Toward Understanding Why Adam Converges Faster Than SGD for Transformers

Yan Pan, Yuanzhi Li
Carnegie Mellon University

Introduction

We explore one explanation of why Adam converges faster than SGD for some neural networks such as transformers.

- We argue that the performance of optimization algorithms is closely related to the **directional sharpness** of the update steps.
- We show that SGD has much worse directional sharpness compared to adaptive algorithms.
- We observe that only a small fraction of the coordinates causes the bad sharpness and slow convergence of SGD.
- We propose to use coordinate-wise clipping to reduce sharpness and speed up convergence of optimization algorithms.

Directional Sharpness

We introduce **directional sharpness** and argue it is closely related to the poor performance of SGD.

- The motivation is the quadratic Taylor expansion of the objective function commonly used in convergence proofs.

$$f(x_{t+1}) \approx f(x_t) + \underbrace{\nabla f(x_t)^\top (x_t - x_{t+1})}_{\text{gradient correlation}} + \underbrace{(x_t - x_{t+1})^\top \nabla^2 f(x_t)(x_t - x_{t+1})}_{\text{directional sharpness}}$$

- Optimization algorithms should minimize the two terms that depends on the update step to decrease function value.
- Typically the second-order term is bounded using L-smooth assumption for any optimization algorithm, but this is not tight.
- **Definition**: directional sharpness in the direction v is $v^\top \nabla^2 f(x) v$.
- It is directly related to minimizing the objective function.
- A lower directional sharpness implies the potential to take a larger step size and possibly lead to a larger local decrement of the objective function.

Empirically, we observe that the directional sharpness is much lower for adaptive algorithms than for SGD.

- We study the update step of different optimization algorithms under the same trajectory and local geometry using pseudo-update steps.
- We compute the directional sharpness of different optimization algorithms and visualize the optimization landscape in the update direction of a variety of optimization algorithms.

Coordinate-wise Clipping

We propose to use **coordinate-wise clipping** to improve sharpness.

- We find the gradient norm and sharpness are concentrated in a small fraction of coordinates, and clipping those coordinates can significantly decrease directional sharpness.
- The use of clipping in optimization algorithms is a trade-off between improving gradient correlation and reducing directional sharpness.

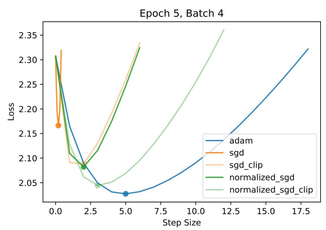
Our experiments show that the performance of optimization algorithms is indeed positively correlated with the directional sharpness of optimization algorithms in the trajectory:

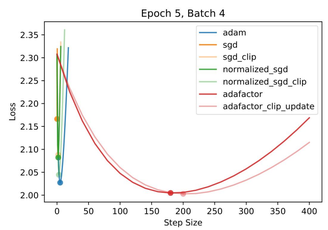
- The update direction of clipped algorithms has better directional sharpness.
- Coordinate-wise clipped algorithms converge faster than their original version.

Experiments

We conducted two types of experiments on two tasks.

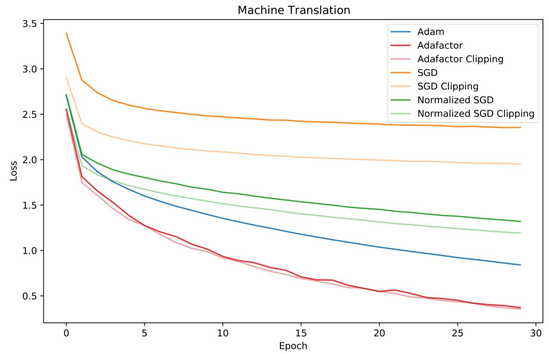
- We compute the directional sharpness of a variety of optimization algorithms and visualize the corresponding loss landscape direction, under different local geometry.
- We implement clipping algorithms and show they converge faster.

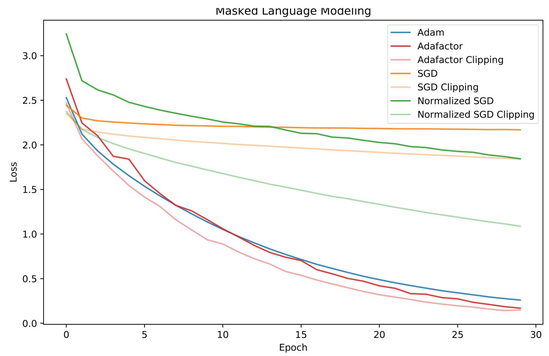




Algorithm	Sharpness
Adam	0.16190993
SGD	31.04433435
SGD Clipping	1.77876506
Normalized SGD	0.77112307
Normalized SGD Clipping	0.38075357
Adafactor	0.00000319
Adafactor Clipping	0.00000253

The loss landscape in different update directions on machine translation and their corresponding sharpness.





Clipped optimization algorithms generally converge faster than the original algorithms.

Conclusion

Our work provides a new important insight of why Adam converges faster than SGD in practice.

- Slow convergence of SGD is related to bad directional sharpness.
- Coordinate-wise clipping and adaptive algorithms can fix it.

References

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. In *International Conference on Learning Representations*, 2018.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33, 2020.