Artificial Neural Networks Deep Learning Machine Learning

Why do the state of the art deep learning models like ResNet and DenseNet use SGD with momentum over Adam for training?

Answer Request Follow 46 Comment Downvote

3 Answers

Conner Davis, Researcher at The University of Texas at Dallas
Answered Jun 13

Adaptive learning methods weren't designed to yield optimal results. They were designed to yield acceptable results regardless of how badly you screw up your hyperparameters.

SGD with momentum is very good at finding high-quality local minima. This is for several reasons, one of which is that momentum prevents it from converging to sharp local minima.

On the other hand, “smart” optimizers tend to assume they are already on a path to a local minima and adjust the learning rate so that they continue moving towards *that* minima and reach it as quickly as possible.

For that reason (and some others), optimizes like Adam are more likely to converge to sharp local minima while SGD with momentum is more likely to converge to “flat” local minima.

Sharp minima are often more the results of what data points you happen to have than they are of the structure of the underlying problem. Flatter minima are more likely to result from the structure of the problem. That tends to mean better generalization as well as lower training loss.

Adaptive optimizers are great for getting viable results quickly, but SGD + momentum is usually better for getting truly state of the art results. It just takes more effort to tune the hyperparameters properly.

All of the above is very much heuristic/intuition-based and hand-wavy. There's some theory to back up parts of it and practical results to verify other parts, but it's not by any means a strict rule.

TLDR: Adam makes it easy to converge quickly and on the first try. SGD+momentum takes more work to tune, but tends to reach better optima.

3.3k Views · 67 Upvotes · Answer requested by Brando Miranda

Upvote 67 Downvote Bookmark

Sam Sinai
I want to echo Charles, can you point me to the research regarding this? Even no form...

1 more comments from Charles H Martin

Tapabrata Ghosh, Founder and CEO at Vathys
Answered Jun 9

Have you ever done a side by side loss/ error plot of SGD + Nesterov momentum? I'm not asking to be condescending, if you plot them side by side you'll often notice that Adam drops rapidly and then plateaus, quite similarly to exponential decay, and SGD displays a very similar behavior, but it often has a second drop which often puts it lower than Adam when it's all done.

To answer your question more directly, Adam is more robust to learning rate changes so when you change hyperparameters, you can use Adam and get reasonably good results. However, a well tuned SGD can do better. SGD requires a well tuned learning rate, but can outperform Adam and other adaptive optimizers if done so.

Related Questions

Why does vanilla SGD and Momentum sometimes outperform optimizers such as AdaDelta in deep learning?

How many GPUs do Google and Facebook use to train all their deep learning models?

Why is AdaDelta not favored in Deep Learning communities while AdaGrad is preferred by many over other SGD variants?

Can I use a PS4 to train deep learning models?

Is there any deep learning model trained on Arabic texts?

What are energy-based models and how can they be used in deep learning?

Can we use SGD to train the mixture model, such as GMM and movMF?

When training a Deep Learning model, does the sequence or order of samples in the training data matter?

Can Dropout be used to avoid overfitting a deep learning model with very small training data?

Is line search used commonly with SGD while learning the parameters for a deep neural networks?

More Related Questions

Question Stats

45 Public Followers
7,305 Views
Last Asked Jun 13
Edits

Recommended All

Deepesh Nathani, Learning ML and Deep Learning
Answered Jun 9

Check out this

> Surprising result: adaptive gradient methods generalize worse than SGD. Time to swap out Adam for carefully tuned SGD? h/t @niru_m https://t.co/BGQhQFS8qF
> — Ben Poole (@poolio) May 24, 2017

The paper mentioned in the tweet tested the generalization capability of adaptive learning methods and SGD. Apparently if we carefully tune the hyperparameters for SGD and Momentum we can obtain less test error compared to Adam or RMSProp

568 Views · 5 Upvotes

Upvote 5 Downvote Bookmark

Recommended All

Top Stories from Your Feed