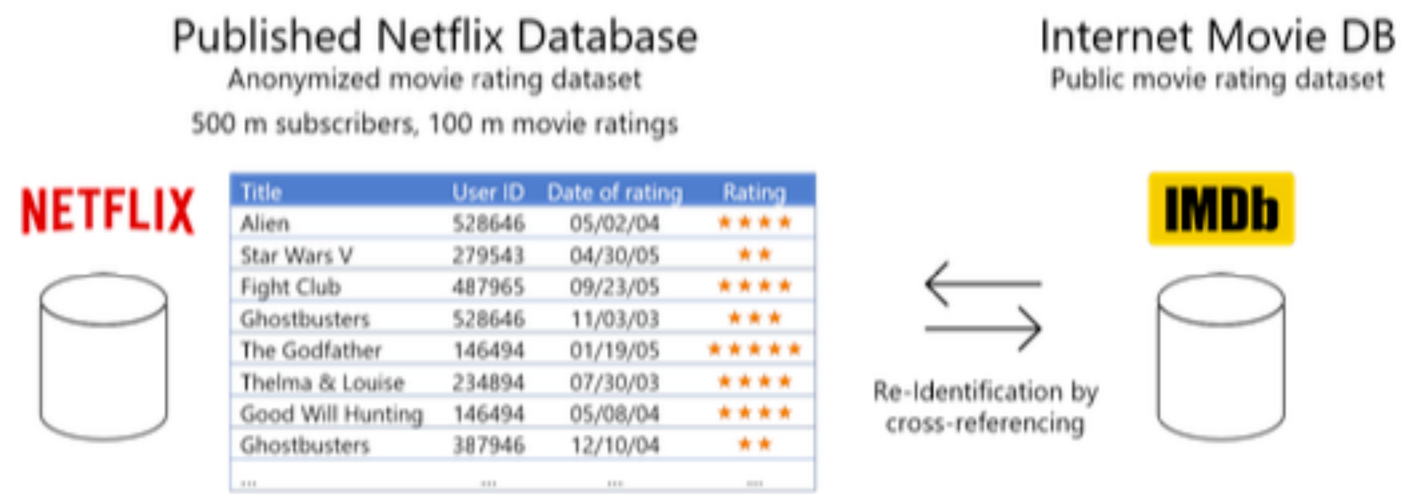
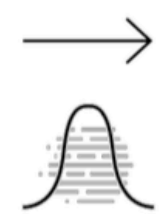


Figure 3: Re-Identification Attack in the Context of the "Netflix Prize" Competition

Original Dataset

Educ.	Age	Gender	Diagnosis	Treatment	Outcome
1	35	0	1	025	0
4	40	0	0	083	2
2	27	1	0	044	1
5	47	1	1	087	0
...	...	...	...	...	...



Synthesize, add noise

Differentially Private Synthetic Dataset

Educ.	Age	Gender	Diagnosis	Treatment	Outcome
4	50	1	1	081	0
3	44	0	1	045	1
1	32	1	1	022	2
3	27	1	0	090	1
...	...	...	...	...	...

Figure 13: Generating a Differentially Private Dataset Based on Original Data

Whitepaper: https://azure.microsoft.com/en-us/resources/microsoft-smartnoisedifferential-privacy-machine-learning-case-studies/

Code: https://github.com/opendp/smartnoise-samples

Multiplicative Weights Exponential Mechanism (MWEM)

- Achieves Differential Privacy by combining Multiplicative Weights and Exponential Mechanism techniques
- A relatively simple but effective approach
- Requires fewer computational resources, shorter runtime

Differentially Private Generative Adversarial Network (DPGAN)

- Adds noise to the discriminator of the GAN to enforce Differential Privacy
- Has been used with image data and electronic health records (HER)

Private Aggregation of Teacher Ensembles Generative Adversarial Network (PATEGAN)

- A modification of the PATE framework that is applied to GANs to preserve Differential Privacy of synthetic data
- Improvement of DPGAN, especially for classification tasks

DP-CTGAN

- Takes the state-of-the-art CTGAN for synthesizing tabular data and applies DPSGD (the same method for ensuring Differential Privacy that DPGAN uses)
- Suited for tabular data, avoids issues with mode collapse
- Can lead to extensive training times

PATE-CTGAN

- Takes the state-of-the-art CTGAN for synthesizing tabular data and applies PATE (the same method for ensuring Differential Privacy that PATEGAN uses)
- Suited for tabular data, avoids issues with mode collapse

Qualified Architecture to Improve Learning (QUAIL)

- Ensemble method to improve the utility of synthetic differentially private datasets for machine learning tasks
- Combines a differentially private synthesizer and an embedded differentially private supervised learning model to produce a flexible synthetic data set with high machine learning utility

News@Northeastern

HUMANS ARE TRYING TO TAKE BIAS OUT OF FACIAL RECOGNITION PROGRAMS. IT'S NOT WORKING–YET.

One likely reason: lack of diversity in the datasets.
Common mitigation approach: provide algorithms with datasets that represent all groups equally and fairly

Does it work? Only for a stereotypical sense of fairness:
Khan: "The people in the images appeared to fit racial stereotypes. For example, an algorithm was more likely to label an individual in an image as “white” if that person had blond hair."

https://news.northeastern.edu/2021/02/22/humans-are-trying-to-take-bias-out-of-facial-recognition-programs-its-not-working-yet/

arXiv.org > cs > arXiv:2102.02320

Computer Science > Computer Vision and Pattern Recognition

[Submitted on 3 Feb 2021]

One Label, One Billion Faces: Usage and Consistency of Racial Categories in Computer Vision

Zaid Khan, Yun Fu

https://arxiv.org/abs/2102.02320

"We find evidence that racial categories encode stereotypes, and exclude ethnic groups from categories on the basis of nonconformity to stereotypes. Representing a billion humans under one racial category may obscure disparities and create new ones by encoding stereotypes of racial systems."

Google AI Blog

The latest news from Google AI

Introducing Model Search: An Open Source Platform for Finding Optimal ML Models

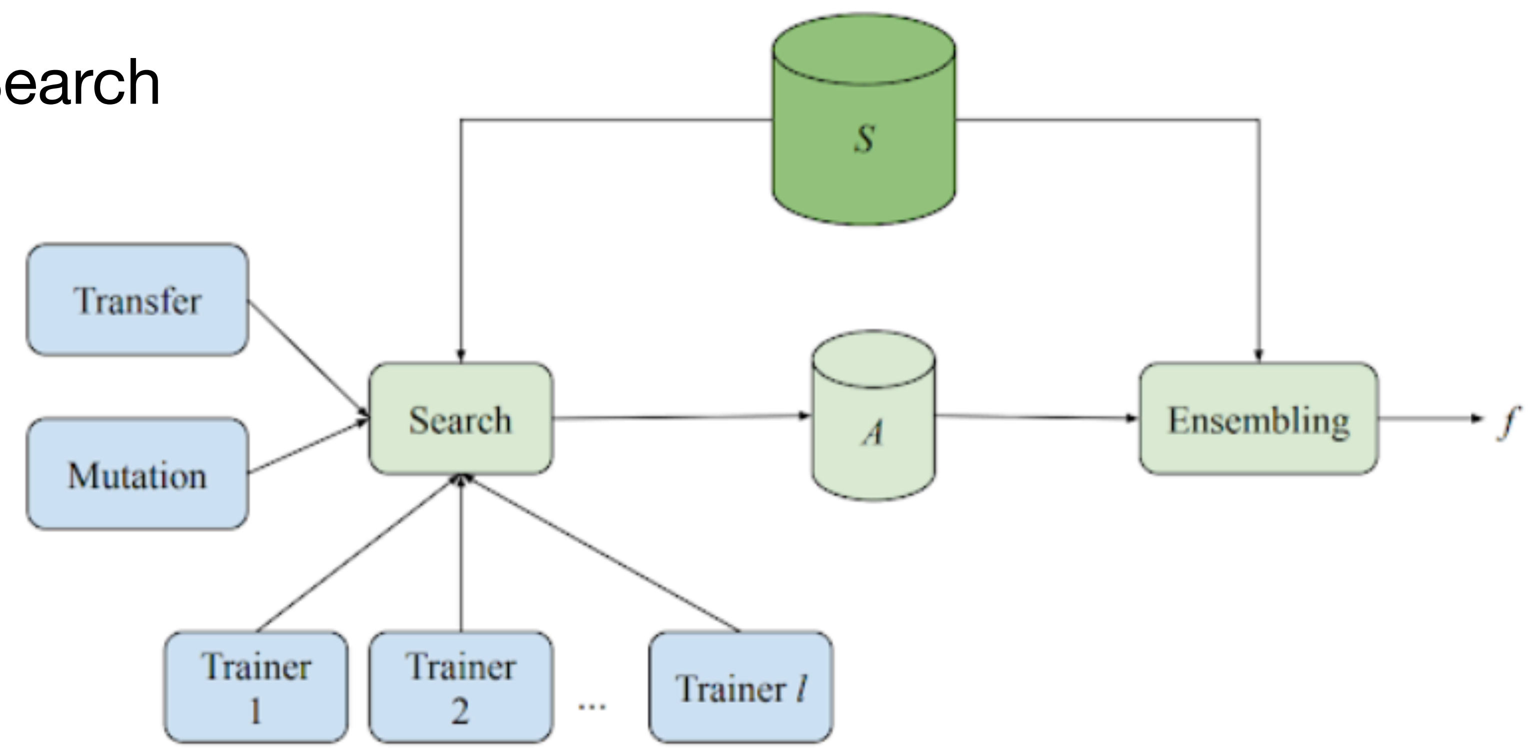
Friday, February 19, 2021

Posted by Hanna Mazzawi, Research Engineer and Xavi Gonzalvo, Research Scientist, Google Research

https://ai.googleblog.com/2021/02/introducing-model-search-open-source.html

AutoML and Neural Architecture Search
- Reinforcement learning
- Evolutionary algorithms
- Combinatorial search

What's new?



Model Search schematic illustrating the distributed search and ensembling. Each trainer runs independently to train and evaluate a given model. The results are shared with the search algorithm, which it stores. The search algorithm then invokes mutation over one of the best architectures and then sends the new model back to a trainer for the next iteration. S is the set of training and validation examples and A are all the candidates used during training and search.

Introducing Model Search: An Open Source Platform for Finding Optimal ML Models

Friday, February 19, 2021

Posted by Hanna Mazzawi, Research Engineer and Xavi Gonzalvo, Research Scientist, Google Research

https://ai.googleblog.com/2021/02/introducing-model-search-open-source.html

What it does

- train models asynchronously (using building blocks)
- use beam search to check completed tries and see what to try next
- mutation of best architectures for next round
- transfer learning & knowledge distillation:
  - match high-performing model's prediction in addition to maximizing prediction accuracy
  - copy suitable weights over to new models

"In a recent paper, we demonstrated the capabilities of Model Search in the speech domain by discovering"

INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

https://pdfs.semanticscholar.org/1bca/d4cdfbc01fbb60a815660d034e561843d67a.pdf

Figure 3: *Language identification accuracy while searching the top 5 architectures and the previous system.*

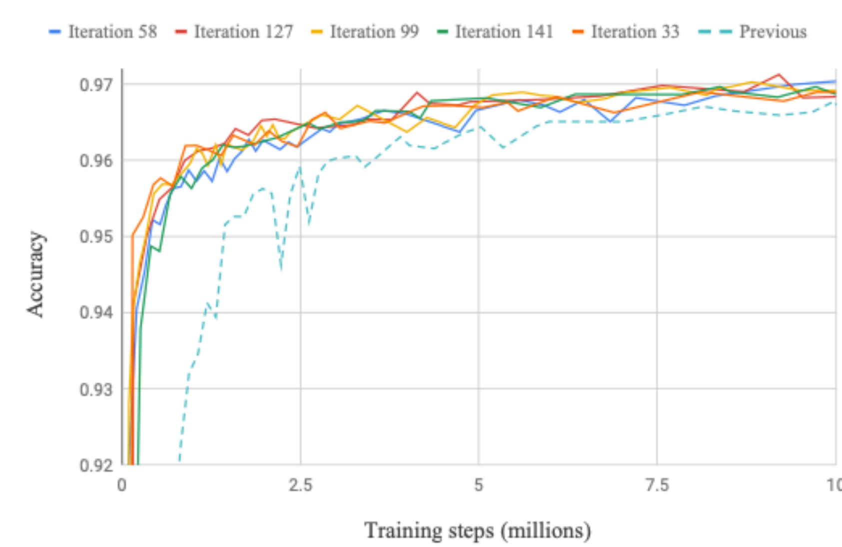


Figure 4: *Keyword spotting accuracy while searching the top 5 architectures and the previous system.*

arXiv.org > cs > arXiv:2102.07988

Search...

Help | Advanced S

Computer Science > Machine Learning

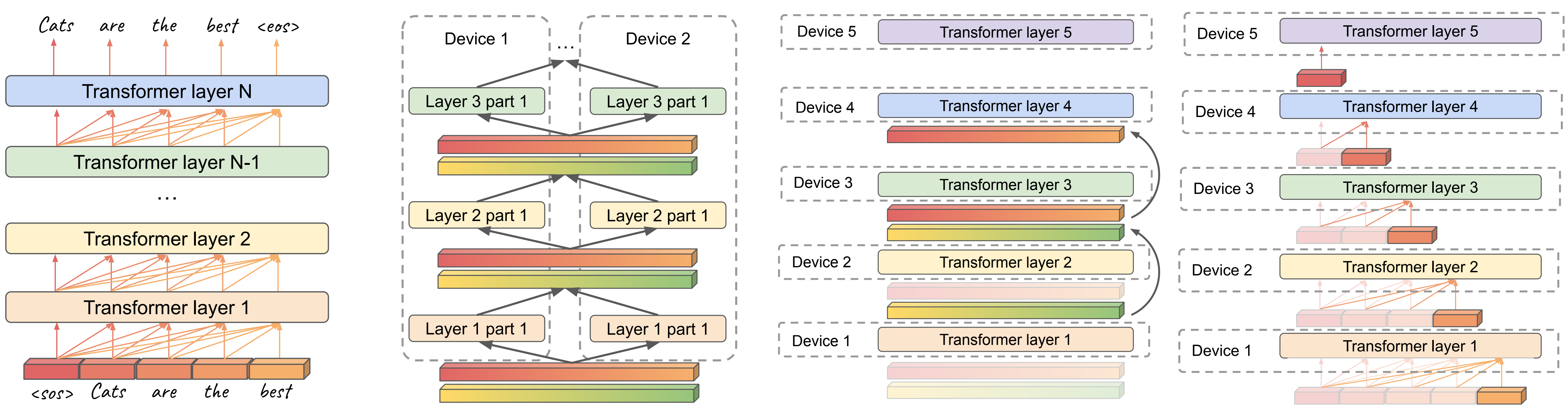
[Submitted on 16 Feb 2021]

TeraPipe: Token-Level Pipeline Parallelism for Training Large-Scale Language Models

Zhuohan Li, Siyuan Zhuang, Shiyuan Guo, Danyang Zhuo, Hao Zhang, Dawn Song, Ion Stoica

https://arxiv.org/abs/2102.07988

We show that TeraPipe can speed up the training by 5.0x for the largest GPT-3 model with 175 billion parameters on an AWS cluster with 48 p3.16xlarge instances compared with state-of-the-art model-parallel methods.



(a) Transformer-based LM

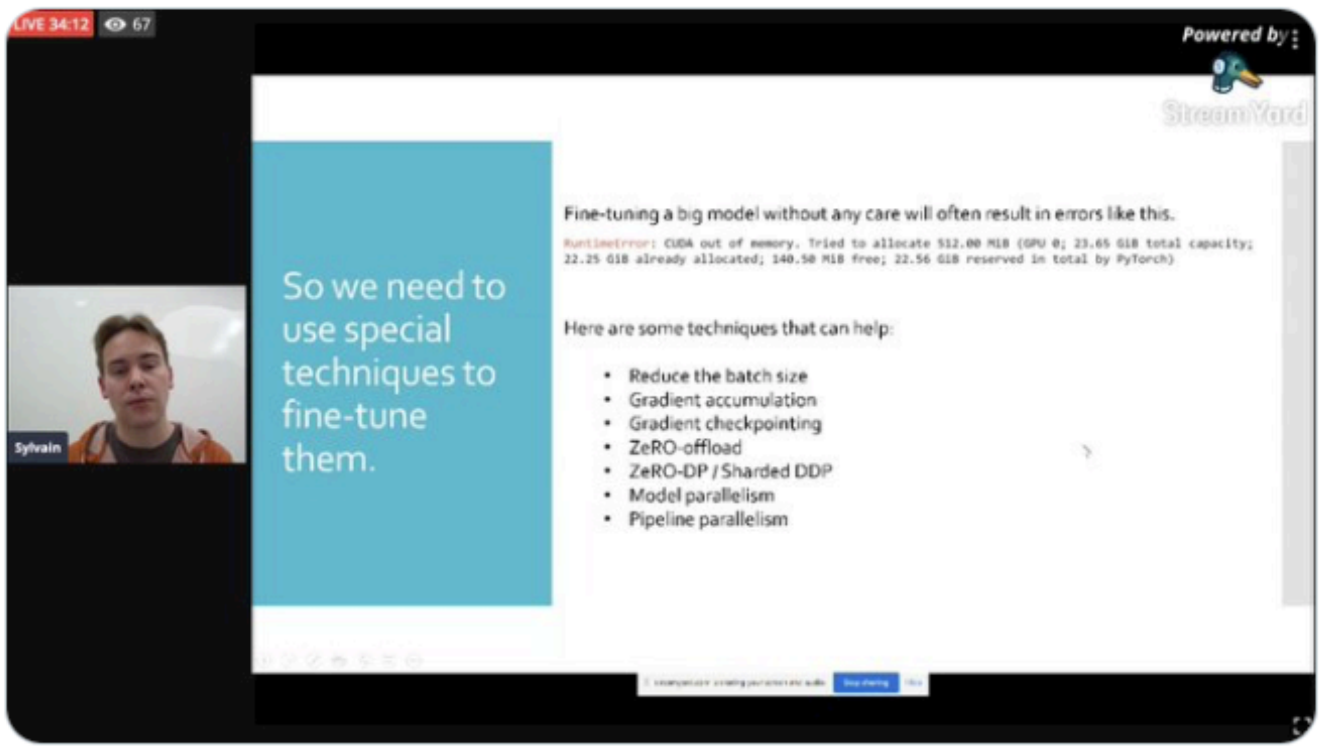
(b) Operation partitioning (Megatron-LM)

(c) Microbatch-based pipeline parallelism (GPipe)

(d) Token-based pipeline parallelism (TeraPipe)

Maria Khalusova
@mariaKhalusova

Techniques to fine-tune a large model on a single GPU, by @GuggerSylvain :

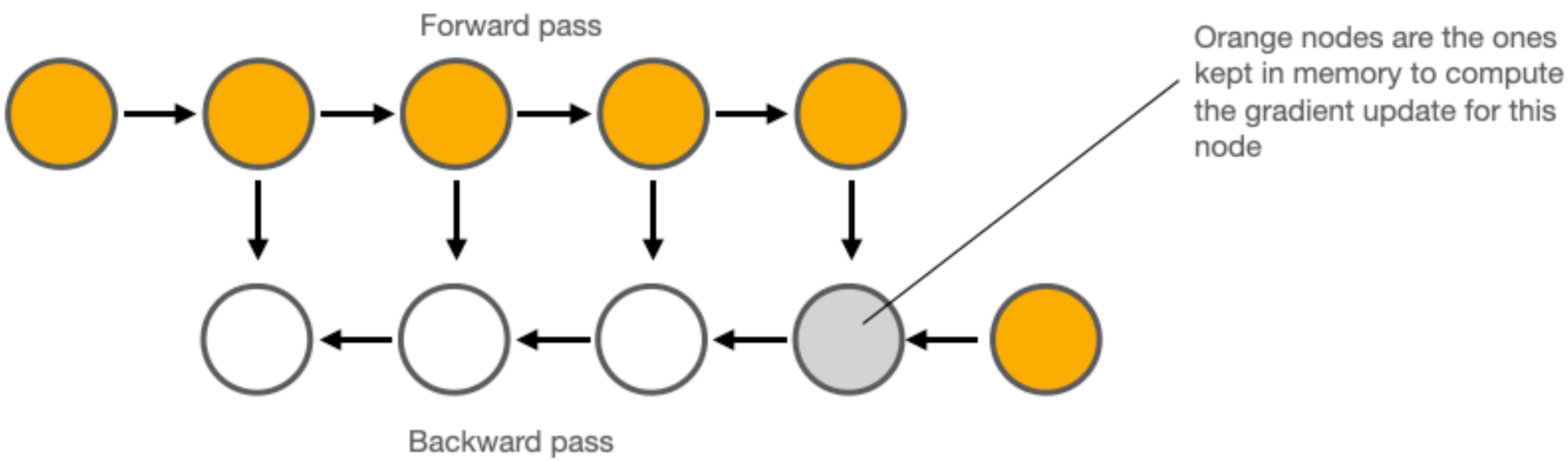


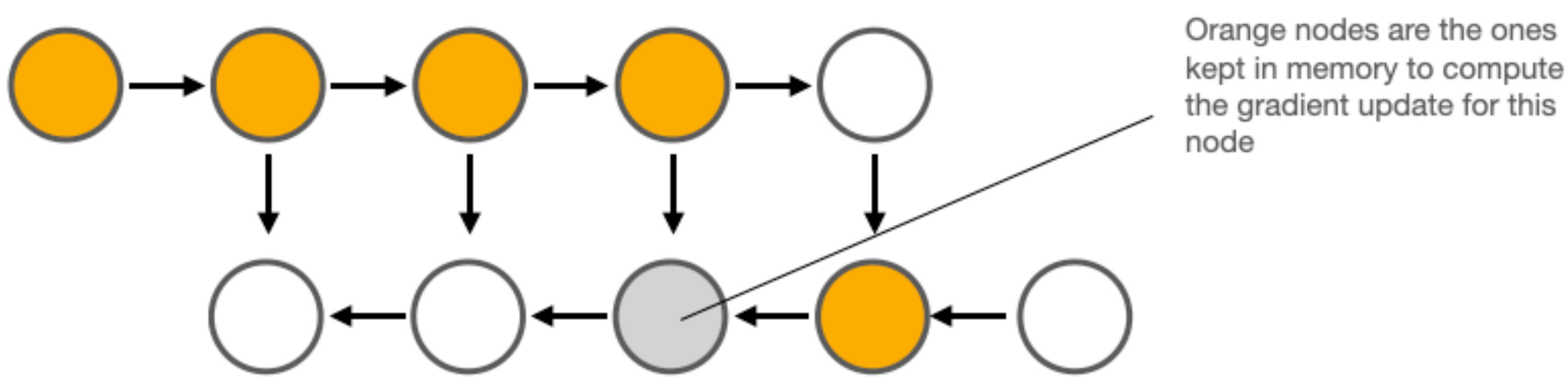
5:05 PM · Feb 25, 2021 · Twitter Web App

- Reduce the batch size
- Gradient accumulation
- Gradient checkpointing
- ZeRO-offload
- ZeRO-DP / Sharded DDP
- Model parallelism
- Pipeline parallelism

Gradient Checkpointing

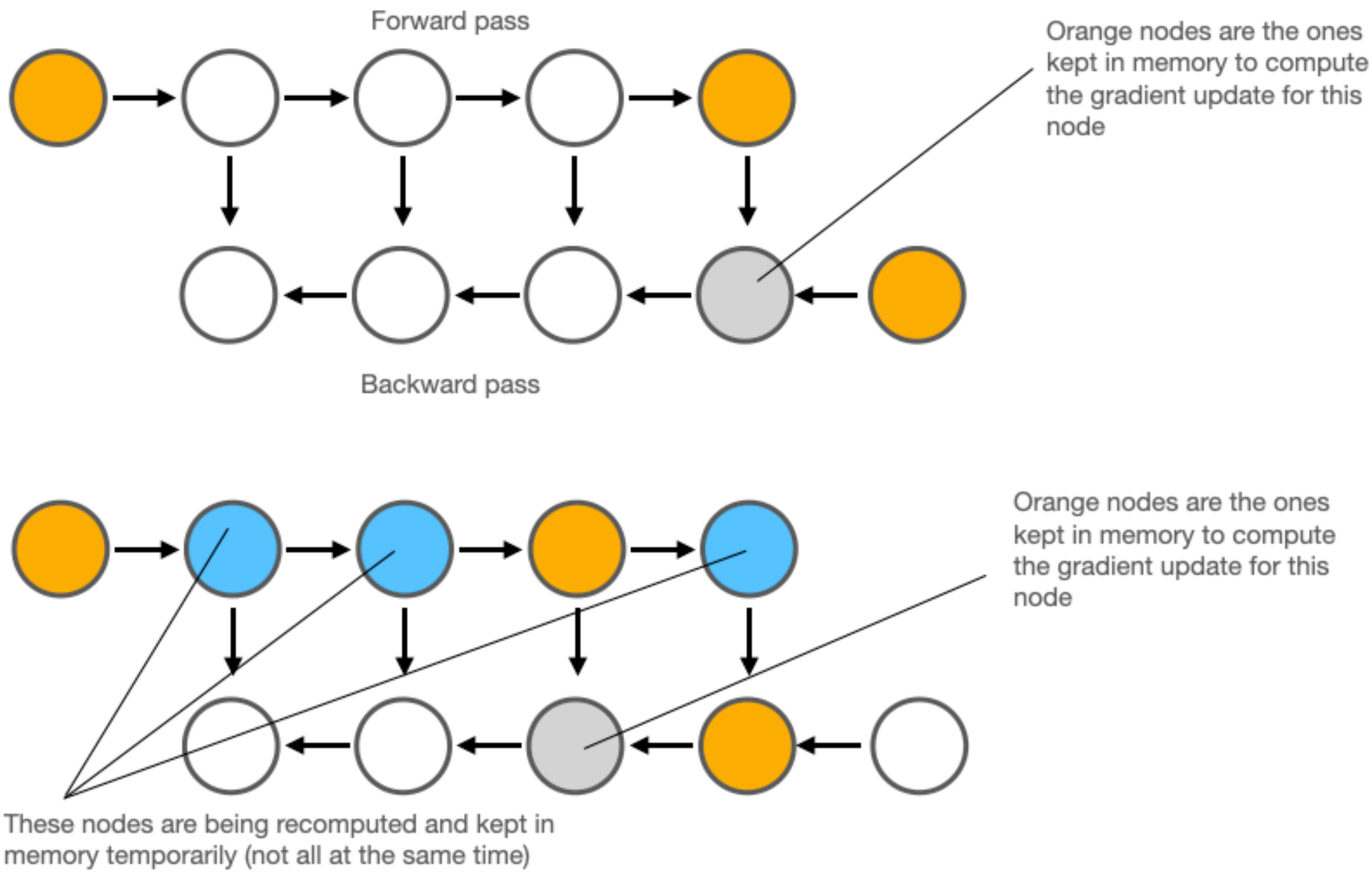
Drawing inspired by https://github.com/cybertronai/gradient-checkpointing





Gradient Checkpointing

Drawing inspired by https://github.com/cybertronai/gradient-checkpointing



Example: https://github.com/rasbt/deeplearning-models/blob/master/pytorch_ipynb/mechanics/gradient-checkpointing-nin.ipynb

Zero Redundancy Optimizer (ZeRO)

https://www.deepspeed.ai/tutorials/zero/

ZeRO is a powerful set of memory optimization techniques that enable effective FP16 training of large models with billions of parameters, such as GPT-2 and Turing-NLG 17B.

Compared to the alternative model parallelism approaches for training large models, a key appeal of ZeRO is that no model code modifications are required.

ZeRO reduces the memory consumption of each GPU by partitioning the various model training states (weights, gradients, and optimizer states) across the available devices (GPUs and CPUs) in the distributed training hardware

ZeRO-Offload

ZeRO-Offload is a ZeRO optimization that offloads the optimizer memory and computation from the GPU to the host CPU

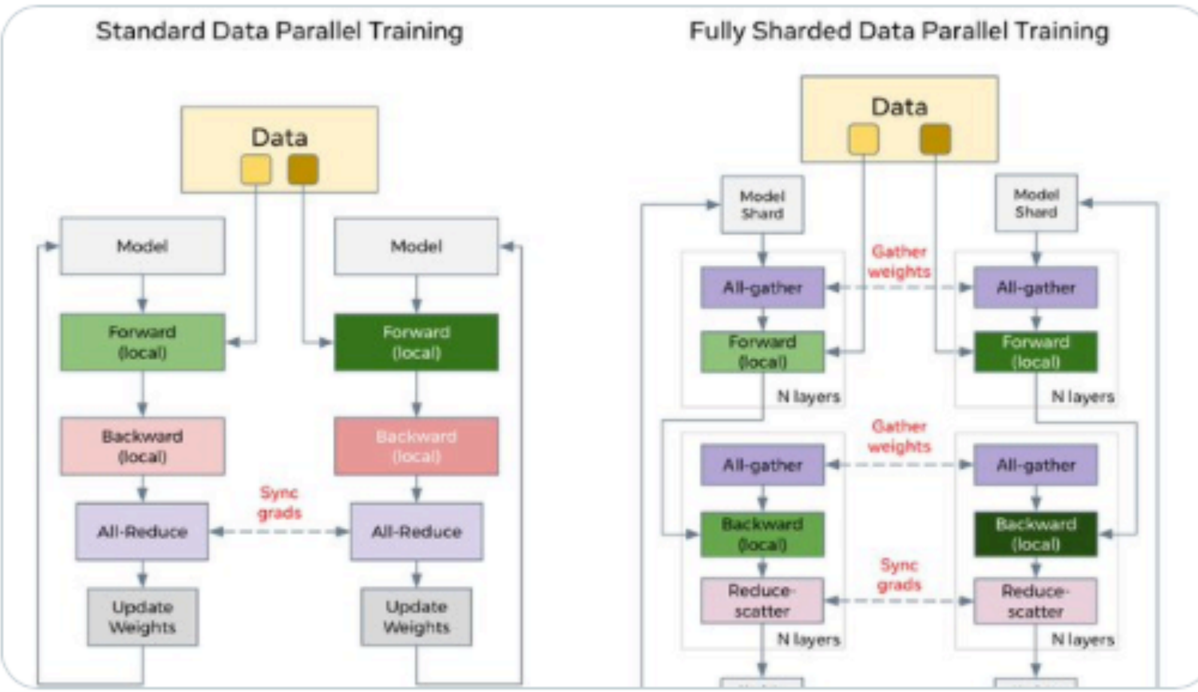
ZeRO-Offload enables large models with up to 13 billion parameters to be efficiently trained on a single GPU.

to prevent the optimizer from becoming a bottleneck, ZeRO-Offload uses DeepSpeed's highly optimized CPU implementation of Adam called DeeSpeedCPUAdam. DeepSpeedCPUAdam is 5X–7X faster than the standard PyTorch implementation

https://www.deepspeed.ai/tutorials/zero-offload/

PyTorch @PyTorch · 6h

FairScale, a PyTorch extension for efficient large scale training, is releasing FullyShardedDataParallel, which shards model params across GPUs (+offload to CPU). Details: github.com/facebookresear.... Inspired by DeepSpeed/@MSFTResearch, and made by @myleott @m1nxu @sam_shleifer



1 79 249

https://github.com/facebookresearch/fairscale/pull/413

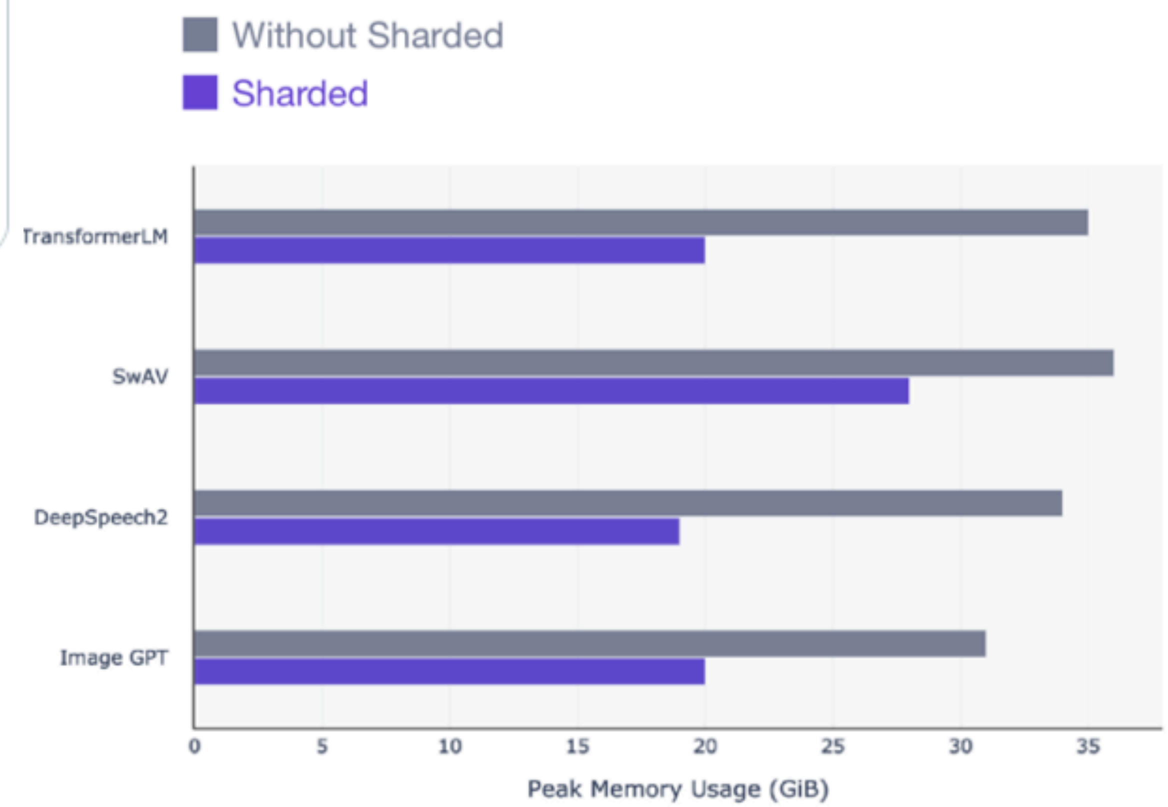


Image source: (Sean Narenth) (with modifications)

https://towardsdatascience.com/sharded-a-new-technique-to-double-the-size-of-pytorch-models-3af057466dba

FairScale

https://github.com/facebookresearch/fairscale

Pipe

Run a 4-layer model on 2 GPUs. The first two layers run on cuda:0 and the next two layers run on cuda:1.

```
import torch

import fairscale

model = torch.nn.Sequential(a, b, c, d)
model = fairscale.nn.Pipe(model, balance=[2, 2], devices=[0, 1], chunks=8)
```