Stochastic Optimization

AdaGrad

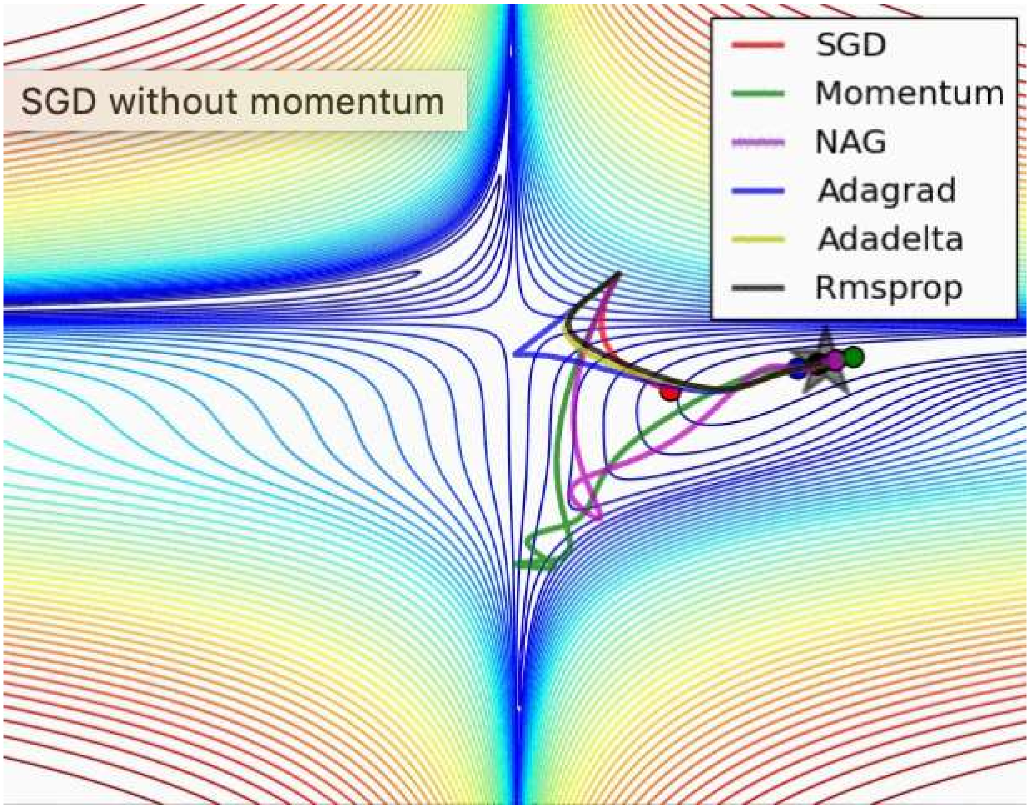
Edit

AdaGrad is a stochastic optimization method that adapts the learning rate to the parameters. It performs smaller updates for parameters associated with frequently occurring features, and larger updates for parameters associated with infrequently occurring features. In its update rule, Adagrad modifies the general learning rate η at each time step t for every parameter θ_i based on the past gradients for θ_i:

$$\theta_{t+1,i} = \theta_{t,i} - \frac{\eta}{\sqrt{G_{t,ii} + \epsilon}} g_{t,i}$$

The benefit of AdaGrad is that it eliminates the need to manually tune the learning rate; most leave it at a default value of 0.01. Its main weakness is the accumulation of the squared gradients in the denominator. Since every added term is positive, the accumulated sum keeps growing during training, causing the learning rate to shrink and becoming infinitesimally small.

Image: Alec Radford



Papers

Search for a paper or author

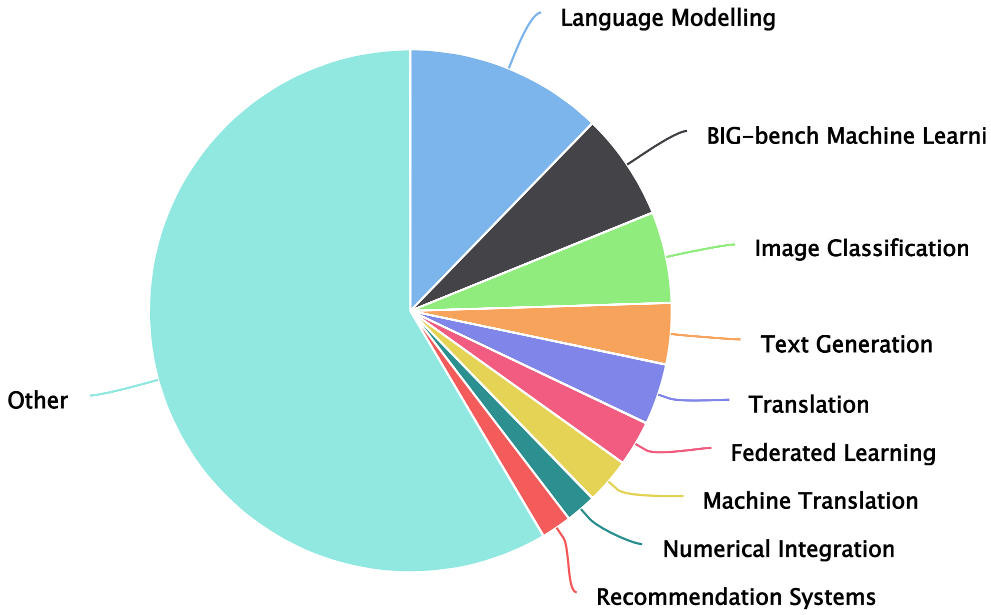
Paper	Code	Results	Date	Stars ↑
Memory Efficient Adaptive Optimization Tomer Koren, Yoram Singer, Vineet Gupta, Rohan Anil	●	—	1 Dec 2019	32,446
Memory-Efficient Adaptive Optimization Tomer Koren, Yoram Singer, Vineet Gupta, Rohan Anil	●	📊	30 Jan 2019	32,437
CTRL: A Conditional Transformer Language Model for Controllable Generation Bryan McCann, Nitish Shirish Keskar, Lav R. Varshney, Caiming Xiong, Richard Socher	●	—	11 Sep 2019	11,166

Paper	Code	Results	Date	Stars ↑
Augmenting Self-attention with Persistent Memory Sainbayar Sukhbaatar, Guillaume Lample, Herve Jegou, Armand Joulin, Edouard Grave	●	📊	2 Jul 2019	3,986
Adaptive Gradient Methods with Dynamic Bound of Learning Rate Yuanhao Xiong, Xu sun, Yan Liu, Liangchen Luo	●	—	26 Feb 2019	2,904
Improving Neural Language Models with a Continuous Cache Nicolas Usunier, Armand Joulin, Edouard Grave	●	📊	13 Dec 2016	2,550
Adaptivity without Compromise: A Momentumized, Adaptive, Dual Averaged Gradient Method for Stochastic Optimization Aaron Defazio, Samy Jelassi	●	—	26 Jan 2021	795
Riemannian Adaptive Optimization Methods Octavian-Eugen Ganea, Gary Bécigneul	●	—	1 Oct 2018	784
Once Detected, Never Lost: Surpassing Human Performance in Offline LiDAR based 3D Object Detection Zhaoxiang Zhang, Yuntao Chen, Naiyan Wang, Yiming Mao, Feng Wang, Lue Fan, Yuxue Yang	●	—	24 Apr 2023	721
YellowFin and the Art of Momentum Tuning Jian Zhang, Ioannis Mitliagkas	●	—	12 Jun 2017	422

Showing 1 to 10 of 162 papers

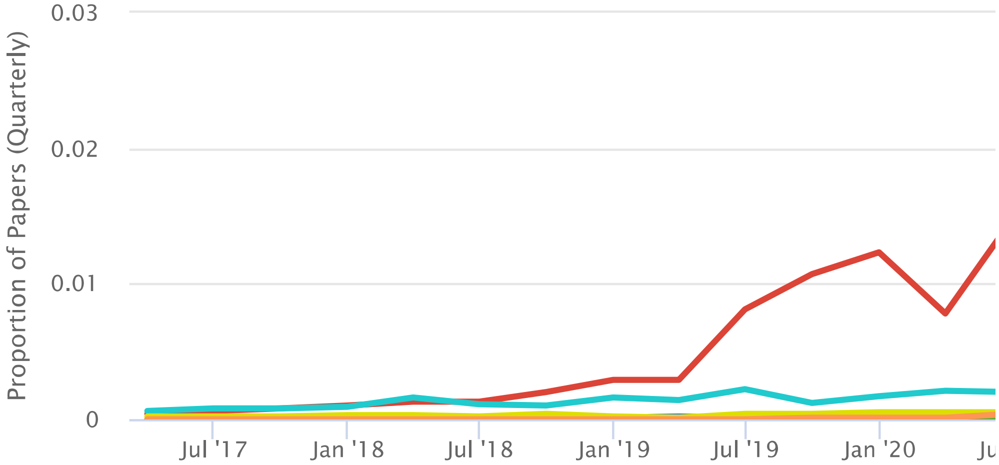
Previous 1 2 3 4 5 ... 17 Next

Tasks



Task	Papers	Share
Language Modelling	13	12.26%
BIG-bench Machine Learning	7	6.60%
Image Classification	6	5.66%
Text Generation	4	3.77%
Translation	4	3.77%
Federated Learning	3	2.83%
Machine Translation	3	2.83%
Numerical Integration	2	1.89%
Recommendation Systems	2	1.89%

Usage Over Time



This feature is experimental; we are continuously improving our matching algorithm.

Components

Component	Type	Edit
No Components Found	You can add them if they exist; e.g. Mask R-CNN uses RoIAlign	

Categories

Edit

Stochastic Optimization

Large Batch Optimization

Contact us on: hello@paperswithcode.com.

Papers With Code is a free resource with all data licensed under CC-BY-SA.

Terms Data policy Cookies policy from Meta AI