

Recurrent neural networks

Giulio Galvan

The model

RNN

Given an input sequences $\{\mathbf{u}\}_{t=1,\dots,T}$, with $\mathbf{u}_t \in \mathbb{R}^p$, the output sequence of a RNN $\{\mathbf{y}\}_{t=1,\dots,T}$, with $\mathbf{y}_t \in \mathbb{R}^o$, is defined by the following:

$$\mathbf{y}^t \triangleq F(W^{out} \cdot \mathbf{a}^t + \mathbf{b}^{out}) \quad (1)$$

$$\mathbf{a}^t \triangleq W^{rec} \cdot \mathbf{h}^{t-1} + W^{in} \cdot \mathbf{u}^t + \mathbf{b}^{rec} \quad (2)$$

$$\mathbf{h}^t \triangleq \sigma(\mathbf{a}^t) \quad (3)$$

$$\mathbf{h}^0 \triangleq \vec{0}, \quad (4)$$

where $\sigma(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$ is a non linear function applied element-wise called activation function.

The optimization problem

Given a dataset D :

$$D \triangleq \{\{\bar{\mathbf{u}}^{(i)}\}_{t=1,\dots,T}, \bar{\mathbf{u}}_t^{(i)} \in \mathbb{R}^p, \{\bar{\mathbf{y}}^{(i)}\}_{t=1,\dots,T}, \bar{\mathbf{y}}_t^{(i)} \in \mathbb{R}^o; i = 1, \dots, N\} \quad (5)$$

we define a loss function $L_D : \mathbb{R}^N \rightarrow \mathbb{R}_{\geq 0}$ over D as

$$L_D(\mathbf{x}) \triangleq \frac{1}{N} \sum_{i=1}^N \sum_{t=1}^T L_t(\bar{\mathbf{y}}_t^{(i)}, \mathbf{y}_t^{(i)}(\mathbf{x})), \quad (6)$$

where $L_t(\cdot, \cdot)$ is an arbitrary loss function for the time step t . The problem is

$$\min_{\mathbf{x} \in \mathbb{R}^N} L_D(\mathbf{x}) \quad (7)$$

Stochastic gradient descent (SGD)

SGD is the standard framework in most of the applications.

Algorithm 1: Stochastic gradient descent

Data:

$D = \{\langle \mathbf{u}^{(i)}, \mathbf{y}^{(i)} \rangle\}$: training set

\mathbf{x}_0 : candidate solution

m : size of each minibatch

Result:

\mathbf{x} : solution

```
1  $\mathbf{x} \leftarrow \mathbf{x}_0$ 
2 while stop criterion do
3    $I \leftarrow$  select  $m$  training example  $\in D$ 
4    $\alpha \leftarrow$  compute learning rate
5    $\mathbf{x} \leftarrow \mathbf{x} - \alpha \nabla_{\mathbf{x}} L(\mathbf{x}; \langle \mathbf{u}^{(i)}, \mathbf{y}^{(i)} \rangle, i \in I)$ 
6 end
```

A pathological problem example

An input sequence:

marker	0	1	0	...	0	1	0	0
value	0.3	0.7	0.1	...	0.2	0.4	0.6	0.9

The predicted output should be the sum of the two one marked positions (1.1).

Why is this a difficult problem?

Because of it's long time dependencies.

A pathological problem example

An input sequence:

marker	0	1	0	...	0	1	0	0
value	0.3	0.7	0.1	...	0.2	0.4	0.6	0.9

The predicted output should be the sum of the two one marked positions (1.1).

Why is this a difficult problem?

Because of it's long time dependencies.

A pathological problem example

An input sequence:

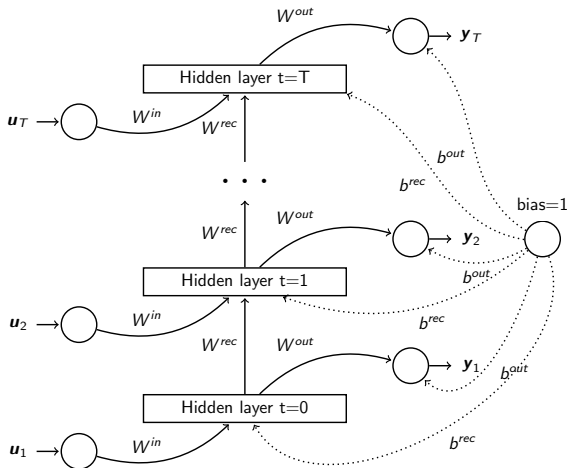
marker	0	1	0	...	0	1	0	0
value	0.3	0.7	0.1	...	0.2	0.4	0.6	0.9

The predicted output should be the sum of the two one marked positions (1.1).

Why is this a difficult problem?

Because of it's long time dependencies.

Unfolding



Understanding the gradient structure

Simply applying the chain rule it's easy to see that

$$\nabla L(\mathbf{x}) = \sum_{t=1}^T \nabla L_{|t}(\mathbf{x}), \quad (8)$$

where $\nabla L_{|t}$ is ...

Vanishing gradient: an upper bound

$$\frac{\partial \mathbf{a}^t}{\partial \mathbf{a}^k} = \prod_{i=t-1}^k \text{diag}(\sigma'(\mathbf{a}^i)) \cdot W^{\text{rec}}. \quad (9)$$

Taking the singular value decomposition of W^{rec} :

$$W^{\text{rec}} = S \cdot D \cdot V^T \quad (10)$$

where S, V^T are squared orthogonal matrices and $D \triangleq \text{diag}(\mu_1, \mu_2, \dots, \mu_r)$ is the diagonal matrix containing the singular values of W^{rec} . Hence:

$$\frac{\partial \mathbf{a}^t}{\partial \mathbf{a}^k} = \prod_{i=t-1}^k \text{diag}(\sigma'(\mathbf{a}^i)) \cdot S \cdot D \cdot V^T \quad (11)$$

Since U and V are orthogonal matrix, hence

$$\|U\|_2 = \|V^T\|_2 = 1,$$

and

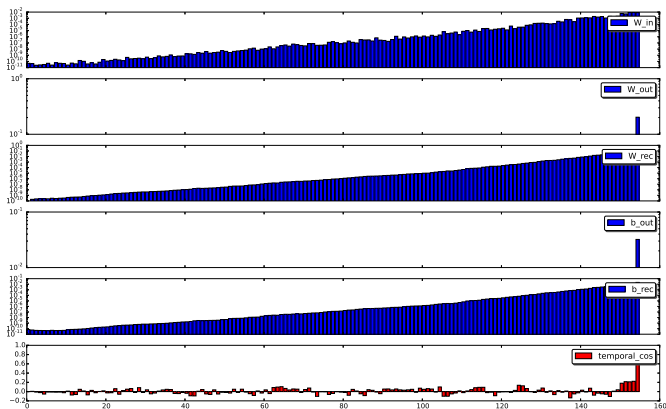
$$\|diag(\lambda_1, \lambda_2, \dots, \lambda_r)\|_2 \leq \lambda_{max},$$

we get

$$\left\| \frac{\partial \mathbf{a}^t}{\partial \mathbf{a}^k} \right\|_2 = \left\| \left(\prod_{i=t-1}^k diag(\sigma'(\mathbf{a}^i)) \cdot S \cdot D \cdot V^T \right) \right\|_2 \quad (12)$$

$$\leq (\sigma'_{max} \cdot \mu_{max})^{t-k-1} \quad (13)$$

Temporal gradient norms: an illustration



The end