Speeding Up Learning and Inference

Danna Gurari

University of Colorado Boulder Spring 2025



https://dannagurari.colorado.edu/course/neural-networks-and-deep-learning-spring-2025/

Review

- Last lecture on model compression:
 - Motivation
 - Pruning
 - Knowledge distillation
 - Final project report: Overleaf tutorial
- Assignments (Canvas):
 - Final project outline due in one week
- Expect a Piazza announcement from Supriya Naidu about course feedback
- Questions?

Today's Topics

- Motivation
- Hardware tricks
- Architectural tricks
- Training tricks
- Programming tutorial

Today's Topics

- Motivation
- Hardware tricks
- Architectural tricks
- Training tricks
- Programming tutorial

Trend: More Compute-Intensive Models

Training Time (more FLOPs) Inference Time (more generated tokens)



e.g., A Common Training Situation



Boss: What did you do last month?

You: Trained the model for one epoch.





Boss: Umm, fine, what is your plan for next month?

You: Train... train the model for one more epoch?





https://hanlab.mit.edu/files/course/slides/MIT-TinyML-Lec13-Distributed-Training-I.pdf

Why Is Extensive Compute Undesirable?

- Time-consuming
- Expensive
- Increased environmental impact from carbon emissions

e.g., A Common Training Situation

Burning up



Energy used to train models, MWh

e.g., Current Inference Situation



"Google estimates that three-fifths of its total data-center energy use goes on billions of inference queries." "On average, a ChatGPT query needs nearly 10 times as much electricity to process as a Google search."

Economist Sep 21, 2024; https://www.goldmansachs.com/insights/articles/AI-poised-to-drive-160-increase-in-power-demand

When Is Extensive Compute Unrealistic?

- Learning in resource-constrained settings; e.g., on-device due to privacy concerns and poor/no Internet
- Inference in resource-constrained environments; e.g.,





https://www.ephotozine.com/article/19-thingsto-look-out-for-in-a-smartphone-camera--31055

https://en.wikipedia.org/wiki/ Wearable_technology



https://www.buzzfeednews.com/article/katienotopo ulos/facebook-is-making-camera-glasses-ha-ha-oh-no



https://aws.amazon.com/blogs/machine-learning/demystifyingmachine-learning-at-the-edge-through-real-use-cases/ How to develop and use AI models with less compute?

Today's Lecture: A Sampler of HW, Architecture, and Training Tricks



https://beyondmeresustenance.com/mexican-charcuterie-board/

Today's Topics

- Motivation
- Hardware tricks
- Architectural tricks
- Training tricks
- Programming tutorial

https://breetapp.com/blog/how-to-maximise-your-crypto-investment-portfolio



MAXIMIZE YOUR RETURNON INVESTMENT WITH THESE STRATEGIES

Ideas

- Quantization: reduce precision
- Operator fusion: reduce memory read/writes
- Caching: reduce computation
- Distributed optimization: parallelize computation

Quantization: Reduce Precision

- Which precision and ranges do different data types offer?
 - 32 bits: e.g., ~ ±3.4×10³⁸ with 2³² or ~4 billion values (default PyTorch type)
 - 8 bits: [-127, 127]
 - 4 bits: [-8, 7]



https://medium.com/@dillipprasad60/qlora-explained-a-deep-dive-intoparametric-efficient-fine-tuning-in-large-language-models-llms-c1a4794b1766

Quantization: Reduce Precision

• e.g.,

- what number of bits are used in this example?
- what bin value should be used for each of the 9 quantized values?



https://xailient.com/blog/4-popular-model-compression-techniques-explained/

Z-values Weights 1.5 0.2 1.0 0.1 0.5 Z-values Weights 0.0 0.0 -0.5 -0.1 -1.0-0.2 -1.5-0.3 h1 h2 h3 h4 h5 h1 h2 h3 h4 h5 Layers Layers Activations Gradients 0.02 0.5 0.01 Activations Gradients 0.0 0.00 -0.01 -0.5h2 h3 h4 h5 h1 h3 h4 h5 h1 h2 Layers Layers

Activation: tanh - Initializer: Glorot Normal - Epoch 0

https://towardsdatascience.com/hyper-parameters-in-action-part-ii-weight-initializers-35aee1a28404

How much precision is needed for training?

e.g., recall batch normalization causes values to center on 0 and range roughly between -1 and 1

e.g., 4 bits means values mapped to 16 floating point bins (i.e., 2⁴ possible values)



[-1., -0.7, -0.53, -0.39, -0.28, -0.18, -0.09, 0., 0.08, 0.16, 0.25, 0.34, 0.44, 0.56, 0.72, 1.] (Normalisation)

https://medium.com/@dillipprasad60/qlora-explained-a-deep-dive-intoparametric-efficient-fine-tuning-in-large-language-models-llms-c1a4794b1766

How much precision is needed for training?

e.g., recall batch normalization causes values to center on 0 and range roughly between -1 and 1



What are risks of using fewer bits?

What are risks and benefits of using fewer bits (e.g., when done with pruning)?

How much precision is needed for training?

e.g., recall batch normalization causes values to center on 0 and range roughly between -1 and 1

	Actual No. Of Parameters/ size	Actual Top-5 Error Rate (%)	Method Type	Para./size after compression	Compression Achieved	Top-5 Error Rate (%) after Compression	Speedup Achieved
AlexNet	61M/240 MB	19.7	Pruning	6.7M	9x	19.67	Зx
			Pruning and quantizatio n	6.9 MB	35x	19.7	Зx
VGG16	138M/512 MB	10.4	Pruning	10.3M	13x	10.88	5x
			Pruning and quantizatio n	11.3 MB	49x	10.91	3x to 4x

https://xailient.com/blog/4-popular-model-compression-techniques-explained/

Quantization Options

٠

e.g.,		A100 80GB PCIe		A100 80GB SXM		
	FP64 (Exceeds default)	9.7 TFLOPS				
	FP64 Tensor Core (Exceeds default)	19.5 TFLOPS				
	FP32 (default)	19.5 TFLOPS How much is the speed-up?				
	Tensor Float 32 (TF32)	(without *sparsity)	156 TFLOPS	312 TFLOPS*		
	BFLOAT16 Tensor Core		312 TFLOPS	624 TFLOPS*		
	FP16 Tensor Core		312 TFLOPS	624 TFLOPS*		
	INT8 Tensor Core		624 TOPS	1248 TOPS*		

https://www.nvidia.com/en-us/data-center/a100/

Quantization Options

• e.g.,



https://images.nvidia.com/aem-dam/en-zz/Solutions/data-center/nvidia-ampere-architecture-whitepaper.pdf

Quantization Options: A100 vs H200

NVIDIA H200 Tensor Core GPU

Form Factor	H200 SXM ¹	
FP64	34 TFLOPS	9.7 TFLOPS
FP64 Tensor Core	67 TFLOPS	19.5 TFLOPS
FP32	67 TFLOPS	19.5 TFLOPS
TF32 Tensor Core	989 TFLOPS ²	156 TFLOPS 312 TFLOPS*
BFLOATI6 Tensor Core	1,979 TFLOPS ²	312 TFLOPS 624 TFLOPS*
FP16 Tensor Core	1,979 TFLOPS ²	312 TFLOPS 624 TFLOPS*
INT8 Tensor Core Use for training?	3,958 TFLOPS ²	624 TOPS 1248 TOPS*

https://www.fibermall.com/blog/why-gpu-require-hbm.htm; https://www.nvidia.com/en-us/data-center/a100/

A100 80GB PCIe



https://developer.nvidia.com/blog/achieving-fp32-accuracy-for-int8-inference-using-quantization-aware-training-with-tensorrt/

DeepSeek-V3 now runs at 20 tokens per second on Mac Studio, and that's a nightmare for OpenAI



Awni Hannun 🤣 @awnihannun · Mar 24 🧭 · · The new Deep Seek V3 0324 in 4-bit runs at > 20 toks/sec on a 512GB M3 Ultra with mlx-lm!

https://venturebeat.com/ai/deepseek-v3-now-runs-at-20-tokens-per-second-on-mac-studio-and-thats-a-nightmare-for-openai/

Ideas

- Quantization: reduce precision
- Operator fusion: reduce memory read/writes
- Caching: reduce computation
- Distributed optimization: parallelize computation

Idea: Fuse Tasks to Reduce Overhead



https://quadric.io/2023/09/13/how-to-unlock-the-power-of-operator-fusion-to-accelerate-ai/

Approach: Fuse Tasks to Reduce Overhead

Avoid shuttling data back-and-forth to external memory by determining what operations to fuse so relevant information stays in limited memory close to the GPU

- When would you prefer L2 cache versus Load Register Memory?



https://quadric.io/2023/09/13/how-to-unlock-the-power-of-operator-fusion-to-accelerate-ai/

Approach: Fuse Tasks to Reduce Overhead

This approach is especially important given HW trends: processor performance (tripling every two years) exceeds that of memory access speed (~half as much)





(Economist, Sep 21, 2024)



https://quadric.io/2023/09/13/how-to-unlock-the-power-of-operator-fusion-to-accelerate-ai/

FlashAttention: Fuses Attention Operations



Dao et al. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. Neurips 2022

FlashAttention: Fuses Attention Operations



Dao et al. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. Neurips 2022

FlashAttention2



An estimated 2x speed-up compared to FlashAttention!

Dao. FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning. arXiv 2023

Ideas

- Quantization: reduce precision
- Operator fusion: reduce memory read/writes
- Caching: reduce computation
- Distributed optimization: parallelize computation

KV Caching: Store Keys and Values for Attention



(Operations repeated in future time steps have results stored in cache)

Limitations: uses memory inefficiently since needed cache size regularly changes and reserves lots of cache memory to support large sizes

https://medium.com/@joaolages/kv-caching-explained-276520203249

vLLM: Overcomes Limitations of KV Caching

- Inspiration: virtual memory with paging used in operating systems to limit memory fragmentation by allocating memory "pages" on demand
- Idea: an LLM serving engine using PageAttention, which allocates KV cache into fixed-size blocks (aka, "pages") that don't have to be contiguous



Figure 4. vLLM system overview.



Figure 5. Illustration of the PagedAttention algorithm, where the attention key and values vectors are stored as non-contiguous blocks in the memory.

Kwon et al. Efficient Memory Management for Large Language Model Serving with PagedAttention. SOSP 2023
Ideas

- Quantization: reduce precision
- Operator fusion: reduce memory read/writes
- Caching: reduce computation
- Distributed optimization: parallelize computation

Recall Multi-Thread Programming



https://www.cs.uic.edu/~jbell/CourseNotes/OperatingSystems/4_Threads.html

Distribute Neural Network Computations Across Multiple Devices (e.g., GPUs, CPUs)

- Key questions:
 - What tasks to delegate to each device?
 - What communication to facilitate across devices and how to do this?
- Popular Methods:
 - Distributed Data Parallel (DDP): each "machine" holds copy of model to perform forward and backward passes on different data subsets, with cross-machine communication (i.e., check-pointing) to identify average gradients across machines
 - If you had 4 GPUs, approximately what amount of speed-up might you expect?
 - Distributed Lower-Communication Training (DiLoCo): distributes training to multiple "islands" of a massive cluster (e.g., 100,000 GPUs), with regular checkpointing in each and less checkpointing across them, enabling devices to come from different sources (e.g., locations, companies, countries)
 - Zero Redundancy Optimizer (ZeRO): supports parallelism without code changes; partitions model training states (e.g., weights, gradients) across devices, enabling training of very large models (e.g., trillions of parameters)

Summary: How to Get a Greater Return on Your HW Investments

- Quantization: reduce precision
- Operator fusion: reduce memory read/writes
- Caching: reduce computation
- Distributed optimization: parallelize computation

Today's Topics

- Motivation
- Hardware tricks
- Architectural tricks
- Training tricks
- Programming tutorial

Ideas

• Mixture of experts: reduce computation

• Multi-token prediction: accelerate computation

Mixture of Experts

• Can activate only a subset of model parameters at inference time



Mixture of Experts: Expert Implementation



Mixture of Experts: Experts

• Focus is on specific tokens, rather than topics (e.g., computer science)



Mixture of Experts: Routers (aka, Gate Network)

• Given an input, the router selects the best-suited expert(s)



Mixture of Experts: Routers

• Given an input, the router selects the best-suited expert(s); e.g.,



Mixture of Expert: Dense vs Sparse Layer





Beneficial for inference time!

Learning challenge: load balancing to apply similar importance to many experts rather than overfitting to few experts

Mixture of Experts: Router Implementation



Mixture of Expert: Inference Time Set-Up



Mixture of Experts: Routers

• Paths can differ at different time steps for autoregressive models



Ideas

- Mixture of experts: reduce computation
- Multi-token prediction: accelerate computation

Multi-Token Prediction

Discarded at inference (or used to speed up model up to 3 times) 4-token 4 5 targets Head 1 Shared Inputs 2 3

Gloeckle et al. Better & Faster Large Language Models via Multi-token Prediction. April 2024

Use N output heads to predict N tokens at each time step:

Summary

- Mixture of experts: reduce computation
- Multi-token prediction: accelerate computation

Today's Topics

- Motivation
- Hardware tricks
- Architectural tricks
- Training tricks
- Programming tutorial

Pioneering Methods: Historical Context



Ideas

- Curriculum learning
- Dataset distillation

Accelerate Learning with a Curriculum

Random Order of Examples



Ordered Examples

Barry Fold 1	Terrent II Terrent II Terren	
terostanti los es falabilitos Tilo Lion (forman es falabilitos Regionaria de la propeitar integritos falaban falabante de la propeitar integritos terostanti los es falabilitos terostanti los este aconsecuentes estas estas problementes terostanti los este aconsecuentes estas estas problementes terostanti los estas estas estas estas problementes terostanti los estas estas estas estas estas problementes terostanti los estas estas estas estas estas problementes terostanti los estas	tero of experiments of the second sec	
Wing Use Polareau in Multi-Ministry Hack Accession at the ProMotion Pailable Research 1 Pailable Research <t< td=""><td>Wing Lip, Dolaring in Automatic 4 444.4 Accession Paidata basis 1 Basis Accession 4 Paidata basis 1 Basis Accession 4 Paidata basis 3 Basis Accession 5 Paidata basis 3 Basis Accession 5 Paidata basis 3 Basis Accession 7 Paidata Basis 3 Basis Accession 7<td></td></td></t<>	Wing Lip, Dolaring in Automatic 4 444.4 Accession Paidata basis 1 Basis Accession 4 Paidata basis 1 Basis Accession 4 Paidata basis 3 Basis Accession 5 Paidata basis 3 Basis Accession 5 Paidata basis 3 Basis Accession 7 Paidata Basis 3 Basis Accession 7 <td></td>	
Details factor Hardine Parties Balang instructions Markin Station If Balang instructions If Markin Station If Advectories If Markin Station If Advectories If Markin Station If Balang Instructions If Markin Station If Advectories If Markin Station If Balang Instructions If <t< td=""><td>Produkt Karin 1 Remote System Bildings Rotting Characterized 1 White, Stagert 10 Bildings Rotting Characterized 1 Region Characterized 10 Bilding Rotting Characterized 3 Region Characterized 10 Bilding Rotting Characterized 4 Region Characterized 10 Bilding Rotting Characterized 4 Region Characterized 10 Bilding Rotting Characterized 4 Region Characterized 10 Bilding Rotting Characterized 3 Region Characterized 10 State Characterized 30 Region Characterized 10 Region Characterized 10 State Characterized 3 Region Characterized 10 Region Characterized 10 State Characterized 30 Regenon Characterized 30 <t< td=""><td>a month a support</td></t<></td></t<>	Produkt Karin 1 Remote System Bildings Rotting Characterized 1 White, Stagert 10 Bildings Rotting Characterized 1 Region Characterized 10 Bilding Rotting Characterized 3 Region Characterized 10 Bilding Rotting Characterized 4 Region Characterized 10 Bilding Rotting Characterized 4 Region Characterized 10 Bilding Rotting Characterized 4 Region Characterized 10 Bilding Rotting Characterized 3 Region Characterized 10 State Characterized 30 Region Characterized 10 Region Characterized 10 State Characterized 3 Region Characterized 10 Region Characterized 10 State Characterized 30 Regenon Characterized 30 <t< td=""><td>a month a support</td></t<>	a month a support
Endong Restauctions Marking Network Marking Network Fedding Restauctions Marking Network Marking Network 9 Prest Fedding Marking Network Restauctions 10 Prest Fedding Prestauctions Restauctions 11 Prest Fedding Prestauctions Restauctions 11 Prest Fedding Prestauctions Restauctions 12 Prest Fedding Prestauctions Restauctions 13 Prest Fedding Prestauctions Restauctions 14 Prest Fedding Prestauctions Prestauctions 14 Prest Fedding Prestauctions Prestauctions 14 Prest Fedding Prestauctions Prestauctions 15 Prestauctions Prestauctions Prestauctions 16 Prestauctions Prestauctions Prestauctions 17 Prestauctions Prestauctions <td>Intermedia Appropriate Perifythe J Barry State Redding Mathematics States J Barry States States 1 State Field Barry States J Barry States States 1 State Field Barry States J Barry States States 1 State Field Barry States J Barry States States 1 State States States J Barry States States States States 1 State States States J Barry States States States States 1 States States States J Barry States States States States 1 States States States J Barry States States States States 1 States States States J Barry States Sta</td> <td>100 C</td>	Intermedia Appropriate Perifythe J Barry State Redding Mathematics States J Barry States States 1 State Field Barry States J Barry States States 1 State Field Barry States J Barry States States 1 State Field Barry States J Barry States States 1 State States States J Barry States States States States 1 State States States J Barry States States States States 1 States States States J Barry States States States States 1 States States States J Barry States States States States 1 States States States J Barry States Sta	100 C
Holding Hartingtonia Hart Field <	Holding Muttherman Barger In Barger Voltagile Docum J Barger In Jones Folds J Barger In Status Barger J Barger In Jones Tob Bard J Barger In Tobust Tobox Volutione J Barger In Barger India J Barger In Barger India J Barger India Barger India J	100 C
Interp Mathematical States A Strate Tybel B Marchall B Marchall B Marchall B Marchall B Marchall B Strate Tybel B Marchall B Strate Tybel B Strat Tybel B	Interpretation J Second Display J J Mart Hykk Second Display J Second Display J Mart Hykk Second Display Second Display Second Display J Mart Hykk Second Display Second Display Second Display Second Display J Person Second Display Text Hould J Person Person Mart Hould J Person Person Mart Hould Second Display Second Display Second Display J Anary Halls Second Display Second Display Second Display Second Display Display Display Second Display Second Display Text Display Display Display Second Display Second Display Text Display Display Display Display Second Display Second Display Text Display Display Display Display Display Second Display Second Display Text Display Display Display Display Display Second Display Second Display Text Display Disp	And the second se
Alter Fraik B Kurl Fraik B Ange Yoldi B Barne Too Book	Part Field Balance To Note Field 8 Note Field 8 Note Field 8 Note To To To To Make 9 Note To Note 1 Note To Note 1 Note To Note 1 Area To Note 1 Area To Note 1 Area To Note 10 Deven To Note 10 Note To Note 10 Deven To Note 10 Note To Not	Administration of the second s
11 Met 1940 2 12 Met 1940 3 13 Met 1940 3 14 Met 1940 3 15 Met 1940 3 15 Met 1940 3 15 Met 1940 3 16 Met 1940 3 16 Met 1940 3 16 Met 1940 3 17 Met 1940 3 18 Met 1940 3 19 Met 1940 3 19 Met 1940 3 10 Met 1940 3 10 Met 1940 3 11 Met 1940 3 11 Met 1940 3 11 Met 1940 3 12 Met 1940 3 13 Met 1940 3 14 Met 1940 3 14 Met 1940 3 14 Met 1940 </td <td>Inter Held Bartinov Bartinov Bartinov Martinov Bartinov Martinov Bartinov Martinov Distantinov Bartinov Bartinov Stata Bartinov Bartinov Bartinov Bartinov</td> <td>and and a second s</td>	Inter Held Bartinov Bartinov Bartinov Martinov Bartinov Martinov Bartinov Martinov Distantinov Bartinov Bartinov Stata Bartinov Bartinov	and and a second s
bit Bok 3- Edited Water Section (Section (Sec	Bart Bark 0 Bart Bark Finde Bark 3 Barts Brand Rock 3 Person Brand Rock 3 Person Jewer Velle 3 Person Jewer Velle 10 Barts Jewer Velle 10 Barts Jewer Velle 11 Barts Neue Barts Park 12 Barts Neue Barts Park 13 Barts Neue Barts Park 14 Barts Neue Barts Park 16 Barts Neue The Start Mark 10 Person Neue The Start Mark 10 Person Neue The Start Mark 10 Person Neue The Start Mark 10 Reson Neue Bart Mark 10 <	adox Sarina Al
Faster Buck 2 Permit 6 Branci Buck 3 Period 6 Marchals 32 Period 8 Marchals 32 Barthals 8 Marchals 32 Barthals 8 Marchals 32 Barthals 8 Marchals 32 Barthals 8 Barthals 12 San all finding 8 Statial Rush 11 Pergents 8 Barthals 12 San all finding 8 Barthals 13 Pergents 8 Barthals 14 Pergents 8 Barthals 15 Barthals 8 Barthals 16 Pergents 8 Barthals 17 Pergents 8 Barthals 18 Barthals 18 Parthals 18 Barthals 18 Barthals 19 Barthals 18 Barthals 19 Barthals 19 Barthals 10 Barthals 10 Barthals 10 Barthals 10 Barthals 10 Barthals 10 Barthals 10 </td <td>Finder Basel 2 Person Roual Rock 3 Person Terr Star Basel 9 Person Marchaek 9 Person Marchaek 10 Rest Marchaek 10 Rest Marchaek 10 Rest Start Rock 10 Rest Start Too Rock 10 Rest Provide Rock 10 Rest Start Too Rock 10 Rest Provide Rock 10 Rest Rock Too Rock 10<!--</td--><td>and some Theory and the state of the</td></td>	Finder Basel 2 Person Roual Rock 3 Person Terr Star Basel 9 Person Marchaek 9 Person Marchaek 10 Rest Marchaek 10 Rest Marchaek 10 Rest Start Rock 10 Rest Start Too Rock 10 Rest Provide Rock 10 Rest Start Too Rock 10 Rest Provide Rock 10 Rest Rock Too Rock 10 </td <td>and some Theory and the state of the</td>	and some Theory and the state of the
Broad Hork J Path A Jong Yolds B Papotas B Jong Yolds D Papotas B Maximal D Papotas B Maximal D Papotas B Maximal D Papotas B Maximal D Papotas B Jone The Book D Papotas Papotas Jone The Book D Papotas Papotas The The Book D Papotas Papotas Papotas Papotas Papotas Papotas Papotas Papotas Papotas Papotas Papotas Papotas Papotas Papotas Papotas Papotas Papotas Papotas Papotas Papotas Papotas Papotas Papotas Papotas Papotas Papotas Papotas Papotas Papotas Papotas Papotas Papotas Papotas Papotas Papotas Papotas <td>Broad Rock J. Personal Jown Tar Mark W Personal Jown Tarkin W Personal Jown Tarkin Personal Bartison Jown Tarkin Personal Bartison Particial Personal Bartison Particial Personal Bartison Particial Personal Bartison Particial Personal Bartison Tarkin Rock Personal Personal Partison Personal Personal Partison Personal Personal Personal Personal Personal Personal Personal Personal Personal Personal Personal Part Park Personal Personal Paronal Paris Personal Personal<!--</td--><td></td></td>	Broad Rock J. Personal Jown Tar Mark W Personal Jown Tarkin W Personal Jown Tarkin Personal Bartison Jown Tarkin Personal Bartison Particial Personal Bartison Particial Personal Bartison Particial Personal Bartison Particial Personal Bartison Tarkin Rock Personal Personal Partison Personal Personal Partison Personal Personal Personal Personal Personal Personal Personal Personal Personal Personal Personal Part Park Personal Personal Paronal Paris Personal Personal </td <td></td>	
The first the second	Ten Dar Bank,	
J. Mary Hells Jamma Hamman 3 Marchank J. Deal Norther System 9 Parket Res 11 Adjetter Paremet and American 9 J. Jack Ham 12 Adjetter Paremet and American 9 J. Jack Ham 12 Adjetter Paremet and American 9 J. Jack Ham 12 Adjetter Paremet and American 9 J. Jack Ham 12 Paremet and American 9 The Table Back 14 Paremet and American 9 The Table Back 16 Paremet and American 9 Paremet Field Into 16 105050 16 Back Table 9 Paremet Field Into 16 105050 16 Back Table 9 Paremet Field Into 16 105050 16 Back Table 9 Paremet Field Into 16 105050 16 Back Table 9 Paremet Field Into 16 16 Paremet and Deparement 9 Paremet Field Into 16 16 Paremet and Into 5 9 Ange Into 16 16 Mack Into 16 9	J. Area rolate. Internant It Macrinauk J. Di Parada Rism J. Di Parada Rism J. Di Status Rism J. Di Theor Tab Rism J. Di Theor Tab Rism J. Di Theor Tab Rism J. Di Parater Tab Rism J. Di Rismedia Parater J. Di Rismedi Parater	
Marcinals J2 Tail Nuclei Sprace N Junct Non 12 Aplicar Parents and Nucleims N Schell Nucl. 12 Schull Statistics Schull Statistics Schull Statistics Schull Nucl. 13 Schull Statistics Schull	Marchank Jo Built Namk Jaster John Jo Mark Name Binnen Fold 22 Alaphnacy Far Scheid Rock 21 Scheid Rock Experiment Scheid Rock 25 Experiment Scheid Rock Phore Tob Sock Veintenen 26 Experiment Scheid Rock Prace Tob Scheid Rock 27 Marstenik Rockope Fold 28 Scheiderer Scheiderer Rockope Fold 29 Scheiderer Scheider Rockope Fold 29 Scheiderer Scheiderer	
Parket Ren 11 Alghter Person and American Dama PAM 12 Statist Renk 12 Alghter Person and American Dama PAM 12 Statist Renk 11 Person in Dama PAM 12 There Tak Benk 13 Equition 16 There Tak Benk 16 Equition 16 Present Fold in Motion 16 Equition 17 Experiment Renk Valuemen 17 Ender Stream 17 Experiment Renk Valuemen 16 Person and Tournam 12 Experiment Renk 16 Person and Tournam 10 Experiment Renk 17 Maximum 16 Auge Renk 17 Maximum 16 Experiment Renk 17 Maximum 16 Experiment Renk 17 Maximum 16 <tr< td=""><td>Provid liem 11 Appliest Part Brain Fold 12 Special Part Their Hold 12 Special Part Their Fold 12 Special Part Their Tak Book 10 Special Part Part Tak Book 10 Special Part Part Tak 10 Special Part Part Tak Book 10 Part Tak Part Tak Book 10 Part Tak Part Tak Book 10 Part Tak Part Tak Book 20 Monsmith Part Tak Book 20 Monsmith Accession Book 20 Monsmith Part Tak Book 20 Monsmith Accession Book 20 Monsmith Part Tak Book 20 Monsmith Part Tak Book 20 Monsmith Part Tak Book 20 Monsmith <td>- Bernar</td></td></tr<>	Provid liem 11 Appliest Part Brain Fold 12 Special Part Their Hold 12 Special Part Their Fold 12 Special Part Their Tak Book 10 Special Part Part Tak Book 10 Special Part Part Tak 10 Special Part Part Tak Book 10 Part Tak Part Tak Book 10 Part Tak Part Tak Book 10 Part Tak Part Tak Book 20 Monsmith Part Tak Book 20 Monsmith Accession Book 20 Monsmith Part Tak Book 20 Monsmith Accession Book 20 Monsmith Part Tak Book 20 Monsmith Part Tak Book 20 Monsmith Part Tak Book 20 Monsmith <td>- Bernar</td>	- Bernar
Base Feld 12 Description 9 Jack Feld 11 Expendent 9 Stars Feld field 11 Expendent 9 Stars Feld field 11 Expendent 9 Stars Feld field 12 Expendent 9 Stars Feld field 12 Expendent 9 Present Fuld field 16 Expendent 9 Large Feld field 16 Matigates 2 Varial of Anti- 20 Massath and Polyment 2 Tass Feld from 20 Massath and Polyment 2 Tage Tell field 20 Massath and Polyment 3 Associated field 20 Massath 30 Associated field 20 Massath 30 Associated field 21 Tass field field 12 Associated field 21 Tass field field 12 Associated field 21 Tass field fi	Bases Feld 12 Sensative Jane Held 10 Sensative Start Held 10 Sensative Pare Held Balancia Balancia Larevic Lett Ford 10 Marine Pare Held Balancia Balancia Balancia 10 Marine Pare Held Balancia Balancia Balancia 10 Marine Balancia 10 Marine Pare Held 10 Marine Mark Held 10 Senative Senative 10 Senative <td>and and furniture</td>	and and furniture
April Malk Sachall Rock Tarloit Rock Tar	Area Male Annual Real Territoria T	
Statist Red. 11 Peperer 2 Stars Vie Red. 51 Inputtion 55 Poissel Field Noted. 55 Inputtion 57 Poissel Field Noted. 56 Inputtion 57 Poissel Field 66 National Poissel 57 Part Field 16 National Poissel 57 Levice Lett. Boh. 17 Managht ref. 57 National Poissel 18 National Poissel 57 National Poissel 18 National Poissel 57 National Poissel 18 National Poissel 57 National Poissel 20 National Poissel 58 National Poissel 21 National Poissel 58 National Poissel 21 National Poissel 58 National Poissel 23 National Poissel 58 National Poissel 24 National Poissel 58 National Poissel 24 National Poissel 58 Poissel Conis an National	Scholl Real 11 Poperate Bare Tak Bond, Valatime 15 Imputtive Possed Fact to Mobili 16 Regaritive Lareve Last, Post 17 Managative Fast Fable Note 16 Separate Russlag Take 26 Separate Russlag Take 27 Massage Pag Tak Bond 20 Separate Associate Real 27 Network Associate Real 28 Poster Associate Real 29 Real Associate De Vestion Real 27 Real Associate De Vestion Real 29 Real Associate De Vestion Real 29 Real Associate De Vestion Real 27 Real Associate De Vestion Real 29 Real Associate De Vestion Real 29 Real Associate	
There Too Bend. 34 Ignition 55 Present Food in Motion 56 Respective. 57 Present Food in Motion 56 Relation and Parameters 57 Present Food in Motion 56 Relation and Parameters 57 Present Food in Motion 56 Maintain and Parameters 57 Present Food in Motion 56 Maintain and Parameters 57 Present Food in Motion 58 Present and Parameters 58 Maintain Parameter 57 Maintain and Parameters 58 Maintain Parameter 57 Maintain 58 Marketer of Para 57 Maintain 58 Parameter Motion 57 Maintain 58 Parameter Marketer of Para 58 Present 58 Parameter Maintain 58 Present 58 Parameter Marketer of Para 58 Praintain 58 Parameter Maintain 58 Praintain 58 Parameter Marketer Marketer 58 Praintain	Theor Table Deck. 34 Hyperities Theor Table Notations 35 Happetties If Ares Table 36 Happetties If Ares Table 36 Happetties If Ares Table 36 Happetties Interpret Date Pool 36 Poorers and Mandache Table Tables 36 Poorers and Mandache Table Tables 36 Superstress Mandache Superstress Mandache Table Tables 36 Mandache Superstress Mandache Superstress Mandache Tables 37 Mandache 37 Mandache Tables 36 Tooles Tooles Happetties Tables 37 Mandache 38 Tooles Tables 16 Superstress Happetties Apple Tables 17 Mandache Apple Apple Tables 18 Mandache Apple Apple Tables 18 Mandache 38 Mandache Tables 18 <td></td>	
There for book Variance 5 Beginnine 8 Process for bit 90000 00 Baladian and Panaman 2 # Part fold 01 Part fold Part fold 2 # Part fold 01 Manage fold 2 Part fold 2 Part fold 01 Manage fold 2 Manage fold 2 Part fold 01 30 Manage fold 3 3 Working fold 20 Manage fold 3 3 3 Part fold 20 Manage fold 3<	Theor for back Values 5 Imagenroe Present Fact in Motion 5 Imagenroe # Per Fall 5 Imagenroe Larver Last Post 67 Material Fast Fall Nota 67 Material Fast Fall Nota 67 Material Fast Fall Nota 68 Person at Fast Fall Nota 68 Person at Fast Fall Nota 68 Person at Fast Fall Nota 20 Material Tay Tai Fast And 20 Material Advertim Brest 34 Person at Fast Fall 50 Account Tay Tai Fast Andres 50 Account Tay Tai Fast Andres 50 Account Tay Tai Fast Andres Prest 50 Account Tailer Tails 50 Account Tailer Tails 50 Account Tailer Tails 50 Account Tailer Tails 50 Account Tailer Tailer Tailer 50 Accoun	
Present Fold is 16500 (h) Balance of Panness 2 # Rev Fold (f) Hadron 2 Balance of Panness (f) Manualia and Polynomial 2 Preserve of Explored (f) Manualia and Polynomial 2 Weinstein Reve (f) Manualia and Polynomial 2 Weinstein Reve (f) Manualia 1 Typ Xie Finds (f) Manualia 1 Anyonither Brest (f) Anyon 1 Particip Controls (f) Anyon 1 Convert Wingh (f) Particip Controls 1 Convert Wingh (f) Particip Controls 1 Weinshare Root (f) Manuali Controls 1 Weinshare Root (f) Ma	Proceed First in Model (b) Balance # Ren Field (b) Feature 1 Levrice Last, Pool (c) Massaine Radia Field (c) Massaine Accessing Field (c) Massaine Accessing Field (c) Massaine Reg Construct of Parti (c) Reg Partie Reg Construct of Parti (c) Reg Partie Reg Construct of Partie (c) Reg Partie Reg Construct Construct One (c) Reg Partie Reg Construct Construct One (c) Reg Partie Reg Construct Construct One (c) Reg Partie	
# New York Pattern 2 Larvin Link Book Print Tell Malagian 2 Pain Tell Book Print Tell Book Print Tell Book 2 Bandage Yold Print Tell Book Print Tell Book 2 Bandage Yold Print Tell Book 2 Print Tell Book 2 Bandage Yold Print Tell Book 2 Print Tell Book 2 Bandage Yold Print Tell Book 2 Print Tell Book 2 Bandage Yold Print Tell Book 2 Print Tell Book 2 Bandage Yold Print Tell Book Print Tell Book 2 Print Tell Book 2 Bandage Yold Print Tell Book Print Tell Book 2 Print Tell Book 2 Print Tell Book Print Tell Book Print Tell Book 2 Print Tell Book 2 Print Tell Book Print Tell Book Print Tell Book 2 Print Tell Book 2 Print Tell Book Print Tell Book Print Tell Book 2 Print Tell Book 2 <t< td=""><td>Part Fuld: Post</td><td>of Postson</td></t<>	Part Fuld: Post	of Postson
Der Viel Call, Bohl IV Malaste. 2 Der Viel Call, Bohl IV Messache all Deparation 2 Der Viel Call, Bohl IV Messache all Deparation 2 Strandard, Frank IV Machine 1 Der Viel Call IV Machine 1 Der Viel Call IV Machine 1 Aussider of Paris IV Machine 1 Der Viel Call IV Apple 1 Der Viel Call IV Apple 1 Problem (Intell IV Apple 1 Converting Both IV Apple 1 Converting Both IV Paris 1 Converting Both IV Paris 1 Vielsbarr, Both IV Vielsbarr, Both 1 Vielsbarr, Both IV Vielsbarr,	Poer Faste Levine Land, Post Le	
La refer Lafa, José Pais Edi Boré San Jali Pole San Jali Pole	Larger Link peer	
Instruction	Number Frank Observer Namedari Frank Sin Ang Samitiro of Parts Barris Dar Systemic Sin Philips frank Sin Namedari Samitiro of Parts Barris Dar Systemic Sin Namedari Samitiro Color of Craffiel Sin Namedari Samitiro Samitiro Sin Namedari Samitiro Sin <	and Polymour and States of States
Stocking Color Dis Sequence Dis Page Tail Rock 20 Markeen Re Taig Tail Rock 20 Markeen Re According Rock 20 Markeen Re According Rock 20 Markeen Re Ray Keenistor of Paries 20 According Rock Re Dating Inter Parie 20 According Rock Re Problem Inter Parie 20 According Rock Re Problem Inter Parie 20 According Rock Re Problem Inter Parie 20 Prolong Re Concup Ring Rock 20 Prolong Tolescontrage Re Transford Rock	Wandag Fale 3a Sequence First-Sour Souri 27 Markent Markent Tag Ta Brack 27 Markent Markent Jag Van Brack 28 Print Markent Jang Van Brack 29 Markent Print Print Jang Van Brack 29 Markent Root Auge Thating sine Filtitititititititititititititititititit	Exponent (9
Free dure times 2: Maximum 8: Top Xie Rush 20 Observing Party Train 10 Any Konster of Paris 20 Paris 10 Train 10 Any Konster of Paris 20 Paris 80 <	Free Save Seven 21 Marcon Tap To B Real 20 Ryseway Accession Breat 34 Frein Ang Savetine Mreat 34 Frein Ang Savetine Mreat 20 Reat De Typ Brief 25 Acapits Print Partie 26 Acapits Print Partie 27 Acapits Print Partie 26 Acapits Print Partie 27 Acapit Breat Print Partie 27 Acapits Print Partie 27 Acapit Breat Print Partie 28 Print Partie Partie 28 Print Partie	and the bootstand of the second
Top Too Read 32 Obviously According Read 34 Friday 8 Say Londow of Press Reay Reag 8 Dar (by Brain 25 Applite 8 Dar (by Brain 25 Applite 8 Problem int PMIn 26 Applite 8 Problem int PMIn 28 Prolagen 8 Concept Right Read 88 Values 8 Values (soft Read 88 Values 8 Values (soft Read 88 Values 8 Concept Right Read 88 Values 8 Values (soft Read 88 Values 8 Values (soft Read 88 Values (soft Readward) 8 Values (soft Read 8 Values (soft Readward) 8 Values (soft Readward) 8 Values (soft Readward) 8 Values (soft Readward) 8 Values (soft Readward) 8 Values (soft Readward) 8 Values (soft Readward) 8	Top Tet Road 20 Overlap According Road 24 Print Jang Mandher of Persi 1000 mol1 Record Day (by Brist 25 Record Thilding sime Pathics 25 Augin Water Chains Chain or Fragity 10 Application Water Chains Chains or Fragity 10 Application Charter Stratts 28 Printing Charter Stratts 28 Printing Charter Stratts 28 Printing	1000 - 000 - 11 - 000 - F
Advection filted 34 Friefm 34 Ang Konster of Paris Loco and Loc Segment 8 Day Ty British 25 Augits 8 Day Ty British 25 Augits 8 Politing for Paris 26 Augits 8 Politing for Paris 26 Augits 8 Converting British 26 Politing 8 Converting British 26 Politish 6 Converting British 26 Politishing 8 Values 26 Politishing 8 Values 26 Politishing 8 Values 26 Politishing 8 Values 26 Politishing 8 Politishing Paris 27 Right Diagrammers 8 Politishing Paris 26 Politishing Paris 8 Politishing Paris 26 Politishing Paris 8 Politishing Paris 26 Politishing Paris 8 Politishi	Accession front	
Jay Kandor of Paris Loc Segment B Dar (1) finit 25 Applin B Toking into PMIn 25 Applin B Toking into PMIn 26 Applin B Toking into PMIn 26 Applin B Toking into PMIn 27 Parist B Converting Brain 28 Toking in Control B Converting Brain 28 Toking in Control B Toking in Control 27 Toking in Control B Toking in Control 27 Toking in Control B Toking in Control 28 Toking in Control B Toking in Control 29 Toking in Control B Toking in Control 20 Digit Toking in Control B Toking in Control 20 Digit Toking in Control B Toking in Control 20 Digit Toking in Control B Toking in Control 20 Digit Toking in Control B Toking in Control 20 <td>Any Konster of Parts Learning Transmission of Parts Book State Sta</td> <td></td>	Any Konster of Parts Learning Transmission of Parts Book State Sta	
Ang Konstor & Pans Bach 3 Day 1% Bins 25 Aaglin 8 Public part Public 21 Aaglin 8 Public part Public 24 Panel 8 Convert Work 24 Panel 8 Convert Work 24 Panel 8 Convert Work 26 Panel 8 Visionality 26 Visionality 6 Visionality 26 Visionality 8 Visionality 26 Visionality 8 Visionality 26 Visionality 8 Visionality 26 Visionality 8 Visionality 27 Visionality 8 Project State 27 Visionality 8 Project State 27 Visionality 8 Project State 28 Visionality 8 Project State 27 Visionality 8 Project State 28 Visionality	Ang Sensor of Yang The (1)% Boots (2)% Augest The large new Tables (2)% Augest The large transmission (2)% Augest The large transmissi	Jac September
Dat of p limit 27 Auglin 18 Polding into Philin 26 Auglin Schlassmithigs 18 Polding into Philin 27 Auglin Schlassmithigs 18 Polding of Tubes Chait with Net 28 Polding of Chait Schlassmithigs 18 Conversition Start Basis 28 Polding of Chait Schlassmithigs 16 Conversition Start Basis 28 Polding of Chait Schlassmithigs 16 Conversition Schlassmithig 28 Polding of Chait Schlassmithigs 16 Polding Schlassmithig 28 Polding of Chait Schlassmithigs 16 Polding Schlassmithigs 28 Polding of Chait Schlassmithigs 16 Polding Schlassmithigs 29 Polding of Chait Schlassmithigs 17 Polding Schlassmithigs 20 Polding of Polding Schlassmithigs 18 Polding Schlassmithigs 21 Polding of Polding Schlassmithigs 18	Teday in the Web and and the Web and the Web and the Web and and the Web and t	A statement of the second second
Description Description <thdescription< th=""> <thdescription< th=""></thdescription<></thdescription<>	Vestor Table Chel or Frage Noted Table Chel or Frage Noting a Carlo law Tealls Chet ringe Chet ringe The Carlo Law Tealls	
Policy and Carl Other Unit Policy 2 Constrained and 2 Policy Constrained Constra	Postog a Conti late Tanta	Gentline
Christiani 28 Polyan 6 Christiani 28 Viagan 6 Viagan 20 Viagan 7 Viagan 1990 Viagan 7 Viagan 1990 Viagan 7 Viagan 1990 Viagan	Civerback 28 Polyan Volgen	
Comparing Real III Tradition III Traditional Real III Registration III Traditional Real IIII Registration IIII Papers Industry Real Community III Selection Property III Selection Real IIII Real IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII	Concern Mar Bank Bridger	
Tophology Bod 21 Bight Topgan A Projects inling fails Dight Topgan Dight Topgan Dight Topgan Projects inling fails Qualification Dight Topgan Dight Topgan Dight Topgan Billional Proper II Projects Collegies Dight Topgan Dight Topgan		201
Poplet Single Toperatory 0 Poplet Single Toperatory 0 Difficult Poplet 2019 Difficult Poplet 21 Poplet Scheduler 0 Difficult Poplet 2019	Washing Red II High Dist	giri A
Figures (concerning frage) (conc	Rept 26a	fa prinement to the
Billion Line and Line	Projects scolag Asian	No
	Stimul Paper 31 Papers 8	statistics and Ground

Big Book of Math; Dinah Zike

How did you learn math?

Accelerate Learning with a Curriculum

Random Order of Examples

Ordered Examples







How did you learn to read?

Key Questions In Creating a Curriculum

- How many levels to include in the curriculum from easy to hard?
- How to define what is "easier" versus "harder"?

How to Evaluate a Curriculum?

- How good is the model? generalization performance on test data
- How long did learning take? training convergence speed

Pioneering Task: Shape Prediction

Classify each shape as rectangle, ellipse, or triangle



Solution: 3-layer neural network

1. Easy (Basic): less shape variability (squares, circles, and equilateral triangles); 10,000 examples

2. Hard (Geom): more shape variability (rectangles, ellipses, and triangles); 10,000 examples

Shape Prediction: Curriculum Learning

Results when training on easier examples for *n* epochs and then training on harder examples until 256 epochs or validation error hits minimum (20 random initializations).

What are the benefits of curriculum learning?

How many epochs should we train the model before introducing hard examples?



No curriculum

Bengio et al., Curriculum Learning, 2009

Ideas

- Curriculum learning
- Dataset distillation

Distill Large Dataset Into a Small Number of Synthetic Images



Yiu, Lu, and Wang. Dataset Distillation: A Comprehensive Review. PAMI 2023

Pioneering Paper



Synthetic training images enable more efficient training (i.e., less training images and so gradient descent steps) while reducing data storage costs and bypassing privacy concerns

Typical Approach: Many Optimization Objectives

```
Algorithm 1: Dataset Distillation Framework
 Input: Original dataset T
 Output: Synthetic dataset S
 Initialize S
                                ▷ Random, real, or core-set
 while not converge do
      Get a network \theta \triangleright Random or from some cache
      Update \theta and cache it if necessary
                               \triangleright Via \mathcal{S} or \mathcal{T}, for some steps
      Update S via \mathcal{L}(S, \mathcal{T})
                   ▷ PerM, ParM, DisM, or their variants
 end
 return S
```

Yiu, Lu, and Wang. Dataset Distillation: A Comprehensive Review. PAMI 2023





e.g., Pioneering 2018 paper "optimize[s] a synthetic dataset such that neural networks trained on it could have the lowest loss on the original dataset" "key idea... train the same network using synthetic datasets and original datasets for some steps, respectively, and encourage the consistency of their trained neural parameter" rather than match impacts on training, "obtain synthetic data whose distribution can approximate that of real data"

https://blog.roboflow.com/what-is-dataset-distillation/

Summary

- Curriculum learning
- Dataset distillation

Today's Topics

- Motivation
- Hardware tricks
- Architectural tricks
- Training tricks
- Programming tutorial

Today's Topics

- Motivation
- Hardware tricks
- Architectural tricks
- Training tricks
- Programming tutorial

