ML Operators

LibML Team

10/1/2023

Contributors

- Rylan W. Yancey

Acknowledgements

- Michael Percy

Introduction

This document formally defines the operators libml implements and their gradients. Tensors are assumed to conform to either the NHCW or NCHW format, where N is batch size, H is features, C is channels, and W is width.

---

1 Add

The Addition Operator is defined as $f(u, v) = u + v$. To find the gradient w.r.t u and v, we will use the sum rule, which is defined as:

$$\frac{\delta}{\delta x}(u + v) = \frac{\delta u}{\delta x} + \frac{\delta v}{\delta x}$$

To find the gradient w.r.t u, we will set u as our x and treat v as a constant, which gives us the following:

$$\frac{\delta}{\delta u}(u + v) = \frac{\delta u}{\delta u} + \frac{\delta v}{\delta u} = 1$$

To find the gradient w.r.t v, we will set v as our x and treat u as a constant, which gives us the following:

$$\frac{\delta}{\delta v}(u + v) = \frac{\delta u}{\delta v} + \frac{\delta v}{\delta v} = 1$$

Therefore, we can say that the gradient w.r.t u is 1, and the gradient w.r.t v is 1.

---

2 Batch Norm

The Batch Normalization Operator is "...a technique to provide any layer in a neural network with inputs that are zero mean/unit variance..". In practice, this helps neural networks converge faster. The Batch Norm function is defined as follows:

$$\overrightarrow{y} \leftarrow \boldsymbol{BN}_{\gamma,\beta}(\overrightarrow{x})$$

Batch Norm is an operation over a vector. If X is an M x N Matrix, and Y is an M x N matrix, $\boldsymbol{BN}$ is an operation over the features of X, which is usually (as in our case), the columns, or the N dimension. The rows, or the M dimension, are the batches. For our purposes, we will treat the $\boldsymbol{BN}$ function as a matrix operation. The γ (gamma) and β (beta) symbols are trainable parameters with size Nx1, one parameter per feature.

2.1 Definition

First, we will define μ, which is an N x 1 vector that is the mean of each feature. In plain english, each element of μ is equal to the sum of feature i in each batch j in X divided by M.

$$\sum_{i=0}^{n} \mu_i = \frac{1}{m} \sum_{j=0}^{m} X_{ij}$$

Next, we will define σ, which is an N x 1 vector that is the variance of each feature.

$$\sum_{i=0}^{n} \sigma_i = \frac{1}{m} \sum_{j=0}^{m} (X_{ij} - \mu_i)^2$$

Next, we will normalize X into $\hat{X}$.

$$\sum_{i=0}^{n} \sum_{j=0}^{m} \hat{X}_{ij} = \frac{X_{ij} - \mu_i}{\sqrt{\sigma_i + \epsilon}}$$

Lastly, we will scale and shift according to γ and β to get Y.

$$\sum_{i=0}^{n} \sum_{j=0}^{m} Y_{ij} = \gamma_i \hat{X_{ij}} + \beta_i$$

2.2 Gradient

I'm not sure how to write this elegantly, so I'm linking a source below.

link to in-depth explanation of batch norm gradient by Frederik Kratzert

---

3 Cos

The Cosine Operator is defined as $f(x) = cos(x)$. It is well known that the gradient of the $cos(x)$ function is $sin(x)$, so that is the gradient we will use here.

---

4 Div

The Division Operator is defined as $f(u, v) = \frac{u}{v}$. To find the gradient w.r.t x and y, we will use the quotient rule of derivatives, which is defined as:

$$\frac{\delta}{\delta x}(\frac{u}{v}) = \frac{(u\frac{\delta}{\delta x})v - u(v\frac{\delta}{\delta x})}{v^2}$$

To find the gradient w.r.t u, we will set u as our x and treat v as constant, which gives us the following:

$$\frac{\delta}{\delta u}(\frac{u}{v}) = \frac{v\frac{\delta u}{\delta u} - u\frac{\delta v}{\delta u}}{v^2} = \frac{v - u\frac{0}{\delta u}}{v^2} = \frac{v}{v^2} = \frac{1}{v}$$

To find the gradient w.r.t v, we will set v as our x and treat u as constant, which gives us the following:

$$\frac{\delta}{\delta v}(\frac{u}{v}) = \frac{v\frac{\delta u}{\delta v} - u\frac{\delta v}{\delta v}}{v^2} = \frac{v\frac{0}{\delta v} - u}{v^2} = -\frac{u}{v^2}$$

Therefore, we can say that the gradient w.r.t u is $\frac{1}{v}$, and the gradient w.r.t v is $-\frac{u}{v^2}$.

---

5 Matmul

When A is an M x N matrix, B is an N x P matrix, and C is an M x P matrix, the product AB for an element in C at some i, j is defined as:

$$C_{ij} = \sum_{k=0}^{N} A_{ik}B_{kj}$$

Visually, this is the vector product of row i in A and column j in B. Lets' begin by finding the derivative of the vector product, defined as. When A is a M x 1 vector, and B is an M x 1 vector, the vector product is defined as:

$$\sum_{i=0}^{M} A_i B_i$$

To find the gradient, we will make use of the sum rule and the product rule.

$$\frac{\delta}{\delta X}\sum_{i=0}^{M}\frac{\delta}{\delta X_i}(B_i(A_i\frac{\delta}{\delta X_i})+A_i(B_i\frac{\delta}{\delta X_i}))$$

To find the gradient w.r.t. A, we will substitute X as A.

$$\frac{\delta}{\delta A}\sum_{i=0}^{M}\frac{\delta}{\delta A_i}(B_i\frac{\delta A_i}{\delta A_i}+A_i\frac{\delta B_i}{\delta A_i})$$

We now distribute $\frac{\delta}{\delta A_i}$ and simplify.

$$\sum_{i=0}^{M}B_i\frac{\delta A_i}{\delta A_i}+\sum_{i=0}^{M}A_i\frac{\delta B_i}{\delta A_i}=\sum_{i=0}^{M}B_i+\sum_{i=0}^{M}0=\sum_{i=0}^{M}B_i$$

Therefore, the gradient of the vector product with respect to A is the sum of the elements of B. The same logic is true for the gradient w.r.t. B, which is the sum of the elements of A. Applying this to the definition of an element of C in matrix multiplication, we can say that the gradient of C_{ij} w.r.t A is the sum of the elements in column B_j, and the gradient w.r.t. B is the sum of the elements in row A_i. For the purposes of gradients, we can conclude that the gradient of C w.r.t. A is B^T, and the gradient of C w.r.t. B. is A^T.

$$\frac{\delta}{\delta A}(AB)=B^T$$

$$\frac{\delta}{\delta B}(AB)=A^T$$

We get B^T and A^T becuase it is convention when working with gradients to transpose the derivatives of matrix multiplication. Under the jacobian convention, this is not so.

---

6 Mul

The Multiplication Operator is defined as $f(u,v)=uv$. To find the gradient w.r.t u and v, we will use the product rule of derivatives, which is defined as:

$$\frac{\delta}{\delta x}(uv)=v\frac{\delta u}{\delta x}+u\frac{\delta v}{\delta x}$$

To find the gradient w.r.t u, we will set u as our x and treat v as a constant, which gives us the following:

$$\frac{\delta}{\delta u}(uv)=v\frac{\delta u}{\delta u}+u\frac{\delta v}{\delta u}=v+u\frac{0}{\delta u}=v$$

To find the gradient w.r.t v, simply do the same, this time setting v as x and treat u as constant.

$$\frac{\delta}{\delta v}(uv) = v\frac{\delta u}{\delta v} + u\frac{\delta u}{\delta u} = v\frac{0}{\delta v} + u = u$$

Therefore, we can say that the gradient w.r.t u is v, and the gradient w.r.t v is u.

---

7 Reduce Mean

The Reduce Mean Operator will remove an axis of an input Tensor X, and calculate the sum of the removed axis. Assuming that the input X is an NxH Matrix and output Y is an 1xH Vector, and we are removing the N-axis, we can define the operation for a single element of Y as:

$$Y_j = \frac{1}{N}\sum_{i=0}^{N} X_{ij}$$

Since summations are linear, we can conclude the gradient w.r.t. any X_{ij} is $\frac{1}{N}$.

---

8 Relu

The Rectified Linear Unit function is defined as follows:

$$\begin{cases} x & x > 0 \\ 0 & otherwise \end{cases}$$

To find the gradient w.r.t. x, we use the gradient of each case. If $x > 0$, the gradient is 1. Otherwise, the gradient is 0.

---

9 Sigmoid

The Sigmoid function is defined as $\sigma(x) = \frac{1}{1+e^{-x}}$ To find the gradient, we will apply the chain rule and simplify.

$$\frac{d}{dx}\sigma(x) = \frac{d}{dx}\left[\frac{1}{1+e^{-x}}\right] \tag{1}$$
$$= \frac{d}{dx}\left(1+\mathrm{e}^{-x}\right)^{-1} \tag{2}$$
$$= -(1+e^{-x})^{-2}(-e^{-x}) \tag{3}$$
$$= \frac{e^{-x}}{(1+e^{-x})^2} \tag{4}$$
$$= \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}} \tag{5}$$
$$= \frac{1}{1+e^{-x}} \cdot \frac{(1+e^{-x})-1}{1+e^{-x}} \tag{6}$$
$$= \frac{1}{1+e^{-x}} \cdot \left(\frac{1+e^{-x}}{1+e^{-x}} - \frac{1}{1+e^{-x}}\right) \tag{7}$$
$$= \frac{1}{1+e^{-x}} \cdot \left(1 - \frac{1}{1+e^{-x}}\right) \tag{8}$$
$$= \sigma(x) \cdot (1-\sigma(x)) \tag{9}$$

Therefore, we say that the gradient of the sigmoid function with respect to x is $\sigma(x) \cdot (1-\sigma(x))$.

---

10 Sin

The Sine Operator is defined as $f(x) = sin(x)$. It is well known that the gradient of the $sin(x)$ function is $cos(x)$, so that is the gradient we will use here.

---

11 Softmax

The Softmax Function is defined as follows:

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=0}^{K} e^{z_j}}$$

Z is some vector of length K.

Not sure how to write this gracefully so for now heres some links:
https://www.mldawn.com/the-derivative-of-softmaxz-function-w-r-t-z/
https://eli.thegreenplace.net/2016/the-softmax-function-and-its-derivative

---

12 Sub

The Subtraction Operator is defined as $f(u, v) = u - v$. To find the gradient w.r.t u and v, we will use the subtraction rule, which is defined as:

$$\frac{\delta}{\delta x}(u - v) = \frac{\delta u}{\delta x} - \frac{\delta v}{\delta x}$$

To find the gradient w.r.t u, we will set u as our x and treat v as a constant, which gives us the following:

$$\frac{\delta}{\delta u}(u - v) = \frac{\delta u}{\delta u} - \frac{\delta v}{\delta u} = 1$$

To find the gradient w.r.t v, we will set v as our x and treat u as a constant, which gives us the following:

$$\frac{\delta}{\delta v}(u - v) = \frac{\delta u}{\delta v} - \frac{\delta v}{\delta v} = -1$$

Therefore, we can say that the gradient w.r.t u is 1, and the gradient w.r.t v is -1.