

# Deep learning: Neural Network from scratch

## *creating a neural network library to solve the mnist dataset*

Adrien PELFRESNE - Alexis VAPAILLE

May 31, 2024

## Contents

<b>1</b>	<b>Abstract</b>	<b>2</b>
<b>2</b>	<b>Motivation</b>	<b>2</b>
<b>3</b>	<b>Sequential neural network</b>	<b>2</b>
3.1	General layer interface . . . . .	2
3.2	Dense layer . . . . .	2
3.3	Activation layer . . . . .	3
<b>4</b>	<b>Cost function</b>	<b>3</b>
<b>5</b>	<b>The learning process</b>	<b>4</b>
5.1	Gradient . . . . .	4
5.2	Minimization . . . . .	4
5.3	backpropagation . . . . .	4
5.4	Dense layer backpropagation . . . . .	4
5.5	Activation layer backpropagation . . . . .	5
<b>6</b>	<b>Optimizers</b>	<b>5</b>
6.1	Adam . . . . .	5
<b>7</b>	<b>Batching</b>	<b>6</b>
<b>8</b>	<b>Initializer</b>	<b>6</b>
<b>9</b>	<b>Convolutional neural network</b>	<b>6</b>
<b>10</b>	<b>Metrics</b>	<b>6</b>
<b>11</b>	<b>Training, Validation, and Test Sets</b>	<b>6</b>
<b>12</b>	<b>Results</b>	<b>6</b>
12.1	Network declaration . . . . .	6
12.2	Vapnik–Chervonenkis tradeoff . . . . .	7
12.3	Multi-layer perceptron . . . . .	7
12.3.1	Overfitting . . . . .	7
12.3.2	Greatfitting . . . . .	8
12.4	Convolutional multi-layer perceptron . . . . .	8
<b>13</b>	<b>Conclusion</b>	<b>8</b>

# 1 Abstract

We have created a deep neural network library from scratch, in rust, without the use of any machine learning or deep learning library.

The goal of this project was to solve the mnist dataset of handwritten digits and to provide a performance comparison between two multi-layer perceptrons, with and without convolution.

Our library lets you create your sequential neural network, with a simple and declarative API and is widely open to extension. This project also includes a drawing mode, where you can draw your digit and see what each model guesses. The project is in a WIP stage, but we plan to improve it further shortly

## 2 Motivation

Our motivation with this *from scratch* approach was to deepen our understanding of neural networks. Because high-level libraries like [Keras](#) and the incredible level of abstraction that comes with it, let us sometimes forget the inner machinery. Nevertheless, this underlying work that those libraries are doing is incredible, and we through it was a shame to be a simple user of those interfaces.

The only (mathematical) helper library that we have used is [ndarray](#) which provides an n-dimensional container for general elements and numerics, as well as providing efficient matrix operations on the GPU.

You can find our project on [GitHub](#), we widely encourage readers to check the source code, as in this report, we won't go into code-specific details and focus more on the abstract explanation.

The workspace is split into multiple crates

1. The neural network library, under the crate [nn\\_lib](#)
2. The mnist example, which is using the *nn\_lib* to define the networks and benchmark the dataset, Under the crate [mnist](#)
3. the GUI code to run an interactive drawing mode is included in the workspace in the app file.
4. this [report](#) source code

## 3 Sequential neural network

The multi-layer perceptron is known as a *Sequential* neural network, meaning that the **output**  $Y_N$  of the layer  $L_N$  is the **input**  $X_{N+1}$  of the next layer  $L_{N+1}$ .

$$Y_N = X_{N+1}$$

and conversely,

$$X_N = Y_{N-1}$$

with subscript  $N$  identifying layers in sequential order. from the input layer, when we pass our data, through the output layer, data flows in the network, passing through different kinds of layers.

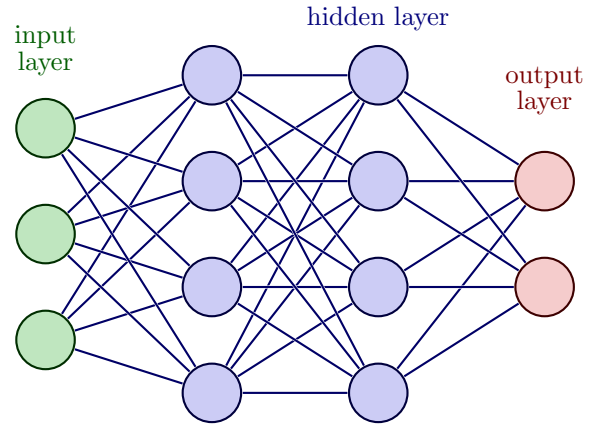


Figure 1: Visual representation of a sequential neural network

### 3.1 General layer interface

We have defined a general *Layer* interface, to be implemented by all the different layers we're going to use.

This interface has two main method : **feed\_forward** and **propagate\_backward**. The *feed\_forward* method takes the input vector  $X \in \mathbb{R}^i$  as a parameter and returns the output vector  $Y \in \mathbb{R}^j$  of the layer. The *propagate\_backward* method take as parameters the layer output gradient  $\frac{\partial C}{\partial Y}$  and return the layer input gradient  $\frac{\partial C}{\partial X}$ . We will cover the use of these functions later, alongside the theoretical explanation of those concepts.

### 3.2 Dense layer

A dense layer, also known as *Linear layer*, is a layer where each input neuron is connected to every output neuron, each connection is called a *weight*.  $w_{ij}$  is the weight connecting neuron  $x_i$  (input) to neuron  $y_j$  (output).

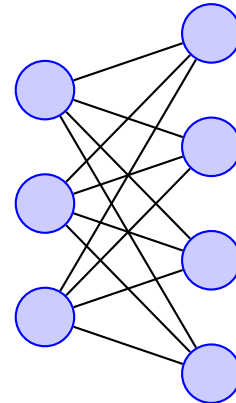


Figure 2: Dense Layer with 3 input nodes and 4 output nodes

We calculate a dense layer neuron output value  $y_j$  as

$$y_j = \sum_{i=1}^n x_i w_{i,j} + b_j$$

which is the bias-corrected weighted sum of all the connected input nodes.  $b_j$  is the bias term for the output

neuron  $j$ . We can rewrite this operation to use matrix operation on whole vectors.

$$Y = XW + B$$

with  $Y \in \mathbb{R}^j$ ,  $X \in \mathbb{R}^i$ ,  $W \in M_{i \times j}(\mathbb{R})$  and  $B \in \mathbb{R}^j$ . In this report, we will annotate vectors :  $X \in \mathbb{R}^N$ , but note that to be able to do the matmul in practice, a vector will be broadcasted to a matrix  $M_{1 \times i}(\mathbb{R})$

### 3.3 Activation layer

Activation functions introduce non-linearity into the neural network. Without them, the network would be a series of linear transformations, regardless of the number of layers, making it equivalent to a single-layer model. Non-linear activation functions allow the network to learn to separate data in a non-linear way

There are two schools of thought for defining activations.

1. activations are used to produce the **output value inside** a dense layer, in this case, the weighted sum of the dense layer is called  $z$ , and  $y$  is calculated by applying an activation  $\sigma$  to  $z$  :  $y : \sigma(z)$ .
2. activations are used **outside** the dense layer in this case, the output of a dense layer  $y$  is equal to the weighted sum  $z$ . An activation layer is placed just after the dense layer, applying the activation function to each neuron.

We implemented our sequential neural network with the second paradigm, because it is easier in code to manage a dense and an activation layer separately, mainly for backpropagation calculation ease.

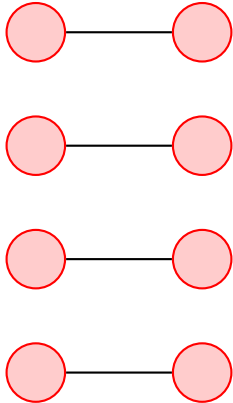


Figure 3: Activation layer with 4 input neuron

The Activation layer takes a vector  $X \in \mathbb{R}^n$  as input and produces a vector  $Y \in \mathbb{R}^n$  as output, with the activation function applied to each neuron.

We must note that some activations function transform a vector by mapping the vector individual components  $x$  through the function, like ReLU

$$ReLU(x) = \max(0, x)$$

While other activation functions, like *softmax* work on whole vectors. Softmax takes a vector  $z \in \mathbb{R}^K$ , and converts it into a probability distribution of  $K$  possible outcomes.

$$\sigma : \mathbb{R}^K \rightarrow (0, 1)^K, K \geq 1$$

is computed by taking a vector  $z = (z_1, \dots, z_k) \in \mathbb{R}^K$  and compute each component of a vector  $\sigma(z) \in (0, 1)^K$

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}}$$

Current activations functions implemented in our library are :

- Sigmoid
- ReLU
- Tanh
- Softmax

## 4 Cost function

To measure how well a neural network performs, we use a cost function, which outputs a real number by comparing the neural network *output*, and the *observed* value. For the mnist dataset, the *output* value is what the neural network guesses for the image we feed him with, and the *observed* value is the real label for that image (the digit the image is).

A simple cost function is the mean-squared error :

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2$$

The MSE function is great for regression neural networks when we want to predict a continuous value from an input.

To do classification, like in the mnist data of handwritten digits, we use another cost function : the cross entropy cost function

$$\text{Cross Entropy} = - \sum_{c=1}^C y_c \log(\hat{y}_c)$$

with  $C$  the number of classes (e.g, 10 for the mnist dataset),  $y_c$  the observed probability, and  $\hat{y}_c$  the neural network output.

This expression simplifies to

$$\text{Cross Entropy} = - \log(\hat{y}_c)$$

if  $y_c$  is one hot encoded (i.e, 1 for the observed class and 0 for others).

## 5 The learning process

### 5.1 Gradient

The gradient of a function  $\nabla f$  is the vector field that represents the direction and rate of the quickest increase of the function from a point  $p$ .

We can visualize the gradient and the curve of a function  $f(x, y) = -(\cos^2 x + \cos^2 y)^2$

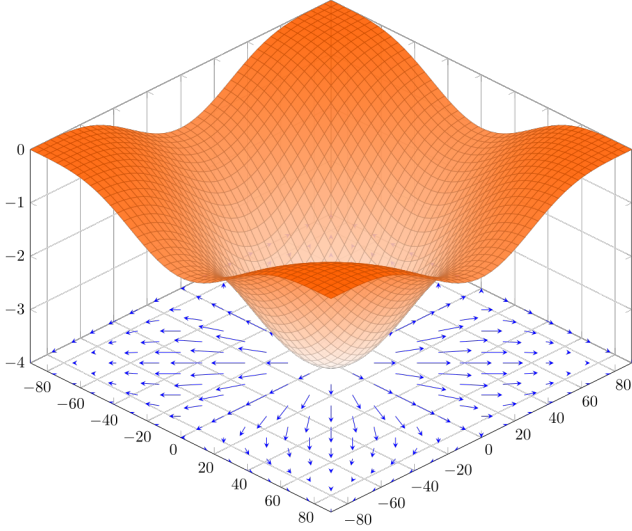


Figure 4: Function representation and gradient field of the function  $-(\cos^2 x + \cos^2 y)^2$

The field vector below the function curve represents the gradient, for each point  $p = (x, y)$  the arrow at  $p$  point toward the direction with the fastest increase of  $f$ . This is useful in optimization problems when we want to find the maximum or the minimum of the function, because it indicates to us which direction to follow at a certain point, to maximize, or minimize the value of the function.

### 5.2 Minimization

To make our neural network lean, we adjust its parameters (i.e, weights and biases for a dense layer) to **minimize the cost function  $C$  of the network**. We can use the cost function gradient, which tells us the direction and magnitude of the fastest **increase** of cost, and proceed to go in the opposite direction, which will make us go toward the fastest **decrease** of the cost function. The algorithm used to minimize the cost function is the *gradient descent*, which will repeatedly descend the curve by going toward the opposite of the gradient of the cost function.

Here is the process for a single gradient descent step :

1. feeds all input data points into the network, calculating the cost for every output.
2. calculates the average cost.

3. computes the gradient of the averaged cost function **with respect to the network parameters**.
4. update the parameters in the opposite direction of the gradient

### 5.3 backpropagation

To do a gradient descent step, we need the gradient of the cost function with respect to the neural network parameters  $\frac{\partial C}{\partial P}$  the algorithm used to calculate this gradient is called **backpropagation**.

Backpropagation precisely calculate (and not only estimate like the finite difference method) the gradient of the cost function  $C$  with respect to the neural network parameters, which we will call parameters gradient  $\frac{\partial C}{\partial P}$ . The key mathematical tool used in backpropagation is the chain rule, which allows the gradient of the cost function to be split into a gradient of simpler functions at each neuron within the network.

For each layer,  $L_n$ , to calculate the parameter's gradient  $\frac{\partial C}{\partial P_n}$ , we need to know the output gradient  $\frac{\partial C}{\partial Y_n}$ , this gradient is given to us by the next layer in the sequential order, because in the sequential architecture  $\frac{\partial C}{\partial X_{n+1}} = \frac{\partial C}{\partial Y_n}$ .

We want to also calculate  $\frac{\partial C}{\partial X_i}$ , the input  $X$  gradient, to transmit it to the previous layer, which needs it to calculate her own parameters gradient  $\frac{\partial C}{\partial P_{i-1}}$ .

You can probably feel why this process is called backpropagation, we are sharing gradients, from layer to layer, starting at the last layer and going back up to the start of the net.

Each layer will have her implementation of **propagate backward method**, in the next sections, we will cover the calculation for the layer we used in our library.

### 5.4 Dense layer backpropagation

The dense layer has two trainable parameters, the weights  $W$ , and the biases  $B$ . Starting with the gradient of the cost function with respect to the weights, we are given the output gradient :

$$\frac{\partial C}{\partial Y} = \begin{bmatrix} \frac{\partial C}{\partial y_1} \\ \frac{\partial C}{\partial y_2} \\ \vdots \\ \frac{\partial C}{\partial y_j} \end{bmatrix} \quad (1)$$

The cost function  $C$  depends on  $Y$  which depend on  $W$  thus  $C$  depend on  $W$  (chain rule)

$$\frac{\partial C}{\partial W} = \frac{\partial C}{\partial Y} \frac{\partial Y}{\partial W}$$

We calculate each weight gradient component  $\frac{\partial C}{\partial W_{ij}}$

$$\frac{\partial C}{\partial W_{ij}} = \sum_{l=1}^N \frac{\partial L}{\partial Y_{lj}} \frac{\partial Y_{lj}}{\partial W_{ij}}$$

because  $W_{ij}$  is used in every example for calculating the  $j^{th}$  column of the output matrix. Let's look at the formula for  $Y_{lj}$  to get the derivative

$$Y_{lj} = X_{l1}W_{1j} + \dots + X_{li}W_{ij} + \dots + X_{lp}W_{pj} + b_j$$

$$\begin{aligned} \frac{\partial Y_{lj}}{\partial W_{ij}} &= X_{li} \\ \Rightarrow \frac{\partial C}{\partial W_{ij}} &= \sum_{l=1}^N \frac{\partial C}{\partial Y_{lj}} X_{li} = \sum_{l=1}^N X_{il}^T \frac{\partial C}{\partial Y_{lj}} \end{aligned}$$

So finally

$$\frac{\partial C}{\partial W} = X^T \frac{\partial C}{\partial Y}$$

for the biases  $B$ .

$$\frac{\partial C}{\partial b} = \frac{\partial C}{\partial Y} \frac{\partial Y}{\partial b} \quad (2)$$

$$\frac{\partial C}{\partial b_i} = \sum_{l=1}^N \frac{\partial C}{\partial Y_{li}} \frac{\partial Y_{li}}{\partial b_i} \quad (3)$$

since the bias term  $b_i$  is used in the evaluation of the entire column of  $Y$ .

$$\frac{\partial Y_{li}}{\partial b_i} = 1$$

$$\frac{\partial C}{\partial b_i} = \sum_{l=1}^N \frac{\partial C}{\partial Y_{li}}$$

$$\frac{\partial C}{\partial b} = \frac{\partial C}{\partial Y}$$

and lastly, we calculate the gradient with respect to  $X$

$$\frac{\partial C}{\partial X} = \frac{\partial C}{\partial Y} \frac{\partial Y}{\partial X} \quad (4)$$

$$\frac{\partial C}{\partial X_{ij}} = \sum_{l=1}^k \frac{\partial C}{\partial Y_{il}} \frac{\partial Y_{il}}{\partial X_{ij}} \quad (5)$$

since the data point  $i$  will only influence the data point  $i$  in  $Y$ . Other data points will not be affected. Further,  $X_{ij}$  is used in the calculation of every dimension of  $Y_{i,:}$ . To calculate the gradient,

$$Y_{il} = \sum_{t=1}^p X_{it} W_{tl}$$

$$\frac{\partial Y_{il}}{\partial X_{ij}} = W_{jl}$$

$$\frac{\partial C}{\partial X_{ij}} = \sum_{l=1}^k \frac{\partial C}{\partial Y_{il}} W_{jl} = \sum_{l=1}^k \frac{\partial C}{\partial Y_{il}} W_{lj}^T$$

$$\Rightarrow \frac{\partial C}{\partial X} = \frac{\partial C}{\partial Y} W^T$$

## 5.5 Activation layer backpropagation

An activation layer doesn't have any trainable parameters. We just need to calculate the input gradient. We are given the output gradient  $\frac{\partial C}{\partial Y}$ , we know that

$$Y = \begin{bmatrix} f(X_1) \\ f(X_2) \\ \vdots \\ f(X_i) \end{bmatrix}$$

applying the chain rule like before,

$$\frac{\partial C}{\partial X} = \frac{\partial C}{\partial Y} \frac{\partial Y}{\partial X}$$

calculating each component

$$\frac{\partial C}{\partial X_{ij}} = \sum_{l=1}^N \frac{\partial C}{\partial Y_{lj}} \frac{\partial Y_{lj}}{\partial X_{ij}}$$

using the same procedure as for the previous calculations

$$\Rightarrow \frac{\partial C}{\partial X} = \frac{\partial C}{\partial Y} \odot Y'$$

with

$$Y' = \begin{bmatrix} f'(X_1) \\ f'(X_2) \\ \vdots \\ f'(X_i) \end{bmatrix}$$

## 6 Optimizers

To update a neural network parameter  $\theta$ , we use the gradient descent formula, with a fixed learning rate  $\eta$

$$\theta := \theta - \eta \frac{\partial C}{\partial \theta}$$

Updating the neural network layers parameters with this method is known, in the Keras API, as the *Stochastic gradient descent optimizer*, but other ways of updating layer parameters exist, we call those alternative methods **optimizers**. One of the best ones is Adam [5]

### 6.1 Adam

Adam accelerates the convergence towards the optimal set of parameters by taking into account past gradient, it will also adapt the learning rate for each parameter.

ADAM initializes two vectors,  $m$  and  $v$ , which store the exponential moving averages of past gradients and past squared gradients, respectively. These vectors are used to scale the gradient updates adaptively.

1. **Initialization:** Initialize two vectors,  $m$  and  $v$ , to store the moving averages of the gradients and their squares, respectively.
2. **Computing Moving Averages of the Gradients:**

$$\begin{aligned} m_t &= \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t, \\ v_t &= \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2, \end{aligned}$$

where  $g_t$  is the gradient at time step  $t$ , and  $\beta_1$  and  $\beta_2$  are factors that control the decay rates.

3. **Bias Correction:** Correct initial bias in  $m$  and  $v$ :

$$\begin{aligned} \hat{m}_t &= \frac{m_t}{1 - \beta_1^t}, \\ \hat{v}_t &= \frac{v_t}{1 - \beta_2^t}. \end{aligned}$$

4. **Update Weights:** Adjust weights based on the corrected moments:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon} \cdot \hat{m}_t,$$

where  $\theta$  are the parameters,  $\eta$  is the learning rate, and  $\epsilon$  is a small constant for numerical stability.

## 7 Batching

We went through several feed-forward implementations in our code. The first one was to pass data points  $x \in \mathbb{R}^i$  one by one, in the network. This was very slow, for example, for the mnist dataset, which has 60k training examples, we were going through the whole neural network: feed forward, cost evaluation, and back-propagation 60k times.

We then learned that we can adapt our code to pass a *batch* of data to the neural network, this batch can be of any length. The input then becomes,  $x \in M_{n \times i}(\mathbb{R})$  with each line of the input matrix representing a single data point. The calculation became much faster because we can now feed  $n$  element at a time in the neural network.

## 8 Initializer

Initializers determine the starting weights of a neural network, it is important to choose the right initializer, depending on our network architecture, to avoid problems like vanishing or exploding gradients, and to optimize convergence. At the moment, we are writing this report, we have implemented two initializers inside our library,

1. the *He* initializer, designed for *ReLU* activation functions, initializes weights from a normal distribution centered at zero with a standard deviation of  $\sqrt{\frac{2}{n}}$  (where  $n$  is the number of input nodes). This setup helps prevent the vanishing gradient problem in deep networks with *ReLU* activations, ensuring that gradients remain large enough to sustain effective learning through many layers.
2. Conversely, the *Glorot Uniform* initializer, also known as Xavier Uniform, is ideal for networks using *sigmoid* or *Tanh* activations. It selects weights from a uniform distribution within  $[-c, c]$ , where  $c = \sqrt{\frac{6}{n_{in} + n_{out}}}$ . This approach maintains a balance in the variance of neurons' outputs across the network, aiding in stable gradient flow in deep networks with saturating activations.

## 9 Convolutional neural network

ALEXIS a toi de travailler

## 10 Metrics

The mnist dataset, is a classification problem, we can implement multiple metrics to evaluate the network per-

formances, we will explain the one we implemented in our library.

1. Accuracy measures the proportion of correct prediction to the total observations in the dataset. for example, if we have correctly classified 950 images out of 1000, the accuracy will be 95 percent.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

Precision and Recall are two metrics acting on a specified class. **Retrieved elements** are all the network predictions for a specific class. **Relevant elements** are all the correct instances of the class [2].

2. Precision answer the question : *How many retrieved items are relevant*

$$\text{Precision} = \frac{\text{Relevant retrieved instance}}{\text{All retrieved instance}}$$

3. Recall answers the question: *How many relevant items are retrieved*

$$\text{Recall} = \frac{\text{Relevant instances}}{\text{All relevant instance}}$$

Those metrics are important to evaluate the network performances, in our implementations and results analysis, we focused on cost and accuracy, but recall, precision, and other classification metrics can be useful to know on which specific digits the neural network is making mistakes.

## 11 Training, Validation, and Test Sets

To monitor the performances of a neural network effectively, it is important to break down the dataset into three distinct subsets: training, validation, and test sets.

- **Training set:** Is used during the learning phase, with training data directly influencing the parameter updates
- **Validation set:** Is used to provide an unbiased evaluation of a model fit on the training dataset while tuning the model's hyperparameters. This set is used during training, to prevent overfitting when the model performs well on the training data but poorly on unseen data.
- **Test set:** Is used only after the model's training and validation phases are complete. It provides the final, unbiased performance metric of the neural network.

## 12 Results

### 12.1 Network declaration

We include a small code snippet, showing our custom library API to declare a sequential neural network. You



can find the API for a sequential network declaration inside the [sequential module](#), and the declaration of the mnist network in the [network declaration module](#)

```
let net = NeuralNetworkBuilder::new()
  .push(DenseLayer::new(
    784,
    32,
    InitializerType::He,
  ))
  .push(
    ActivationLayer::from(Activation::ReLU)
  )
  .push(
    DenseLayer::new(32, 10,
      ↪ InitializerType::He)
  )
  .push(
    ActivationLayer::from(
      Activation::Softmax
    )
  );
.net.watch(MetricsType::Accuracy)
Ok(net.compile(GradientDescent::new(0.01),
  ↪ CostFunction::CrossEntropy)?)
```

The layers are added in sequential order with the push method. the *watch* method will add a *MetricsType* to the list of metrics that the program needs to calculate, in the example we watched the accuracy metrics. those metrics are given by epochs for the training and validation dataset. and at the end of the test dataset.

## 12.2 Vapnik–Chervonenkis tradeoff

To create a great network, we will try to reach the *VC trade-off principle* [7] [9], this principle states that:

1. A more complex model might fit the training data very well (low training error) but risks overfitting, meaning it could perform poorly on new data because it is too tailored to the training set (high variance).
2. A less complex model might not fit the training data as closely (higher training error) but can generalize better on new data because it does not capture the noise and specific details of the training set (low variance).

Our challenge will thus be, to find the right level of complexity that minimizes both the error on the training data.

## 12.3 Multi-layer perceptron

### 12.3.1 Overfitting

We will analyze the result of an overfitting model, to see how this behavior translates into metrics. To make a model overfit, we have multiple options

- increase the model complexity (number of hidden layers, number of neurons per layer)
- reduce the batch size to make the learning process very granular and increase the chance of the network learning too many specific patterns

In our case we have chosen to lower the batch size, here are our network hyperparameters

hyperparameters	epochs	batch size
Value	30	5

Table 1: hyperparameters overfitting model

Here is the cost comparison, over 30 epochs between the training cost, and the validation cost.

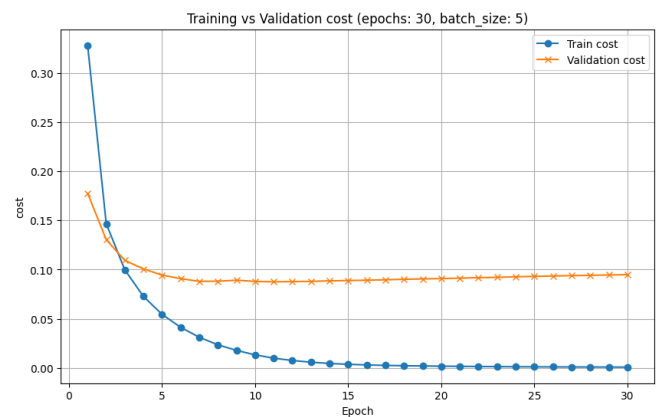


Figure 5: Training vs. Validation cost overfitting model

The overfitting stands out, our train cost gets very low, while the validation cost remains stable this means that our neural network has learned too many specific patterns in the training data, making the error go very low on this set, but on the validation set, the network doesn't generalize well

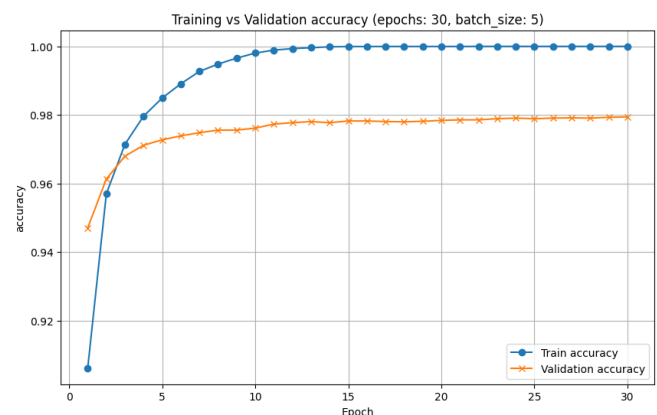


Figure 6: Training vs. Validation cost overfitting model

The same observation applies to the evolution of accuracy over epochs. Accuracy reached a stunning 100

percent on our training data, meaning that it doesn't make any mistakes, but the validation accuracy struggles to get over 98 percent.

### 12.3.2 Greatfitting

We have achieved our best generalization with the following hyperparameters and network definitions.

hyperparameters	epochs	batch size
Value	15	100

Table 2: hyperparameters great fitting model

```
let net = SequentialBuilder::new()
    .push(
        DenseLayer::new(784, 256,
            InitializerType::He))
    .push(
        ActivationLayer::from(
            Activation::ReLU)
    )
    .push(
        DenseLayer::new(256, 256,
            InitializerType::He))
    .push(
        ActivationLayer::from(
            Activation::ReLU)
    )
    .push(
        DenseLayer::new(256, 10,
            InitializerType::GlorotUniform))
    .push(
        ActivationLayer::from
        (
            Activation::Softmax)
        )
    .watch(MetricsType::Accuracy);
Ok(net.compile(GradientDescent::new(0.04),
↪ CostFunction::CrossEntropy?))
```

Here is the comparison between the training cost and the validation cost :

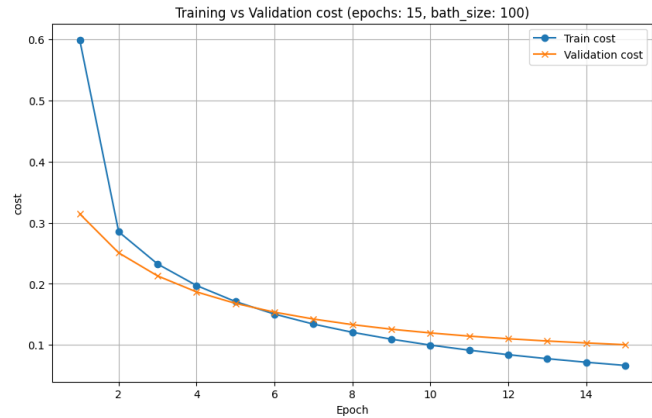


Figure 7: Training vs. Validation cost great fitting model

Here, our model seems to have generalized patterns, because the training cost and the validation cost follows the same path, slowly increasing, with the gap between curves never blowing off the roof. We reach a validation cost of 0.097, and the cost for the testing data set is 0.091.

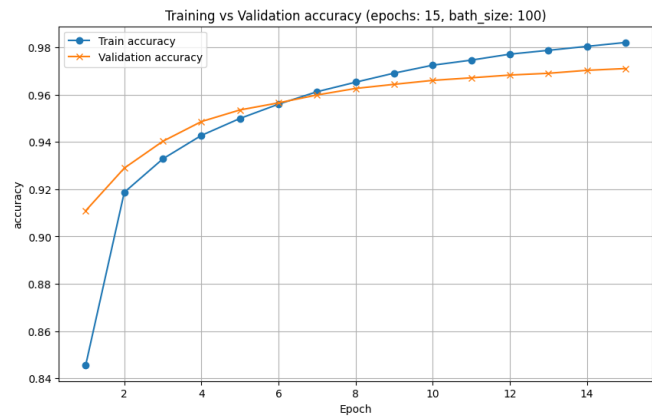


Figure 8: Training vs. Validation cost great fitting model

We reach a validation accuracy of 97.10 percent for the final epochs, and the accuracy for the testing data set is 97.31 percent. which is not a *state-of-the-art* accuracy, but is a good result.

## 12.4 Convolutional multi-layer perceptron

Alexou fais les resultats !!

## 13 Conclusion

This project of creating our own *from scratch* library has brought us a lot of knowledge, It has made us greatly understand the whole neural network process. We have planned to add more functionality to our library and optimize the code base even after the end of this project. We want to add more optimizers, the dropout layer, and a learning rate scheduler. and over the longer term, we'd



like to add other neural network paradigms, such as Recurrent neural networks, and use our library for our experiments with deep learning!

## References

- [1] Papers with Code. *He Initialization*. n.d. URL: <https://paperswithcode.com/method/he-initialization>.
- [2] Wikipedia contributors. *Precision and recall* — *Wikipedia, The Free Encyclopedia*. [Online; accessed 28-May-2024]. 2024. URL: [https://en.wikipedia.org/wiki/Precision\\_and\\_recall](https://en.wikipedia.org/wiki/Precision_and_recall).
- [3] Stack Exchange. *Backpropagation with Softmax Cross Entropy*. 2016. URL: <https://stats.stackexchange.com/questions/235528/backpropagation-with-softmax-cross-entropy>.
- [4] Google. *Cross Entropy*. 2024. URL: <https://www.google.com/search?q=cross+entropy&sourceid=chrome&ie=UTF-8>.
- [5] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: 1412.6980 [cs.LG].
- [6] V7 Labs. *A Guide to Cross-Entropy Loss*. n.d. URL: <https://www.v7labs.com/blog/cross-entropy-loss-guide>.
- [7] Yann LeCun. *Quand la machine apprend: La révolution des neurones artificiels et de l'apprentissage profond*. Odile Jacob, 2019.
- [8] Towards Data Science. *Derivative of the Softmax Function and the Categorical Cross Entropy Loss*. 2019. URL: <https://towardsdatascience.com/derivative-of-the-softmax-function-and-the-categorical-cross-entropy-loss-ffceefc081d1>.
- [9] Vladimir Vapnik and Alexey Chervonenkis. “Theory of pattern recognition”. In: (1974).
- [10] Wikipédia. *Matrice Jacobienne*. 2024. URL: [https://fr.wikipedia.org/wiki/Matrice\\_jacobienne](https://fr.wikipedia.org/wiki/Matrice_jacobienne).
- [11] Wikipedia. *Hadamard product (matrices)*. 2024. URL: [https://en.wikipedia.org/wiki/Hadamard\\_product\\_\(matrices\)](https://en.wikipedia.org/wiki/Hadamard_product_(matrices)).
- [12] Wikipedia. *One-hot*. 2024. URL: <https://en.wikipedia.org/wiki/One-hot>.
- [13] Wikipedia. *Precision and recall*. 2024. URL: [https://en.wikipedia.org/wiki/Precision\\_and\\_recall](https://en.wikipedia.org/wiki/Precision_and_recall).
- [14] Wikipedia. *Softmax function*. 2024. URL: [https://en.wikipedia.org/wiki/Softmax\\_function](https://en.wikipedia.org/wiki/Softmax_function).