Demon: Momentum Decay for Improved Neural Network Training by John Chen

Abstract: Momentum is a popular technique in deep learning for gradient-based optimizers. We propose a decaying momentum (Demon) rule, motivated by decaying the total contribution of a gradient to all future updates. Applying Demon to Adam leads to significantly improved training, notably competitive to momentum SGD with learning rate decay, even in settings in which adaptive methods are typically non-competitive. Similarly, applying Demon to momentum SGD improves over momentum SGD with learning rate decay in most cases. Notably, Demon momentum SGD is observed to be significantly less sensitive to parameter tuning than momentum SGD with learning rate decay schedule, critical to training neural networks in practice. Results are demonstrated across a variety of settings and architectures, including image classification, generative models, and language models. Demon is easy to implement and tune, and incurs limited extra computational overhead, compared to the vanilla counterparts. Code is readily available.