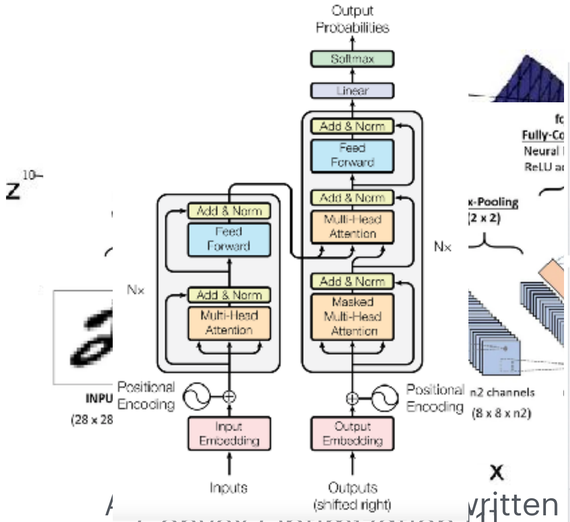
Training Large Models

Philipp Krähenbühl, UT Austin

A brief history

- Pre 2012: Small CPU-only models, convex optimization, limited performance
- 2012-2018: Single GPU models, better non-convex optimizers, better architectures
- 2019-2022: Multi-GPU models, multi-dataset models
- 2023-: Frontier models, internet-scale datasets

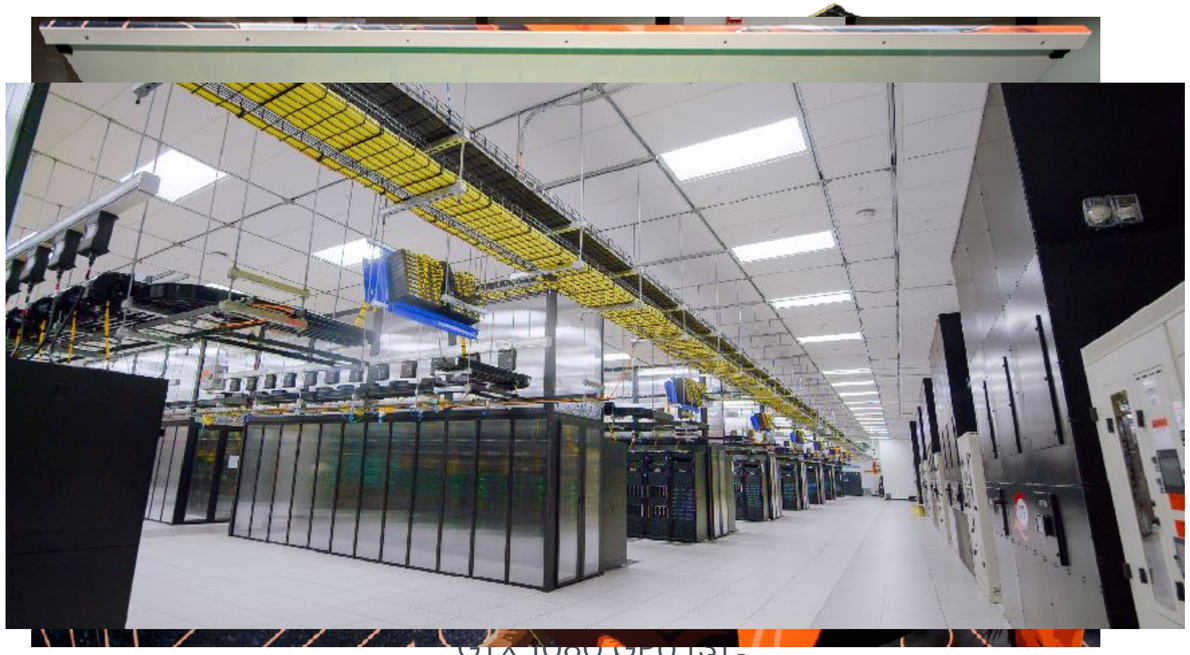


Convex Optimization [1]

Transformer Architecture [1]

Petabytes of Web Archive [2]

LAION 5B Dataset [2]



GTX 1080 GPU [3]

Meta's AI Cluster [3]

Lonestar6 system at TACC [3]

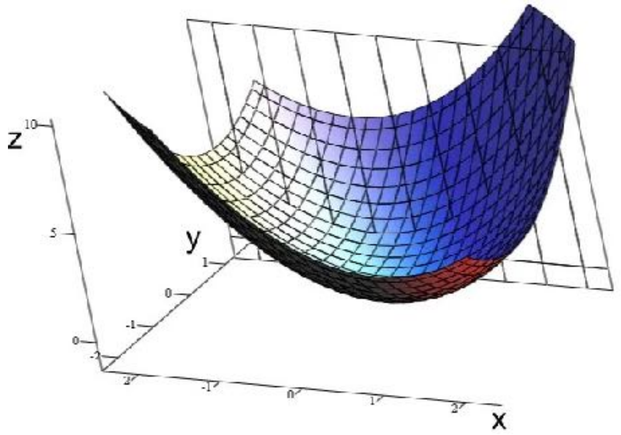
[1] Ashish Vaswani et al. Attention Is All You Need. 2017. Advances in Neural Information Processing Systems, 2017, 30. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[2] Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. 2009. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

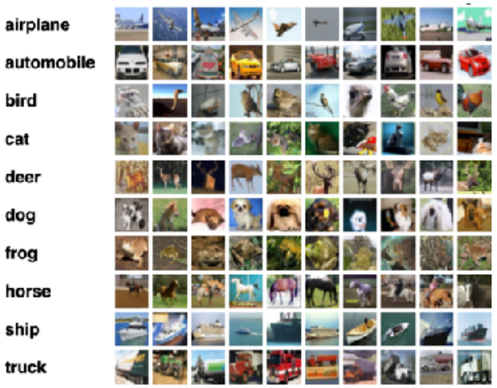
[3] Meta. Introducing the AI Research SuperCluster — Meta's cutting-edge AI supercomputer for AI research, 2022. https://ai.facebook.com/blog/ai-rsc/

Pre 2012

- Hand engineered features
- Small datasets
- Mostly convex optimization
- Resource Limitations:
  - Time: Human engineering
  - CPU compute



Convex Optimization [1]



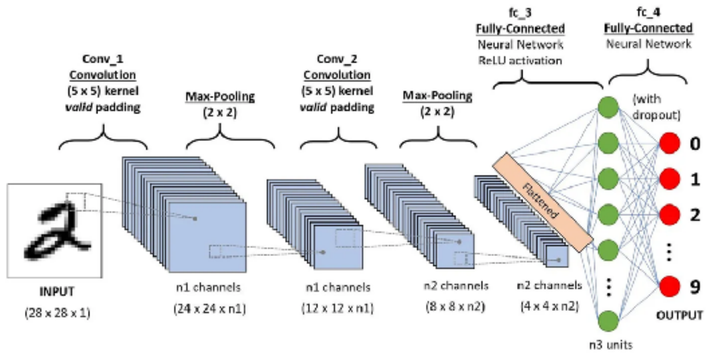
CIFAR10 Dataset [2]

CPUs [3]

[1] https://towardsdatascience.com/understand-convexity-in-optimization-db87653bf920
[2] Alex Krizhevsky, Learning Multiple Layers of Features from Tiny Images, 2009
[3] https://www.pcmag.com/news/cpu-showdown-intel-core-i3-vs-i5

2012-2018

- GPU-based deep networks
- human engineering -> GPU compute
  - Proliferation of large datasets
- Better optimization, network structures
- Resource Limitations
  - GPU compute
  - Good ideas



A CNN to classify handwritten digits [1]

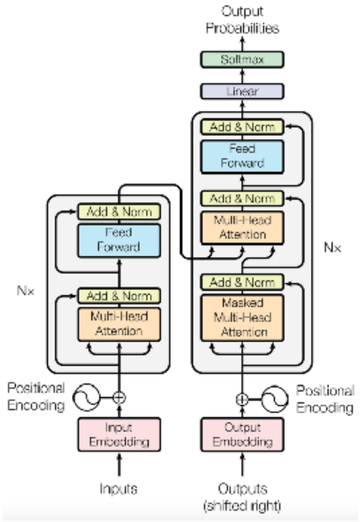
ImageNet images [2]

GTX 1080 GPU [3]

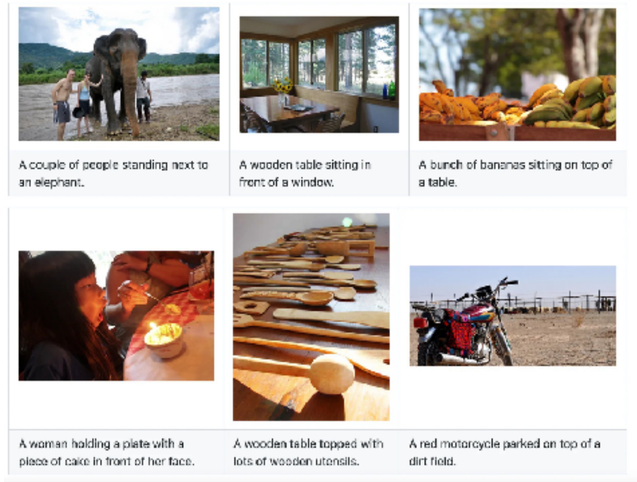
[1] Sumit Saha, A Comprehensive Guide to Convolutional Neural Networks, 2018
[2] Jia Deng et al., ImageNet: A large-scale hierarchical image database, 2009
[3] https://www.nvidia.com/en-us/geforce/news/nvidia-geforce-gtx-1080-ti/

2019-2022

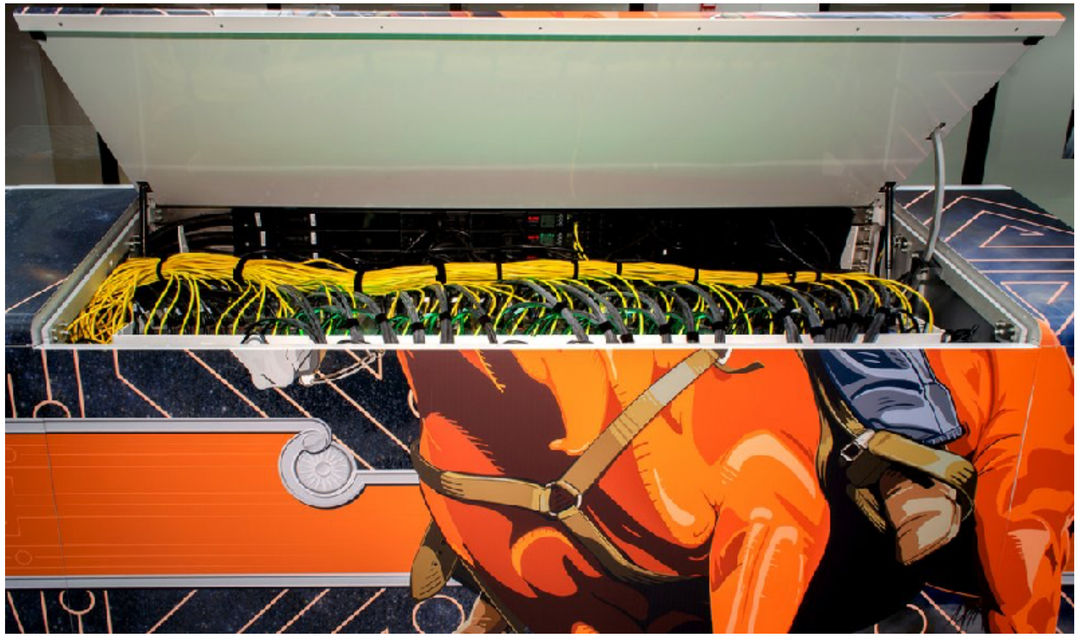
- Multi-GPU models
- Multi-dataset models
- Attention-based models
- Resource Limitations
  - GPU compute + memory
  - Good ideas



Transformer Architecture [1]



LAION 5B Dataset [2]

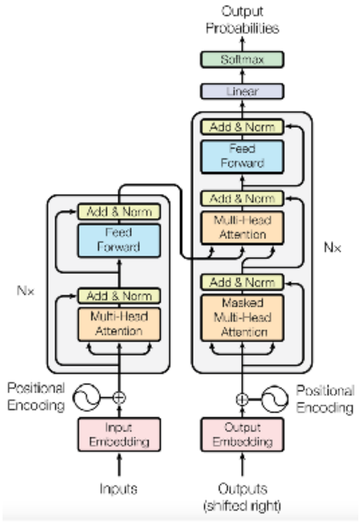


Lonestar6 system at TACC [3]

[1] Ashish Vaswani et al., Attention is all you need, 2017.
[2] Ludwig Schmidt et al., Laion-5b. 2022.
[3] TACC, https://tacc.utexas.edu/systems/lonestar6/

2023-

- Massive models (8B-400B parameters)
- Multi-Node training
- Internet-scale data
- Most basic architectures and infrastructure explored
- Resource Limitation
  - GPU memory



Transformer Architecture [1]

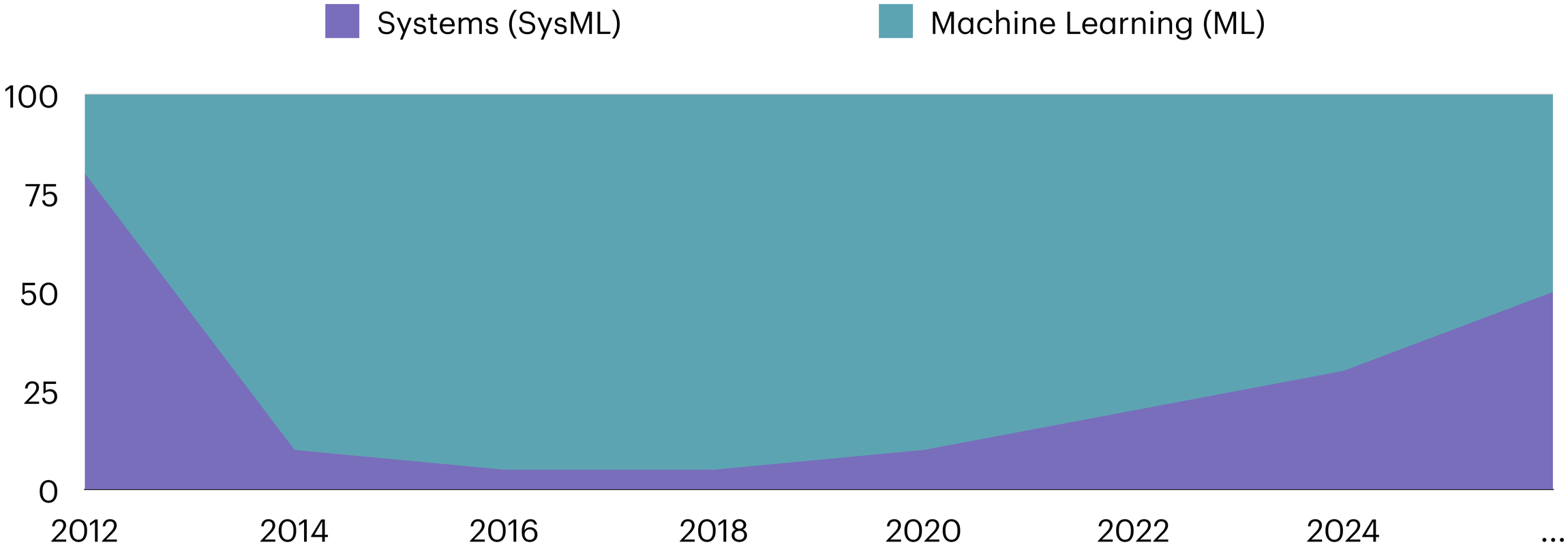
Common Crawl

Petabytes of Web Archive [2]

Meta's AI Cluster [3]

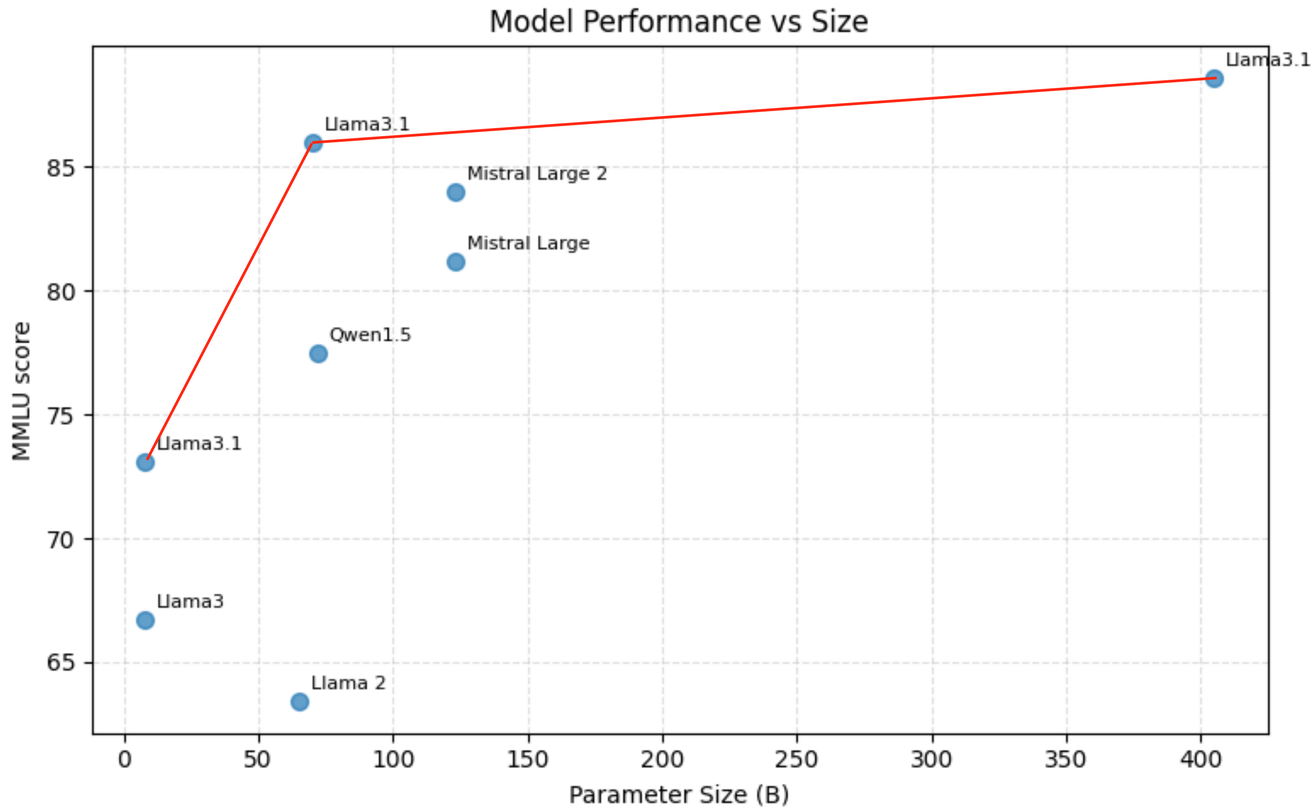
[1] Ashish Vaswani et al., Attention is all you need, 2017.
[2] Common Crawl. https://commoncrawl.org/.
[3] Meta, Introducing the AI Research SuperCluster — Meta's cutting-edge AI supercomputer for AI research, 2022

2012-now


Systems (SysML)
Machine Learning (ML)
100
75
50
25
0
2012
2014
2016
2018
2020
2022
2024
...

Training large models

- Mostly a systems issue
- GPU memory is expensive and limited
- Models are large
  - Larger models empirically work much better



[1] Llama Team, The Llama 3 Herd of Models, 2024
[2] Dan Hendrycks, et al. Measuring massive multitask language understanding, 2020

Training large models

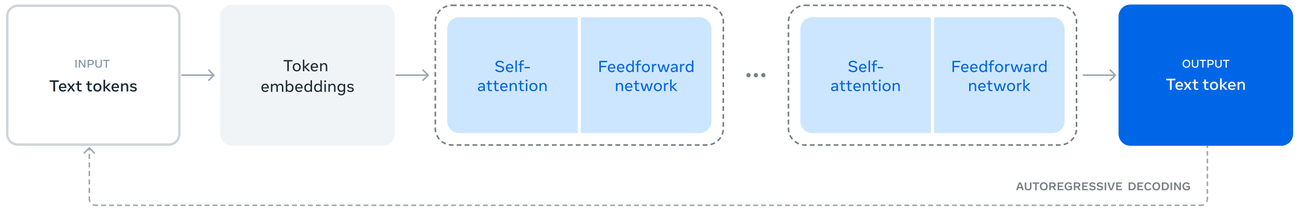
Memory requirements

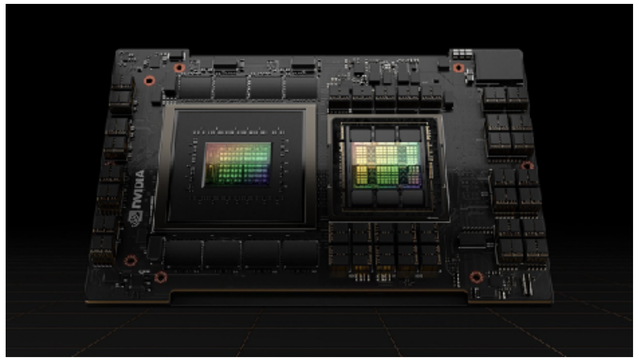
- Without optimization:
  - Model parameters: N
  - Weights: N floats
  - Gradients: N floats
  - Momentum: N floats
  - 2nd momentum (ADAM): N floats
- 16N bytes without counting activations

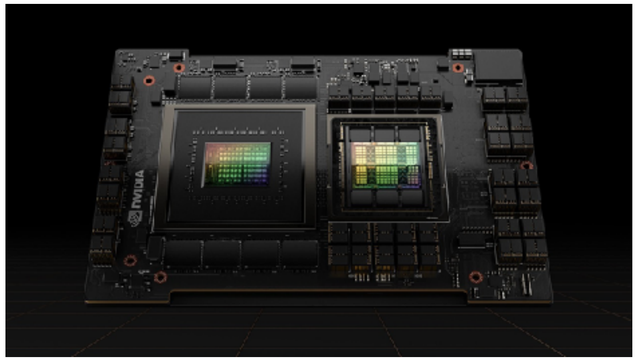
Meta

The Llama 3 Herd of Models

Llama Team, AI @ Meta[1]
[1]A detailed contributor list can be found in the appendix of this paper.





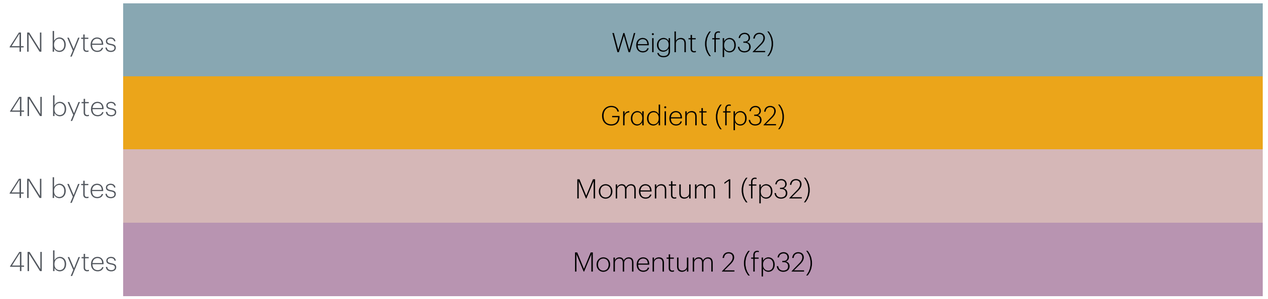


[1] Llama Team, The Llama 3 Herd of Models, 2024
[2] NVIDIA, https://www.nvidia.com/en-us/data-center/h100/

Training large models

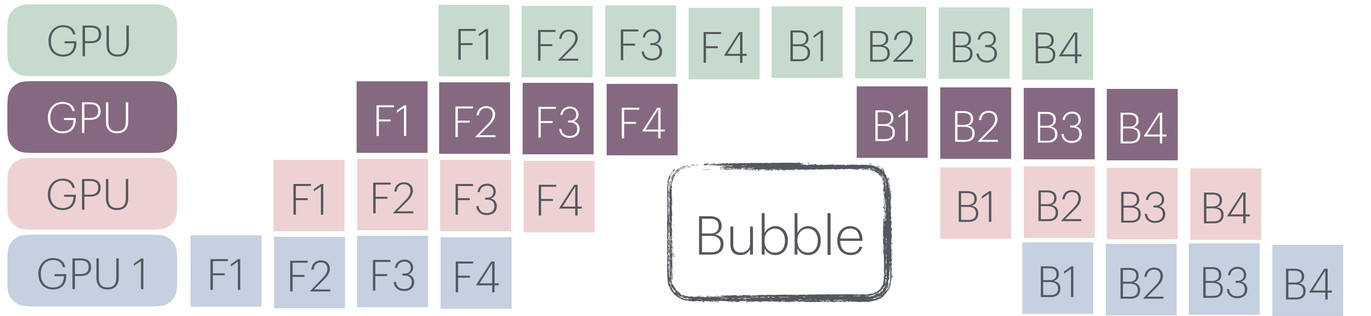
Memory requirements

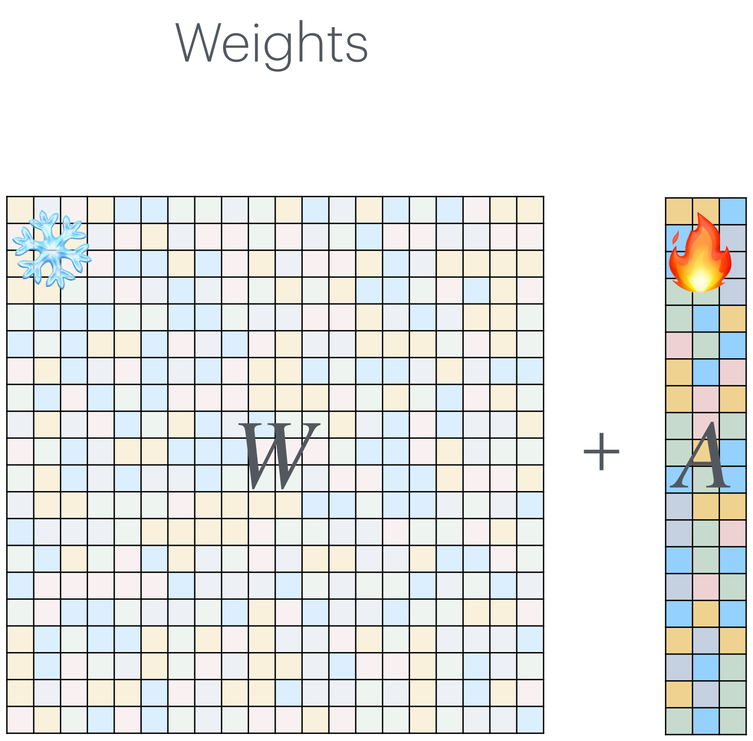
- Without optimization:
  - Model parameters: N
  - Weights: N floats
  - Gradients: N floats
  - Momentum: N floats
  - 2nd momentum (ADAM): N floats
- 16N bytes without counting activations



Training large models

- Mixed precision training
- Distributed Training
- Zero redundancy training
- Low-rank adapters
- Quantization
- Quantized Low-rank adapters
- Low-rank projections
- Checkpointing
- FlashAttention
- Open-source Infrastructure for model training





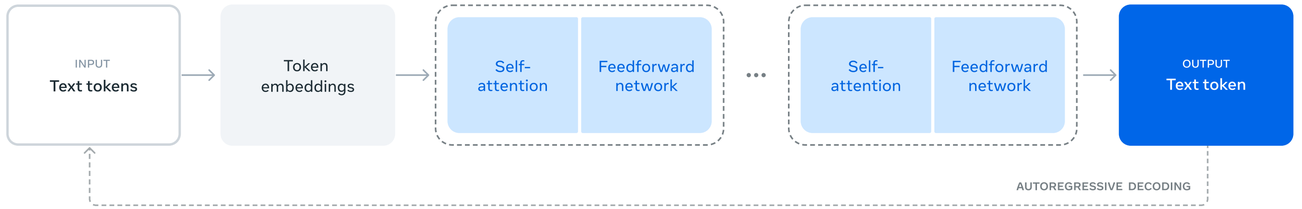
Training large models

- Teaser
  - Without optimizations 16N+ bytes
  - With all optimizations 1-2N bytes

Meta

The Llama 3 Herd of Models

Llama Team, AI @ Meta[1]
[1] A detailed contributor list can be found in the appendix of this paper.



[1] Llama Team, The Llama 3 Herd of Models, 2024
[2] https://www.nvidia.com/en-us/geforce/news/nvidia-geforce-gtx-1080-ti/

References

- [1] NVS Yashwanth, Why convexity is the key to optimization, 2020 (link)
- [2] Alex Krizhevsky, Learning Multiple Layers of Features from Tiny Images, 2009 (link)
- [3] Sumit Saha, A Comprehensive Guide to Convolutional Neural Networks, 2018 (link)
- [4] Jia Deng et al., ImageNet: A large-scale hierarchical image database, 2009 (link)
- [5] Ashish Vaswani et al., Attention is all you need, 2017 (link)
- [6] Christoph Schuhmann et al., LAION-5B. 2022 (link)
- [7] TACC, https://tacc.utexas.edu/systems/lonestar6/
- [8] Meta, Introducing the AI Research SuperCluster — Meta's cutting-edge AI supercomputer for AI research, 2022 (link)
- [9] Llama Team, The Llama 3 Herd of Models, 2024 (link)
- [10] Dan Hendrycks, et al. Measuring massive multitask language understanding, 2020 (link)