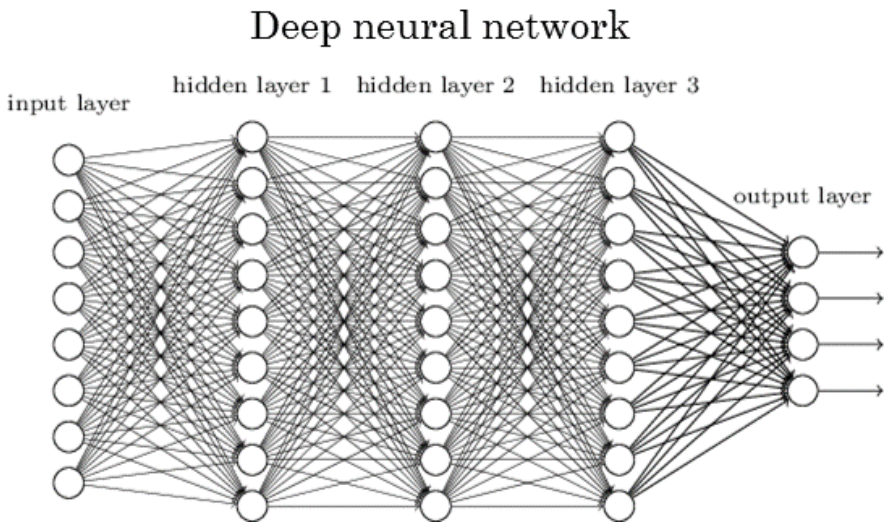
Chapitre 6.3 –Le réseau de neurones et l'activation par *batch*



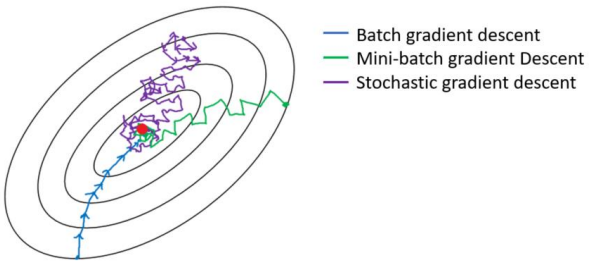
https://datawarrior.wordpress.com/2017/10/31/interpretability-of-neural-networks/

La descente du gradient par mini-batch

La descente du gradient par mini-batch consiste à faire une somme des gradients sur l'activation de *m* données et diviser l'ensemble des champs par la taille de la mini-batch. Cette technique permet à la fonction d'erreur d'avoir plus de précision que la méthode stochastique en réduisant le bruit. Elle est tout de même plus turbulente que la descente par batch entière, mais elle est beaucoup plus rapide.

Mathématiquement, elle est représentée par l'équation suivante :

$$\nabla C = \frac{1}{m}\sum_{b=0}^{m-1} \nabla C\left(W, x_{[b]}^{(*)}, y_{[b]}\right)$$



https://towardsdatascience.com/gradient-descent-algorithm-and-its-variants-10f652806a3

Illustration d'une descente du gradient en employant différente descente du gradient (batch entière, mini-batch et stochastique).

où W : Représente l'ensemble des paramètres d'un réseau.

$x_{[b]}^{(*)}$: Représente un vecteur d'entrée d'indice de batch *b*.

$y_{[b]}$: La valeur attendue de l'activation d'un vecteur $x_{[b]}^{(*)}$.

m : Le nombre totale de vecteur d'entrée qui sera activé dans la *batch*,

C : La fonction d'erreur utilisée pour évaluer l'erreur entre toutes les activations des vecteurs d'entrée $x_{[b]}^{(*)}$ et leur valeur attendue $y_{[b]}$.

L'apprentissage avec moment

L'apprentissage avec moment consiste à évaluer un gradient de la fonction d'erreur en utilisant une valeur ancienne de celui-ci afin de réorienter moins brusquement le gradient à appliquer sur la correction des poids et biais d'un réseau. Dans un jargon en physique, on dirait que l'on a ajouté une inertie à la valeur du gradient.

Voici la séquence de la mise à jour du gradient de la fonction d'erreur :

1) Choisir un paramètre d'inertie β tel que $0 < \beta < 1$:

   Exemple : $\beta \approx 0.9$

2) Calculer le gradient actuel sur les paramètres d'un réseau :

   $$\nabla C$$

3) Calculer le gradient avec moment de l'itération *n* afin de générer un gradient influencé par la valeur actuelle ainsi qu'avec une fraction de la valeur précédente :

   $$\nabla C^{n+1} = \beta \nabla C^{n} + (1-\beta)\nabla C$$

4) Réaliser l'apprentissage avec le gradient avec moment de l'itération *n* + 1 :

   $$\tilde{W} = W - \alpha \nabla C^{n+1}$$

où W sont les paramètres du réseau, $\tilde{W}$ sont les paramètres améliorés et α est le taux d'apprentissage.

L'activation du *batch-network*

En construction …

La normalisation

La fonction de moyenne

La fonction de la moyenne μ consiste à évaluer la moyenne de différentes valeurs que peut prendre un neurone. Habituellement, le type de valeur choisie à évaluer en moyenne est l'agrégation $z_{i[B]}^{(k)}$ du neurone i d'une couche k lors de l'activation d'un vecteur d'entré B dans le réseau.

La fonction de moyenne

$$\mu_i^{(k)} = \mu_i^{(k)}\left(z_{i[0]}^{(k)}, z_{i[1]}^{(k)}, \ldots, z_{i[B]}^{(k)}, \ldots, z_{i[m-1]}^{(k)}\right)$$

pour la *batch* d'agrégation $z_{i[B]}^{(k)}$ où $B \in [0, m-1]$ dont m est le nombre de vecteur d'entrée dans la batch est égale à l'expression suivante :

$$\mu_i^{(k)} = \frac{1}{m}\sum_{B=0}^{m-1} z_{i[B]}^{(k)}$$

L'expression de la différentielle de la fonction de la moyenne est égale à l'expression suivante :

Fonction de la moyenne	Différentielle de la fonction de la moyenne
$\mu_i^{(k)} = \frac{1}{m}\sum_{B=0}^{m-1} z_{i[B]}^{(k)}$	$\mathrm{d}\mu_i^{(k)} = \frac{1}{m}\sum_{B=0}^{m-1} \mathrm{d}z_{i[B]}^{(k)}$

où $i \in [0 \ldots N^{(k)} - 1]$

Preuve :

$$\mu_i^{(k)} = \frac{1}{m}\sum_{B=0}^{m-1} z_{i[B]}^{(k)} \quad \Rightarrow \quad \mathrm{d}\mu_i^{(k)} = \mathrm{d}\left(\frac{1}{m}\sum_{B=0}^{m-1} z_{i[B]}^{(k)}\right)$$

$$\Rightarrow \quad \mathrm{d}\mu_i^{(k)} = \frac{1}{m}\sum_{B=0}^{m-1} \mathrm{d}\, z_{i[B]}^{(k)} \quad \blacksquare$$

La fonction de la variance

La fonction de la variance σ consiste à évaluer l'écart entre différentes valeurs que peut prendre un neurone et la moyenne des valeurs du neurone. Habituellement, le type de valeur choisie à évaluer en variance est l'agrégation $z_{i[B]}^{(k)}$ du neurone i d'une couche k lors de l'activation d'un vecteur d'entré B dans le réseau.

La fonction de variance

$$\sigma_i^{(k)} = \sigma_i^{(k)}\left(z_{i[0]}^{(k)}, z_{i[1]}^{(k)}, \ldots, z_{i[B]}^{(k)}, \ldots, z_{i[m-1]}^{(k)}, \mu_i^{(k)}\right)$$

pour la *batch* d'agrégation $z_{i[B]}^{(k)}$ où $B \in [0, m-1]$ dont m est le nombre de vecteur d'entrée dans la batch est égale à l'expression suivante :

$$\sigma_i^{(k)} = \frac{1}{m}\sum_{B=0}^{m-1}\left(z_{i[B]}^{(k)} - \mu_i^{(k)}\right)^2 \qquad \text{où} \qquad \mu_i^{(k)} = \frac{1}{m}\sum_{B=0}^{m-1} z_{i[B]}^{(k)}$$

L'expression de la différentielle de la fonction de la variance est égale à l'expression suivante :

Fonction de la variance	Différentielle de la fonction de la variance
$\sigma_i^{(k)} = \frac{1}{m}\sum_{B=0}^{m-1}\left(z_{i[B]}^{(k)} - \mu_i^{(k)}\right)^2$	$\mathrm{d}\sigma_i^{(k)} = \frac{2}{m}\sum_{B=0}^{m-1}\left(z_{i[B]}^{(k)} - \mu_i^{(k)}\right)\mathrm{d}z_{i[B]}^{(k)}$

où $\quad i \in [0 \ldots N^{(k)} - 1]$

Preuve :

$$\sigma_i^{(k)} = \frac{1}{m}\sum_{B=0}^{m-1}\left(z_{i[B]}^{(k)} - \mu_i^{(k)}\right)^2 \quad \Rightarrow \quad \mathrm{d}\sigma_i^{(k)} = \mathrm{d}\left(\frac{1}{m}\sum_{B=0}^{m-1}\left(z_{i[B]}^{(k)} - \mu_i^{(k)}\right)^2\right)$$

$$\Rightarrow \quad \mathrm{d}\sigma_i^{(k)} = \frac{1}{m}\sum_{B=0}^{m-1}\mathrm{d}\left(z_{i[B]}^{(k)} - \mu_i^{(k)}\right)^2$$

$$\Rightarrow \quad \mathrm{d}\sigma_i^{(k)} = \frac{1}{m}\sum_{B=0}^{m-1} 2\left(z_{i[B]}^{(k)} - \mu_i^{(k)}\right)\mathrm{d}\left(z_{i[B]}^{(k)} - \mu_i^{(k)}\right)$$

$$\Rightarrow \quad \mathrm{d}\sigma_i^{(k)} = \frac{2}{m}\sum_{B=0}^{m-1}\left(z_{i[B]}^{(k)} - \mu_i^{(k)}\right)\left(\mathrm{d}z_{i[B]}^{(k)} - \mathrm{d}\mu_i^{(k)}\right)$$

$$\Rightarrow \quad \mathrm{d}\sigma_i^{(k)} = \frac{2}{m}\sum_{B=0}^{m-1}\left(z_{i[B]}^{(k)} - \mu_i^{(k)}\right)\mathrm{d}z_{i[B]}^{(k)} - \frac{2}{m}\sum_{B=0}^{m-1}\left(z_{i[B]}^{(k)} - \mu_i^{(k)}\right)\mathrm{d}\mu_i^{(k)}$$

Analysons plus en détail le calcul du terme de droite :

$$\mathrm{T} = -\frac{2}{m}\sum_{B=0}^{m-1}\left(z_{i[B]}^{(k)} - \mu_i^{(k)}\right)\mathrm{d}\mu_i^{(k)} \quad \Rightarrow \quad \mathrm{T} = -\frac{2}{m}\mathrm{d}\mu_i^{(k)}\sum_{B=0}^{m-1}\left(z_{i[B]}^{(k)} - \mu_i^{(k)}\right)$$

$$\Rightarrow \quad \mathrm{T} = -\frac{2}{m}\mathrm{d}\mu_i^{(k)}\left(\sum_{B=0}^{m-1} z_{i[B]}^{(k)} - \sum_{B=0}^{m-1}\mu_i^{(k)}\right)$$

$$\Rightarrow \quad \mathrm{T} = -\frac{2}{m}\mathrm{d}\mu_i^{(k)}\left(\sum_{B=0}^{m-1} z_{i[B]}^{(k)} - m\mu_i^{(k)}\right)$$

$$\Rightarrow \quad \mathrm{T} = -\frac{2}{m}\mathrm{d}\mu_i^{(k)}\left(m\mu_i^{(k)} - m\mu_i^{(k)}\right) \qquad \left(\text{car } \mu_i^{(k)} = \frac{1}{m}\sum_{B=0}^{m-1} z_{i[B]}^{(k)}\right)$$

$$\Rightarrow \qquad \mathrm{T} = 0$$

Ainsi

$$\mathrm{d}\sigma_i^{(k)} = \frac{2}{m}\sum_{B=0}^{m-1}\left(z_{i[B]}^{(k)} - \mu_i^{(k)}\right)\mathrm{d}z_{i[B]}^{(k)} \quad . \ \blacksquare$$

La fonction de la normalisation

La fonction de normalisation $n_{i[B]}^{(k)}$ est une fonction ayant pour but de normaliser des agrégations favorisant une accélération de l'apprentissage. Cette fonction dépendant d'une agrégation $z_{i[B]}^{(k)}$ pour plein d'éléments de la batch B ainsi que de deux nouveaux paramètres $\gamma_i^{(k)}$ et $\beta_i^{(k)}$ décrivant le réseau et prendra la forme de

$$n_{i[B]}^{(k)} = n_{i[B]}^{(k)}\left(z_{i[0]}^{(k)}, z_{i[1]}^{(k)}, \ldots, z_{i[B]}^{(k)}, \ldots, z_{i[m-1]}^{(k)}, \gamma_i^{(k)}, \beta_i^{(k)}\right)$$

Où $\gamma_i^{(k)}$ et $\beta_i^{(k)}$ sont des nouveaux paramètres du réseau nécessaire au fonctionnement de la fonction étant le *scale* et *shift*.

On peut reformuler la dépendance entre les agrégations $z_{i[B]}^{(k)}$ du neurone i pour l'ensemble de la batch B en incluant des éléments statistiques sous la forme suivante :

$$n_{i[B]}^{(k)} = n_{i[B]}^{(k)}\left(\hat{z}_{i[B]}^{(k)}, \gamma_i^{(k)}, \beta_i^{(k)}\right) \qquad \text{avec} \qquad \hat{z}_{i[B]}^{(k)} = \hat{z}_{i[B]}^{(k)}\left(z_{i[B]}^{(k)}, \mu_i^{(k)}, \sigma_i^{(k)}\right)$$

(Normalisation du neurone) (Agrégation normalisée)

tel que

$$\sigma_i^{(k)} = \frac{1}{m}\sum_{B=0}^{m-1}\left(z_{i[B]}^{(k)} - \mu_i^{(k)}\right)^2 \qquad \mu_i^{(k)} = \frac{1}{m}\sum_{B=0}^{m-1} z_{i[B]}^{(k)}$$

(Variance du neurone i) Moyenne du neurone i)

L'expression de la différentielle de la fonction de la normalisation sera égale à l'expression générale suivante :

$$\begin{aligned}\mathrm{d}n_{i[B]}^{(k)} = {} & \frac{\partial n_{i[B]}^{(k)}}{\partial \gamma_i^{(k)}}\mathrm{d}\gamma_i^{(k)} + \frac{\partial n_{i[B]}^{(k)}}{\partial \beta_i^{(k)}}\mathrm{d}\beta_i^{(k)} \\ & + \frac{\partial n_{i[B]}^{(k)}}{\partial \hat{z}_{i[B]}^{(k)}}\left(\frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial z_{i[B]}^{(k)}}\mathrm{d}z_{i[B]}^{(k)} + \frac{1}{m}\frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial \mu_{i[B]}^{(k)}}\sum_{a=0}^{m-1}\mathrm{d}z_{i[a]}^{(k)} + \frac{2}{m}\frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial \sigma_i^{(k)}}\sum_{a=0}^{m-1}\left(z_{i[a]}^{(k)} - \mu_i^{(k)}\right)\mathrm{d}z_{i[a]}^{(k)}\right)\end{aligned}$$

Preuve :

$$n_{i[B]}^{(k)} = n_{i[B]}^{(k)}\left(\hat{z}_{i[B]}^{(k)},\, \gamma_i^{(k)},\, \beta_i^{(k)}\right)$$

$$\Rightarrow \qquad \mathrm{d}n_{i[B]}^{(k)} = \frac{\partial n_{i[B]}^{(k)}}{\partial \hat{z}_{i[B]}^{(k)}}\mathrm{d}\hat{z}_{i[B]}^{(k)} + \frac{\partial n_{i[B]}^{(k)}}{\partial \gamma_i^{(k)}}\mathrm{d}\gamma_i^{(k)} + \frac{\partial n_{i[B]}^{(k)}}{\partial \beta_i^{(k)}}\mathrm{d}\beta_i^{(k)}$$

Si l'on se concentre sur la différentielle de l'agrégation normalisée $\hat{z}_{i[B]}^{(k)}$, nous avons :

$$\mathrm{d}\hat{z}_{i[B]}^{(k)} = \frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial z_{i[B]}^{(k)}}\mathrm{d}z_{i[B]}^{(k)} + \frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial \mu_i^{(k)}}\mathrm{d}\mu_i^{(k)} + \frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial \sigma_i^{(k)}}\mathrm{d}\sigma_i^{(k)}$$

$$\Rightarrow \qquad \mathrm{d}\hat{z}_{i[B]}^{(k)} = \frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial z_{i[B]}^{(k)}}\mathrm{d}z_{i[B]}^{(k)} + \frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial \mu_i^{(k)}}\left(\frac{1}{m}\sum_{a=0}^{m-1}\mathrm{d}z_{i[a]}^{(k)}\right) + \frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial \sigma_i^{(k)}}\mathrm{d}\sigma_i^{(k)}$$

$$\Rightarrow \qquad \mathrm{d}\hat{z}_{i[B]}^{(k)} = \frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial z_{i[B]}^{(k)}}\mathrm{d}z_{i[B]}^{(k)} + \frac{1}{m}\frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial \mu_{i[B]}^{(k)}}\sum_{a=0}^{m-1}\mathrm{d}z_{i[a]}^{(k)} + \frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial \sigma_i^{(k)}}\left(\frac{2}{m}\sum_{a=0}^{m-1}\left(z_{i[a]}^{(k)} - \mu_i^{(k)}\right)\mathrm{d}z_{i[a]}^{(k)}\right)$$

$$\Rightarrow \qquad \mathrm{d}\hat{z}_{i[B]}^{(k)} = \frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial z_{i[B]}^{(k)}}\mathrm{d}z_{i[B]}^{(k)} + \frac{1}{m}\frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial \mu_{i[B]}^{(k)}}\sum_{a=0}^{m-1}\mathrm{d}z_{i[a]}^{(k)} + \frac{2}{m}\frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial \sigma_i^{(k)}}\sum_{a=0}^{m-1}\left(z_{i[a]}^{(k)} - \mu_i^{(k)}\right)\mathrm{d}z_{i[a]}^{(k)}$$

C'est ici que l'on constate que $\mathrm{d}\hat{z}_{i[B]}^{(k)}$ influencera l'agrégation $z_{i[B]}^{(k)}$ par le terme $\mathrm{d}z_{i[B]}^{(k)}$ de la batch B, mais également tous les autres agrégations $z_{i[a]}^{(k)}$ où $a \in [0, m-1]$ de la batch B en raison de la fonction de la moyenne $\mu_i^{(k)}$ et la variance $\sigma_i^{(k)}$ qui eux dépendent de toutes les agrégation $z_{i[B]}^{(k)}$ de la batch B.

Ainsi, nous obtenons

$$\begin{aligned}\mathrm{d}n_{i[B]}^{(k)} = {} & \frac{\partial n_{i[B]}^{(k)}}{\partial \gamma_i^{(k)}}\mathrm{d}\gamma_i^{(k)} + \frac{\partial n_{i[B]}^{(k)}}{\partial \beta_i^{(k)}}\mathrm{d}\beta_i^{(k)} \\ & + \frac{\partial n_{i[B]}^{(k)}}{\partial \hat{z}_{i[B]}^{(k)}}\left(\frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial z_{i[B]}^{(k)}}\mathrm{d}z_{i[B]}^{(k)} + \frac{1}{m}\frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial \mu_{i[B]}^{(k)}}\sum_{a=0}^{m-1}\mathrm{d}z_{i[a]}^{(k)} + \frac{2}{m}\frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial \sigma_i^{(k)}}\sum_{a=0}^{m-1}\left(z_{i[a]}^{(k)} - \mu_i^{(k)}\right)\mathrm{d}z_{i[a]}^{(k)}\right)\end{aligned}$$ ■

La différentielle de la fonction d'erreur avec fonction de normalisation

En construction …

$$\mathrm{d}C = \frac{1}{m}\sum_{B=0}^{m-1}\sum_{i=0}^{N^{(k)}-1}\left(\begin{array}{l}\Delta_{i[B]}^{(k)}\dfrac{\partial n_{i[B]}^{(k)}}{\partial \gamma_i^{(k)}}\mathrm{d}\gamma_i^{(k)} + \Delta_{i[B]}^{(k)}\dfrac{\partial n_{i[B]}^{(k)}}{\partial \beta_i^{(k)}}\mathrm{d}\beta_i^{(k)} \\ + \left(\Delta_{i[B]}^{(k)}\dfrac{\partial \hat{z}_{i[B]}^{(k)}}{\partial z_{i[B]}^{(k)}} + \displaystyle\sum_{a=0}^{m-1}\Delta_{i[a]}^{(k)}\frac{1}{m}\frac{\partial \hat{z}_{i[a]}^{(k)}}{\partial \mu_{i[a]}^{(k)}} + \sum_{a=0}^{m-1}\Delta_{i[a]}^{(k)}\frac{2}{m}\frac{\partial \hat{z}_{i[a]}^{(k)}}{\partial \sigma_i^{(k)}}\left(z_{i[B]}^{(k)} - \mu_i^{(k)}\right)\right)\mathrm{d}z_{i[B]}^{(k)}\end{array}\right)$$

Preuve :

Soit la différentielle de la fonction d'erreur C dont l'erreur propagée à la fonction de normalisation $n_{i[B]}^{(k)}$ de niveau de propagation $k = L - n$ est égale à l'expression

$$\mathrm{d}C = \frac{1}{m}\sum_{B=0}^{m-1}\sum_{i=0}^{N^{(k)}-1}\Delta_{i[B]}^{(k)}\mathrm{d}n_{i[B]}^{(k)}$$

où

$$\begin{aligned}\mathrm{d}n_{i[B]}^{(k)} &= \frac{\partial n_{i[B]}^{(k)}}{\partial \gamma_i^{(k)}}\mathrm{d}\gamma_i^{(k)} + \frac{\partial n_{i[B]}^{(k)}}{\partial \beta_i^{(k)}}\mathrm{d}\beta_i^{(k)} \\ &\quad + \frac{\partial n_{i[B]}^{(k)}}{\partial \hat{z}_{i[B]}^{(k)}}\left(\frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial z_{i[B]}^{(k)}}\mathrm{d}z_{i[B]}^{(k)} + \frac{1}{m}\frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial \mu_{i[B]}^{(k)}}\sum_{a=0}^{m-1}\mathrm{d}z_{i[a]}^{(k)} + \frac{2}{m}\frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial \sigma_i^{(k)}}\sum_{a=0}^{m-1}\left(z_{i[a]}^{(k)} - \mu_i^{(k)}\right)\mathrm{d}z_{i[a]}^{(k)}\right)\end{aligned}$$

Développons l'expression afin d'évaluer l'erreur qui sera propagée par la fonction $n_{i[B]}^{(k)}$ pour chaque neurone i de cette fonction à la fonction d'activation d'agrégation $z_{i[B]}^{(k)}$:

$$\mathrm{d}C = \frac{1}{m}\sum_{B=0}^{m-1}\sum_{i=0}^{N^{(k)}-1}\Delta_{i[B]}^{(k)}\mathrm{d}n_{i[B]}^{(k)}$$

$$\Rightarrow \qquad \mathrm{d}C = \frac{1}{m}\sum_{B=0}^{m-1}\sum_{i=0}^{N^{(k)}-1}\Delta_{i[B]}^{(k)}\left(\begin{array}{l}\dfrac{\partial n_{i[B]}^{(k)}}{\partial \gamma_i^{(k)}}\mathrm{d}\gamma_i^{(k)} + \dfrac{\partial n_{i[B]}^{(k)}}{\partial \beta_i^{(k)}}\mathrm{d}\beta_i^{(k)} \\ + \dfrac{\partial n_{i[B]}^{(k)}}{\partial \hat{z}_{i[B]}^{(k)}}\left(\dfrac{\partial \hat{z}_{i[B]}^{(k)}}{\partial z_{i[B]}^{(k)}}\mathrm{d}z_{i[B]}^{(k)} + \dfrac{1}{m}\dfrac{\partial \hat{z}_{i[B]}^{(k)}}{\partial \mu_{i[B]}^{(k)}}\displaystyle\sum_{a=0}^{m-1}\mathrm{d}z_{i[a]}^{(k)} + \dfrac{2}{m}\dfrac{\partial \hat{z}_{i[B]}^{(k)}}{\partial \sigma_i^{(k)}}\displaystyle\sum_{a=0}^{m-1}\left(z_{i[a]}^{(k)} - \mu_i^{(k)}\right)\mathrm{d}z_{i[a]}^{(k)}\right)\end{array}\right)$$

$$\Rightarrow$$

$$\mathrm{d}C = \frac{1}{m}\sum_{B=0}^{m-1}\sum_{i=0}^{N^{(k)}-1}\left(\begin{array}{l}\Delta_{i[B]}^{(k)}\dfrac{\partial n_{i[B]}^{(k)}}{\partial \gamma_i^{(k)}}\mathrm{d}\gamma_i^{(k)} + \Delta_{i[B]}^{(k)}\dfrac{\partial n_{i[B]}^{(k)}}{\partial \beta_i^{(k)}}\mathrm{d}\beta_i^{(k)} \\ + \dfrac{\partial n_{i[B]}^{(k)}}{\partial \hat{z}_{i[B]}^{(k)}}\left(\Delta_{i[B]}^{(k)}\dfrac{\partial \hat{z}_{i[B]}^{(k)}}{\partial z_{i[B]}^{(k)}}\mathrm{d}z_{i[B]}^{(k)} + \Delta_{i[B]}^{(k)}\dfrac{1}{m}\dfrac{\partial \hat{z}_{i[B]}^{(k)}}{\partial \mu_{i[B]}^{(k)}}\displaystyle\sum_{a=0}^{m-1}\mathrm{d}z_{i[a]}^{(k)} + \Delta_{i[B]}^{(k)}\dfrac{2}{m}\dfrac{\partial \hat{z}_{i[B]}^{(k)}}{\partial \sigma_i^{(k)}}\displaystyle\sum_{a=0}^{m-1}\left(z_{i[a]}^{(k)} - \mu_i^{(k)}\right)\mathrm{d}z_{i[a]}^{(k)}\right)\end{array}\right)$$

$$\Rightarrow$$

$$\begin{aligned}\mathrm{d}C &= \frac{1}{m}\sum_{B=0}^{m-1}\sum_{i=0}^{N^{(k)}-1}\Delta_{i[B]}^{(k)}\frac{\partial n_{i[B]}^{(k)}}{\partial \gamma_i^{(k)}}\mathrm{d}\gamma_i^{(k)} + \frac{1}{m}\sum_{B=0}^{m-1}\sum_{i=0}^{N^{(k)}-1}\Delta_{i[B]}^{(k)}\frac{\partial n_{i[B]}^{(k)}}{\partial \beta_i^{(k)}}\mathrm{d}\beta_i^{(k)} \\ &\quad + \frac{1}{m}\sum_{B=0}^{m-1}\sum_{i=0}^{N^{(k)}-1}\frac{\partial n_{i[B]}^{(k)}}{\partial \hat{z}_{i[B]}^{(k)}}\left(\Delta_{i[B]}^{(k)}\frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial z_{i[B]}^{(k)}}\mathrm{d}z_{i[B]}^{(k)} + \Delta_{i[B]}^{(k)}\frac{1}{m}\frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial \mu_{i[B]}^{(k)}}\sum_{a=0}^{m-1}\mathrm{d}z_{i[a]}^{(k)} + \Delta_{i[B]}^{(k)}\frac{2}{m}\frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial \sigma_i^{(k)}}\sum_{a=0}^{m-1}\left(z_{i[a]}^{(k)} - \mu_i^{(k)}\right)\mathrm{d}z_{i[a]}^{(k)}\right)\end{aligned}$$

À cette étape de la démonstration, nous allons trouver une stratégie afin de factoriser un terme $\mathrm{d}z_{i[B]}^{(k)}$ pour permettre à cette différentielle d'avoir une structure mathématique permettant la propagation de l'erreur $\Delta_{i[B]}^{n(k)}$ de la fonction de normalisation $n_{i[B]}^{(k)}$. Pour ce faire, nous allons changer l'ordre des sommations sur les indices *a* et *B* :

$$\mathrm{d}C = \ldots + \frac{1}{m}\sum_{B=0}^{m-1}\sum_{i=0}^{N^{(k)}-1}\frac{\partial n_{i[B]}^{(k)}}{\partial \hat{z}_{i[B]}^{(k)}}\left(\Delta_{i[B]}^{(k)}\frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial z_{i[B]}^{(k)}}\mathrm{d}z_{i[B]}^{(k)} + \Delta_{i[B]}^{(k)}\frac{1}{m}\frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial \mu_{i[B]}^{(k)}}\sum_{a=0}^{m-1}\mathrm{d}z_{i[a]}^{(k)} + \Delta_{i[B]}^{(k)}\frac{2}{m}\frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial \sigma_{i}^{(k)}}\sum_{a=0}^{m-1}\left(z_{i[a]}^{(k)} - \mu_{i}^{(k)}\right)\mathrm{d}z_{i[a]}^{(k)}\right)$$

$$\Rightarrow \begin{aligned}\mathrm{d}C &= \ldots + \frac{1}{m}\sum_{B=0}^{m-1}\sum_{i=0}^{N^{(k)}-1}\Delta_{i[B]}^{(k)}\frac{\partial n_{i[B]}^{(k)}}{\partial \hat{z}_{i[B]}^{(k)}}\frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial z_{i[B]}^{(k)}}\mathrm{d}z_{i[B]}^{(k)} \\ &+ \frac{1}{m}\sum_{B=0}^{m-1}\sum_{i=0}^{N^{(k)}-1}\Delta_{i[B]}^{(k)}\frac{\partial n_{i[B]}^{(k)}}{\partial \hat{z}_{i[B]}^{(k)}}\frac{1}{m}\frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial \mu_{i[B]}^{(k)}}\sum_{a=0}^{m-1}\mathrm{d}z_{i[a]}^{(k)} + \frac{1}{m}\sum_{B=0}^{m-1}\sum_{i=0}^{N^{(k)}-1}\Delta_{i[B]}^{(k)}\frac{\partial n_{i[B]}^{(k)}}{\partial \hat{z}_{i[B]}^{(k)}}\frac{2}{m}\frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial \sigma_{i}^{(k)}}\sum_{a=0}^{m-1}\left(z_{i[a]}^{(k)} - \mu_{i}^{(k)}\right)\mathrm{d}z_{i[a]}^{(k)}\end{aligned}$$

$$\Rightarrow \begin{aligned}\mathrm{d}C &= \ldots + \frac{1}{m}\sum_{B=0}^{m-1}\sum_{i=0}^{N^{(k)}-1}\Delta_{i[B]}^{(k)}\frac{\partial n_{i[B]}^{(k)}}{\partial \hat{z}_{i[B]}^{(k)}}\frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial z_{i[B]}^{(k)}}\mathrm{d}z_{i[B]}^{(k)} \\ &+ \frac{1}{m}\sum_{a=0}^{m-1}\sum_{B=0}^{m-1}\sum_{i=0}^{N^{(k)}-1}\Delta_{i[B]}^{(k)}\frac{\partial n_{i[B]}^{(k)}}{\partial \hat{z}_{i[B]}^{(k)}}\frac{1}{m}\frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial \mu_{i[B]}^{(k)}}\mathrm{d}z_{i[a]}^{(k)} + \frac{1}{m}\sum_{a=0}^{m-1}\sum_{B=0}^{m-1}\sum_{i=0}^{N^{(k)}-1}\Delta_{i[B]}^{(k)}\frac{\partial n_{i[B]}^{(k)}}{\partial \hat{z}_{i[B]}^{(k)}}\frac{2}{m}\frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial \sigma_{i}^{(k)}}\left(z_{i[a]}^{(k)} - \mu_{i}^{(k)}\right)\mathrm{d}z_{i[a]}^{(k)}\end{aligned}$$

(flip $a \leftrightarrow B$)

$$\Rightarrow \begin{aligned}\mathrm{d}C &= \ldots + \frac{1}{m}\sum_{B=0}^{m-1}\sum_{i=0}^{N^{(k)}-1}\Delta_{i[B]}^{(k)}\frac{\partial n_{i[B]}^{(k)}}{\partial \hat{z}_{i[B]}^{(k)}}\frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial z_{i[B]}^{(k)}}\mathrm{d}z_{i[B]}^{(k)} \\ &+ \frac{1}{m}\sum_{B=0}^{m-1}\sum_{i=0}^{N^{(k)}-1}\sum_{a=0}^{m-1}\Delta_{i[a]}^{(k)}\frac{\partial n_{i[a]}^{(k)}}{\partial \hat{z}_{i[a]}^{(k)}}\frac{1}{m}\frac{\partial \hat{z}_{i[a]}^{(k)}}{\partial \mu_{i[a]}^{(k)}}\mathrm{d}z_{i[B]}^{(k)} + \frac{1}{m}\sum_{B=0}^{m-1}\sum_{i=0}^{N^{(k)}-1}\sum_{a=0}^{m-1}\Delta_{i[a]}^{(k)}\frac{\partial n_{i[a]}^{(k)}}{\partial \hat{z}_{i[a]}^{(k)}}\frac{2}{m}\frac{\partial \hat{z}_{i[a]}^{(k)}}{\partial \sigma_{i}^{(k)}}\left(z_{i[B]}^{(k)} - \mu_{i}^{(k)}\right)\mathrm{d}z_{i[B]}^{(k)}\end{aligned}$$

$$\mathrm{d}C = \ldots + \frac{1}{m}\sum_{B=0}^{m-1}\sum_{i=0}^{N^{(k)}-1}\left(\Delta_{i[B]}^{(k)}\frac{\partial n_{i[B]}^{(k)}}{\partial \hat{z}_{i[B]}^{(k)}}\frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial z_{i[B]}^{(k)}} + \sum_{a=0}^{m-1}\Delta_{i[a]}^{(k)}\frac{\partial n_{i[a]}^{(k)}}{\partial \hat{z}_{i[a]}^{(k)}}\frac{1}{m}\frac{\partial \hat{z}_{i[a]}^{(k)}}{\partial \mu_{i[a]}^{(k)}} + \sum_{a=0}^{m-1}\Delta_{i[a]}^{(k)}\frac{\partial n_{i[a]}^{(k)}}{\partial \hat{z}_{i[a]}^{(k)}}\frac{2}{m}\frac{\partial \hat{z}_{i[a]}^{(k)}}{\partial \sigma_{i}^{(k)}}\left(z_{i[B]}^{(k)} - \mu_{i}^{(k)}\right)\right)\mathrm{d}z_{i[B]}^{(k)}$$

Si l'on réintroduit nos éléments, nous obtenons l'expression suivante :

$$\begin{aligned}\mathrm{d}C &= \frac{1}{m}\sum_{B=0}^{m-1}\sum_{i=0}^{N^{(k)}-1}\Delta_{i[B]}^{(k)}\frac{\partial n_{i[B]}^{(k)}}{\partial \gamma_{i}^{(k)}}\mathrm{d}\gamma_{i}^{(k)} + \frac{1}{m}\sum_{B=0}^{m-1}\sum_{i=0}^{N^{(k)}-1}\Delta_{i[B]}^{(k)}\frac{\partial n_{i[B]}^{(k)}}{\partial \beta_{i}^{(k)}}\mathrm{d}\beta_{i}^{(k)} \\ &+ \frac{1}{m}\sum_{B=0}^{m-1}\sum_{i=0}^{N^{(k)}-1}\left(\Delta_{i[B]}^{(k)}\frac{\partial n_{i[B]}^{(k)}}{\partial \hat{z}_{i[B]}^{(k)}}\frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial z_{i[B]}^{(k)}} + \sum_{a=0}^{m-1}\Delta_{i[a]}^{(k)}\frac{\partial n_{i[a]}^{(k)}}{\partial \hat{z}_{i[a]}^{(k)}}\frac{1}{m}\frac{\partial \hat{z}_{i[a]}^{(k)}}{\partial \mu_{i[a]}^{(k)}} + \sum_{a=0}^{m-1}\Delta_{i[a]}^{(k)}\frac{\partial n_{i[a]}^{(k)}}{\partial \hat{z}_{i[a]}^{(k)}}\frac{2}{m}\frac{\partial \hat{z}_{i[a]}^{(k)}}{\partial \sigma_{i}^{(k)}}\left(z_{i[B]}^{(k)} - \mu_{i}^{(k)}\right)\right)\mathrm{d}z_{i[B]}^{(k)}\end{aligned}$$

$$\mathrm{d}C = \frac{1}{m}\sum_{B=0}^{m-1}\sum_{i=0}^{N^{(k)}-1}\left(\begin{aligned}&\Delta_{i[B]}^{(k)}\frac{\partial n_{i[B]}^{(k)}}{\partial \gamma_{i}^{(k)}}\mathrm{d}\gamma_{i}^{(k)} + \Delta_{i[B]}^{(k)}\frac{\partial n_{i[B]}^{(k)}}{\partial \beta_{i}^{(k)}}\mathrm{d}\beta_{i}^{(k)} \\ &+\left(\Delta_{i[B]}^{(k)}\frac{\partial n_{i[B]}^{(k)}}{\partial \hat{z}_{i[B]}^{(k)}}\frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial z_{i[B]}^{(k)}} + \sum_{a=0}^{m-1}\Delta_{i[a]}^{(k)}\frac{\partial n_{i[a]}^{(k)}}{\partial \hat{z}_{i[a]}^{(k)}}\frac{1}{m}\frac{\partial \hat{z}_{i[a]}^{(k)}}{\partial \mu_{i[a]}^{(k)}} + \sum_{a=0}^{m-1}\Delta_{i[a]}^{(k)}\frac{\partial n_{i[a]}^{(k)}}{\partial \hat{z}_{i[a]}^{(k)}}\frac{2}{m}\frac{\partial \hat{z}_{i[a]}^{(k)}}{\partial \sigma_{i}^{(k)}}\left(z_{i[B]}^{(k)} - \mu_{i}^{(k)}\right)\right)\mathrm{d}z_{i[B]}^{(k)}\end{aligned}\right) \blacksquare$$

La fonction de *batch-normalization*

L'objection de la fonction de *batch-normalization* consiste à créer à partir de valeurs possibles[1] d'agrégations $z_i^{(k)}$ d'une couche k des nouvelles valeurs d'agrégation $\hat{z}_i^{(k)}$ dîtes normalisées avant que celle-ci soient activée par la fonction d'activation à la sortie de la couche k.

Soit les agrégations

$$z_i^{(k)} = \left\{ z_{i[0]}^{(k)}, z_{i[1]}^{(k)}, \ldots, z_{i[B]}^{(k)}, \ldots, z_{i[m-1]}^{(k)} \right\}$$

où $z_{i[B]}^{(k)}$ est l'agrégation du neurone i d'une couche de neurones k à partir d'un vecteur d'entrée de batch B ($B \in [0, m-1]$), alors :

La moyenne des agrégations de la batch	$\mu_i^{(k)} = \frac{1}{m} \sum_{B=0}^{m-1} z_{i[B]}^{(k)}$
La variance des agrégations de la batch	$\sigma_i^{(k)} = \frac{1}{m} \sum_{B=0}^{m-1} \left(z_{i[B]}^{(k)} - \mu_i^{(k)} \right)^2$
L'agrégation normalisée	$\hat{z}_{i[B]}^{(k)} = \frac{z_{i[B]}^{(k)} - \mu_i^{(k)}}{\sqrt{\sigma_i^{(k)} + \varepsilon}}$ où $\varepsilon \to 0^+$
La normalisation	$n_{i[B]}^{(k)} = \gamma_i^{(k)} \hat{z}_{i[B]}^{(k)} + \beta_i^{(k)}$

Cette approche propose l'ajout de deux nouveaux paramètres aux neurones du réseau :

Le changement d'échelle (*scale*)	Le décalage (*shift*)
$\gamma_i^{(k)}$	$\beta_i^{(k)}$

Ainsi, il faudra entraîner ces deux nouveaux paramètres lors de la descente du gradient afin de réduire l'erreur du réseau.

Les dérivées partielles des paramètres de la fonction *batch-normalization*

En construction …

$$\frac{\partial n_{i[B]}^{(k)}}{\partial \hat{z}_{i[B]}^{(k)}} = \frac{\partial}{\partial \hat{z}_{i[B]}^{(k)}} \left(\gamma_i^{(k)} \hat{z}_{i[B]}^{(k)} + \beta_i^{(k)} \right) \quad \Rightarrow \quad \frac{\partial n_{i[B]}^{(k)}}{\partial \hat{z}_{i[B]}^{(k)}} = \gamma_i^{(k)}$$

$$\frac{\partial n_{i[B]}^{(k)}}{\partial \gamma_i^{(k)}} = \frac{\partial}{\partial \gamma_i^{(k)}} \left(\gamma_i^{(k)} \hat{z}_{i[B]}^{(k)} + \beta_i^{(k)} \right) \quad \Rightarrow \quad \frac{\partial n_{i[B]}^{(k)}}{\partial \gamma_i^{(k)}} = \hat{z}_{i[B]}^{(k)}$$

[1] Les valeurs possibles dépendent directement des données admissibles à l'entraînement. Plus les vecteurs d'entraînement sont nombreux et variés, plus cette méthode de normalisation est efficace.

$$\frac{\partial n_{i[B]}^{(k)}}{\partial \beta_i^{(k)}} = \frac{\partial}{\partial \beta_i^{(k)}}\left(\gamma_i^{(k)}\,\hat{z}_{i[B]}^{(k)} + \beta_i^{(k)}\right) \quad \Rightarrow \quad \frac{\partial n_{i[B]}^{(k)}}{\partial \beta_i^{(k)}} = 1$$

$$\frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial z_{i[B]}^{(k)}} = \frac{\partial}{\partial z_{i[B]}^{(k)}}\left(\frac{z_{i[B]}^{(k)} - \mu_i^{(k)}}{\sqrt{\sigma_i^{(k)} + \varepsilon}}\right) \quad \Rightarrow \quad \frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial z_{i[B]}^{(k)}} = \frac{1}{\sqrt{\sigma_i^{(k)} + \varepsilon}}$$

$$\frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial \mu_i^{(k)}} = \frac{\partial}{\partial \mu_i^{(k)}}\left(\frac{z_{i[B]}^{(k)} - \mu_i^{(k)}}{\sqrt{\sigma_i^{(k)} + \varepsilon}}\right) \quad \Rightarrow \quad \frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial \mu_i^{(k)}} = \frac{-1}{\sqrt{\sigma_i^{(k)} + \varepsilon}}$$

$$\frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial \sigma_i^{(k)}} = \frac{\partial}{\partial \sigma_i^{(k)}}\left(\frac{z_{i[B]}^{(k)} - \mu_i^{(k)}}{\sqrt{\sigma_i^{(k)} + \varepsilon}}\right) \quad \Rightarrow \quad \frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial \sigma_i^{(k)}} = -\frac{\left(z_{i[B]}^{(k)} - \mu_i^{(k)}\right)}{\left(\sqrt{\sigma_i^{(k)} + \varepsilon}\right)^2}\frac{\partial\left(\sqrt{\sigma_i^{(k)} + \varepsilon}\right)}{\partial \sigma_i^{(k)}}$$

$$\Rightarrow \quad \frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial \sigma_i^{(k)}} = -\frac{\left(z_{i[B]}^{(k)} - \mu_i^{(k)}\right)}{\left(\sqrt{\sigma_i^{(k)} + \varepsilon}\right)^2}\frac{1}{2\sqrt{\sigma_i^{(k)} + \varepsilon}}$$

$$\Rightarrow \quad \frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial \sigma_i^{(k)}} = -\frac{\left(z_{i[B]}^{(k)} - \mu_i^{(k)}\right)}{2\left(\sigma_i^{(k)} + \varepsilon\right)^{3/2}}$$

$$\frac{\partial \mu_i^{(k)}}{\partial z_{i[B]}^{(k)}} = \frac{\partial}{\partial z_{i[B]}^{(k)}}\left(\frac{1}{m}\sum_{B=0}^{m-1} z_{i[B]}^{(k)}\right) \quad \Rightarrow \quad \frac{\partial \mu_i^{(k)}}{\partial z_{i[B]}^{(k)}} = \frac{1}{m}$$

$$\frac{\partial \sigma_i^{(k)}}{\partial z_{i[B]}^{(k)}} = \frac{\partial}{\partial z_{i[B]}^{(k)}}\left(\frac{1}{m}\sum_{B=0}^{m-1}\left(z_{i[B]}^{(k)} - \mu_i^{(k)}\right)^2\right) \quad \Rightarrow \quad \frac{\partial \sigma_i^{(k)}}{\partial z_{i[B]}^{(k)}} = \frac{2}{m}\left(z_{i[B]}^{(k)} - \mu_i^{(k)}\right)$$

$$\frac{\partial \sigma_i^{(k)}}{\partial \mu_i^{(k)}} = \frac{\partial}{\partial \mu_i^{(k)}}\left(\frac{1}{m}\sum_{B=0}^{m-1}\left(z_{i[B]}^{(k)} - \mu_i^{(k)}\right)^2\right) \quad \Rightarrow \quad \frac{\partial \sigma_i^{(k)}}{\partial \mu_i^{(k)}} = \frac{1}{m}\sum_{B=0}^{m-1} \frac{\partial}{\partial \mu_i^{(k)}}\left(z_{i[B]}^{(k)} - \mu_i^{(k)}\right)^2$$

$$\Rightarrow \quad \frac{\partial \sigma_i^{(k)}}{\partial \mu_i^{(k)}} = \frac{1}{m}\sum_{B=0}^{m-1} 2\left(z_{i[B]}^{(k)} - \mu_i^{(k)}\right)\frac{\partial}{\partial \mu_i^{(k)}}\left(z_{i[B]}^{(k)} - \mu_i^{(k)}\right)$$

$$\Rightarrow \quad \frac{\partial \sigma_i^{(k)}}{\partial \mu_i^{(k)}} = -\frac{2}{m}\sum_{B=0}^{m-1}\left(z_{i[B]}^{(k)} - \mu_i^{(k)}\right)$$

$$\Rightarrow \quad \frac{\partial \sigma_i^{(k)}}{\partial \mu_i^{(k)}} = -\frac{2}{m}\sum_{B=0}^{m-1} z_{i[B]}^{(k)} + \frac{2}{m}\sum_{B=0}^{m-1}\mu_i^{(k)}$$

$$\Rightarrow \quad \frac{\partial \sigma_i^{(k)}}{\partial \mu_i^{(k)}} = -2\mu_i^{(k)} + \frac{2}{m}\left(m\mu_i^{(k)}\right)$$

$$\Rightarrow \quad \frac{\partial \sigma_i^{(k)}}{\partial \mu_i^{(k)}} = 0$$

L'équation de la propagation de l'erreur de la fonction *batch-normalization*

Pour la fonction *batch-normalization* tel que

$$n_{i[B]}^{(k)} = \gamma_i^{(k)}\,\hat{z}_{i[B]}^{(k)} + \beta_i^{(k)} \quad \text{avec} \quad \hat{z}_{i[B]}^{(k)} = \frac{z_{i[B]}^{(k)} - \mu_i^{(k)}}{\sqrt{\sigma_i^{(k)} + \varepsilon}} \quad \text{où } \varepsilon \to 0^+,$$

nous avons les expressions suivantes pour les dérivées partielles de la fonction d'erreur par rapport aux paramètres de *scale* $\gamma_u^{(k)}$ et de *shift* $\beta_u^{(k)}$ ainsi que pour la propagation de l'erreur $\Delta_{u[B]}^{n(k)}$ de la fonction de la fonction $n_{i[B]}^{(k)}$:

Propagation de l'erreur	Gradient du *scale*	Gradient du *shift*
$\Delta_{u[B]}^{n(k)} = \frac{\gamma_i^{(k)}}{m\sqrt{\sigma_u^{(k)} + \varepsilon}}\left(m\Delta_{u[B]}^{(k)} - \sum_{c=0}^{m-1}\Delta_{u[c]}^{(k)} - \hat{z}_{u[B]}^{(k)}\sum_{d=0}^{m-1}\Delta_{u[d]}^{(k)}\hat{z}_{u[d]}^{(k)}\right)$	$\frac{\partial C}{\partial \gamma_u^{(k)}} = \frac{1}{m}\sum_{B=0}^{m-1}\Delta_{u[B]}^{(k)}\,\hat{z}_{u[B]}^{(k)}$	$\frac{\partial C_{[B]}}{\partial \beta_u^{(k)}} = \frac{1}{m}\sum_{B=0}^{m-1}\Delta_{u[B]}^{(k)}$

Preuve :

Débutons avec la différentielle générale d'une fonction de normalisation

$$\mathrm{d}C = \frac{1}{m}\sum_{B=0}^{m-1}\sum_{i=0}^{N^{(k)}-1}\left(\begin{array}{l}\Delta_{i[B]}^{(k)}\frac{\partial n_{i[B]}^{(k)}}{\partial \gamma_i^{(k)}}\mathrm{d}\gamma_i^{(k)} + \Delta_{i[B]}^{(k)}\frac{\partial n_{i[B]}^{(k)}}{\partial \beta_i^{(k)}}\mathrm{d}\beta_i^{(k)} \\ + \left(\Delta_{i[B]}^{(k)}\frac{\partial n_{i[B]}^{(k)}}{\partial \hat{z}_{i[B]}^{(k)}}\frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial z_{i[B]}^{(k)}} + \sum_{a=0}^{m-1}\Delta_{i[a]}^{(k)}\frac{\partial n_{i[a]}^{(k)}}{\partial \hat{z}_{i[a]}^{(k)}}\frac{1}{m}\frac{\partial \hat{z}_{i[a]}^{(k)}}{\partial \mu_{i[a]}^{(k)}} + \sum_{a=0}^{m-1}\Delta_{i[a]}^{(k)}\frac{\partial n_{i[a]}^{(k)}}{\partial \hat{z}_{i[a]}^{(k)}}\frac{2}{m}\frac{\partial \hat{z}_{i[a]}^{(k)}}{\partial \sigma_i^{(k)}}\left(z_{i[B]}^{(k)} - \mu_i^{(k)}\right)\right)\mathrm{d}z_{i[B]}^{(k)}\end{array}\right)$$

Nous pouvons débuter par remplacer $\frac{\partial n_{i[B]}^{(k)}}{\partial \hat{z}_{i[B]}^{(k)}} = \gamma_i^{(k)}$ donnant le même résultat pour l'ensemble des vecteurs de la batch ce qui réduit à ceci après la factorisation du terme $\gamma_i^{(k)}$:

$$\mathrm{d}C = \frac{1}{m}\sum_{B=0}^{m-1}\sum_{i=0}^{N^{(k)}-1}\left(\begin{array}{l}\Delta_{i[B]}^{(k)}\frac{\partial n_{i[B]}^{(k)}}{\partial \gamma_i^{(k)}}\mathrm{d}\gamma_i^{(k)} + \Delta_{i[B]}^{(k)}\frac{\partial n_{i[B]}^{(k)}}{\partial \beta_i^{(k)}}\mathrm{d}\beta_i^{(k)} \\ +\gamma_i^{(k)}\left(\Delta_{i[B]}^{(k)}\frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial z_{i[B]}^{(k)}} + \sum_{a=0}^{m-1}\Delta_{i[a]}^{(k)}\frac{1}{m}\frac{\partial \hat{z}_{i[a]}^{(k)}}{\partial \mu_{i[a]}^{(k)}} + \sum_{a=0}^{m-1}\Delta_{i[a]}^{(k)}\frac{2}{m}\frac{\partial \hat{z}_{i[a]}^{(k)}}{\partial \sigma_i^{(k)}}\left(z_{i[B]}^{(k)} - \mu_i^{(k)}\right)\right)\mathrm{d}z_{i[B]}^{(k)}\end{array}\right)$$

Par la suite, ajoutons les expressions

$$\frac{\partial n_{i[B]}^{(k)}}{\partial \gamma_i^{(k)}} = \hat{z}_{i[B]}^{(k)} \quad \text{et} \quad \frac{\partial n_{i[B]}^{(k)}}{\partial \beta_i^{(k)}} = 1$$

ce qui donnera

$$\mathrm{d}C = \frac{1}{m}\sum_{B=0}^{m-1}\sum_{i=0}^{N^{(k)}-1}\left(\begin{array}{l}\Delta_{i[B]}^{(k)}\hat{z}_{i[B]}^{(k)}\mathrm{d}\gamma_i^{(k)} + \Delta_{i[B]}^{(k)}\mathrm{d}\beta_i^{(k)} \\ +\gamma_i^{(k)}\left(\Delta_{i[B]}^{(k)}\frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial z_{i[B]}^{(k)}} + \sum_{a=0}^{m-1}\Delta_{i[a]}^{(k)}\frac{1}{m}\frac{\partial \hat{z}_{i[a]}^{(k)}}{\partial \mu_{i[a]}^{(k)}} + \sum_{a=0}^{m-1}\Delta_{i[a]}^{(k)}\frac{2}{m}\frac{\partial \hat{z}_{i[a]}^{(k)}}{\partial \sigma_i^{(k)}}\left(z_{i[B]}^{(k)} - \mu_i^{(k)}\right)\right)\mathrm{d}z_{i[B]}^{(k)}\end{array}\right).$$

Ensuite, introduisons les calculs

$$\frac{\partial \hat{z}_{i[B]}^{(k)}}{\partial z_{i[B]}^{(k)}} = \frac{1}{\sqrt{\sigma_i^{(k)} + \varepsilon}} \quad , \quad \frac{\partial \hat{z}_{i[a]}^{(k)}}{\partial \mu_i^{(k)}} = \frac{-1}{\sqrt{\sigma_i^{(k)} + \varepsilon}} \quad \text{et} \quad \frac{\partial \hat{z}_{i[a]}^{(k)}}{\partial \sigma_i^{(k)}} = -\frac{\left(z_{i[a]}^{(k)} - \mu_i^{(k)}\right)}{2\left(\sigma_i^{(k)} + \varepsilon\right)^{3/2}}$$

ce qui donnera le résultat suivant après avoir factorisé le terme $\frac{1}{\sqrt{\sigma_i^{(k)} + \varepsilon}}$:

$$\mathrm{d}C = \frac{1}{m}\sum_{B=0}^{m-1}\sum_{i=0}^{N^{(k)}-1}\left(\begin{array}{l}\Delta_{i[B]}^{(k)}\hat{z}_{i[B]}^{(k)}\mathrm{d}\gamma_i^{(k)} + \Delta_{i[B]}^{(k)}\mathrm{d}\beta_i^{(k)} \\ +\frac{\gamma_i^{(k)}}{\sqrt{\sigma_i^{(k)} + \varepsilon}}\left(\Delta_{i[B]}^{(k)} - \frac{1}{m}\sum_{a=0}^{m-1}\Delta_{i[a]}^{(k)} - \frac{1}{m}\sum_{a=0}^{m-1}\Delta_{i[a]}^{(k)}\frac{\left(z_{i[a]}^{(k)} - \mu_i^{(k)}\right)}{\left(\sigma_i^{(k)} + \varepsilon\right)}\left(z_{i[B]}^{(k)} - \mu_i^{(k)}\right)\right)\mathrm{d}z_{i[B]}^{(k)}\end{array}\right).$$

Nous allons réécrire $\left(\sigma_i^{(k)} + \varepsilon\right) = \sqrt{\sigma_i^{(k)} + \varepsilon}\sqrt{\sigma_i^{(k)} + \varepsilon}$ pour reconstruire des termes

$$\hat{z}_{i[a]}^{(k)} = \frac{z_{i[a]}^{(k)} - \mu_i^{(k)}}{\sqrt{\sigma_i^{(k)} + \varepsilon}} \qquad \text{et} \qquad \hat{z}_{i[B]}^{(k)} = \frac{z_{i[B]}^{(k)} - \mu_i^{(k)}}{\sqrt{\sigma_i^{(k)} + \varepsilon}}$$

et obtenir

$$\mathrm{d}C = \frac{1}{m}\sum_{B=0}^{m-1}\sum_{i=0}^{N^{(k)}-1}\left(\begin{array}{l}\Delta_{i[B]}^{(k)}\hat{z}_{i[B]}^{(k)}\mathrm{d}\gamma_i^{(k)} + \Delta_{i[B]}^{(k)}\mathrm{d}\beta_i^{(k)} \\ +\frac{\gamma_i^{(k)}}{\sqrt{\sigma_i^{(k)} + \varepsilon}}\left(\Delta_{i[B]}^{(k)} - \frac{1}{m}\sum_{a=0}^{m-1}\Delta_{i[a]}^{(k)} - \frac{1}{m}\sum_{a=0}^{m-1}\Delta_{i[a]}^{(k)}\hat{z}_{i[a]}^{(k)}\hat{z}_{i[B]}^{(k)}\right)\mathrm{d}z_{i[B]}^{(k)}\end{array}\right).$$

En forçant la factorisation du facteur 1 / m ainsi que le terme constant $\hat{z}_{i[B]}^{(k)}$ dans la sommation sur l'indice *a*, nous avons

$$\mathrm{d}C = \frac{1}{m}\sum_{B=0}^{m-1}\sum_{i=0}^{N^{(k)}-1}\left(\begin{array}{l}\Delta_{i[B]}^{(k)}\hat{z}_{i[B]}^{(k)}\mathrm{d}\gamma_i^{(k)} + \Delta_{i[B]}^{(k)}\mathrm{d}\beta_i^{(k)} \\ + \dfrac{\gamma_i^{(k)}}{m\sqrt{\sigma_i^{(k)}+\varepsilon}}\left(m\Delta_{i[B]}^{(k)} - \sum_{a=0}^{m-1}\Delta_{i[a]}^{(k)} - \hat{z}_{i[B]}^{(k)}\sum_{a=0}^{m-1}\Delta_{i[a]}^{(k)}\hat{z}_{i[a]}^{(k)}\right)\mathrm{d}z_{i[B]}^{(k)}\end{array}\right) \quad \blacksquare$$

L'activation de la fonction *batch-normalization* avec l'usage d'un seul vecteur d'entrée

En mode « *inference* », l'activation de la fonction *batch-normalization* sera égale à l'expression suivante :

$$n_u^{(k)} = \frac{\gamma_u^{(k)}}{\sqrt{\widetilde{\sigma}_u^{(k)}+\varepsilon}}z_u^{(k)} + \left(\beta_i^{(k)} - \frac{\gamma_u^{(k)}\widetilde{\mu}_u^{(k)}}{\sqrt{\widetilde{\sigma}_u^{(k)}+\varepsilon}}\right) \quad \text{avec} \quad \varepsilon \to 0^+$$

où $\widetilde{\sigma}_u^{(k)}$: Variance moyennée sur plusieurs activations par *batch* précédente.

$\widetilde{\mu}_u^{(k)}$: Moyenne moyennée sur plusieurs activations par *batch* précédente.