Classification de sentiments par réseau de neurones récursif tensoriel

Thomas Moreau
Ecole polytechnique
thomas.moreau@polytechnique.edu

Bertrand Rondepierre
Ecole polytechnique
bertrand.rondepierre@polytechnique.edu

***Résumé*—Dans ce travail on étudie un modèle de réseau de neurones récursif tensoriel utilisé pour faire de la prédiction de sentiments.**

I. Entrainement du modèle

A. Critère d'entrainement

En notant $\theta = (W, W_s, V, L)$ les paramètres du modèle on définit l'énergie suivante à partir de la distance de Kullback-Leibler (le signe moins est bien présent contrairement à la formule donnée dans l'article) :

$$E(\theta) = -\sum_{i \in Noeuds} \sum_{j=1}^{C} t_j^i \log y_j^i + \lambda ||\theta||^2$$

Pour simplifier les notations, on réservera l'index i aux noeuds du réseau.

B. Equations de backpropogation

On notera en préliminaire avec le softmax :

$$\sigma_j(y) = \frac{e^{y_j}}{\sum_k e^{y_k}}$$

que la dérivée est simplement :

$$\partial_i \sigma_j(y) = (\delta_{i,j} - \sigma_i(y))\sigma_j(y)$$

1) Gradient par rapport à W_s: On fait donc le calcul de la dérivée par rapport à $[W_s]_{kl}$ qui est différent des suivants puisqu'il ne s'agit pas encore de rétropropagation $y_j^i = \sigma_j(W_s x^i)$:

$$\begin{aligned}
\frac{\partial E}{\partial [W_s]_{kl}} &= -\sum_i \sum_j \frac{t_j^i}{y_j^i} \frac{\partial y_j^i}{\partial [W_s]_{kl}} \\
&= -\sum_i \sum_j \sum_{c=1}^{C} t_j^i \delta_{kc} (\delta_{cj} - \sigma_c(W_s x^i)) x_l^i \\
&= -\sum_i \sum_j t_j^i (\delta_{kj} - y_k^i) x_l^i \\
&= \sum_i (y_k^i - t_k^i) x_l^i
\end{aligned}$$

Donc chaque nœud i contribue au gradient par un terme $(y^i - t^i)(x^i)^T$.

2) Gradients par rapport aux autre θ: Pour comprendre le terme de rétropropagation, on va dériver par rapports aux paramètres autres que W_s :

$$\begin{aligned}
\frac{\partial E}{\partial \theta} &= -\sum_i \sum_j \frac{t_j^i}{y_j^i} \frac{\partial y_j^i}{\partial \theta} \\
&= -\sum_i \sum_j \sum_c t_j^i (\delta_{pj} - y_p^i) \left[W_s \frac{\partial x^i}{\partial \theta} \right]_c \\
&= \sum_i \sum_j (y_j^i - t_j^i) \left[W_s \frac{\partial x^i}{\partial \theta} \right]_j \\
&= \sum_i (y^i - t^i) W_s \frac{\partial x^i}{\partial \theta}
\end{aligned}$$

Pour le noeud i on note i_g et i_d les indices des fils gauche et droite. On a donc :

$$x_j^i = f\left(\begin{bmatrix} x^{i_g} \\ x^{i_d} \end{bmatrix}^T V^{[j]} \begin{bmatrix} x^{i_g} \\ x^{i_d} \end{bmatrix} + W_{j:} \begin{bmatrix} x^{i_g} \\ x^{i_d} \end{bmatrix} \right)$$

On voit donc que $\partial x^i / \partial \theta$ va comporter deux parties : une partie qui constitue le gradient en propre pour le noeud i, et une partie en facteur de $\frac{\partial}{\partial \theta} \begin{bmatrix} x^{i_g} \\ x^{i_d} \end{bmatrix}$ qui va se rétropropager aux fils x^{i_g} et x^{i_d}. On note $\delta^{i,s} = W_s^T (y^i - t^i)$. On a alors en considérant que la fonction d'activation a une dérivée simple en fonction d'elle même comme c'est le cas pour $\tanh$ $f'(x) = F(f(x))$

$$\begin{aligned}
\frac{\partial x_j^i}{\partial \theta} = F(x_j^i) & \left(2 \begin{bmatrix} x^{i_g} \\ x^{i_d} \end{bmatrix}^T V^{[j]} + W_{j:} \right) \frac{\partial}{\partial \theta} \begin{bmatrix} x^{i_g} \\ x^{i_d} \end{bmatrix} \\
&+ \begin{bmatrix} x^{i_g} \\ x^{i_d} \end{bmatrix}^T \frac{\partial V^{[j]}}{\partial \theta} \begin{bmatrix} x^{i_g} \\ x^{i_d} \end{bmatrix} + \frac{\partial W_{j:}}{\partial \theta} \begin{bmatrix} x^{i_g} \\ x^{i_d} \end{bmatrix}
\end{aligned}$$

On voit apparaître en première ligne la partie qui se rétropropage aux fils et en deuxième ligne la partie propre au nœud i. Ainsi le gradient total va s'écrire :

$$\frac{\partial E}{\partial \theta} = \sum_i (\delta^{i,comp})^T \mathcal{E}^{i,\theta}$$

En notant $\mathcal{E}_j^{i,\theta}$ la partie propre au nœud i, et $\delta_{j,:}^{i,prop}$ la partie rétropropagée, soit :

$$\mathcal{E}_j^{i,\theta} = \begin{bmatrix} x^{i_g} \\ x^{i_d} \end{bmatrix}^T \frac{\partial V^{[j]}}{\partial \theta} \begin{bmatrix} x^{i_g} \\ x^{i_d} \end{bmatrix} + \frac{\partial W_{j:}}{\partial \theta} \begin{bmatrix} x^{i_g} \\ x^{i_d} \end{bmatrix}$$

$$\delta_{j,:}^{i,prop} = F(x_j^i) \left(2 \begin{bmatrix} x^{i_g} \\ x^{i_d} \end{bmatrix}^T V^{[j]} + W_{j:} \right)$$

Cette dernière partie rétropropagée se somme sur les j pour donner un total :

$$\begin{aligned} \delta^{i,down} &= \sum_j \delta_j^{i,s} \delta_{j:}^{i,prop} \\ &= \sum_j 2(\delta_j^{i,s} F(x_j^i)) V^{[j]} \begin{bmatrix} x^{i_g} \\ x^{i_d} \end{bmatrix} + W^T(\delta^{i,s} \otimes F(x^i)) \end{aligned}$$

Et on a donc les updates suivants pour calculer les $\delta^{i,comp}$ en notant p_i le père du noeud i qui qualifient l'algorithme de rétropropagation :

$$\begin{aligned} \delta^{i,comp} &= \delta^{i,s} + \delta^{p_i,down}[1:d] \text{ Si i est le fils gauche} \\ &= \delta^{i,s} + \delta^{p_i,down}[d+1:2d] \text{ Si i est le fils droit} \end{aligned}$$

Reste donc à calculer les différentes parties propres pour chaque paramètre.

3) Par rapport au biais W: La partie propre de la dérivée par rapport à W_{kl} s'écrit :

$$\frac{\partial W_{j:}}{\partial W_{kl}} \begin{bmatrix} x^{i_g} \\ x^{i_d} \end{bmatrix} = \delta_{kj} \begin{bmatrix} x^{i_g} \\ x^{i_d} \end{bmatrix}_l$$

D'où la partie propre (contribution des noeuds pères uniquement) :

$$\delta_k^{i,comp} \begin{bmatrix} x^{i_g} \\ x^{i_d} \end{bmatrix}_l$$

Soit pour le noeud i une contribution :

$$\frac{\partial E^i}{\partial W} = \delta^{i,comp} \begin{bmatrix} x^{i_g} \\ x^{i_d} \end{bmatrix}^T$$

4) Par rapport au tenseur V: De même la partie propre pour le paramètre $V_{kl}^{[j]}$:

$$\frac{\partial E^i}{\partial V_{kl}^{[j]}} = \delta_j^{i,comp} \begin{bmatrix} x^{i_g} \\ x^{i_d} \end{bmatrix}_k \begin{bmatrix} x^{i_g} \\ x^{i_d} \end{bmatrix}_l$$

Soit un total :

$$\frac{\partial E^i}{\partial V^{[j]}} = \delta_j^{i,comp} \begin{bmatrix} x^{i_g} \\ x^{i_d} \end{bmatrix} \begin{bmatrix} x^{i_g} \\ x^{i_d} \end{bmatrix}^T$$

5) Par rapport au dictionnaire L: Enfin pour les coefficients de L, seuls les feuilles font une contribution lorsque $x^i = L_k$, qui se lit directement :

$$\frac{\partial E^i}{\partial L_k} = \delta^{i,comp} \delta(x^i = L_k)$$