Calcul du gradient automatisé

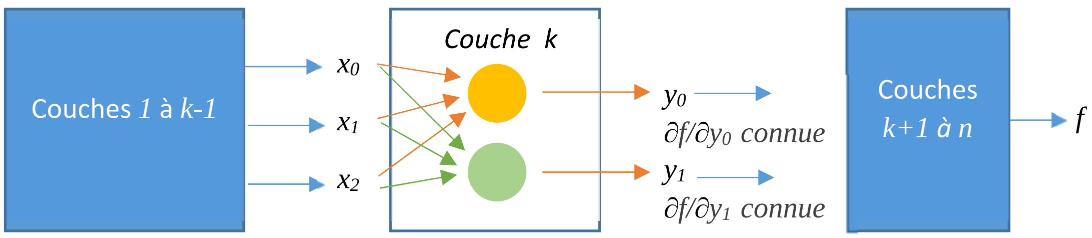
Les librairies de réseaux de neurones (Tensorflow, Pytorch) comme la librairie Numpy (avec Autograd) permettent de calculer le gradient d'une fonction de manière automatique. Pour cela, on utilise on système en trois phases :

- Dans la première phase appelée **Forward**, on calcule la valeur de la fonction. Ici les calculs sont effectués de manière progressive comme on les ferait à la main. Parallèlement, la librairie construit un graph de calcul qui représente le flot d'opérations effectuées pour arriver au résultat final. Dans ce graph sont stockés les résultats intermédiaires de chaque calcul.
- Dans la deuxième phase appelée **Backward** ou aussi **rétropropagation**, la librairie va parcourir le graph de calcul en sens inverse. Cette phase permet de construire toutes les dérivées partielles par rapport aux paramètres internes. Cette phase utilise les formules classiques de dérivation.
- Une troisième phase consiste à remettre à zéro les informations stockées pour le calcul du gradient lors de la rétropropagation. A ce niveau, les résultats intermédiaires stockés dans le graph doivent être mis à zéro pour éviter qu'il soit pris en compte lors de la prochaine itération.

Le calcul du gradient par rétropropagation

Les réseaux de neurones avec lesquels nous allons travailler sont décrits sous forme de couches (layer). Chaque couche prend les informations en entrée, effectue des calculs sur ces données et fournit des résultats pour la couche suivante. Le réseau complet constitue la fonction f et nous cherchons à déterminer les dérivées partielles de f par rapport à chaque paramètre interne. Une fois les dérivées partielles déterminés, nous connaissons le gradient nécessaire à l'algorithme de descente.

Durant la passe Backward, nous allons partir de la couche terminale (à droite) et reculer à chaque itération pour déterminer les dérivées partielles de la couche précédente. Ainsi, à l'itération n-k, nous pouvons modéliser notre système comme ceci :



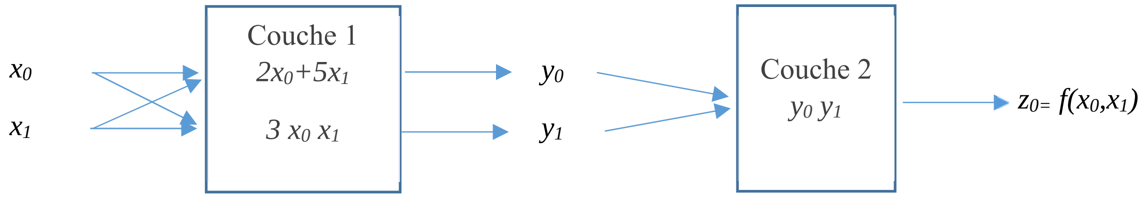
Les itérations précédentes ont permis de déterminer les valeurs de $\partial f/\partial y_0$ et $\partial f/\partial y_1$ et maintenant nous cherchons à déterminer les valeurs de $\partial f/\partial x_0$, $\partial f/\partial x_1$ et $\partial f/\partial x_2$. Pour cela, nous utilisons le théorème général des dérivations :

$$\frac{\partial f}{\partial x_i} = \sum_{j=0}^{m} \frac{\partial f}{\partial y_j} \times \frac{\partial y_j}{\partial x_i}$$

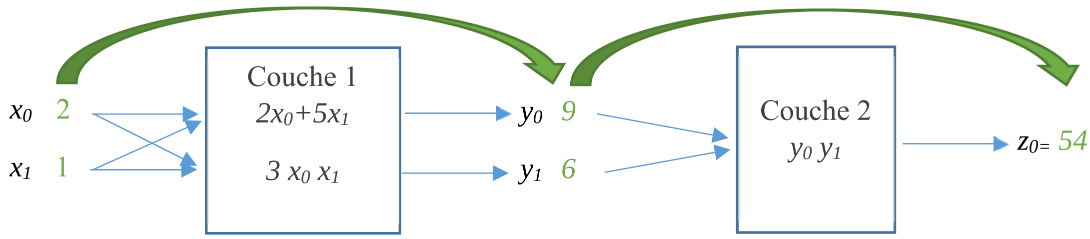
Les valeurs de $\partial f/\partial y_j$ ont été déterminées à l'itération précédente. Les termes $\partial y_j/\partial x_i$ correspondent aux dérivées partielles des fonctions utilisées dans la couche courante k. Ces fonctions sont connues et par exemple, on peut avoir $y_3 = ln(x_1)$. On peut donc exprimer leurs dérivées en utilisant les formules de dérivations : $\partial y_3/\partial x_1 = ln'(x_1) = 1/x_1$. Comme les valeurs des x_i sont connues depuis la passe Forward, il peut évaluer l'expression : $1/x_1$ pour connaître la valeur de $\partial y_3/\partial x_1$. Il reste donc à évaluer la formule du théorème général de dérivation pour trouver la valeur de $\partial f/\partial x_i$.

Exemple : $f(x_0, x_1) = (2x_0+5x_1)\times(3\, x_0\, x_1)$

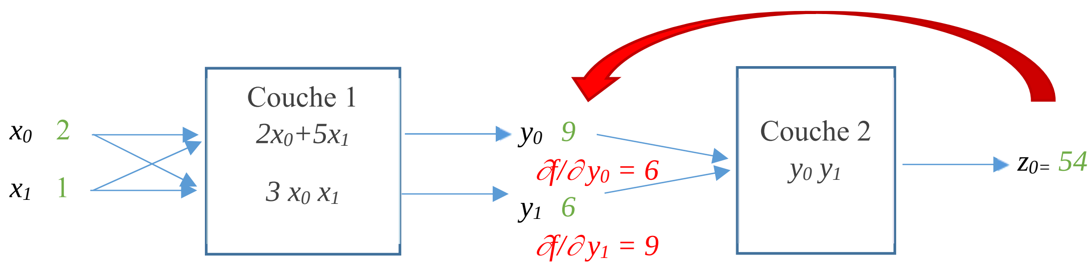
Voici le graph de calcul que nous pouvons associer à cette fonction :



Prenons $x_0=2$ et $x_1=1$. Nous allons tout d'abord effectuer la passe Forward :



La passe Forward est maintenant terminée. Nous commençons la rétropropagation en démarrant par la droite. La couche 2 effectue le calcul suivant $z_0(y_0 y_1) = y_0 \times y_1$. Les formules classiques de dérivation nous donnent $\partial z_0/\partial y_0 = \partial(y_0\, y_1)/\partial y_0 = y_1$. Or la valeur y_1 est connue car elle a été déterminée durant la passe Forward : $y_1=6$. Nous connaissons donc la valeur de $\partial f/\partial y_0$. Le raisonnement est identique pour y_1 : $\partial f/\partial y_1 = \partial z_0/\partial y_1 = \partial(y_0\, y_1)/\partial y_1 = y_0 = 9$.



Nous allons maintenant passer à l'étape suivante où nous allons rétropropager les valeurs des dérivées $\partial f/\partial y_0$ et $\partial f/\partial y_1$ pour trouver $\partial f/\partial x_0$ et $\partial f/\partial x_1$. Dans cette configuration plus complexe où $\partial f/\partial x_0$ dépend de $\partial f/\partial y_0$ et $\partial f/\partial y_1$, nous devons utiliser le théorème général :

$$\frac{\partial f}{\partial x_0} = \frac{\partial f}{\partial y_0} \times \frac{\partial y_0}{\partial x_0} + \frac{\partial f}{\partial y_1} \times \frac{\partial y_1}{\partial x_0}$$

Il reste à calculer :

- $\partial y_0/\partial x_0= \partial(2x_0+5x_1)\ /\partial x_0=2$
- $\partial y_1/\partial x_0=\partial(3\ x_0\ x_1)\ /\partial x_0=3\ x_1=3\times 1=3$

Ainsi : $\partial f/\partial x_0=6\times 2+9\times 3 = 39$

Nous calculons $\partial f/\partial x_1$ de la même manière :

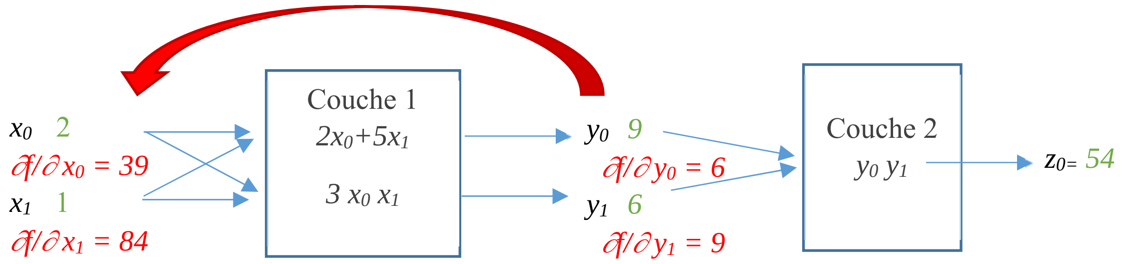
$$\frac{\partial f}{\partial x_1} = \frac{\partial f}{\partial y_0} \times \frac{\partial y_0}{\partial x_1} + \frac{\partial f}{\partial y_1} \times \frac{\partial y_1}{\partial x_1}$$

Il reste à calculer :

- $\partial y_0/\partial x_1= \partial(2x_0+5x_1)\ /\partial x_1=5$
- $\partial y_1/\partial x_1=\partial(3\ x_0\ x_1)\ /\partial x_1=3\ x_0=6$

Ainsi : $\partial f/\partial x_1=6\times 5+9\times 6 = 84$

Nous obtenons finalement :



Quel est l'intérêt de cette approche ? Elle est relativement simple à mettre en place, informatiquement parlant. La méthode de la rétropropagation permet de faire les calculs de dérivation localement dans chaque bloc, ce qui est beaucoup plus simple que de dériver une formule complexe avec des centaines de variables.

D'ailleurs, nos résultats sont-ils justes ? Pour vérifier, plutôt que de dériver $f(x_0, x_1)$ nous allons plutôt utiliser la formule d'approximation pour vérifier notre calcul :

$\partial f/\partial x_0 = [\,f(x_0+\varepsilon, x_1)- f(x_0, x_1)]\ /\ \varepsilon$

$\partial f/\partial x_1 = [\,f(x_0, x_1+\varepsilon)- f(x_0, x_1)]\ /\ \varepsilon$ avec $\varepsilon = 0.001$