The **AlgoPerf**: Training Algorithms Benchmark Faster neural network training through better training algorithms

Frank Schneider

September 14, 2023



Neural Networks are extremely useful models...

In practice, neural networks are...

...but!

- ► ...**slow** to train (training can easily take days or weeks)
- ► ...tedious (demanding human intervention and fiddeling)
- ► ...expensive (requiring dozens of trial runs)

Neural Networks are extremely useful models...

In practice, neural networks are...

...but!

- ▶ ...**slow** to train (training can easily take days or weeks)
- ▶ ...tedious (demanding human intervention and fiddeling)
- ► ...expensive (requiring dozens of trial runs)

We all want neural network training to be faster, more automatic, and more efficient!

The state of deep learning training methods

A confusingly crowded field of methods & hyperparameters

A huge number of training methods...

Optimized Addam CTI CTI CTI CTI CTI CTI CTI CTI CTI CTI	Ref.
Adules [11] Card Monitation [16] No. Kain Adultation [17] [17] [16] No. Kain Adultation [17] [17] [16] No. Kain Adultation [16] [16] No. Kain [16] No. Kain Adultation [16] [16] [16] No. Kain	014
Addem III Offen III Peld Addraft III Peld III Peld Addraft III Pedd III Peld Addraft III Pedd III Peld Addraft III Pedd Pedd Pedd Addraft III Pedd Pedd Pedd Addraft III Pedd Pedd Pedd Addraft III Pedd Pedd Pedd Pedd Addraft III Pedd Pedd III Pedd	12.4
Addition Till Carinelial Till Person 20 - Angel Addition Till Darwey Till Notes Addition Till	123-1
Addata 1-2 Data 1-2 Data Addotad 1-2 DAta 1-2 Data Addota 1-3 DAta 1-2 Data Addota 1-4 DATa 1-2 Data Addota 1-3 Data 1-2 Data Addota 1-4 Data 1-2 Data Data Addota 1-3 Data 1-3 Data Data Addota 1-4 Data 1-4 Data Dat	1241
Addition Control Control Finite State Finite State Addition Control Control Finite State Addition Finite State Finite State Kontrol Additiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii	64 T221
Addition Chief (h) Persphere (h) Col Persphere (h) Addition Chief (h) Persphere (h) Col Addition Addition Chief (h) Col Col Addition Addition Chief (h) Col Col Col Addition Col Col Col Col Col Addition Col Col Col Col Col Col Addition Col	0.0
Adding Ti Billion (2014) Ti Outlane (2014) Addin Ti Different Ti Outlane (2014) Addin Ti Outlane (2014) Ti Solid Addin Ti Outlane (2014)	1211
ADAMA First Exclusion First Exclusion ADAMA First	10.00
ADAMA First Exclusion First Exclusion ADAMA First	100
Adds 11 Fm 1 Register Avaperty 11 Presco 11 Register Addem 111 Cathery 11 Register Avaperty 111 Cathery 11 Register	100
Addem 110 Description 110 Methods Addrift, L. 111 Red addrift, L. 111 Methods Addrift, L. 111 Red addrift, Red	1100
Aut/TML TML MaxAbility TML MaxAbility TML MaxAbility Aut/Markata 100 100,	15.00
Aut/TML TML MaxAbility TML MaxAbility TML MaxAbility Aut/Markata 100 100,	120-0
Adapta T.P. PROD F.E. F.G.B. Add and C.S.M.M. F.E. S.G.B. F.E. S.G.B. Add and C.S.M.M. F.E. S.G.B. F.E. S.G.B. S	11.0
CONTRACT Control Contro Control <thcontrol< th=""> <th< td=""><td>1.000</td></th<></thcontrol<>	1.000
Add UP GMML UP Labor/MELDIN Addram UP GAM UP Math Addram UP GAM UP Math Addram UP GAM UP Math Addram UP GAMP UP Math Addram UP GAMP UP Status Addram UP GAMPY UP Status Addram UP Math UP Status Addram UP Math UP St	1111
Addem Disk Calabity Disk Adde Abary III Calabity III Mail Abary IIII Calabity IIII Mail Abary IIIII Calabity IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII	Long Land
Adda Imp QUAL Imp SUM Adda Imp CCA Imp SCA Imp SCA Imp Imp <td>1232</td>	1232
Alter Im G2G31 F1 K-August CMP Astad, Gradyal Im Gradyal Im Gradyal Astad, Im Gradyal Im Gradyal Im Gradyal Astad, Im Gradyal Im Stadyal Im Stadyal Astadyal Im Im Stadyal Im Stadyal Im Stadyal Astadyal Im Im Stadyal Im Stadyal Im Stadyal Astadyal Im Stadyal Im	101
Alamidi. Um Gasidagi Um Uffler Alamidi Um Gasidagi Um Schult Alamidi Um Gasidagi Um Schult Alamidi Um Gasidagi Um Schult Alamidi Um Holton Um Schult Alamidi Um Holton Um Schult Alamidi Um Holton Um Schult Alamidi Um Kolton Um Schult Alamidi Um	
Anders 1= GRAPT Fel SCIP Anders 1= Grant 1 Science Science Analistic 1= Grant 1 Science Science Analistic 1= Grant 1= Science Science Analistic 1= Holes 1= Science Science Analistic 1= Holes 1= Science Science Analistic 1= Science Science Science Science Analistic 1= Science Science Science Science Analistic 1= Science Science Science Science	100
Alter 111 Generation 111 SCD-84 Astabab 111 Kod-84 111 SCD-84 <td< td=""><td>Dist.</td></td<>	Dist.
Address Linit Gavers 111 StD Ed. Address 114 Kitobal 114 StD Ed. Address 114 Kitobal 114 StD Ed. Address 1171 Kitobal 114 Kitobal Address 114 Kitobal 114 Kitobal	1201
Addition 1*1 Holen 1*1 KCRM Adaud*/KCR 1*1 KCRM 1*1 KCRM Adaud*/ 1*1 KCRM 1*1 KcMagnad	11.0
Asian2*(8CDF 111 Hyper-Mass 111 KER from Admit* 111 KER from 111 KER from	10 M
Advant (1) Kolleson-BRCRE) (1) KOLLAN Advant (1) KECKACKIN (1) SCER Advant (1) KECKACKIN (1) SCER Advant (1) KECKACKIN (1) SUbdapad Advant (1) SCERER (1) Barryon	
Adaptiv [10] NEQSAMI [20] SCH Adaptiv [20] KDC [20] Sch Adaptiv [20] KDC [20] Sch Adaptiv [Filmed.
Achanik (700 K700 1000 K40Appen) ACM6 (74 K51A/K70A (74) Humpers	0.104
ADAS [1] Hanger	10.3
	1114
	111,141
	811.18
Addede I'll LAMB [111] BardeD	1.11
AddKZ) [14] LiPup [14] SIQNUSIQN	110
Addelik [11] LABI [11] IMO	111
Aduliant [1] [14097 [1] SMG	1141
Adation: 1214 Look Aband 1214 SNON	ETCH.
ASS/Add/WW UNI MISTAG UNI SOBAdam	1.1
ABCD MADCRAD 11 8892D	12:43
ALC [1] MM [7] Repland 920	1.1
ASSOciated 111 MERA 111 MATTE	1111
ASSGnd 174 MTAbee 114 SWITE	111
Applified 127 MVR-DMVR-2 14 Mdan	1141
Available 1999 Nodam P.4 TERAC	191
ABRE 11 NAMER/SIXME 11 Under	1114
AGAM 101 ND-Adamy 1011 VESCO	E20
Analas 111 New 111 190206/19024/1902	
Availabil [14] Notavia [17] (ND-M	Z- 1254
Bhilten 1111 Notes Adam/News KRAC 1112 HPMCand	212
Ribden III Sodden IIII Mikette	1740
IPCed Dil Neugad Dil Val	1340
MANSProg E1 NESCE LT	
skip but pain but	

from "Descending through a Crowded Valley" (Schmidt, Schneider, Hennig; 2021)

The state of deep learning training methods

A confusingly crowded field of methods & hyperparameters

A huge number of training methods...

Name	Ref	Name	Red.	Kome	Ref.
Accelerated	104	CothAld	19/4	INCE.	04
ACCap	1219	CADA	104	104.	124
Adaliller	1.121	Cord Manageminant	101	Pub, Kalana	124
Actidiatette	LTR.	CPrepr	C HE	Tomat	12+1
Addition/ Adultmon/68	11	Cartoball	10.0	Francisco Francisco III	1223
Addition	10.0	Distant	12118	Protobilistic Polyale	1214
Actualities	10.04	DeepMenner	1111	Prottin	1211
Addinard	division in the local division of the local	DOM:NO.	1111	Princes	11.00
Add Come	E I R	CMRCrowl.	TTT .	OPERALM/OTHER	110
Actachetta	Diff	Exdera	104	B-fulare	100
Alterior	dist.	ERENC	1-1	Rorger.	100
Adulto	11-8	A	LOF C	Reportant	1141
Achiltone	110	toportunal .	111	RAD Price	12-1
AMOUNT	1121	From Anda Balland	TTNE 1	#MTRonov	10.0
Adami	1.14	THE BOD	111	5-9CD	1.00
ADADEBIGAN	100	G-Add and	111	SALLAN.	111-11
Arity	100	CROAM	100	Judger, NAME and	1000
Addiese	12148	Calum	DOT .	1018	1224
Adam	1.1	COALS	LOI .	SAM	101
Adam"	1000	COL54	100	SC Adapted SC 40457 rep-	100
Adapt.	100	Grad-Ann	in the	Char.	100
Adda .	1 mil	CHAPTE	1.1	9.20 9.20	100
					1201
Autore05	0.14	Gravidore	DUVE.	3031-84. SCID-62	
AdminiSte	122.00	Gravity	1.11	SCD GL	101
Address (SCOR)	114	HAlten	1111	W.DEGM	1244
	114	Hyper-Aslam	1000		1 Lord
AutoraT	1008	KOPUSOR-BROKE)	1110	302054	0.101
Adan/W	1116	NE-QN-CNN	Della	90100	111
Adarah	1004	REPORT.	10110	Si Oxforgenti	1114
AUA6	Diff	NFLR/RFBA	111	Hamper	10,000
148B	1111	LL/olam/LtMenoriter	CT III	Thursdoment	1111
Addealer	1111	LAME	1111	Spritcip	111
AGGK(Z)	1110	LaProp	1.11.2	SNONU SACRU	1101
AUDINE	10.00	LAR	010	1043	111
Arbdupt	1114	LINCOT	1.1	SHACL.	1141
Achieve	1214	Louid About	THE .	SNEM .	1101
Aches / Add 8-88	Arrest	MISTAC	1.4	SoltAdam	10.1
ABOD	10.00	MADCRAD	1.10	1896GD	1541
AG C	11/2	5555	1000	they filmed \$211	1.1
AADDoord	1111	LTD:	110	MANTE	100
ASSIGNA	12.94	MTALma	1118	MADITE-	111
According	11.73	MVRC-UMPRC-2	114	Tendam	1141
Areand Si	COM.	bind into	100	TENTAL	1.1
Aller	112	NAME NUMBER	10	Liddense.	111
AGAM	1.1	ND-Marr	100	VENCE	120
Acadas	1.0	Nette	DEAL	PRESS PRESS PRESS	1
Acadhiat	1.0	Newborn	12.0	VALUE AND DESCRIPTION OF THE PARTY OF THE PA	1000
Riden.	1.0	Name Adam, New KIPAC	Dur.	Whitead	10.0
SChieve	110	Nexcludem	LUNE .	To Bran Fire	1240
Richten	100	Novolucione Newsgened	111	Notice Part	1240
#ANSPICE		NUSGI		1.0	1.00
BANGPICE	11	Philan	122		

from "Descending through a Crowded Valley" (Schmidt, Schneider, Hennig; 2021)

...and training tricks

- OneCycle scheduler, gradient checkpointing
- Genetic Algorithm for Hyperparameters
- Avoid batches that lead to NaN/inf losses
- One cycle, low fidelity training, SGD with restarts
- Proximal optimization for regularizers
- Line searches for the maximum learning rate
- Normalized updates
- Distributed Shampoo, Normformer, GLU
- Weight averaging

from the NeurIPS "HITY Workshop" (Schneider et al.; 2022)

- FreezeOut
- A different epsilon value!
- Check hyperparameter performance over multiple seeds
- Lowering the learning rate!
- Normalizing data works better than batch or layer norm
- Mixed precision training
- Train with a small subset
- Cyclic and one cycle LR
- Label smoothing

►

We desperately need new benchmarks for neural network training algorithms.

The state of "benchmarking" in current deep learning optimizer papers

No standardized procedure to follow

- ► Each paper "invents" their own evaluation protocol.
- Unreasonably hard to perform a convincing, informative, and practically relevant comparison with strong baselines.
- ► Lot's of subtle pitfalls with tuning, problem-specification, etc.



Example of pitfalls when comparing training algorithms

Which algorithms trains the fastest depends on what it means for training to be complete

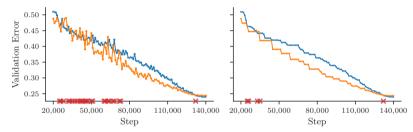


Figure 1: *Left*: Validation error for two different runs (—, —) of ADAM on RESNET-50 on IMAGENET. *Right*: The *best* validation error obtained so far. The runs intersect multiple times (**¥**).

Example of pitfalls when comparing training algorithms

Which algorithms trains the fastest depends on what it means for training to be complete

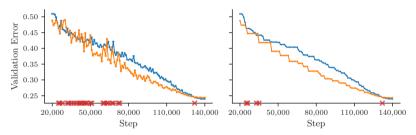


Figure 1: *Left:* Validation error for two different runs (—, —) of ADAM on RESNET-50 on IMAGENET. *Right:* The *best* validation error obtained so far. The runs intersect multiple times (*****).

- Directly comparing training curves is ill-posed.
- Without defining the target in advance, we can champion any method (moving the goal post after the experiment).
- Tuning goals must align.

MLCommons introduces the AlgoPerf: Training Algorithm Benchmark

An unprecedented effort to find faster deep learning training algorithms

ML Commons Algorithms Working Group

AlgoPerf: Training Algorithm Benchmark A standardized benchmark competition to measure neural network training speedups due to algorithmic changes. An open large-scale effort by 25+ researchers from Google, University of Tübingen, University of Toronto, Meta Al, etc.

Chaired by



George Dahl

Google



Frank Schneider University of Tübingen

A standardized competitive time-to-results benchmark

► A competitive time-to-results benchmark.

 \rightarrow Everyone's submissions are everyone's strong baselines.

A standardized competitive time-to-results benchmark

► A competitive time-to-results benchmark.

 \rightarrow Everyone's submissions are everyone's strong baselines.

► Fixed hardware, workloads, and process.

 \rightarrow Submissions need to provide training algorithm improvements.

A standardized competitive time-to-results benchmark

► A competitive time-to-results benchmark.

 \rightarrow Everyone's submissions are everyone's strong baselines.

► Fixed hardware, workloads, and process.

 \rightarrow Submissions need to provide training algorithm improvements.

► Aggregate across a variety of realistic workloads using performance profiles.

 \rightarrow No specialized solutions but **general-purpose methods**.

A standardized competitive time-to-results benchmark

► A competitive time-to-results benchmark.

- \rightarrow Everyone's submissions are everyone's strong baselines.
- ► Fixed hardware, workloads, and process.
 - \rightarrow Submissions need to provide training algorithm improvements.
- ► Aggregate across a variety of realistic workloads using performance profiles. → No specialized solutions but general-purpose methods.
- ► Explicitly accounts for hyperparameter tuning by providing search spaces. → Runable training algorithms, not algorithm templates.

The **AlgoPerf** Training Algorithms Benchmark

We need you!

Read the Rules	github.com/mlcommons/algorithmic-efficiency/blob/main/RULES.md
Read the Paper	arxiv.org/abs/2306.07179
Submit!	Call for Submission coming soon

- ▶ Benchmark experts Join the effort and tell us how to improve!
- ML community Help us spread the word!
- ► Algorithm researchers Submit! AlgoPerf is the easiest way to convincingly demonstrate the capabilities of your training algorithm.