Dissertation Preparation

José Pedro Castro Fonseca

U. PORTO

FEUP **FACULDADE DE ENGENHARIA**
UNIVERSIDADE DO PORTO

Integrated Masters in Electrical and Copmuter Engineering

Supervisor: João Canas Ferreira

Second Supervisor: Ivo Timóteo

February 8, 2016

Dissertation Preparation

José Pedro Castro Fonseca

Integrated Masters in Electrical and Copmuter Engineering

February 8, 2016

Resumo

Este documento ilustra o formato a usar em dissertações na Faculdade de Engenharia da Universidade do Porto. São dados exemplos de margens, cabeçalhos, títulos, paginação, estilos de índices, etc. São ainda dados exemplos de formatação de citações, figuras e tabelas, equações, referências cruzadas, lista de referências e índices. Este documento não pretende exemplificar conteúdos a usar. É usado o *Loren Ipsum* para preencher a dissertação.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam vitae quam sed mauris auctor porttitor. Mauris porta sem vitae arcu sagittis facilisis. Proin sodales risus sit amet arcu. Quisque eu pede eu elit pulvinar porttitor. Maecenas dignissim tincidunt dui. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec non augue sit amet nulla gravida rutrum. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Nunc at nunc. Etiam egestas.

Donec malesuada pede eget nunc. Fusce porttitor felis eget mi mattis vestibulum. Pellentesque faucibus. Cras adipiscing dolor quis mi. Quisque sagittis, justo sed dapibus pharetra, lectus velit tincidunt eros, ac fermentum nulla velit vel sapien. Vestibulum sem mauris, hendrerit non, feugiat ac, varius ornare, lectus. Praesent urna tellus, euismod in, hendrerit sit amet, pretium vitae, nisi. Proin nisl sem, ultrices eget, faucibus a, feugiat non, purus. Etiam mi tortor, convallis quis, pharetra ut, consectetuer eu, orci. Vivamus aliquet. Aenean mollis fringilla erat. Vivamus mollis, purus at pellentesque faucibus, sapien lorem eleifend quam, mollis luctus mi purus in dui. Maecenas volutpat mauris eu lectus. Morbi vel risus et dolor bibendum malesuada. Donec feugiat tristique erat. Nam porta auctor mi. Nulla purus. Nam aliquam.

ii

Abstract

Here goes the abstract written in English.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Sed vehicula lorem commodo dui. Fusce mollis feugiat elit. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Donec eu quam. Aenean consectetuer odio quis nisi. Fusce molestie metus sed neque. Praesent nulla. Donec quis urna. Pellentesque hendrerit vulputate nunc. Donec id eros et leo ullamcorper placerat. Curabitur aliquam tellus et diam.

Ut tortor. Morbi eget elit. Maecenas nec risus. Sed ultricies. Sed scelerisque libero faucibus sem. Nullam molestie leo quis tellus. Donec ipsum. Nulla lobortis purus pharetra turpis. Nulla laoreet, arcu nec hendrerit vulputate, tortor elit eleifend turpis, et aliquam leo metus in dolor. Praesent sed nulla. Mauris ac augue. Cras ac orci. Etiam sed urna eget nulla sodales venenatis. Donec faucibus ante eget dui. Nam magna. Suspendisse sollicitudin est et mi.

Fusce sed ipsum vel velit imperdiet dictum. Sed nisi purus, dapibus ut, iaculis ac, placerat id, purus. Integer aliquet elementum libero. Phasellus facilisis leo eget elit. Nullam nisi magna, ornare at, aliquet et, porta id, odio. Sed volutpat tellus consectetuer ligula. Phasellus turpis augue, malesuada et, placerat fringilla, ornare nec, eros. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Vivamus ornare quam nec sem mattis vulputate. Nullam porta, diam nec porta mollis, orci leo condimentum sapien, quis venenatis mi dolor a metus. Nullam mollis. Aenean metus massa, pellentesque sit amet, sagittis eget, tincidunt in, arcu. Vestibulum porta laoreet tortor. Nullam mollis elit nec justo. In nulla ligula, pellentesque sit amet, consequat sed, faucibus id, velit. Fusce purus. Quisque sagittis urna at quam. Ut eu lacus. Maecenas tortor nibh, ultricies nec, vestibulum varius, egestas id, sapien.

Donec hendrerit. Vivamus suscipit egestas nibh. In ornare leo ut massa. Donec nisi nisl, dignissim quis, faucibus a, bibendum ac, diam. Nam adipiscing hendrerit mi. Morbi ac nulla. Nullam id est ac nisi consectetuer commodo. Pellentesque aliquam massa sit amet tellus. Vivamus sodales aliquam leo.

Agradecimentos

Aliquam id dui. Nulla facilisi. Nullam ligula nunc, viverra a, iaculis at, faucibus quis, sapien. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Curabitur magna ligula, ornare luctus, aliquam non, aliquet at, tortor. Donec iaculis nulla sed eros. Sed felis. Nam lobortis libero. Pellentesque odio. Suspendisse potenti. Morbi imperdiet rhoncus magna. Morbi vestibulum interdum turpis. Pellentesque varius. Morbi nulla urna, euismod in, molestie ac, placerat in, orci.

Ut convallis. Suspendisse luctus pharetra sem. Sed sit amet mi in diam luctus suscipit. Nulla facilisi. Integer commodo, turpis et semper auctor, nisl ligula vestibulum erat, sed tempor lacus nibh at turpis. Quisque vestibulum pulvinar justo. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Nam sed tellus vel tortor hendrerit pulvinar. Phasellus eleifend, augue at mattis tincidunt, lorem lorem sodales arcu, id volutpat risus est id neque. Phasellus egestas ante. Nam porttitor justo sit amet urna. Suspendisse ligula nunc, mollis ac, elementum non, venenatis ut, mauris. Mauris augue risus, tempus scelerisque, rutrum quis, hendrerit at, nunc. Nulla posuere porta orci. Nulla dui.

Fusce gravida placerat sem. Aenean ipsum diam, pharetra vitae, ornare et, semper sit amet, nibh. Nam id tellus. Etiam ultrices. Praesent gravida. Aliquam nec sapien. Morbi sagittis vulputate dolor. Donec sapien lorem, laoreet egestas, pellentesque euismod, porta at, sapien. Integer vitae lacus id dui convallis blandit. Mauris non sem. Integer in velit eget lorem scelerisque vehicula. Etiam tincidunt turpis ac nunc. Pellentesque a justo. Mauris faucibus quam id eros. Cras pharetra. Fusce rutrum vulputate lorem. Cras pretium magna in nisl. Integer ornare dui non pede.

O Nome do Autor

"You should be glad that bridge fell down.
I was planning to build thirteen more to that same design"

Isambard Kingdom Brunel

Contents

List of Figures

List of Tables

Abreviaturas e Símbolos

ANN   Artificial Neural Networks
WWW   *World Wide Web*

Chapter 1

Theoretical Background

1.1 Basic Concepts of Machine Learning

Machine Learning is a field of Computer Science that studies the development of mathematical techniques that allow software to learn autonomously, without an explicit description of each rule of operation. Its goal is to extract latent features from the data that allow an immediate classification of each input data into a particular class – the catch is that there is no previous rule formulation, but instead we have an adaptive model that adjusts is parameters accordingly with the input data it receives, improving the estimates it yields as it receives new input samples.

Let us consider a practical example. For instance, suppose we want to build a program that given an input audio waveform representation of a spoken word, it matches it into a particular word of a dictionary. We could, of course, devise a set of rules and exceptions for each word analysing some of its features (perhaps the Fourier representation of each one, and, from it, manually finding the appropriate rules for each), but apart from it being a very complex task, it wouldn't be a scalable solution, given the enormous number of words in each language. The approach taken by Machine Learning is diametrically different, and instead of manually processing each waveform, we build a large dataset, of size N, containing the waveforms of several words $[\mathbf{x}_1 \ \mathbf{x}_2 \ \mathbf{x}_2 \ \ldots \mathbf{x}_N]$ – we call this dataset the **training dataset** – and we feed it to our model. Each of the i-th data point was previously labelled, and in fact we feed each training data point $\mathbf{x}_i$ *along* with its corresponding label t_i, so that the model can adapt its parameters accordingly to the *target value* it is supposed to classify. This set of labels $\mathbf{t} = [t_1 \ t_2 \ t_3 \ \ldots \ t_N]$ is called the **target vector**.

We are, then, left with the following question: how can the model quantitatively evaluate the quality of its current set of parameters? That could be achieved in a number of ways, but the most usual is using a **Cost Function** that, as the name suggests, measures the cost of each wrong classification of the model. The model then evolves in a way that minimizes the cost function. A usual choice for the cost function is the **sum of squares error**. Mathematically, if $y_\theta^i = y_\theta(x_i)$ is the prediction for the input data point x_i with label t_i, given the current set of parameters θ, the

cost function using this metric is given by

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{N} \left(y_\theta^i - t_i \right)^2. \tag{1.1}$$

Sometimes, instead of applying the raw data to the model, we can apply some sort of *prepro-cessing* to the data to extract the relevant features from it. For instance, instead of just feeding a raw image, we can perform several operations like edge detection or low-pass filtering, and apply them in parallel. In cases of highly dimensional data (i.e. each data vector has a very high number of features), we can apply techniques like **Principal Component Analysis** to reduce the feature space to a smaller dimension one, where the previous features were combined into two or three new features that pose themselves as the most relevant.

The problems described above are, in fact, a subset of the problems that Machine Learning tries to address. These problems are called **classification problems**, because for each input data point, our model tries to fit it into the most appropriate class. But we can also address **regression problems** where the output is not limited to a discrete set of values but rather a continuous interval. On the other hand, the Neural Network that this work will implement addresses a special kind of classification problem, where the classification decision is influenced not only by the current input sample, but also by a given *window of samples* that trail the current sample.

In summary, the most typical setting for a Machine Learning problem is having a large *input dataset* which we use to *train* our model (i.e. allowing him to dynamically adapt its set of parameters θ), in order to produce an output label y_i for each of them that minimizes a quality metric, the *Cost Funcion*, which can be chosen to the sum of squared differences, the log-loss, or any other appropriate mathematical relationship between the estimate y_i and the correct label t_i.

Now that the basic Machine Learning concepts have been presented, I will discuss, henceforth, one of the most important algorithms that address the supervised classification problems, the *Artificial Neural Networks* that will be discussed in Section 1.2 as somewhat of a contextual intro-duction to the main theme of the thesis, which will be Recurrent Neural Networks (Section 1.3), namely the **Long Short-Term Memory Networks** (Section 1.4), both of which are improvements over the initial formulation of the ANNs. These two last networks branch even further from these set of problems, and are usually employed in *Deep Learning* tasks, where we try to extract even higher level information from data at the expense of increased model complexity.

1.2   Artificial Neural Networks

Artificial Neural Networks (ANN) are mathematical structures that, as the name suggests, try to mimic the Human Brain. ANN's building blocks, like their biological counterpart, model the high-level behaviour of biological neurons, in the sense that they neglect unnecessary biological aspects (such as modeling all the voltages across the neuron and all its electromagnetic interactions), and only retain its fundamental underlying mathematical function, which is a weighted linear

combination of its inputs subject to a *activation function*, i.e. a function that outputs a decision value depending on its inputs. Mathematically, we have

$$y = f(\mathbf{w}^T \mathbf{x} + b_0) \tag{1.2}$$

where $\mathbf{w} = [w_1 \; w_2 \; w_3 \; \ldots \; w_n]$ is the input weight vector, $\mathbf{x} = [x_1 \; x_2 \; x_3 \; \ldots \; x_n]$ the input vector, b_0 is the bias factor and $f(\cdot)$ is the chosen activation function. Furthermore, we call the scalar quantity $a = \mathbf{w}^T \mathbf{x} + b_0$ the **activation**, since its value determines how the activation function will behave. Figure 1.1 exemplifies the roles of these variables within our neuron model, and compares each part of it with the biological counterpart.

As far as the activation function is concerned, we can have several types. An immediate choice would be the **Binary Step Function** that outputs -1 if the activation is **below** a given threshold and 1 otherwise. There can also be **real valued activation functions**, whose output is not binary, but rather that of a continuously differentiable function, such as the logistic sigmoid $\sigma(a) = \frac{1}{1+e^{-a}}$ or the hyperbolic tangent $\tanh(a)$. This aspect will prove useful for the usual training methods, that involve the computation of derivatives. In Figure 1.2 these activation functions are plotted.

A neuron by itself can be thought of as a simple linear regression, where we optimize the weight of each feature according to a target value, or function. While important in some applications, the main interest in ANN is to evaluate increasingly more complex models, and not a simple linear regression. This is achieved by *chaining* nodes to one another, connecting the output of a given node to one of the inputs of another. We call *layers* to a group of these nodes that occupy the same hierarchical position. There can be any number of layers, with any number of nodes, but most implementations generally have 3 layers: the *initial* layer, the *hidden* layer (in the middle) and the *output* layer. Figure 1.3 suggests a possible structure for a 3 layer ANN.

Regarding the training of ANNs, is is performed through a two-step process: first, a **feed-forward** step where the input is applied, and the activations are evaluated in succession up to the output neurons; then, the perform the **backpropagation** step, where we calculate the errors in each of the nodes, but now from the output to the input: the weights are updated and optimized using an iterative method called *Gradient Descent*, where if τ is the current time step, the next update on the weight matrix $\mathbf{w}^{(\tau+1)}$ is given by

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \nabla E(\mathbf{w}^{(\tau)}) \tag{1.3}$$

where $E(\cdot)$ is the error function. As we can see, the weight matrix is moved in the direction that minimizes the error function the most, and η controls how fast this is achieved, being the reason why it is called the **learning rate**.

The computation of gradient of the error function comprises the evaluation of its derivatives with respect to each weight of all network connections, w_{ij}. They are

$$\frac{\partial E}{\partial w_{ji}} = \delta_j f(a_i) \tag{1.4}$$

where $f(\cdot)$ is the activation function of the neuron and

$$\delta_j = f'(a_j)\sum_k w_{kj}\delta_k. \tag{1.5}$$

The interpretation of these equations is simple. If w_{ji} is the weight of the connection between the neuron j we are considering and a neuron i in a previous layer, then the sum over k relates to all the neurons in the *next* layer to which j connects: this way, since the update on w_{ji}, according to 1.3, is given by

$$w_{ji}^{(\tau+1)} = w_{ji}^{(\tau)} - \eta\frac{\partial E}{\partial w_{ji}} \tag{1.6}$$

we see that, from 1.4 it simply is the product of the error of the current neuron, δ_j, with the output of the previous neuron $f(a_i)$. In turn, from 1.5, we see the recurrence relationship between it and the weighted sum of all posterior neurons that connect to it. Hence, the name backpropagation is now clear: we are, in fact, propagating the errors backwards into the neuron of interest, weighted by the corresponding weight, but now *backwards* instead of forward, as before. For the output units, the δ_j is simply the difference between the produced output and the corresponding label for that sample. This two-step process is performed for every data point in out dataset. For a complete proof of the above formulas, see (Bishop, 2006, chap. 5.3.1).

1.3   Recurrent Neural Networks

A Recurrent Neural Network (RNN) is, essentially, a regular ANN where some neurons (especially in the hidden layer) have *feedback connections* to themselves, i.e. their outputs are fed as inputs. The relevance of this different structure is the possibility to retain *sequence* information about the data. Before, each incoming data point only contributed to the training of the network, but no information about the correlation between themselves and the data points that preceded them did not influence the training step. They were regarded as if no temporal relationship existed, and therefore each data point is conditionally independent of any other. This is obviously not necessarily truth, and in fact there are many cases, where the correlation between data points is high for those closely spaced in time, in which it is actually completely false, as in video signals, audio signals, or other kinds of *temporal sequences* of data. Therefore, the feedback connection of the neuron to himself acts as a kind of *memory element* that takes into account in the present decision, the history of decisions previously taken, and hence the previous data. Figure **??** suggests a possible structure for a neuron of a hidden layer in an RNN, and also an alternate representation, where the structure is unfold through time.

The training of an RNN is usually performed using a variation of Backpropagation, called **Backpropagation Through Time** (BPTT), that as the name implies, performs the same backpropagation procedure discussed for the ANNs, but now taking into account the unfolding of the

network through a fixed training epoch T like Figure **??**. Due to this very fact, this training procedure is memory and performance consuming, and so it will not be used in my final work, but instead a novel approach, the **Simulatneous Perturbation Stochastic Approximation**.

Even though RNNs outperform static ANNs in sequence recognition problems Bengio (1991), they fail to retain long-term dependencies. Of course that the weight training process is itself a form of memory, but the problem is that the weight update is much slower than the activations Hochreiter et al. (2001), and therefore this memory only retains short-term dependencies. This is because of the so-called **Vanishing Gradients Problem** Bengio et al. (1994); Hochreiter et al. (2001), where the error decays exponentially through time, and the impact of previous incoming data points on the training of the weights, and thus the current decision, quickly decreases.

1.4 Long Short-Term Memory Networks

To overcome the issue of failing to remember long-term dependencies, Hochreiter and Schmidhüber proposed, in 1997, a novel approach to the RNNs called the *Long-Short Term Network* Hochreiter and Schmidhuber (1997). This section explains the main idea of this approach 1.4.1, and also how it is trained 1.4.2, serving as a support basis for the work of this thesis. Although originally formulated in 1997, its formulation has been incrementally updated in **?** and **?**, and the most current version is the one in **?**. One of the inital proposers of LSTM, Prof. Jürgen Schmidhüber, did a survey on the most common variations of the structure **?** last year, and this will the basis of this short theoretical presenation, as well as the work that will be developed in this thesis.

1.4.1 Structure, Operation and Equations

A single LSTM neuron is presented in Figure 1.4. As we can see from the picture, we still have the recurrent connections from the regular RNNs, but now there are multiple entry points that control the flow of information through the network. Although ommited from the picture, all the gates are biased, as is suggested in Equations **??**. The main components, their role and relevance, are explained as follows

- **Input Gate** – this is the input gate, where the importance of each feature of the input vector at time t, $\mathbf{x}^t$, and the output vector at the previous time step $\mathbf{y}^{t-1}$ is weighed in, producing an output $\mathbf{i}^t$.

- **Block Input Gate** – as the name implies, this gate controls the flow of information from the input gate to the memory cell. It also receives the input vector and the previous output vector as inputs, but it does not have peephole connections and its dynamics are controlled by a different set of weights. The **activation function of this gate can be any**, but the most common choice is the **Hyperbolic Tangent**.

- **Forget Gate** – its role is to control the contents of the Memory Cell, either to set them or reset them, using the *Hadamard Elementwise* matrix multiplication of its output at time t,

$\mathbf{c}^{(t)}$, with the contents of the memory unit at the previous time step, $\mathbf{c}^{(t-1)}$. The activation function of this gate is **always sigmoid**.

- **Output Block Gate** – this gate has a role very similar to that of the Block Input Gate, but now it controls the information flow *out* of the LSTM neuron, namely the activated Memory Cell output.

- **Memory Cell** – the cornerstone of the LSTM neuron. This is the memory element of the neuron, where the previous state is kept, and updated accordingly to the dynamics of the gates that connect to it. Also, this is where the peephole connections come from:

- **Output Activation** – the output of the Memory Cell goes through this activation function that, as the gate activation function, can be any, but the *hyperbolic tangent* is the most common choice.

- **Peepholes** – direct connections *from* the memory cell that allow for gates to 'peep' the states of the memory cell. They were added after the initial 1997 formulation, and their absence was proven to have a minimal performance impact **?**.

After this small conceptual definitions, that allow us to grasp some intuition on the operation of a *single* LSTM cell, I can present the overview of a *layer* of LSTM neurons and their formal mathematical formulation, that will be needed both for the high-level model and the HDL description. The operation of each gate of the layer is given by the following set of equations

$$\mathbf{z}^{(t)} = g(\mathbf{W}_z\mathbf{x}^{(t)} + \mathbf{R}_z\mathbf{y}^{(t-1)} + \mathbf{b}_z) \tag{1.7}$$

$$\mathbf{i}^{(t)} = \sigma(\mathbf{W}_i\mathbf{x}^{(t)} + \mathbf{R}_i\mathbf{y}^{(t-1)} + \mathbf{p}_i \odot \mathbf{c}^{(t-1)} + \mathbf{b}_i) \tag{1.8}$$

$$\mathbf{f}^{(t)} = \sigma(\mathbf{W}_f\mathbf{x}^{(t)} + \mathbf{R}_f\mathbf{y}^{(t-1)} + \mathbf{p}_f \odot \mathbf{c}^{(t-1)} + \mathbf{b}_f) \tag{1.9}$$

$$\mathbf{c}^{(t)} = \mathbf{i}^{(t)} \odot \mathbf{z}^{(t)} + \mathbf{f}^{(t)} \odot \mathbf{c}^{(t-1)} \tag{1.10}$$

$$\mathbf{o}^{(t)} = \sigma(\mathbf{W}_o\mathbf{x}^{(t)} + \mathbf{R}_o\mathbf{y}^{(t-1)} + \mathbf{p}_o \odot \mathbf{c}^{(t)} + \mathbf{b}_o) \tag{1.11}$$

$$\mathbf{c}^{(t)} = \mathbf{o}^{(t)} \odot h(\mathbf{z}^{(t)}) \tag{1.12}$$

The *i*-th element of the previous vectors in bold corresponds to the value of the gate of the *i*-th neuron of the layer, which is a very convenient and compact representation of the whole layer. Although useful for a high-level description in a programming language, the matrix representation may not be suitable for a direct HDL port. In that case, all the above equations still hold, but instead of *vectors* of outputs, we will have a single output, and we only use the appropriate value in the weight matrices $\mathbf{W}$. and $\mathbf{R}$.

1.4.2 Training – SPSA

Since this work aims to perform on-chip learning, it is important to find a suitable learning scheme. Since I am aiming at an hardware implementation, and although the memory resources of nowadays FPGAs are abundant, one must find an algorithm that uses the less memory as possible (at the smallest performance cost possible), in order to use the additional memory, for instance, to add additional LSTM cells that can improve the performance of our system. That being said, we see that the usual training algorithms for LSTM cells – i.e. Real-Time Recurrent Learning (RTRL), Truncated Backpropagation Through Time (BPTT) or a mixture of both **?** Hochreiter and Schmidhuber (1997) **?** – usually involve the storage of the deltas of each layer for every time instant in the training epoch (from 0 to T), which is a highly non-scalable solution both in terms of memory and performance. A most efficient approach to training times series dependant structures like LSTM is the use of **Simultaneous Perturbation Stochastic Approximation** (SPSA) **?**. The weight update for the i-th weight at the time step t is given by

$$\Delta w_i^{(t)} = \frac{J(\mathbf{w}^{(t)} + \beta \mathbf{s}^{(t)}) - J(\mathbf{w}^{(t)})}{\beta s_i^{(t)}} \tag{1.13}$$

where β is the magnitude of the perturbation to be introduced, $\mathbf{s}^{(t)}$ is a *sign vector* whose i-th element, $s_i^{(t)}$, is either -1 or 1. This way, we see that every weight is **randomly** incremented either by $-\beta$ or β, and we only need to keep a duplicate of the weight matrix with the perturbations, and we only need to evaluate the cost function twice per incoming sample. As for the *update rule* is concerned, we have

$$w_i^{(t+1)} = w_i^{(t)} - \eta \Delta w_i^{(t)} \tag{1.14}$$

where η is the learning rate, and $\Delta w_i^{(t)}$ is the update for the i-th weight evaluated in 1.13. According to the analysis performed in **?**, performing a second-order Taylor expansion at $\mathbf{w} = \mathbf{w}^{(t)}$, and taking the expected value of the equation, we get that

$$E(\Delta w_i^{(t)}) = \frac{\partial J(\mathbf{w}^{(t)})}{\partial w_i^{(t)}} \tag{1.15}$$

that is, the weight update approximates the gradient of the cost function relative to that weight, and so the learning rule is a special form of *Stochastic Gradient Descent*.

One last comment to be made concerning the update rule, is the hipothetical need for **limits** to the weight values, when the update rule exceeds $|w_{\max}|$ (in that case, we set $w_i^{(t+1)} = \pm w_{\max}$), which sometimes might be needed if the behaviour of the weight update is not appropriate **?**.
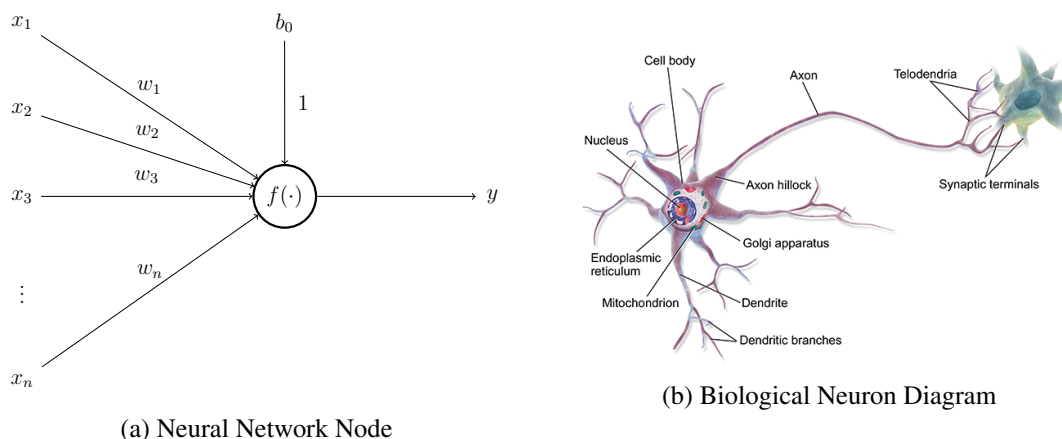
(a) Neural Network Node

(b) Biological Neuron Diagram

Figure 1.1: In Figure 1.1a. each input feature x_i is weighted by its corresponding weight w_i. During the training procedure, these weights are adjusted so that the output y approaches the target value. In Figure 1.1b, we see the diagram of an actual multi polar neuron. The dendrites, where the stimuli are received, plays a role similar to that of the input nodes. The axon transmits the signal to the synaptic terminals, that are similar to the y output
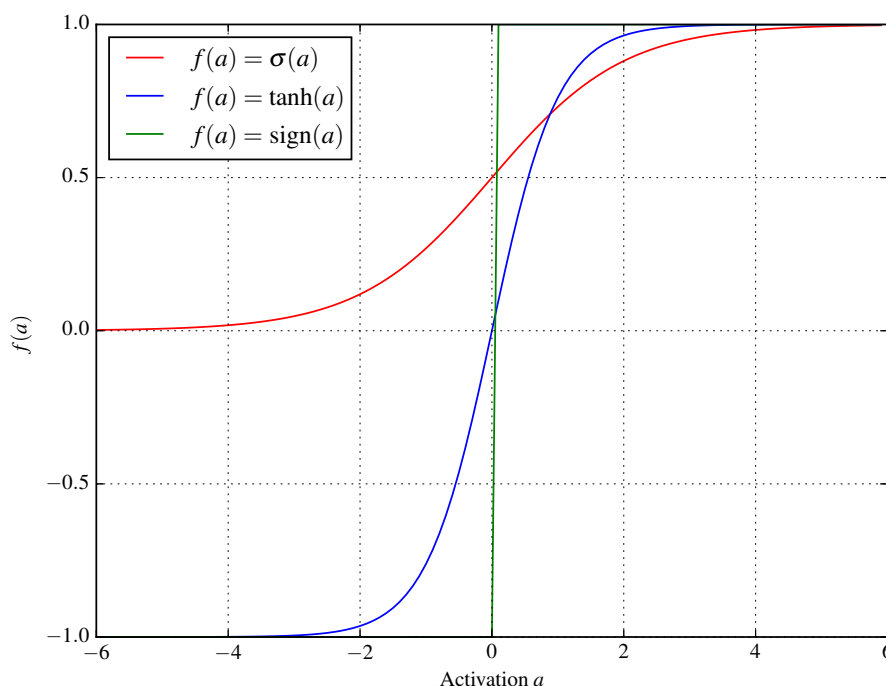

Figure 1.2: Three different activation functions. As you can see, the hyperbolic tangent has the same extreme value as the sign step function, but has a smooth transition between them, which can be interpreted as a *soft decision* in the more ambiguous middle region, reflecting the degree of uncertainty on the decision. On the other hand, the sigmoid function goes from zero to one, and is also smooth like the hyperbolic tangent
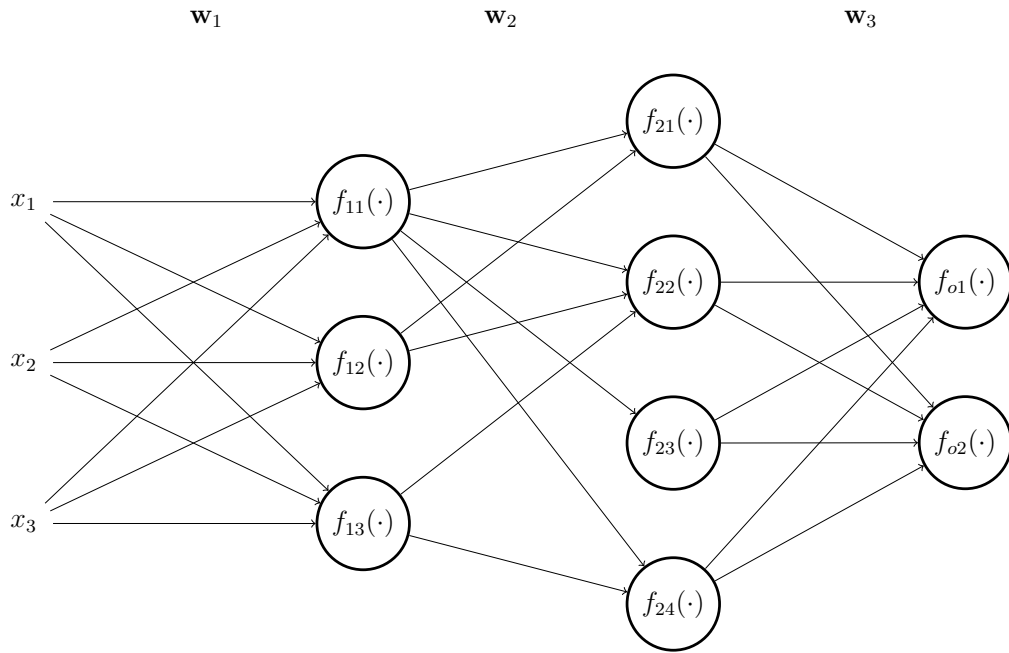
.

Figure 1.3: A three layer ANN. We have omitted some of the connections in the hidden layer, for simplification purposes. $\mathbf{w}_1$ represents the weight matrix of the input layer, $\mathbf{w}_2$ the weight matrix of the connections between the input layer and the hidden layer, and $\mathbf{w}_3$ the weight connections between the hidden and the output layer. $f_{ij}(\dots)$ is the activation function of the j-th neuron of the i-th layer. Since they can be different, I chose different indexes to each.
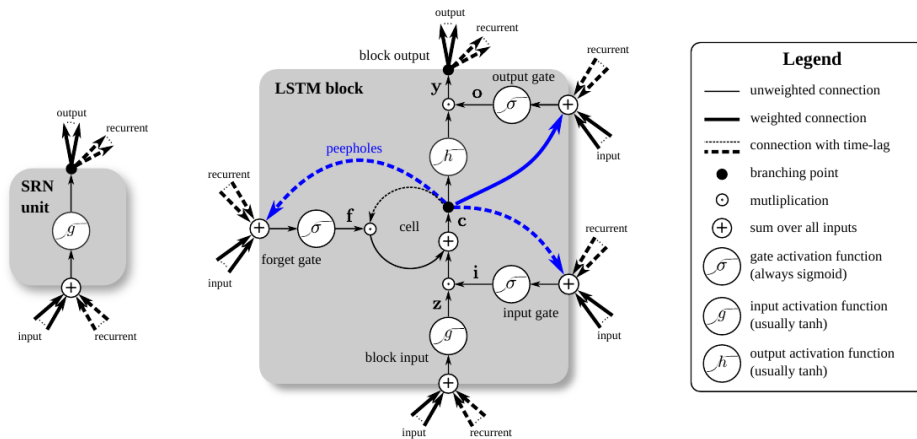
Figure 1.4: A complete LSTM neuron, with all the features as described in **?**

Appendix A

Loren Ipsum

Depois das conclusões e antes das referências bibliográficas, apresenta-se neste anexo numerado o texto usado para preencher a dissertação.

A.1 O que é o *Loren Ipsum*?

Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book. It has survived not only five centuries, but also the leap into electronic typesetting, remaining essentially unchanged. It was popularised in the 1960s with the release of Letraset sheets containing Lorem Ipsum passages, and more recently with desktop publishing software like Aldus PageMaker including versions of Lorem Ipsum (**?**).

A.2 De onde Vem o Loren?

Contrary to popular belief, Lorem Ipsum is not simply random text. It has roots in a piece of classical Latin literature from 45 BC, making it over 2000 years old. Richard McClintock, a Latin professor at Hampden-Sydney College in Virginia, looked up one of the more obscure Latin words, consectetur, from a Lorem Ipsum passage, and going through the cites of the word in classical literature, discovered the undoubtable source. Lorem Ipsum comes from sections 1.10.32 and 1.10.33 of "de Finibus Bonorum et Malorum" (The Extremes of Good and Evil) by Cicero, written in 45 BC. This book is a treatise on the theory of ethics, very popular during the Renaissance. The first line of Lorem Ipsum, "Lorem ipsum dolor sit amet...", comes from a line in section 1.10.32.

The standard chunk of Lorem Ipsum used since the 1500s is reproduced below for those interested. Sections 1.10.32 and 1.10.33 from "de Finibus Bonorum et Malorum" by Cicero are also reproduced in their exact original form, accompanied by English versions from the 1914 translation by H. Rackham.

A.3 Porque se usa o Loren?

It is a long established fact that a reader will be distracted by the readable content of a page when looking at its layout. The point of using Lorem Ipsum is that it has a more-or-less normal distribution of letters, as opposed to using "Content here, content here", making it look like readable English. Many desktop publishing packages and web page editors now use Lorem Ipsum as their default model text, and a search for "lorem ipsum" will uncover many web sites still in their infancy. Various versions have evolved over the years, sometimes by accident, sometimes on purpose (injected humour and the like).

A.4 Onde se Podem Encontrar Exemplos?

There are many variations of passages of Lorem Ipsum available, but the majority have suffered alteration in some form, by injected humour, or randomised words which don't look even slightly believable. If you are going to use a passage of Lorem Ipsum, you need to be sure there isn't anything embarrassing hidden in the middle of text. All the Lorem Ipsum generators on the Internet tend to repeat predefined chunks as necessary, making this the first true generator on the Internet. It uses a dictionary of over 200 Latin words, combined with a handful of model sentence structures, to generate Lorem Ipsum which looks reasonable. The generated Lorem Ipsum is therefore always free from repetition, injected humour, or non-characteristic words etc.

References

Yoshua Bengio. *Artificial Neural Networks and Their Application to Sequence Recognition*. PhD thesis, McGill University, Montreal, Que., Canada, Canada, 1991. UMI Order No. GAXNN-72116 (Canadian dissertation).

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157 – 166, 1994. ISSN 10459227. URL http://dx.doi.org/10.1109/72.279181. Gradient based learning algorithms;Information latching;Input/output sequences;Learning algorithms;Parametric dynamical system;Recurrent neural network training;Temporal contingencies;.

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 0387310738.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735 – 80, 1997. ISSN 0899-7667. URL http://dx.doi.org/10.1162/neco.1997.9.8.1735.

Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, and Jürgen Schmidhuber. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. 2001. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.24.7321.

References

Yoshua Bengio. *Artificial Neural Networks and Their Application to Sequence Recognition.* PhD thesis, McGill University, Montreal, Que., Canada, Canada, 1991. UMI Order No. GAXNN-72116 (Canadian dissertation).

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157 – 166, 1994. ISSN 10459227. URL http://dx.doi.org/10.1109/72.279181. Gradient based learning algorithms;Information latching;Input/output sequences;Learning algorithms;Parametric dynamical system;Recurrent neural network training;Temporal contingencies;.

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics).* Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 0387310738.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735 – 80, 1997. ISSN 0899-7667. URL http://dx.doi.org/10.1162/neco.1997.9.8.1735.

Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, and Jürgen Schmidhuber. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. 2001. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.24.7321.