On penalty terms

Giulio Galvan

28 settembre 2015

We have seen that:

$$\frac{\partial g}{\partial W^{rec}} = \sum_{t=1}^{T} \frac{\partial g_t}{\partial \boldsymbol{a}^t} \cdot \sum_{k=1}^{t} \frac{\partial \boldsymbol{a}^t}{\partial \boldsymbol{a}^k} \cdot \frac{\partial \boldsymbol{a}^k}{\partial W^{rec}}, \tag{1}$$

where

$$\frac{\partial \boldsymbol{a}^t}{\partial \boldsymbol{a}^k} = \prod_{i=t-1}^{k} diag(\sigma'(\boldsymbol{a}^i)) \cdot W^{rec}. \tag{2}$$

tends to vanish.

The idea is to develop a penalty term which express a preference for solutions where the components $\frac{\partial \boldsymbol{a}^t}{\partial \boldsymbol{a}^k}$ are far from zero, hence, learning a model which exhibit long memory.

A first attempt could be:

$$\Gamma \triangleq \sum_{t=1}^{T} \sum_{k=1}^{t} \frac{1}{\left\| \frac{\partial \boldsymbol{a}^t}{\partial \boldsymbol{a}^k} \right\|^2}. \tag{3}$$

Γ treats all temporal steps equally: we could assign more importance to distant temporal steps, which are the most critical, modifying *Gamma* as follows:

$$\Gamma \triangleq \sum_{t=1}^{T} \sum_{k=1}^{t} \frac{\sigma(t-k)}{\left\| \frac{\partial \boldsymbol{a}^t}{\partial \boldsymbol{a}^k} \right\|^2}, \tag{4}$$

where $\sigma(\cdot)$ assign different weights depending on the temporal distance $t-k$, for example $\sigma(h) = exp\{h\}$

We can compute the derivative of Γ w.r.t. W^{rec} as follows. Let $A \triangleq \frac{\partial \boldsymbol{a}^t}{\partial \boldsymbol{a}^k}$, for ease of notation and $\|A\|_F$ it's Frobenius norm.

$$\begin{aligned} \frac{\partial \|A\|_F^2}{\partial W_{ij}^{rec}} &= -\frac{1}{\|A\|_F^2} \cdot \frac{\partial}{\partial w_{ij}} \sum_{xy} A_{xy}^2(w_i j) \quad (5) \\ &= -\frac{1}{\|A\|_F^2} \cdot 2 \sum_{xy} A_{xy} \cdot \frac{\partial A_{xy}}{\partial w_{ij}}, \quad (6) \end{aligned}$$

where

$$A_{xy} = \frac{\partial \boldsymbol{a}_x^t}{\partial \boldsymbol{a}_y^k} = \sum_{q \in P(y)} \sum_{l \in P(q)} \dots \sum_{h: x \in P(h)} w_{qy} \dots w_{yh} \cdot \sigma'(a_y^k) \sigma'(a_q^{k+1}) \dots \sigma'(a_x^{t-1}) \tag{7}$$

For efficiency purposes and because 2nd derivatives are not always available, for example when using ReLU units, $\sigma'(a_i^k)$ can be considered constant w.r.t to W^{rec}.