On the Difficulty of Training RNNs

R. Pascanu, T. Mikolov, Y. Bengio

ICML 2013

Presented by OUR NAMES: Hanna M. Dettki

1. Introduction & Background

1.1 Context

- **Importance of sequence modeling**
  - e.g., language, time-series in finance
- Identifying gradient problems (Bengio et al., 1994)
  - **Vanishing gradient problem**: impossible to learn long-term dependencies
  - **Exploding gradient problem**: numerical instabilities $\rightarrow$ unstable training
- $\rightarrow$ Why stable gradient flow is critical for learning temporal dependencies (paper's contribution)

1.2 Schematic & formal def. of RNN

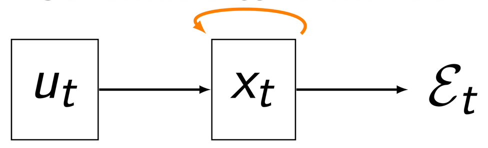

Fig. 1

$$x_t = F(x_{t-1}, u_t, \theta) \qquad (1) \text{ General}$$
$$x_t = W_{rec}\,\sigma(x_{t-1}) + W_{in} u_t + \mathbf{b} \qquad (2)$$

where u_t: input, x_t: state, t: time step, $\mathbf{b}$: bias, $\mathcal{E}_t = \mathcal{L}(x_t)$ (error)

The recurrent connections in the hidden layer allow information to persist from one input to another.

1.3 Training RNNs: Backprop Through Time (BPTT) on Unrolled RNN

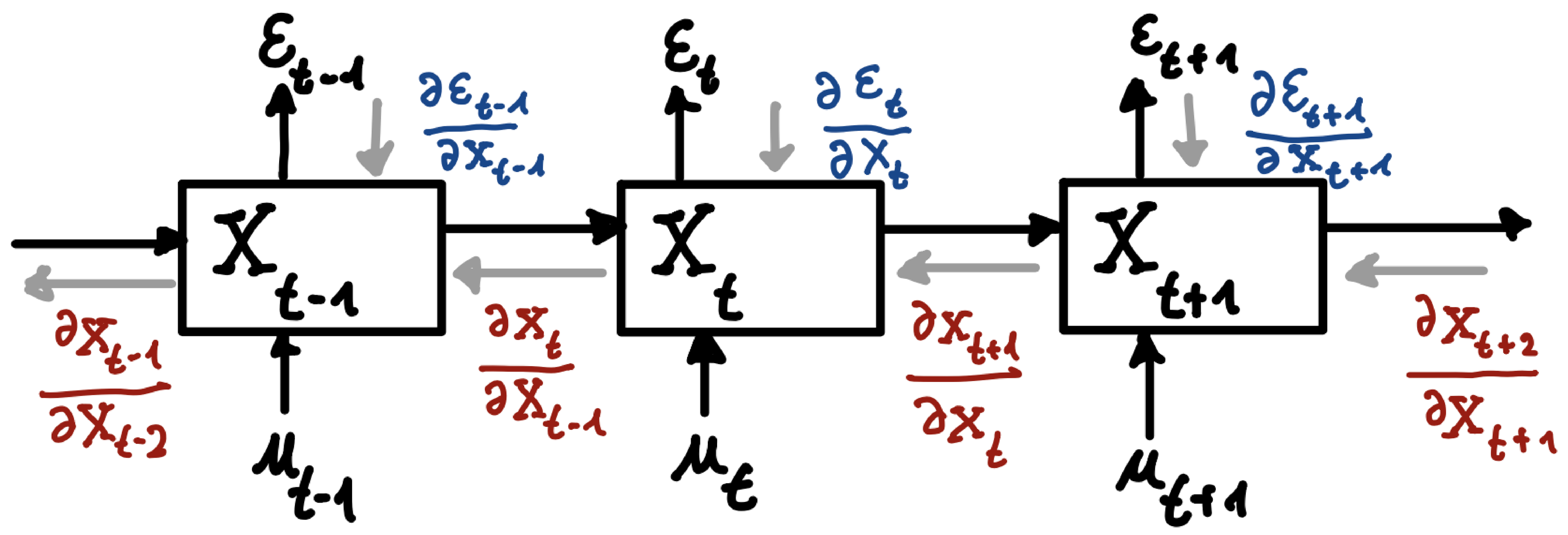

Fig. 2: Unrolled RNN

$$\frac{\partial \mathcal{E}}{\partial \theta} = \sum_{t=1}^{T} \frac{\partial \mathcal{E}_t}{\partial \theta} \qquad (3)$$

$$\frac{\partial \mathcal{E}_t}{\partial \theta} = \sum_{k=1}^{t} \left(\frac{\partial \mathcal{E}_t}{\partial x_t} \frac{\partial x_t}{\partial x_k} \frac{\partial^+ x_k}{\partial \theta} \right) \qquad (4)$$

$$\frac{\partial x_t}{\partial x_k} = \prod_{i=k+1}^{t} W_{rec}^{\top} \cdot \mathrm{diag}(\sigma'(x_{i-1})) \qquad (5)$$

where $\frac{\partial^+ x_k}{\partial \theta}$ denotes the "immediate" partial derivative (treating x_{k-1} as constant).

Blue: total gradient over time. Red: temporal error contribution.

2. The Problem

2.1 Vanishing Gradient problem (VG)

Sufficient condition:

$$\lambda_1 < \frac{1}{\gamma}$$

where: λ_1: largest singular value of W_{rec}
γ: bound on derivative of activation function

(proof: see eq. 6 & 7)

2.2 Exploding Gradient problem (EGs)

- gradients grow exponentially during backprop
- *Necessary condition:*

$$\lambda_1 > \frac{1}{\gamma}$$

2.2.1 Dynamical systems interpretation:

- EGs create steep wall-like structures that are perpendicular to exploding direction in error surface

placeholder figure

Fig. 3

3. Solution & Experiments

3.1 Gradient Clipping

- see Fig. 3 for *motivation*
- Pseudo-code for norm-clipping

3.2 VG-Regularization

- see paper eq. 9 & 10

3.3 MSGD-CR

(combines 3.1 & 3.2)

3.4 Initialization Strategies

- see experiments in Sec. 4

placeholder figure

Fig. 4

placeholder figure

Fig. 5

4. Relevance today & SOTA techniques

- Clipping still relevant!
- Instead of regularization:
  - residual connections
  - gradient checkpointing
  - gating mechanisms
  - layer normalization
  - attention mechanism
  - positional encoding