University of Applied Sciences Aachen
Campus Jülich

Faculty: Medical Engineering and Technomathematics
Course of Study: Scientific Programming

Autonomous Fault Detection Using Artificial Intelligence Applied to CLAS12 Drift Chamber Data

A Bachelor's Thesis by Christian Peters

Jülich, July 17, 2018

Contents

1 Introduction

2 The CLAS12 Particle Detector

3 Deep Learning Fundamentals

3.1 Artificial Neural Networks

Artificial Neural Networks (ANNs) are a class of machine learning algorithms that are loosely inspired by the structure of biological nervous systems [Hay08]. To be precise, each ANN consists of a collection of artificial neurons that are connected with each other, enabling them to exchange signals. The structure of an ANN can be described by a directed graph, i.e. a collection of nodes as well as directed edges. The nodes of the graph represent the neurons, the edges denote their connections. A common way to arrange artificial neurons within a network is to organize them in layers as depicted in Fig. 3.1.

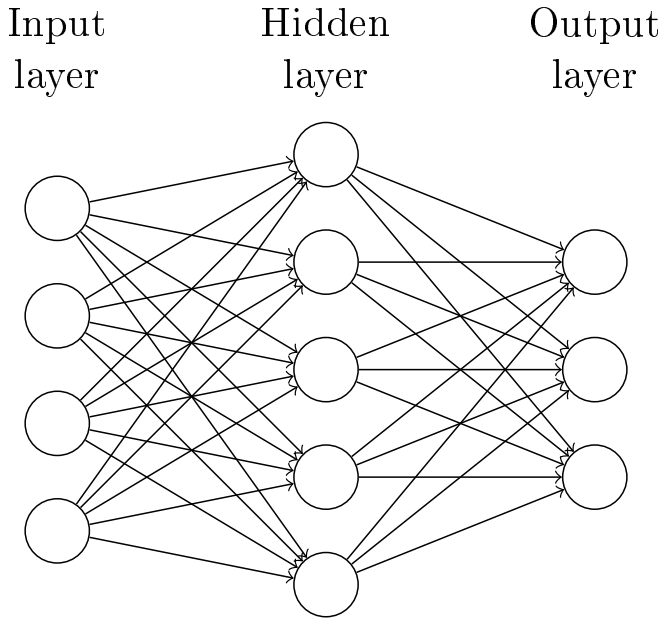

Figure 3.1: A common ANN-structure consisting of three layers of neurons represented by a directed graph.

When an artificial neuron receives a signal from its incoming connections, it may elect

to become active based on the input it collects, see section 3.1.1. In this state, it also influences all neurons it has an outgoing connection to by passing a signal along their channel. These other neurons in turn may also elect to become active, this way a signal can propagate through the network along the connecting edges as seen in Fig. 3.1.

Usually, each ANN consists of at least one layer of neurons that is responsible for receiving signals from the environment, this special type of layer is called an *input layer*. When the neurons in this layer receive a signal, they propagate it to their connected neighbors in the next layer. This process repeats until the *output layer* is reached. The neurons in the output layer represent the result of the whole network. Each layer in between the input and the output layer is called a *hidden layer* because there is no direct communication between the neurons in such a layer and the environment. Networks that satisfy this basic architectural model, where each layer is fully connected with its following layer and signals only flow in one direction without cycles, are called *fully connected feedforward networks*.

The goal behind artificial neural networks is to convert an input signal into a meaningful output by feeding it through the network. If the network is able to detect relevant features or patterns in the input signal, it can be used to perform tasks such as classification or regression, i.e. approximate discrete or continuous functions. In order for this to be possible, some kind of learning has to take place which enables the network to capture the the patterns present in the data it is confronted with. We will take a further look at these aspects as well as the mathematical model of a neural network in the following sections.

3.1.1 Modeling Artificial Neurons

To fully understand how each neuron processes the signals it receives, it is necessary to develop a mathematical model that describes all the operations taking place. The following descriptions are partially based on the explanations that are provided in [Hay08].[1] As shown in Fig. 3.2, the neural model basically consists of three components:

1. **A set of weighted inputs:** Each connection that is leading into a neuron has a weight w_{kj} associated with it where k denotes the neuron in question and j denotes the index of the neuron that delivers its input to the current neuron k. The signal that passes the connection is multiplied by the related weight of that connection before arriving at the next component. There might arise the question why the indexing of a weight from neuron j to neuron k is w_{kj} and *not* w_{jk}. This is the case

[1] See chapter I.3: *Models of a Neuron* for more details.

because in practical implementations of neural networks, the weights are usually stored in matrices where each row corresponds to a neuron k and each column corresponds to an input j which allows for much faster computations by heavily utilizing matrix-multiplication.

2. **A summation unit:** This component adds up all the weighted signals that arrive at the neuron as well as a constant bias value b_k that is independent of the inputs. The reason for adding the bias term is explained in section 3.1.3.

3. **An activation function:** The activation function $\phi(\cdot)$ applies a transformation to the output of the summation unit that is usually non-linear. The value of the activation function is the output of the neuron which will travel further through the network alongside the corresponding connections. In section 3.1.2, a more detailed explanation of activation functions as well as some commonly used examples will be provided.

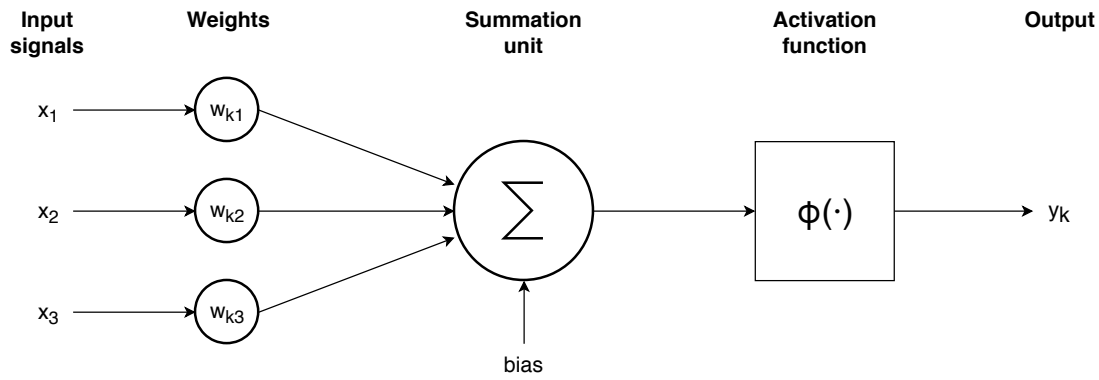

Figure 3.2: The components of the neural model. This neuron labeled k receives three input signals that are first multiplied by the associated weights, summed up including a bias and then fed into an activation function that will determine the ouput signal.

Transforming this model into mathematical equations, the output of the summation unit of a particular neuron k with n input signals x_j can be described by the following formula:

$$z_k = \sum_{j=1}^{n} x_j \cdot w_{kj} + b_k \tag{3.1}$$

where b_k denotes the bias term of neuron k and z_k describes the result of the summation

unit.

As a consequence, the output signal y_k of neuron k can be computed by applying the activation function $\phi(\cdot)$ to the output of the summation unit which can be described by the following expression:

$$y_k = \phi(z_k) \tag{3.2}$$

As we shall see in section 3.2.3, the entire knowledge of the network is stored within the weights as well as the biases. Adjusting this configuration in order to better match the data that the network receives is the main goal of training, see section 3.2.3.

3.1.2 Activation Functions

The basic task of an activation function is to determine the level of activity that a neuron emits based on the input it receives. Because the incoming signals are first weighted and summed up by the summation unit, they arrive at the activation function as a single value z_k. Since the output y_k of neuron k is also a scalar, each activation function can be described as $\phi : \mathbb{R} \to \mathbb{R}$. A very important property of an activation function is *nonlinearity*. This is due to the fact that chaining together multiple linear functions, as would be the case if each artificial neuron had a linear activation function, collapses into just a single linear transformation, which would make it impossible for the network to learn any concepts beyond simple linear relationships. In the following paragraphs, an overview of the most popular activation functions will be presented that is based on the descriptions found in [PG17].[2]

The Sigmoid Function This activation function transforms an input z into a range between 0 and 1 based on the following equation:

$$\phi(z) = \frac{1}{1 + e^{-\theta \cdot z}} \tag{3.3}$$

The θ parameter is used to adjust the sensitivity of the sigmoid function with respect to its input signal. High values of θ lead to steep slopes around $z = 0$ while smaller values will lead to smoother slopes. An illustration of this relationship is presented in Fig. 3.3. One important reason why the sigmoid function was often used at the time neural networks were first developed, is that it reduces the impact of outliers in the data without removing them. When the input of a neuron is large, it is reduced to a number near one, when it is very negative, the activation evaluates to a number near zero. This

[2] See section *Activation Functions* in chapter two.

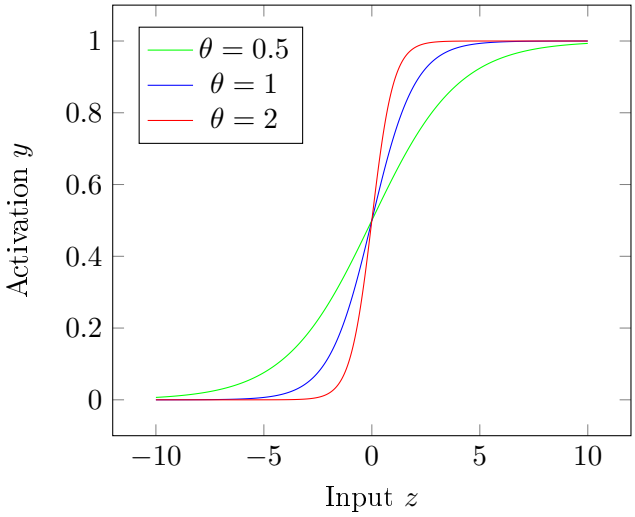

Figure 3.3: The sigmoid activation function plotted for different values of θ.

behaviour adds to the overall robustness of the network.

In the early days of neural networks, people were also seeking for biological inspirations when constructing ANNs. The graph of the sigmoid function can also be interpreted as the firing rate of a biological neuron that saturates for big inputs, which contributed to its popularity in the past. However, caution is advised when putting too much weight onto these interpretations, since real neurons are much more complex in practice than simple mathematical equations.

The Rectified Linear Unit (ReLU) Because it is not always desirable to reduce large signals to a smaller scale, the ReLU function will only replace negative values with zero and leave positive values untouched. This behaviour can be modeled by the following expression:

$$\phi(z) = \max(0, z) \tag{3.4}$$

When building deep neural networks, one of the problems that sometimes arise is that a signal will fade out when propagating through many hidden layers. This issue is remedied to some degree by using the ReLU function because big signals are not cut down. The fact that all negative values are set to zero when the ReLU function is used leads to sparsity among the neuron activations which promotes simpler and possibly richer representations. This analogy can also be found in real biological neurons, which makes it an interesting thing to note that ReLUs are actually more biologically inspired than the sigmoid function [GBB11]. Another benefit of the ReLU function is that its

derivative is either 1 or 0, which makes it very simple to compute. This will turn out to be important when looking into the training of neural networks. Because of all these advantages, ReLUs are one of the state-of-the-art activation functions in deep neural networks. A plot of the ReLU function is presented in Fig. 3.4.

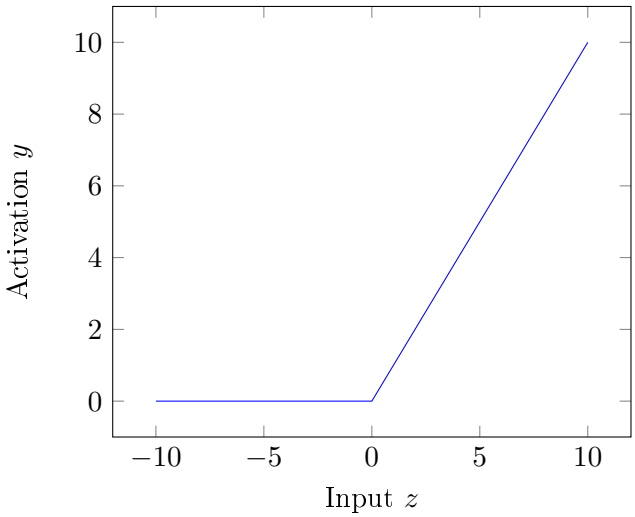

Figure 3.4: The ReLU activation function.

The Softmax Activation Function This activation function is usually applied to the output neurons of a network. When a neural network is used to perform classification tasks, each output neuron is commonly associated with a specific class. In classification tasks, it is highly desirable to assign a probability to each class that represents how likely it is that the input data belongs to that class. The softmax activation function is used to achieve this by setting up the output neurons to represent a probability distribution over all possible classes. In an output layer consisting of n output neurons, the softmax function for each neuron i of that layer can be described by the following equation, where z_i denotes the summation units' output of the i^{th} neuron:

$$\phi(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} \tag{3.5}$$

The softmax activation function represents, loosely speaking, the percentage of the current neurons activation with respect to the compound activation of all neurons in the layer. The inputs z_i of the ouput layer can thus be interpreted as the *unnormalized log-probabilities* for each class.

There might arise the question why each input z_i is first fed into the exponential function e^x before translating the activations into probabilities. This is done to further amplify the strongest signals and attenuate the weaker ones which results in more clear-cut values which makes learning for the network easier, as shown in the following example:

Imagine the z_i inputs of the output layer are given by the following vector: $(2, 4, 2, 1)^T$. If we just normalize these values to obtain a probability for each neuron, we get $(0.22, 0.44, 0.22, 0.11)^T$. Using the exponential function first, we roughly get $(0.1, 0.76, 0.1, 0.04)^T$ which amplifies the most likely outcomes and attenuates the less likely ones.

3.1.3 The Role of the Bias Value

There still remains the question why in each artificial neuron there is a bias value b_k added to the weighted sum of the inputs. The reason for this is related to the activation function: The bias term acts like a parameter that determines how to shift the activation function along the x-axis. We already know from Eq. 3.1 that for a neuron k with n inputs the toal input signal z_k adds up to:

$$z_k = \sum_{j=1}^{n} x_j \cdot w_{kj} + b_k$$

Let us denote the weighted sum of the input signals as a separate value $a_k = \sum_{j=1}^{n} x_j \cdot w_{kj}$ that describes the raw input of the neuron. This means that $z_k = a_k + b_k$ and using the sigmoid function (see section 3.1.2) as an example to demonstrate the effects of the bias value, we can slightly rewrite it as

$$\phi(a_k) = \frac{1}{1 + e^{-(a_k + b_k)}}$$

also setting $\theta = 1$ for demonstration purposes. Plotting the activation function for different values of b_k immediately reveals the effect of the bias value as a shift-parameter which can be seen in Fig. 3.5.

What this shift means is that the bias term acts like a threshold that has to be overcome in order for the neuron to become active. Positive bias values lead to activity even when the raw input a_k is still negative and negative bias values require bigger input signals in order for the neuron to fire.

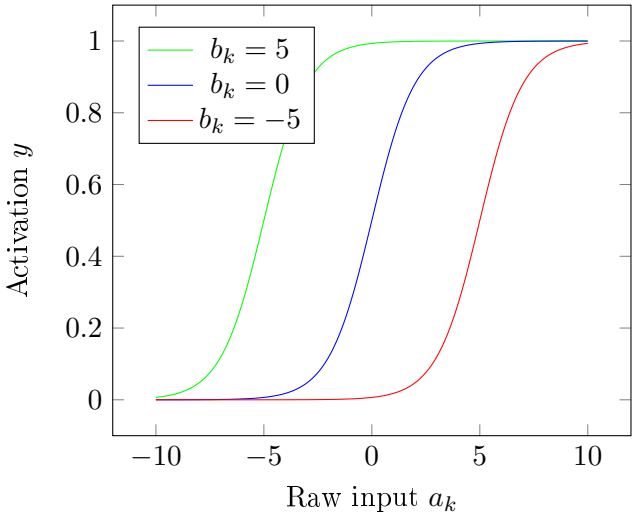

Figure 3.5: The sigmoid activation function plotted for different bias values.

3.2 Neural Networks as Classifiers

After establishing a mathematical model that helps us to describe a neural network, there is still one problem to be solved: How to train the network to be able to successfully perform tasks such as classification? In order to figure this out, we will first take a look at classification tasks in general and then explore how to set up and train a neural network to perform classification.

3.2.1 Classification

The basis of a classification task is usually formed by a dataset that consists of features as well as labels. The goal of the classification algorithm is to predict the label of an instance of the dataset by only looking at its features. In order to achieve this, the classifier first has to build a model based on a training dataset. This procedure is called *training* (see section 3.2.3). In the next step, called *testing*,the classifier is presented with some new examples that it did not see during training. The classifier is tested on these new examples to estimate its performance and to see if it was able to learn any concepts from the data, i.e. to *generalize*. Because the classifier infers a function from labeled data, classification is an example of a broader domain called *supervised learning*.

3.2.1.1 Evaluating a Classifier

In order to find out how well a classifier generalizes after training, the results of the testing phase can be entered into a *confusion matrix* that is structured as shown in Table 3.1.

	Class Positive (Predicted)	**Class Negative (Predicted)**
Class Positive (Actual)	True Positives (TP)	False Negatives (FN)
Class Negative (Actual)	False Positives (FP)	True Negatives (TN)

Table 3.1: The structure of a confusion matrix for a classification task with two classes "Positive" and "Negative".

Each entry in this matrix describes how often the classifier was presented with an example of the row-class during testing and predicted that the example belongs to the column-class. For instance, the value FP (False Positives) expresses how often the classifier saw an example of class "Negative" and predicted that it belongs to class "Positive". The same principle also applies to the other cells in the confusion matrix. It should be noted that this concept can be extended to classification tasks with more than two classes as well by simply adding new rows and columns for each new class. The resulting measurements of true positives, true negatives, false positives and false negatives can be used to compute the following evaluation metrics:[3]

Accuracy This measurement determines the percentage of examples in the testing set that the classifier predicted correctly. It can be denoted by the following equation:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.6}$$

This metric works well if there is roughly an equal amount of examples for each class. However, if one of the classes makes up most of the examples, the classifier can reach a high degree of accuracy by just predicting the label of the dominant class every single time. This impairs the significance of this metric when imbalances among the classes are present.

[3] A collection of these metrics can also be found in [PG17], see chapter *Evaluating Models*.

Precision The precision score shows the percentage of examples that were correctly classified as positive among all examples that the classifier labeled positive:

$$Precision = \frac{TP}{TP + FP} \tag{3.7}$$

This metric can also be interpreted as an estimate of the conditional probability that the classifier is right given that it predicted a positive class:

$$Precision = P(\text{Classifier is right}|\text{Classifier predicted POSITIVE})$$

Recall This measurement remedies the imbalance issues of the accuracy metric by determining the percentage of correctly classified examples for each separate class. It can be denoted by the following expression:

$$Recall = \frac{TP}{TP + FN} \tag{3.8}$$

The recall score can also be interpreted as an estimate of the conditional probability that the classifier is right given a specific class:

$$Recall = P(\text{Classifier is right}|\text{Class is POSITIVE})$$

F1 Score This metric combines precision and recall to calculate their so called *harmonic mean*. It is often used when evaluating classification models, thus its equation is also displayed here:

$$\text{F1 Score} = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3.9}$$

It should be noted that all these measurements can also be extended to classification tasks with more than two classes. This is done by first computing the metrics for each class separately and then taking the average of these values to estimate a global score.

3.2.2 Network Architecture for Classification

The architecture of the neural network that will be used to perform the classification task is highly dependent on the structure of the dataset. Remembering that a neural network consists of an *input layer* as well as *hidden layers* and an *output layer*, the question is how to assemble these layers to fit the task well.

The first consideration is that each feature in the dataset will correspond to an input

signal that is fed into the network, thus the amount of neurons in the input layer must be equal to the amount of features in the dataset. Because the only responsibility of the input units is to receive a signal from the environment and pass it on to the next layer, these neurons don't have a special activation function that transforms the input. The activation of these neurons is simply the identity of the incoming signal. In order to avoid features on larger numerical scales dominate features with smaller values, the data is usually normalized and scaled to equal ranges first before being fed into the network.

The number of hidden layers that are inserted between the input and the output layer highly depends on the complexity of the task. As the number of hidden neurons grows, there are more parameters (weights and biases) left to be adjusted during training which means more capacity for the network to learn. However, the danger lays in the fact that if there are too many hidden neurons and hidden layers, the network will just use this capacity to memorize the training examples and not extract general concepts from them which will lead to low accuracy on unseen examples. This problem can also be described by the more general term *overfitting*. On the contrary, if there are not enough hidden neurons, the network won't be able to capture all concepts that are present in the data which will lead to an opposite effect: *underfitting*. Both overfitting and underfitting harm the ability of the network to generalize well beyond the training data. In practice however, it is recommended to rather have too many hidden neurons and hidden layers than an insufficient amount, because there are other techniques such as *regularization* that punish increasing model complexity to prevent overfitting [Ben12]. The first choice of activation function that is used in the hidden layers is usually the ReLU (see section 3.1.2) because of its various beneficial properties.

As already hinted at in section 3.1.2 about the softmax activation function, it is highly useful if the network is able to not only predict the correct label but also to indicate how certain it is about it. This is why the output layer will consist of as many neurons as there are classes in the dataset which will enable us to use the softmax function on this layer to retrieve a set of probabilities for each example that is presented to the network. The neuron that shows the highest degree of activity, i.e. assigns the highest probability, determines the label the network will assign to the example as a result of the classification process.

3.2.3 Training the Network

In order to be able to improve the quality of the networks' predictions, i.e. training the network, we first have to introduce a way of measuring the performance of the network with respect to the training examples it is presented with.

Let x be an example input from the training dataset and $y'(x)$ be the desired output of the network that corresponds to the example. Both x and $y'(x)$ are vectors. The element x_i represents the input signal of the i^{th} input neuron and the element $y'_i(x)$ represents the desired activation of the i^{th} output neuron. To measure how close the actual output $y(x)$ of the network is to the desired output $y'(x)$, we can use the *sum of the squared errors*:

$$L_x = \sum_{i=1}^{n} (y_i(x) - y'_i(x))^2 = \left\| y(x) - y'(x) \right\|^2 \tag{3.10}$$

where n is the number of output neurons and L_x resembles the *loss of the network* for a single example x.

The *average total loss* of the network over all examples in the dataset (the total number of examples will be denoted by N) can be computed by averaging the losses of every single example:

$$L = \frac{1}{N} \cdot \sum_{x} L_x \tag{3.11}$$

We can also express this value in terms of the current configuration of the neural network that is represented by the set of weights w and the set of biases b that is currently used as the *loss function* $L(w, b)$. Now being able to measure the training performance of the network with respect to its configuration by calculating the average loss $L(w, b)$, we can define the training problem as follows:

Find a set of weights w and biases b such that $L(w, b) \rightarrow min$

This implies that training the network is an optimization problem where the weights and biases of the network are adjusted to find the minimum of the loss function $L(w, b)$.

The most common approach to solve the optimization problem is a technique called *gradient descent*. In each step of this procedure, the gradient $\nabla L(w, b)$ of the loss function L with respect to the weights w as well as the biases b is computed. This is done because the gradient always points in the direction of the steepest ascent of a function. In order to minimize L, one can simply take tiny successive steps in the direction of the *negative* gradient to arrive at a local minimum of L resulting in the following algorithm describing how to adjust the weights w and biases b in each step t:

$$(w, b)^T_{t+1} = (w, b)^T_t - \alpha \cdot \nabla L(w, b) \tag{3.12}$$

The α parameter in this equation describes the size of the steps that are taken in the direction of the negative gradient and is also called the *learning rate* of the network.

Choosing a reasonable value for α is essential for a successful trainig phase. If the learning rate is too big, the steps taken will also be too big resulting in skipping and not finding the minimum. Too small values of α will lead to slow convergence.

If the surface of L is convex, i.e. there is only one global minimum, the algorithm is guaranteed to converge for a sufficiently small learning rate. In practical application however, this property is usually not present due to the complexity of L. Despite of this circumstance, gradient descent usually still works well and converges to a local minimum of L that is usually sufficient for the network to solve the classification task.

There still remains the question how to compute the gradient $\nabla L(w, b)$ in each step of gradient descent. The answer to this is a procedure called *backpropagation* [RHW86].

3.2.3.1 The Backpropagation Algorithm

In order to derive the backpropagation algorithm that enables us to compute the gradient of the loss function, a little expansion of the current notation is necessary. The output of the k^{th} neuron of layer l in the network will now be denoted by $y_k^{(l)}$. This is done to indicate in which layer the described neuron resides. Likewise, the input z, bias b and weights w of neuron k in layer l will also receive a superscript denoting the current layer. Using n_l to describe how many neurons there are in layer l, we can slightly rewrite the equations that describe the input z and the output y of a neuron k like this:

$$z_k^{(l)} = \sum_{j=1}^{n_{l-1}} w_{kj}^{(l)} \cdot y_j^{(l-1)} + b_k^{(l)} \tag{3.13}$$

$$y_k^{(l)} = \phi(z_k^{(l)}) \tag{3.14}$$

Because the total loss of the network is just the average of the losses for each single example (see Eq. 3.11), the gradient of L can be computed like this:

$$\nabla L(w, b) = \nabla \left(\frac{1}{N} \cdot \sum_x L_x(w, b) \right) = \frac{1}{N} \cdot \sum_x \nabla L_x(w, b) \tag{3.15}$$

This means that computing the gradient of the total loss L is the same as computing the gradients of the losses for every single training example x and then taking the average.

The next step is to find a way to compute the partial derivatives that make up the components of the gradient. What this means is to find out how sensitive the loss function reacts to changes in a single weight $w_{kj}^{(l)}$ or a single bias $b_k^{(l)}$. Because all the weights as well as the bias of a neuron are combined with the inputs in its summation unit, it

is helpful to take an intermediate step: Rather than computing the partial derivatives directly, it makes sense to think about how changes in the input $z_k^{(l)}$ of a particular neuron in the network impact the loss L_x. This sensitivity of the loss function with respect to the input of a particular neuron will be dentoted by the following equation:

$$\delta_k^{(l)} = \frac{\partial L_x}{\partial z_k^{(l)}} \tag{3.16}$$

where $\delta_k^{(l)}$ describes the sensitivity of the loss function with respect to changes in the input of neuron k in layer l.

Utilizing the *chain rule* of calculus, one can now write the partial derivatives of the loss function with respect to the weights as well as the biases like this:

$$\frac{\partial L_x}{\partial w_{kj}^{(l)}} = \frac{\partial L_x}{\partial z_k^{(l)}} \cdot \frac{\partial z_k^{(l)}}{\partial w_{kj}^{(l)}} = \delta_k^{(l)} \cdot \frac{\partial z_k^{(l)}}{\partial w_{kj}^{(l)}} \tag{3.17}$$

$$\frac{\partial L_x}{\partial b_k^{(l)}} = \frac{\partial L_x}{\partial z_k^{(l)}} \cdot \frac{\partial z_k^{(l)}}{\partial b_k^{(l)}} = \delta_k^{(l)} \cdot \frac{\partial z_k^{(l)}}{\partial b_k^{(l)}} \tag{3.18}$$

Computing the terms $\frac{\partial z_k^{(l)}}{\partial w_{kj}^{(l)}}$ and $\frac{\partial z_k^{(l)}}{\partial b_k^{(l)}}$ is fairly straightforward:

$$\frac{\partial z_k^{(l)}}{\partial w_{kj}^{(l)}} = \frac{\partial}{\partial w_{kj}^{(l)}} \sum_{i=1}^{n_{l-1}} w_{ki}^{(l)} \cdot y_i^{(l-1)} + b_k^{(l)} = y_j^{(l-1)} \tag{3.19}$$

$$\frac{\partial z_k^{(l)}}{\partial b_k^{(l)}} = \frac{\partial}{\partial b_k^{(l)}} \sum_{i=1}^{n_{l-1}} w_{ki}^{(l)} \cdot y_i^{(l-1)} + b_k^{(l)} = 1 \tag{3.20}$$

Now the only component that is left to be calculated is the sensitivity of the loss with respect to the input of each neuron, $\delta_k^{(l)}$. In order to compute this value, two cases have to be distinguished: First, if l is the output layer, $\delta_k^{(l)}$ will only influence the loss through one single neuron k. Keeping this in mind, calculating $\delta_k^{(l)}$ for the output layer goes as follows:

$$
\begin{aligned}
\delta_k^{(l)} &= \frac{\partial L_x}{\partial z_k^{(l)}} \overset{\textit{chain rule}}{=} \frac{\partial L_x}{\partial y_k^{(l)}} \cdot \frac{\partial y_k^{(l)}}{\partial z_k^{(l)}} \\
&= \left(\frac{\partial}{\partial y_k^{(l)}} \sum_{i=1}^{n} (y_i^{(l)}(x) - y_i'(x))^2 \right) \cdot \frac{\partial}{\partial z_k^{(l)}} \phi(z_k^{(l)}) \\
&= 2 \cdot (y_k^{(l)} - y_k'(x)) \cdot \phi'(z_k^{(l)})
\end{aligned}
\tag{3.21}
$$

The second case is l being a hidden layer. In this scenario, $\delta_k^{(l)}$ will influence the output of the loss function through all the neurons in layer $l+1$. Taking this into consideration, $\delta_k^{(l)}$ for each hidden neuron can be computed like this:

$$
\begin{aligned}
\delta_k^{(l)} &= \frac{\partial L_x}{\partial z_k^{(l)}} \overset{\textit{chain rule}}{=} \frac{\partial L_x}{\partial y_k^{(l)}} \cdot \frac{\partial y_k^{(l)}}{\partial z_k^{(l)}} \\
&= \frac{\partial L_x}{\partial y_k^{(l)}} \cdot \phi'(z_k^{(l)}) \overset{\textit{infl. on next layer}}{=} \left(\sum_{i=1}^{n_{l+1}} \frac{\partial L_x}{\partial z_i^{(l+1)}} \cdot \frac{\partial z_i^{(l+1)}}{\partial y_k^{(l)}} \right) \cdot \phi'(z_k^{(l)}) \\
&= \left(\sum_{i=1}^{n_{l+1}} \delta_i^{(l+1)} \cdot \frac{\partial z_i^{(l+1)}}{\partial y_k^{(l)}} \right) \cdot \phi'(z_k^{(l)}) \\
&= \left(\sum_{i=1}^{n_{l+1}} \delta_i^{(l+1)} \cdot \frac{\partial}{\partial y_k^{(l)}} \left(\sum_{j=1}^{n_l} w_{ij}^{(l+1)} \cdot y_j^{(l)} + b_i^{(l+1)} \right) \right) \cdot \phi'(z_k^{(l)}) \\
&= \left(\sum_{i=1}^{n_{l+1}} \delta_i^{(l+1)} \cdot w_{ik}^{(l+1)} \right) \cdot \phi'(z_k^{(l)})
\end{aligned}
\tag{3.22}
$$

Putting it all together, we can formulate the backpropagation algorithm for a single training example x as shown in Algorithm 1.

This procedure is repeated for every example x and the average of the computed gradients determines the direction of each step during gradient descent.

It should be noted that there are several extensions to the algorithm of gradient descent. One very popular variation called *stochastic gradient descent* [Bot12] does not compute the gradient with respect to every single example, but divides the whole dataset into separate randomly sampled batches instead. Each batch is then used to take a step of gradient descent by computing the average gradient of all the examples in the batch. This is done to speed up the process of learning by not having to iterate over the whole dataset to take one step in the parameter space.

Algorithm 1 Backpropagation

$x \leftarrow$ current example
for each layer $l = 2, \ldots, n$ **do** ▷ Feed the input through the network
  for each neuron k in l **do**
    $z_k^{(l)} \leftarrow \sum_{j=1}^{n_{l-1}} w_{kj}^{(l)} \cdot y_j^{(l-1)} + b_k^{(l)}$ ▷ Compute the input of each neuron
    $y_k^{(l)} \leftarrow \phi(z_k^{(l)})$ ▷ Compute the activation
  end for
end for
for each layer $l = n, \ldots, 2$ **do** ▷ Backward pass
  for each neuron k in l **do**
    if l is output layer **then** ▷ Compute delta
      $\delta_k^{(l)} \leftarrow 2 \cdot (y_k^{(l)} - y'_k(x)) \cdot \phi'(z_k^{(l)})$ ▷ See 3.21
    else
      $\delta_k^{(l)} \leftarrow \left(\sum_{i=1}^{n_{l+1}} \delta_i^{(l+1)} \cdot w_{ik}^{(l+1)} \right) \cdot \phi'(z_k^{(l)})$ ▷ See 3.22
    end if
    for each neuron j in $l-1$ **do**
      $\frac{\partial L_x}{\partial w_{kj}^{(l)}} \leftarrow \delta_k^{(l)} \cdot y_j^{(l-1)}$ ▷ Calculate gradient w.r.t. weight, see 3.17
    end for
    $\frac{\partial L_x}{\partial b_k^{(l)}} \leftarrow \delta_k^{(l)}$ ▷ Calculate gradient w.r.t. bias, see 3.18
  end for
end for
return $\nabla L_x(w, b)$ ▷ Return the gradient of L_x

3.2.4 When to Stop Training

Being able to compute the weight and bias updates during each step of optimization, there still remains the question when to stop the training procedure. In practice it is very common to only train for a predefined number of passes through the whole dataset. Such a pass through the training data, during which every single training example is presented to the network once, is called an epoch.

It is also common to have a third dataset next to the training and testing set which is used to validate that the network still improves its generalization ability during training. This dataset is called the *validation set*. Usually, during each epoch, the error rate on the validation set is computed. If this error does not decrease further during training for a predefined number of epochs, the training is stopped. This technique is also called *early stopping*.

3.2.5 Initializing the Network

The last step that has to be taken before training a neural network to solve a classification problem is initializing the weights and biases. At first glance it may be tempting to initialize all the parameters with the same number, for instance zero. Investigating further on this idea yields that this way of initializing the weights and biases makes it impossible for the network to learn. If all the weights in a layer have the same value, then every neuron in that layer will receive the same compounded input and thus will emit the same signal. Looking back at the backpropagation algorithm, it becomes evident that this leads to equal weight updates as well. This causes the neurons in a layer to continuously show equal amounts of activity which prevents the network from learning any meaningful concepts.

To overcome this symmetry between neuron activations that is induced by initializing the weights equally, one has to employ initialization techniques that lead to *symmetry breaking* among the neurons. This is best achieved by initializing each weight with a random number. A very common approach is to use either a gaussian or a uniform distribution with a mean of zero and a variance of $\frac{2}{n_{in}+n_{out}}$, where n_{in} is the number of neurons in the preceding layer and n_{out} is the number of neurons in the following layer of the weight. This procedure is also called "Xavier Initialization" and leads to improved learning beyond just breaking the symmetry between neuron activations by achieving a balance of the weight values that works well in practice [GB10].

4 Convolutional Neural Networks

While fully connected feed forward networks work well on data of moderate dimensionality, training them becomes increasingly more difficult once the number of inputs grows. In the case of image classification for example, even an image with just a resulution of 256x256 pixels produces $256 \cdot 256 = 65536$ inputs. Considering that colored images usually have at least three color channels (take the popular **R**ed **G**reen **B**lue color model as an example), this multiplies the amount of input dimensions by a factor of 3, yielding $65536 \cdot 3 = 196608$ dimensions total. If we wanted to use a fully connected hidden layer with only half as many hidden neurons, we would have to optimize $196608 \cdot 98304 = 19,327,352,832$ weights only for the first layer. Let us assume that storing a single weight in floating point format costs 4 bytes. This would lead to spacial requirements of $19,327,352,832 \cdot 4 = 77,309,411,328$ bytes or 77.31 gigabytes! One should quickly notice that using this kind of neural networks to classify images is beyond infeasible. But how is it possible that state-of-the-art classifiers achieve human like performance in image classification, also relying on artificial neural networks [Rus+14]? In the following sections, we will explain how to modify our current feed forward architecture in order to cope with these challenges and how these astonishing results are possible.

4.1 Overview

One characteristic of fully connected feed forward networks is that every neuron in the first hidden layer is connected to every input node. While this allows every neuron to make use of every piece of information accross the whole input, it also increases the complexity of the task that each neuron tries to solve. If the data has a specific spatial structure, as would be the case for images, this knowledge is not incorporated into the fully connected architecture. In Convolutional Neural Networks (CNNs), the goal is to reduce the model complexity by making use of the spacial structure of the input. This is done by connecting each neuron of the first hidden layer only to a locally constrained area of the image. We call this area the *local receptive field* of the neuron. Each neuron

in the first hidden layer will receive its own local receptive field of inputs that it manages. This way, each hidden neuron is no longer fully connected to every input, instead it is only connected to a specific part of it. To reduce the complexity of the model even further, every hidden neuron of a layer will share its weights and biases with all the other neurons of that same layer. Intuitively, this means that each neuron will look for the same input feature in the data, the only difference being the local receptive field the individual neurons watch. This concept, also called a *convolution layer*, is the main foundation of the CNN architecture.

4.2 The Convolutional Architecture

The convolutional architecture is composed of three main parts [PG17][1]:

1. The input layer
2. Feature extraction layers
3. Classification layers

These components are stacked on top of each other to successively break down the classification task into smaller problems, by first extracting the relevant features from the data and then performing classification on the basis of these high level features. All the relevant layer types are described in more detail within the following sections.

4.2.1 The Input Layer

Because CNNs are based on spatial assumptions on the input data, it has to be arranged in a way that makes optimal use of its structure. This is why it is most common to organize the input neurons in the form of a two dimensional grid, where each cell resembles a part of the input. In the case of image classification for example, one would typically set the x and y dimensions of the grid equal to the numer of pixels in the input image in the corresponding directions. This results in each input neuron in the grid resembling one pixel in the image. As was the case with ordinary feed forward networks, the input layer does not compute an activation function and just passes its signal further into the network.

[1] See *CNN Architecture Overview* in chapter 4.

4.2.2 Feature Extraction Layers

4.2.2.1 Convolution Layers

The convolution layer, which was briefly described above, is the main component of the feature extraction block. It is used to detect features across the input by having each neuron watch a specific part of it. This can be visualized best by imagining the neurons be arranged in a two dimensional grid where each neuron in the grid watches a corresponding area in the input (see Fig . . .). To describe this behaviour more precisely, we can slightly extend the neural model we have already developed for feed forward networks.

Recalling that the first component of the basic neural model was the weighted sum of the inputs with a set of weights, we can transfer this concept to convolutional layers by interpreting the weights as a filter that each neuron applies to its local receptive field. This filter, that we will also refer to as a kernel, can be described by a stack of matrices of equal dimensions, where each matrix consists of the weights that are applied to the local region in the input within one channel. The amount of matrixes stacked corresponds to the depth of the input data. When dealing with images, this would be the same as the amount of channels (e.g. RGB has three color channels, thus the kernel would consist of three stacked matrices). To compute the weighted sum of the input with the filter, the dot product between the local area of the input and the filter is applied and a bias value is added as usual.

This procedure can also be expressed in terms of mathematical equations as follows: Let the width of the kernel K be x_K, the height be y_K and the depth be z, where z is usually equal to the amount of different channels when dealing with images. Let us further assume that the local receptive field of the current neuron k can be described by the coordinates (i, j) with respect to the input X, where (i, j) indicates the shift of the field in x and y direction. The result of applying the filter and adding the bias value b can be denoted by the following formula:

$$z_k = \sum_{n=1}^{z} \sum_{m=1}^{y_K} \sum_{l=1}^{x_K} X[i+l, j+m, n] \cdot K[l, m, n] + b \tag{4.1}$$

This operation is very similar to the concept of convolution in the area of signal processing, which is the origin of the name *convolutional* neural networks.

After calculating the summation result z_k for every neuron in the grid, an activation function is applied just like in feed forward networks. Because of its various advantages, the most common choice is the ReLU function which results in the following expression

describing the output y_k of every neuron k in the grid:

$$y_k = max(0, z_k) \tag{4.2}$$

Carrying out this computation on every neuron in the grid yields a new pattern of activations across the layer, which is called a *feature map* and can be interpreted intuitively as follows: When computing the dot product between the kernel K and the local receptive field of the neuron, this dot product acts like a similarity measure between the local input and K. When the dot product is large, this indicates that the kernel and the local input area very similar, when it is very negative, the input and the kernel are contrasting. If the dot product is close to zero, this means that kernel and input have almost nothing in common. Thus, areas of high activity in the feature map indicate, where the input is most similar to the kernel. Thinking further about this relationship, it follows that the kernel K itself represents a *feature* in the input that every neuron watches out for, which is why the convolutional layer is used as a *feature extractor*.

To leverage this procedure even further, each convolution layer does not only produce a single activation map for a single feature, but repeats the process multiple times with different kernels, yielding a stack of activation maps that display the presence of multiple features. In addition to the amount of different features to detect, there are a few other parameters to adjust in a convolutiona layer. A brief overview of all these tunable parameters will be provided in the following.

Kernel Size The kernel size can be described by the width and the height of each neurons' local receptive field. In Eq. 4.1, these dimensions are denoted by the parameters x_K and y_K. Note that the depth of the kernel is always predefined as the depth of the input from the previous layer.

Kernel Stride Up to this point, we did not yet define how to arrange the local receptive fields of the neurons. Introducing a stride parameter, we can set the first local field to the region described by a $(0, 0)$ coordinate shift (see the i and j parameters in Eq. 4.1). Each consecutive neuron will receive a receptive field that is shifted by some amount. To arrange these fields, we will first only shift in the x direction of the input. The amount of shifting in each step is described by the x_stride of the kernel. After shifting across the x axis is no longer possible, the x position of the local receptive field is reset to zero and a shift happens in the y direction. This shift can be adjusted by the parameter y_stride. The same procedure is repeated until every area of the input is covered by a neuron.

Number of Kernels This parameter describes the number of kernels to apply in a convolution layer and, as we saw earlier, corresponds to the amount of features to detect. Because every kernel results in its own feature map, the depth of the output of a convolution layer is equal to its number of kernels.

4.2.2.2 Pooling Layers

Pooling layers are usually inserted after convolution layers to capture the most essential patterns of the feature maps in order to reduce their complexity. This is done by downsampling the activations, resulting in a smaller input grid for the next layer. Quite similar to convolution layers, a pooling layer can also be imagined as a filter that runs over the input, but instead of computing dot products, this filter only returns the maximum value of the region it is currently looking at. This technique is also known as *max-pooling*. Unlike convolution layers however, pooling layers do not influence the depth of the input because each feature map is processed separately. There are two parameters that can be adjusted in each pooling layer:

- **Size:** This parameter controls the size of the filter that is used for downsampling.
- **Stride:** Similar to the stride parameter of a convolution layer, this value adjusts the step size in each direction that is taken during pooling.

Due to the fact that the downsampling operation purposefully discards information regarding the exact location where the feature was detected, this procedure contributes to the *spacial invariance* of convolutional neural network. What this means it that the network is able to recognize certain features, no matter where exactly they are in the image. This is one of the most important traits of the CNN architecture which greatly contributes to its effectiveness in practical applications.

4.2.3 Classification Layers

After the feature extraction has been performed, the complexity of the input data has usually been reduced significantly by stacking multiple convolution and pooling layers. On the basis of this simpler and also richer representation of the data, an ordinary fully connected feed forward network can be used to perform the classification task. This is done by inserting a fully connected hidden layer after the last layer of the feature extraction part. This hidden layer is connected to each neuron in the grid. Depending on the complexity of the problem, it is also possible to insert multiple fully connected hidden layers right after each other to complete the classification task. As usual, the

last layer consists of as many output neurons as there are classes, employing the softmax activation function to compute the classification result.

5 Implementing and Testing a CNN-Model in DL4J

6 Discussion

7 Conclusion

Bibliography

[Ben12] Y. Bengio. "Practical recommendations for gradient-based training of deep architectures". In: *ArXiv e-prints* (June 2012). arXiv: 1206.5533 [cs.LG].

[Bot12] Léon Bottou. "Stochastic Gradient Descent Tricks". In: *Neural Networks: Tricks of the Trade*. Springer, Berlin, Heidelberg, 2012. ISBN: 978-3-642-35288-1.

[GB10] Xavier Glorot and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks". In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Yee Whye Teh and Mike Titterington. Vol. 9. Proceedings of Machine Learning Research. PMLR, 13–15 May 2010, pp. 249–256.

[GBB11] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. "Deep Sparse Rectifier Neural Networks". In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Geoffrey Gordon, David Dunson, and Miroslav Dudík. Vol. 15. Proceedings of Machine Learning Research. Fort Lauderdale, FL, USA: PMLR, Nov. 2011, pp. 315–323.

[Hay08] Simon Haykin. *Neural Networks and Learning Machines*. 3rd ed. Prentice Hall International, 2008. ISBN: 978-0131471399.

[PG17] Josh Patterson and Adam Gibson. *Deep Learning: A Practitioner's Approach*. 1st ed. O'Reilly Media, 2017. ISBN: 978-1491914250.

[RHW86] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors". In: *nature* 323 (1986).

[Rus+14] O. Russakovsky et al. "ImageNet Large Scale Visual Recognition Challenge". In: *ArXiv e-prints* (Sept. 2014). arXiv: 1409.0575 [cs.CV].

List of Figures