The Shattered Gradients Problem: If resnets are the answer, then what is the question?

Balduzzi *et al.*, ICML 2017

03/08/2017

Background

- training deep networks isn't easy
- problems can be mitigated by:
  - unsupervised pre-training
  - correct initialization of weights
  - batch normalization
  - **skip connections** i.e., activation functions of the form:

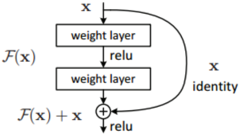
$$f(x) = \underbrace{\rho(x)}_{\text{non linearity}} + x$$

- SOTA results achieved by architectures which include **skip connections**: this paper tries of understand **why**

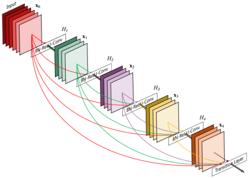
Background

some examples of *skip connections*:

- res-nets [He *et al.*,, 2015]



- denseNets [Huang *et al.*,, 2017]



- also highway nets [Srivastava *et al.*, 2015)], etc

Background

current ideas/intuitions about why *skip connections* are so effective:

- Raiko *et al.*, [2012] show that such maps help to make Fisher info matrix *more diagonal*
- improves gradient flow (due to linear nature)
- (this paper) avoids **shattered gradients**

shattered gradients

the gradients in deep networks behave like white noise. Bad news for first-order algorithms that assume that gradients at nearby points are similar!

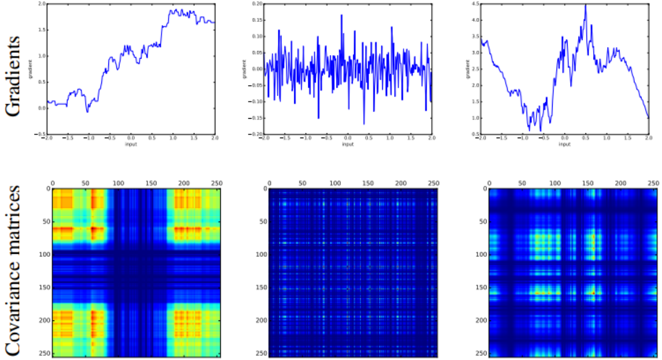
Shattered gradients

- they focus on the following class of networks:

$$\begin{aligned} f_{w,b}(x) &= w^T \mathrm{ReLU}(x \cdot v - b) \\ &= w^T \max(0, x \cdot v - b) \end{aligned}$$

  where $v = (1, \ldots, 1)$ and initialize $w, b \sim \mathcal{N}(0, \sigma^2)$
- study $\frac{\partial f_w}{\partial x}$ for different values of x
- not realistic, but provide a sandbox where gradients can be isolated and studied.
- all results studied at **initialization** (i.e., no training)

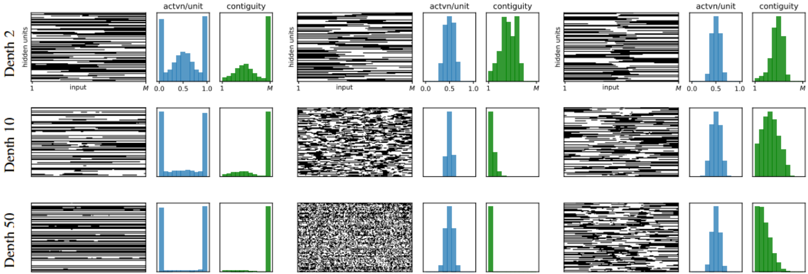
Shattered gradients



(a) 1-layer feedforward. (b) 24-layer feedforward. (c) 50-layer resnet.

$\frac{\partial f_w}{\partial x}$ behaving like white noise makes neuron's effect on output very unstable

Shattered gradients



(a) Feedforward net without batch norm. (b) Feedforward with batch normalization. (c) Resnet with batch normalization.

Batch norm without skip connections makes gradients *less Lipschitz*

Theory

- define $\nabla_i = \frac{\partial f_w}{\partial n}\left(x^{(i)}\right)$ and study $\mathcal{R}(i,j) = \frac{\mathbb{E}[\nabla_i \nabla_j]}{\sqrt{\mathbb{E}[\nabla_i^2]\cdot\mathbb{E}[\nabla_j^2]}}$

  recall $x^{(i)} \in \mathbb{R}$ is the univariate input

- **Theorem 1**: in feed-forward networks with weights initialized with variance $\sigma^2 = \frac{2}{N}$ (e.g., Xavier) then:

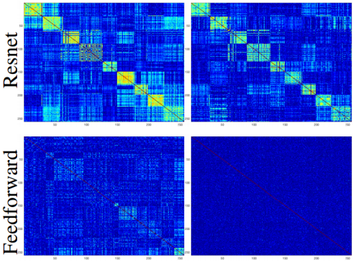
$$\mathcal{R}^{fnn}(i,j) = \frac{1}{2^L}$$

- **Theorem 2**: in a network with skip connections with batch norm and weights initialized with variance $\sigma^2 = \frac{2}{N}$ then:

$$\mathcal{R}^{skip}(i,j) \sim \frac{1}{\sqrt{L}}$$

Empirical example: CIFAR-10

- covariance matrices for resnet (with skip connections) and feedforward network
- training examples ordered based on *k*-means clustering



(a) Depth = 2 (b) Depth = 50

Looks-linear (LL) initialization

- propose to train networks with the activation function:

  $$f(x) = W_1^T \text{ReLU}(x) + W_2^T \text{ReLU}(-x)$$

  and initialize with $W_1 = W_2$ so that $f(x) = x$
- claim LL initialization avoids shattering