

# Normalización por Batches

Jorge Enciso

November 9, 2024

## Definición

Sea  $X \in \mathbb{R}^{b \times i}$  un batch de datos que se descompone en  $m$  mini-batches  $x_i$ .

$$\mu = \frac{1}{m} \sum_{n=1}^m x_n \quad (1)$$

$$\sigma^2 = \frac{1}{m} \sum_{n=1}^m (\mu - x_n)^2 \quad (2)$$

$$\hat{x}_i = \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}} \quad (3)$$

$$y_i = \gamma x_i + \beta \quad (4)$$

## Retropropagación

Para poder implementar la retropropagación a través de la normalización por batches, debemos entender que  $\gamma$  y  $\beta$  son parámetros del modelo que serán aprendidos. Por lo tanto, necesitamos calcular  $\frac{\partial y_i}{\partial \gamma}$  y  $\frac{\partial y_i}{\partial \beta}$ .

Es importante recordar que estamos retropropagando a través de mini-batches, por lo que las derivadas parciales deben ser tratadas con cuidado.

$$y_i = \gamma \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta \quad (5)$$

Debemos acumular las derivadas parciales de todos los mini-batches, ya que esto es fundamental en la retropropagación, similar a cómo acumulamos gradientes cuando trabajamos con batches y reducción por sumas:

$$\frac{\partial L}{\partial \gamma} = \sum_{n=1}^m \frac{\partial L}{\partial y_n} \hat{x}_n \quad (6)$$

$$\frac{\partial L}{\partial \beta} = \sum_{n=1}^m \frac{\partial L}{\partial y_n} \quad (7)$$

A continuación, calculamos la derivada parcial con respecto a los valores de entrada en mini-batches:

$$\frac{\partial L}{\partial x_n} = \frac{\partial L}{\partial y_i} \gamma \frac{\partial \hat{x}_n}{\partial x_n} \quad (8)$$

$$\frac{\partial \hat{x}_n}{\partial x_n} = \frac{1 - \frac{\partial \mu}{\partial x_n}}{\sqrt{\sigma^2 + \epsilon}} - \frac{\frac{\partial \sigma^2}{\partial x_n} (1 + \mu)}{2(\sigma^2 + \epsilon)^{\frac{-3}{2}}} \quad (9)$$

$$\frac{\partial \sigma^2}{\partial x_i} = \frac{1}{m} \left[ \sum_{\substack{n=1 \\ n \neq i}}^m 2(x_n - \mu) \frac{\partial \mu}{\partial x_i} \right] + \frac{2(x_i - \mu) \left( 1 - \frac{\partial \mu}{\partial x_i} \right)}{m} \quad (10)$$

$$\frac{\partial \mu}{\partial x_i} = \frac{1}{m} \quad (11)$$

Si reescribimos estas expresiones podemos llegar a las ecuaciones originales del paper:

$$\frac{\partial L}{\partial x_i} = \frac{\partial L}{\partial \hat{x}_i} \frac{1}{\sqrt{\sigma^2 + \epsilon}} + \frac{\partial L}{\partial \sigma^2} \frac{2(x_i - \mu)}{m} + \frac{\partial L}{\partial \mu} \frac{1}{m} \quad (12)$$