
Data Momentum

Fartash Faghri

Department of Computer Science
University of Toronto
faghri@cs.toronto.edu

Abstract

We propose optimization methods with data weighting.

1 Introduction

2 Related Works

3 Method

SGD:

$$g_t = \hat{g}_t^{\text{SGD}} = \left. \frac{\partial}{\partial \theta} \mathbb{E}[L(\theta)] \right|_{\theta=\theta_t} \quad (1)$$

SGD with momentum:

$$\hat{g}_t^{\text{Mom}} = \beta \hat{g}_{t-1}^{\text{Mom}} + (1 - \beta) g_t \quad (2)$$

$$= \sum_{t=1}^T (1 - \beta) \beta^{T-t} g_t \quad (3)$$

$$\left. \frac{\partial \hat{L}^{\text{Mom}}}{\partial \theta} \right|_{\theta=\theta_t} = \hat{g}_t^{\text{Mom}} \quad (4)$$

$$\hat{L}^{\text{Mom}} = \int \sum_{t=1}^T (1 - \beta) \beta^{T-t} g_t d\theta \quad (5)$$

$$= (1 - \beta) \sum_{t=1}^T \beta^{T-t} \int g_t d\theta \quad (6)$$

$$= (1 - \beta) \sum_{t=1}^T \beta^{T-t} \mathbb{E}[L(\theta_t)] \quad (7)$$

SGD with momentum is minimizing an exponential average of the loss of the model. So it is stabilizing learning.

3.a Minimum viable project

3.b Nice to haves