CHAPTER 2

An Introduction to Neural Networks

Abstract|Deep Neural Networks are at the forefront of many state-of-the art approaches to Natural Language Processing (NLP). The field of NLP is currently awash with papers building on this method, to the extent that it has quite aptly been described as a tsunami (Manning, 2015). While a large part of the field is familiar with this family of learning architectures, it is the intention of this thesis to be available for a larger audience. Hence, although the rest of this thesis assumes familiarity with neural networks, this chapter is meant to be a foundational introduction for those with limited experience in this area. The reader is assumed to have some familiarity with machine learning and NLP, but not much beyond that.

We begin by exploring the basics of neural networks, and look at the three most commonly used general architectures for neural networks in NLP: Feed-forward Neural Networks, Recurrent Neural Networks, and Convolutional Neural Networks. Some common NLP scenarios are then outlined together with suggestions for suitable architectures.

2.1 Introduction

The term *deep learning* is used to refer to a family of learning models, which represent some of the most powerful learning models available today. The power of this type of model lies in part in its intrinsic hierarchical processing of input features, which allows for learning representations at multiple levels of abstraction (LeCun et al., 2015). This type of model is commonly referred to by several umbrella terms, such as *deep learning*, and *(deep) neural networks*.[1] In this chapter, I aim to introduce the basic concepts of NNs, at a level sufficient to understand the work in this thesis. The following sections are meant to cover the most basic workings in an intuitive, and theoretically supported, manner. In addition to this background, I give an overview of the scenarios in which different recurrent NN architectures might be suitable in NLP (Section 2.4.2).

History

Neural networks have a long history, and have been popular in three main waves.[2] In the first wave, roughly between the $40s - 60s$, they appeared under the moniker of *cybernetics* (e.g., Wiener, 1948), inspired by the Hebbian learning rule (Hebb, 1949). In this wave, the *perceptron* was first outlined (Rosenblatt, 1957), which is still relatively popular today (see Section 2.3). Next, in the $80s$ and $90s$, *connectionism* was on the rise. In this wave, the algorithm for backwards propagation of errors (Rumelhart et al., 1985) was described, which is at the core of how neural networks are trained (see Section 2.3.3).

[1] While some make a distinction between *deep* and non-deep neural networks (NNs) depending on the amount of layers used, there is no real consensus on where the line between these models should be. For the sake of consistency, I attempt to refer to this family of models as NNs, or some specification thereof, as consistently as possible.

[2] Only a very brief overview of the history is given here. For more details, the reader is referred to, e.g., Wang et al., 2017, or Goodfellow et al., 2016.

After an AI winter lasting roughly from 97 to 06, we finally arrive at the current wave (or tsunami), in which the term *deep learning* is favoured. This wave was initiated by works on deep belief networks (Hinton et al., 2006), and has been the subject of much attention after successes in, e.g., reducing error rates in some tasks by more than 50% (LeCun et al., 2015). The recent advances made in the current wave further include breakthroughs in both recognition (He et al., 2016) and generation (**?**) of images, in NLP tasks such as machine translation (Bahdanau et al., 2014; Wu et al., 2016) and parsing (Chen and Manning, 2014), as well as in the strategic board game Go (**?**), and the first-person shooter Doom (**?**).

2.2 Representation of NNs, terminology, and notation

Before embarking upon this journey and exploring the wondrous world of NNs, it is necessary to equip ourselves with some common ground in terms of terminology, notation, and how NNs are generally represented in this thesis. Figure 2.1 contains a NN, which we will go through in detail.

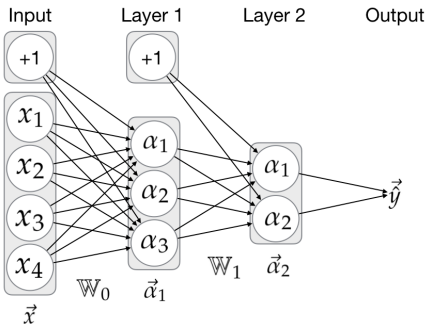

Figure 2.1: A basic Neural Network

First, note that the network is divided into three vertical slices. Each such slice represents a *layer*, marked by a light grey field. Each layer contains one or more white circles, each representing a *unit*, or *neuron*.[3] In this network, each unit has a connection to every unit in the following layer. These connections are represented by arrows, which denote some weighting of the output of the unit at the start of the arrow, for the input of the unit at the end of the arrow. Each layer can be described mathematically as a vector of *activations*. In the case of the first layer (the input layer), these activations are equal to the input (i.e. $\vec{\alpha}_0 = \vec{x}$). The final layer encodes the output of the network, which is denoted by $\hat{y}$. Each layer up until the final layer also contains a special unit, marked by $+1$, which is called the *bias* unit.[4] The collection of all arrows between two layers, can be described mathematically as a matrix of weights (e.g., $\mathbb{W}_0$).[5] The application of the weight matrix to the input of the network can be described in linear algebraic notation as

$$\vec{z}_1 = \mathbb{W}_0\vec{x} = \mathbb{W}_0\vec{\alpha}_0, \tag{2.1}$$

where $\vec{z}_1$ is the vector (i.e. a series of numbers) resulting from this linear transformation. The $\vec{z}$-vectors can be referred to as pre-activation vectors. Each hidden layer thus first encodes the sum of the multiplications of each of the activations in the previous layer by some weight. For instance, if we set $\mathbb{W}_0$ to be matrix of ones, then the pre-activation value of the first unit in the first hidden layer $z_0^1 = \Sigma_{i=0}^4 \vec{x}_i$. The final piece of the puzzle is to calculate the output of each unit in the layer, by applying an activation function to the

[3] In this thesis, the term *unit* is preferred. While this is conventional in much of NLP, it is also debatable whether borrowing terminology for neural networks from neuroscience is motivated at all (this is discussed further in Section 2.7).

[4] Although this can be discussed at length, suffice it to say that including bias units facilitates learning.

[5] We will cover ways in which to learn these weights later in this chapter (Section 2.3.3).

pre-activation vector,

$$\vec{\alpha}_1 = \sigma(\vec{z}_1), \tag{2.2}$$

where σ is some non-linear activation function.[6] Essentially, this is all that a basic FFNN is – a series of matrix-vector multiplications, with non-linearities applied to it. Now, how can this be used to solve problems, and how does the network learn to do this? The answers to these questions will be made clear in the course of the following few pages.

Notation

Before we continue, a brief note on the notation used in this thesis. Scalars are represented with lower case letters (x, y), vectors are represented with lower case letters with arrows ($\vec{x}, \vec{y}$), and matrices are represented with blackboard upper case letters ($\mathbb{X}, \mathbb{Y}$). Subscripts are used to denote the layer number, and where necessary, a superscript is used to denote indexation, to indicate the unit number in the case of deep networks. For instance, α_i^j indicates the activation of unit j in layer i, and $\mathbb{W}_i$ indicates the weight matrix for layer i. Activation functions (Section 2.3.2) are denoted with σ, occasionally subscripted with the actual function used (σ_{ReLU}).

2.3 Feed-forward Neural Networks

A Feed-forward Neural Network (FFNN), also known as a multilayer perceptron, is perhaps the most basic variant of neural networks, and is the kind depicted in the previous figure. As mentioned, a neural network can be seen as a collection of non-linear functions applied to a collection of matrices and vectors, thus mapping from one

[6] Traditionally the activation function (σ) used is some *sigmoidal* function, such as the logistic function. However, many functions are suitable, given that they satisfy certain properties (see Section 2.3.2).

domain (e.g., words) to another (e.g., PoS tags). Let us consider a concrete example, in which $\mathbb{X}$ contains information about the current weather, and $\mathbb{Y} = \{0, 1\}$ denotes human-annotated labels denoting whether or not the weather is considered good. In this case, $\mathbb{X}$ contains several variables, each representing a certain type of weather (e.g. x_1 = calm weather, and x_2 = sunny weather).[7] Table 2.1 represents the weather judgements of this example, where $y = 1$ indicates good weather, and $y = 0$ indicates not-so-good weather.

Table 2.1: Weather appraisal (mimicking the logical AND function).

Calm (x_1)	**Sunny (x_2)**	**Label (y)**
0	0	0
0	1	0
1	0	0
1	1	1

The table shows the judgements of someone who considers weather to be good (i.e. $y = 1$) only when it is both calm *and* sunny (i.e. the logical AND function). Let us now consider our second neural network, which can solve the problem of determining whether the weather is good, based on this person's judgements, in Figure 2.2.

Table 2.2: Weather appraisal by the neural network in Figure 2.2.

Calm (x_1)	**Sunny (x_2)**	z_1	$\sigma(z_1) = \sigma(\alpha_1) = \hat{y}$	**Label (y)**
0	0	-15	≈ 0	**0**
0	1	-5	≈ 0	**0**
1	0	-5	≈ 0	**0**
1	1	5	≈ 1	**1**

Applying the calculations detailed in the previous section to this net-

[7] Note that we begin numbering of features with 1, as the index 0 is reserved for the bias terms.

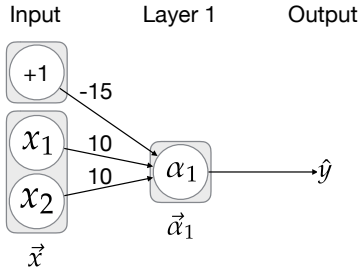

Figure 2.2: A Neural network coding the AND logical function.

work yields the results shown in Table 2.2. If only one of x_1, x_2 is active, the activation α_1 is approximately 0, while the activation is 1 if both x_1 and x_2 are active. We get these values, by applying the perhaps most commonly used activation function to z. This is the logistic function, defined as

$$f(z) = \frac{1}{1 + e^{-z}}, \tag{2.3}$$

where e is Euler's number. Plotting this function, yields the graph in Figure 2.3. Hence, the value of $f(x)$ approaches 0 when $x < 0$, and approaches 1 when $x > 0$.

The simple neural network considered here, is what is also referred to as a *perceptron*, although, technically, a perceptron uses the *step* function as its activation function – in other words, if $x < \lambda$ where λ is some threshold, then $f(x) = 0$, and if $x > \lambda$, then $f(x) = 1$ (Rosenblatt, 1957). This is a very simple and useful architecture, but there are many problems which can not be easily solved by a perceptron, such as those in which the decision boundary to be learned is nonlinear. Take, for instance, the problem given in Table 2.3.

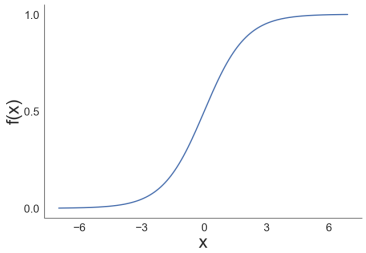

Figure 2.3: Plot of the logistic function.

Table 2.3: Weather appraisal (mimicking the logical XOR function).

Snowy (x_1)	**Sunny (x_2)**	**Label (y)**
0	0	0
1	0	1
0	1	1
1	1	0

This table shows the labels provided by some annotator who considers weather to be good (i.e. $y = 1$) if it is *either* snowy *or* sunny – but not both (i.e. the logical XOR function). As mentioned, a single unit (i.e. a perceptron) is not able to learn this decision boundary (Minsky and Papert, 1988).[8] Let us now have a look at a neural network which encodes this function. This is depicted in Figure 2.4. For the sake of clarity, all weights between x_1, x_2 and a_1, a_2 are set to 5, but only one of these weights is shown.

Applying the calculations detailed in the previous section to this

[8] This is frequently cited as a potential catalyst for the *the AI winter*, in which funding and interest in artificial intelligence was at a low. Nonetheless, it was known at the time that the XOR problem could be solved by a neural network with more hidden units (Rumelhart et al., 1985).

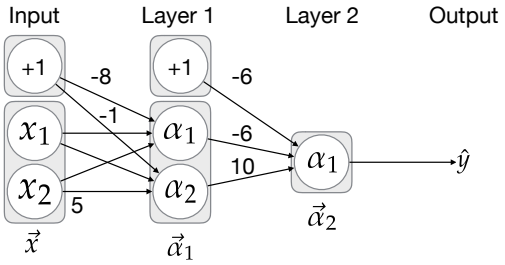

Figure 2.4: A Neural network coding the XOR logical function.

network yields the results shown in Table 2.4. If both or none of x_1, x_2 are active, then the network will output 0, whereas if one and only one of x_1, x_2 are active, the network will output 1. Exactly what we want!

Table 2.4: Weather appraisal by the neural network in Figure 2.4.

Calm (x_1)	**Sunny (x_2)**	z_1^1	z_1^1	z_2^1	$\sigma(z_2^1) = \alpha_1^2 = \hat{y}$	y
0	0	-8	-1	-6	≈ 0	**0**
1	0	-3	4	4	≈ 1	**1**
0	1	-3	4	4	≈ 1	**1**
1	1	3	9	-2	≈ 0	**0**

The last two examples have shown how neural networks can encode certain simple functions. It turns out that neural networks can do much more than this, and are in fact *universal function approximators* (Cybenko, 1989; Hornik et al., 1989). What this means, is that no matter the function, there is guaranteed to be a neural network with a single layer and a finite number of hidden units, such that for each potential input x, (a close approximation of) the value $f(x)$ is output

from the network.

The class of networks discussed here is useful for many tasks in NLP, and can be used as a simple replacement for other classifiers. Furthermore, they can be expanded by adding more units to each layer, or by adding more layers. This allows such networks to learn to solve interesting NLP problems, like language modelling (Bengio et al., 2003; Vaswani et al., 2013), and sentiment classification (Iyyer et al., 2015).

Before going into other NN architectures, we will first consider some of the inner workings of NNs. This includes how we represent our input, how weights are obtained, and finally some limitations which motivate the use of more complex architectures than FFNNs.

2.3.1 Feature representations

In many NLP problems, we are interested in mapping from some textual language representation (x) to some label (y). This textual representation can take many forms, both depending on the problem at hand, and on the choices made when approaching the problem. As an example, say we are interested in doing sentiment analysis, i.e., given a text (x), predict whether the text is positive or negative in sentiment (y). The perhaps simplest way of representing the text is to count the occurrences of each word in the text. The intuition behind this is that if a text contains many negative words (*horrible, bad, appalling*), it is more likely to be negative in sentiment than if it contains many positive words (*wonderful, good, exquisite*). Since these *features* (i.e. counts of each word) need to be passed to an FFNN, they need to be represented as a single fixed-length vector $\vec{x}$. What one might then do, is to assign an index to each unique word, and assign the count of each word to that index in the vector. This can be referred to as a *bag-of-words* model.

Although this type of feature representation is sufficient for some problems, and is traditionally used extensively, more recent develop-

ments include using other types of representations based on distributional semantics. This is covered in more detail in the next chapter, in Section 3.2.5.

2.3.2 Activation Functions

As stated in Section 2.3, each hidden unit applies an activation function to the sum of its weighted inputs. While many functions might be used, an activation function should have certain properties. One such property is that the function needs to be non-linear. It is for this kind of function that it has been proven that a two-layer neural network is a universal function approximator (Cybenko, 1989). Additionally, the function should be monotonic, as the error surface associated with a single-layer model will then be convex (Wu, 2009).[9] There are several other important properties, which are not covered here. Some of the more commonly used activation functions are listed in Table 2.5.

Table 2.5: Commonly used activation functions in neural networks

Name	**Function**
Logistic (aka. sigmoid)	$f(x) = \frac{1}{1+e^{-x}}$
Hyperbolic Tangent (tanh)	$f(x) = \frac{2}{1+e^{-2x}} - 1$
Rectified Linear Unit (ReLU)	$f(x) = \begin{cases} 0 & \text{for} \quad x < 0 \\ x & \text{for} \quad x \geq 0 \end{cases}$
Leaky ReLU	$f(x) = \begin{cases} 0.01x & \text{for} \quad x < 0 \\ x & \text{for} \quad x \geq 0 \end{cases}$
Softmax	$f(\vec{x})_i = \frac{e^{x_i}}{\sum_{k=1}^{K} e^{x_k}}$ for $i = 1, \ldots, K$

[9] A monotonic function is either non-increasing or non-decreasing in its entirety.

The traditionally popular logistic function was already described in Figure 2.3. We will now consider some other commonly used activation functions. Activation functions turn out to be one of the areas in which biological inspiration has been directly applicable to the development of neural networks. The Rectified Linear Unit (ReLU) is in fact remarkably similar to what happens in a biological neuron (Hahnloser et al., 2000; Hahnloser and Seung, 2001). That is to say, when the input is below a certain threshold, the neuron does not fire, and when the input is above this threshold, the neuron fires with a current proportional to its input. ReLUs have been found to make it substantially easier to train deep networks (Nair and Hinton, 2010), and are currently very widely used. One disadvantage of ReLUs is that they can wind up in a state in which they are inactive for almost all inputs, meaning that no gradients flow backward through the unit. This, in turn, means that the unit is perpetually stuck in an inactive state, which at a large scale can decrease the network's overall capacity. This is mitigated by using leaky ReLUs, for which even input < 0 leads to some activity, allowing for error propagation given any input value.

The softmax function is generally only used at the final layer in classification problems, as it yields a probability distribution based on its input.

2.3.3 Learning

Learning in an FFNN happens in two phases. First, in the forward propagation pass, the network sends a given input through the network, and produces some output. Then, the *error* of this output is calculated, as compared to some target label, and this error is sent back through the network, updating the weights of the network so as to make a more accurate prediction given the input in the next

forward pass.[10]

We have already seen the largest part of the forward pass, as in the examples with the AND and XOR functions in Section 2.3. The only remaining part of the forward pass, is how the error of the network is calculated – for this, a loss function is necessary.

Loss functions

The loss functions used in NNs, generally fall into two classes – those used for classification problems (i.e. when attempting to predict some discrete class label, out of a finite set of labels), and those used for regression problems (i.e. when attempting to predict some continuous score). Classification is one of the most common cases in NLP (e.g. in POS tagging, NER, language identification, and so on). In such cases, the activation function of the final layer is the softmax function, which allows for interpreting the layer's activations as a probability distribution over the labels under consideration. Most often, the cross-entropy between this predicted probability distribution and the target probability distribution is used to calculate the error, or loss L, such that

$$L_{cross-entropy}(\vec{\hat{y}}, \vec{y}) = -\sum_i \vec{y}_i \log \vec{\hat{y}}_i, \tag{2.4}$$

where L denotes the loss function, $\vec{y}$ is the target probability distribution over labels, $\vec{\hat{y}}$ is the model's predicted model distribution given an input x. A high error thus indicates that the predicted probability distribution is not consistent with the target probability distribution, and therefore changes should be made accordingly in the backward propagation pass.

[10] This is referred to as *backward propagation of errors*, and is covered later in this section.

Another loss function, common in regression, is the squared error function, defined as

$$L_{squared} = (\hat{y} - y)^2, \tag{2.5}$$

where $\hat{y}$ is the predicted label, and y is the true label. This function is commonly used in regression, and is especially handy for explaining backpropagation, as in the next section.

Backpropagation

Backward propagation of errors, or *backprop* (Rumelhart et al., 1985; LeCun et al., 1998b), is an algorithm for calculating the gradient of the loss function, for each weight. The gradient can, in turn, be used to update the weights by using an optimisation algorithm, such as gradient descent (discussed further in Section 2.3.3). Intuitively seen, gradient-based methods operate by viewing the errors as a geometric area, and use the slope of the area in which they are (i.e., the gradient) in order to shift weights towards obtaining an error in a minimum of this area, as in Figure 2.5.

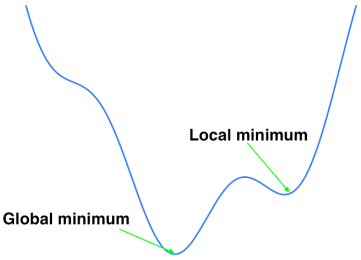

Figure 2.5: Non-convex error surface

Backprop relies on the fact that the partial derivative of the error of a certain weight $\mathbb{W}_i^j$, with respect to the loss function, can be

easily calculated if we know the partial derivative of the outputs in the layer following that weight. It turns out that this is, indeed, the case, as the derivative of the output layer is quite easily obtained. The output error of a given unit, δ_i, is calculated as

$$\delta_i = \begin{cases} \alpha_i(1-\alpha_i)(\alpha_i - y_i) & \text{for output units } i, \\ \alpha_i(1-\alpha_i)(\alpha_i \sum_{\ell \in L} \delta_\ell \mathbb{W}_{i\ell}) & \text{for other units } i, \end{cases} \tag{2.6}$$

where α_i is the activation of the current unit, y_i is the target output, L is the collection of all units receiving input from the current unit, and $\mathbb{W}_{i\ell}$ is the weight from the current unit to unit ℓ. Let us consider a concrete example, and go through the forward pass, calculation of the error, and the backward pass. The network in Figure 2.6 shows a neural network with its weights.

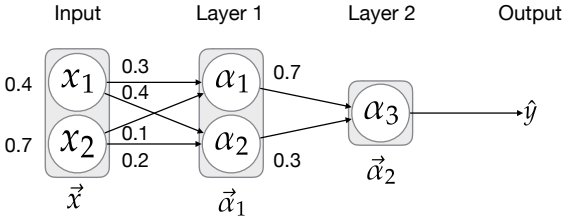

Figure 2.6: A neural network with weights for our backpropagation example

Assuming that the activation function used is the logistic function, the following calculations hold:

$$\begin{aligned} \alpha_1 &= \sigma(x_1 \times 0.3 + x_2 \times 0.1) = \sigma(0.4 \times 0.3 + 0.7 \times 0.1) = 0.547, \\ \alpha_2 &= \sigma(x_1 \times 0.4 + x_2 \times 0.2) = \sigma(0.4 \times 0.4 + 0.7 \times 0.2) = 0.574, \\ \alpha_3 &= \sigma(\alpha_1 \times 0.7 + \alpha_2 \times 0.3) = \sigma(0.19 \times 0.7 + 0.3 \times 0.3) = 0.635. \end{aligned} \tag{2.7}$$

Assuming that the target output is $y = 0.4$, we can now calculate the error. Applying equation 2.6, we can obtain the error of the output,

namely

$$\begin{aligned}\delta_3 &= \alpha_3(1-\alpha_3)(\alpha_3 - y),\\ &= 0.635(1-0.635)(0.635-0.4),\\ &= 0.054.\end{aligned} \tag{2.8}$$

The errors of the two hidden units can also be calculated, yielding

$$\begin{aligned}\delta_2 &= \alpha_2(1-\alpha_2)(\alpha_2\delta_3\mathbb{W}_{2\ell}),\\ &= 0.547(1-0.547)(0.547\times 0.027\times 0.3),\\ &= 0.001,\end{aligned} \tag{2.9}$$

and

$$\begin{aligned}\delta_1 &= \alpha_1(1-\alpha_1)(\alpha_1\delta_3\mathbb{W}_{1\ell}),\\ &= 0.547(1-0.547)(0.547*0.027*0.7),\\ &= 0.002.\end{aligned} \tag{2.10}$$

We now need to update the weights used, via gradient descent. This can be done by shifting the weights with some constant with respect to the error obtained,

$$\Delta\mathbb{W}_{ij} = -\gamma\alpha_i\delta_j, \tag{2.11}$$

where $\Delta\mathbb{W}_{ij}$ is the amount with which to change $\mathbb{W}_{ij}$ (i.e., the weight between the firing and receiving unit), γ is some learning rate, α_i is the activation of the firing unit, and and δ_j is the error of the receiving unit. Hence, if we set $\gamma = 1$, the changes of the weights are calculated as

$$\begin{aligned}\Delta\mathbb{W}_{x_1,\alpha_1} &= -\gamma x_1\delta_1 = -1\times 0.4\times 0.002 = -0.0008,\\ \Delta\mathbb{W}_{x_1,\alpha_2} &= -\gamma x_1\delta_2 = -1\times 0.4\times 0.001 = -0.0004,\\ \Delta\mathbb{W}_{x_2,\alpha_1} &= -\gamma x_2\delta_1 = -1\times 0.7\times 0.002 = -0.0014,\\ \Delta\mathbb{W}_{x_2,\alpha_2} &= -\gamma x_2\delta_2 = -1\times 0.7\times 0.001 = -0.0007,\\ \Delta\mathbb{W}_{\alpha_1,\alpha_3} &= -\gamma\alpha_1\delta_3 = -1\times 0.547\times 0.054 = -0.030,\\ \Delta\mathbb{W}_{\alpha_2,\alpha_3} &= -\gamma\alpha_2\delta_3 = -1\times 0.574\times 0.054 = -0.031.\end{aligned} \tag{2.12}$$

Using the new weights yields the following activations in the next forward pass, given the same input:

$$\begin{aligned}
\alpha_1 &= \sigma(0.4 \times (0.3 + \Delta\mathbb{W}_{x_1,\alpha_1}) + 0.7 \times (0.1 + \Delta\mathbb{W}_{x_2,\alpha_1}) = 0.547, \\
\alpha_2 &= \sigma(0.4 \times (0.4 + \Delta\mathbb{W}_{x_2,\alpha_1}) + 0.7 \times (0.2 + \Delta\mathbb{W}_{x_2,\alpha_2}) = 0.574, \\
\alpha_3 &= \sigma(0.547 \times (0.7 + \Delta\mathbb{W}_{\alpha_1,\alpha_3}) + 0.574 \times (0.3 + \Delta\mathbb{W}_{\alpha_2,\alpha_3}) = 0.627,
\end{aligned} \tag{2.13}$$

and the output error

$$\begin{aligned}
\delta_3 &= \alpha_3(1 - \alpha_3)(\alpha_3 - y), \\
&= 0.318(1 - 0.318)(0.318 - 0.4), \\
&= 0.053,
\end{aligned} \tag{2.14}$$

which is smaller than the previous error where $\delta_3 = 0.054$. This process is repeated with other training examples, until some criterion is reached, such as a sufficiently low average loss.

Optimisation Methods

Backpropagation, as described in the previous section, can provide us with the derivatives of the error surface. This can be used in a variety of ways to update the weights. What we just saw, in Equation 2.11, is known as gradient descent. One of the most commonly used optimisation methods is Stochastic Gradient Descent (SGD). In SGD, a minibatch of n samples is drawn from the training set, the gradient is calculated based on this batch, and the weights are then updated accordingly (Bottou, 1998). Other algorithms, such as AdaGrad (Duchi et al., 2011) and RMSProp (Hinton, 2012), learn and adapt the learning rate (γ) for each weight. Modifying the learning rate in this manner can both increase the rate at which the error decreases, and lead to lower overall errors. A recent and increasingly popular optimisation method is Adam, which is similar to RMSProp and yields better results on a many problems (Kingma and Ba, 2014). The choice

of optimisation method is not all that straightforward, and no real consensus exists for how this should be done (Schaul et al., 2014). Hence, commonly, trial-and-error is applied in order to make this choice, by experimentally investigating performance on a development set.

Finding the global minimum

The goal of an optimisation algorithm is to find the global minimum, as shown in Figure 2.5. What so-called *gradient-based* optimisation algorithms do, is to calculate the derivative with respect to this error surface, and shift the weights so as to move towards the closest of all local minima. Such local minima can, however, be the source of a host of problems, if the loss is high compared to the global minimum. This is a frequently occurring issue, and it is possible to construct small neural networks in which this scenario appears (Sontag and Sussmann, 1989; Brady et al., 1989; Gori and Tesi, 1992). It turns out, however, that practically speaking, when considering larger neural networks, it is not particularly important to find the global minimum. This has to do with the fact that, in the case of supervised learning with deep neural networks, most local minima appear to have a low loss function value, roughly equivalent to that of the true global minimum (Saxe et al., 2013; Dauphin et al., 2014; Goodfellow et al., 2015; Choromanska et al., 2015).

Parameter Initialisation

There are several methods for initialising the weights in a neural network. Naively, one might think to set the all weight matrices $\mathbb{W} = 1$, however due to how backpropagation works, this will result in all hidden units representing the same function, and receiving the exact same weight updates. Therefore, some random process is required. Common methods include those introduced by Glorot and

Bengio (2010), and Saxe et al. (2013). When employing the ReLU activation function, He et al. (2015b) show that weights should be initialised based on a Gaussian distribution with standard deviation $\sqrt{\frac{2}{d_{in}}}$, where d_{in} is the input dimensionality. In the case of recurrent neural networks, which are covered in Section 2.4, particular care needs to be taken, and weight initialisation is often done using orthogonal matrices (cf. Goodfellow et al.[p.404], 2016).

Regularisation in Neural Networks

One of the most common problems when training an ML system in general, is that of overfitting – and neural networks are no exception. Overfitting occurs when the network does not generalise to data outside of the training set, while having a low loss on the training set itself. Generalisation is one of the most important parts of learning, as learning without generalisation is simply the memorisation of a training set. A model which has only memorised the training set is of little practical value, as it will most likely fail miserably on unseen examples. In order to avoid overfitting, regularisation techniques are typically employed. The probably most common regularisation technique used today, is dropout (Srivastava et al., 2014). In dropout, every activation has a probability p of not being included in the forward and backward passes, during training. This procedure leads to significantly lower generalisation error, as the network needs to be more robust, and less reliant on specific units. In the case of recurrent neural networks, which are covered next, specific variants of dropout exist. such as recurrent dropout (Semeniuta et al., 2016), or variational dropout (Gal and Ghahramani, 2016) in which the same dropout mask is used for each time step. Another commonly used manner of regularisation is *weight decay*, in which the magnitudes of weights are decreased according to some criterion (Krogh and Hertz, 1992).

2.4 Recurrent Neural Networks

Although FFNNs are suitable for many problems, they do not take the structure of the input into account. Although it is possible to attempt to enforce this in such a network, this has several disadvantages, such as the fact that the amount of parameters which need to be tuned can become prohibitively large. Luckily, there are architectures for dealing with structure, as this is not entirely unimportant when considering natural language. Two such approaches are covered in the following sections.

Recurrent Neural Networks (RNNs) are an extension of feed-forward neural networks, which are designed for sequential data (Elman, 1990). They can be thought of as a sequence of copies of the same FFNN, each with a connection to the following time step in the sequence, sharing parameters between time steps. RNNs take a sequence of *arbitrary length* as input ($x_1, x_2, \ldots, x_t$), and return another sequence ($\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_t$). Each x_t in the input sequence is a vector representation of element n_t in the sequence. Each $\hat{y}_t$ in the output sequence can take advantage of information in the sequence up to step t in the input sequence. This is illustrated in Figure 2.7. Each layer is shown as containing only one unit, which is here meant as an abstraction depicting the entire internal representation of the RNN. The left side of the figure depicts an FFNN with a loop, whereas the right side shows the *unrolled* version of the network. The output of the hidden layer is passed as an input to the hidden layer in the next time step.

An RNN is essentially a group of FFNNs with connections to one another. This connection is a sort of loop, going from the hidden layer of the network at time x_t to the hidden layer at x_{t+1}. In other words, $\vec{y_t}$ is calculated as

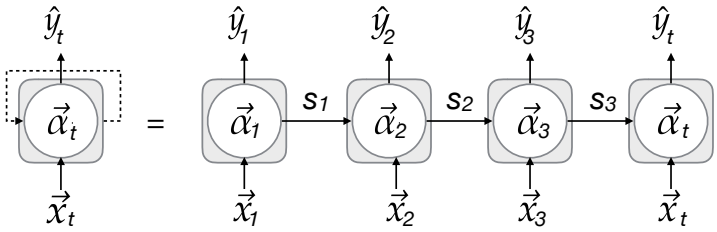

Figure 2.7: A simple RNN with a connection from the hidden state of the previous time step to the current side step. Left side shows the FFNN with the loop, whereas the right side shows the unrolled network.

$$\begin{aligned} \vec{z}_t &= \mathbb{W}_s \vec{x}_t + \mathbb{U}\vec{s}_{t-1}, \\ \vec{s}_t &= \sigma_s(\vec{z}_t), \\ \vec{\hat{y}}_t &= \sigma_y(\mathbb{W}_y \vec{s}_t), \end{aligned} \tag{2.15}$$

where $\mathbb{W}_s$ is the matrix of weights for the current time step's input ($\vec{x}_t$), $\mathbb{U}$ is a weight matrix for the connections from the previous time step, $\vec{s}_t$ is a state vector representing the history of the sequence, t is the index of the current time step, $\mathbb{W}_y$ is the matrix of weights for the output, and the rest is defined as for FFNNs. This is what is also referred to as an Elman net, or a Simple RNN (Elman, 1990). The advantage of having access to $\vec{s}$, is that the network can take advantage of preceding information when outputting $\vec{\hat{y}}_t$. For instance, in the case of POS tagging, if the current input is *fly*, and the state vector shows that the previous word was *to*, we most likely want to output the tag *verb*. Hence, in this way, the prediction at each time step is conditioned on the inputs in the entire preceding sequence. There are also variants of this, in which the net's outputs are used to cal-

culate the state vector, as in the case of Jordan nets (Jordan, 1997), which is defined such that

$$\vec{z}_t = \mathbb{W}_s \vec{x}_t + \mathbb{U}\vec{\vec{y}}_{t-1}. \tag{2.16}$$

Although RNNs, in theory, can learn long dependencies (i.e. that an output at a certain time step is dependant on the state at a time step far back in the history), and can handle input sequences of arbitrary length, they are in practice heavily biased to the most recent items in the given sequence, and thus difficult to train on long sequences with long dependencies (Bengio et al., 1994). For instance, in the case of language modelling, given a sentence such as *My mother is from Finland, so I speak fluent* ..., it is quite likely that the omitted word should be *Finnish*. However, as the distance between such dependencies grows, it becomes increasingly difficult for an RNN to make use of such contextual information. In general, this is because deep neural networks suffer from having *unstable* gradients, as the gradients calculated by backprop (Section 2.3.3) are dependant on the output of the network, which can be quite far away from the first layers in the network. One problem with this is that this can lead to *vanishing* gradients (i.e. the gradient becomes very small). This happens since the gradient in early layers of the network are the result of a large number of multiplication operators on numbers < 1. One might consider the fact that, since these multiplications involve the weights in the network, we might just set the weights to be really large. Although this might seem like a good idea, this will likely lead to the converse of the issue one is trying to avoid, namely that of *exploding* gradients. Since most common optimisation methods are gradient-based, this is problematic (see Section 2.3.3 for optimisation details).

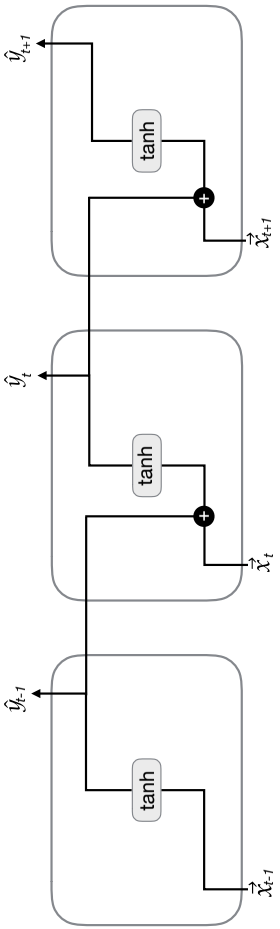

Figure 2.8: Internal view of an RNN in three time steps.

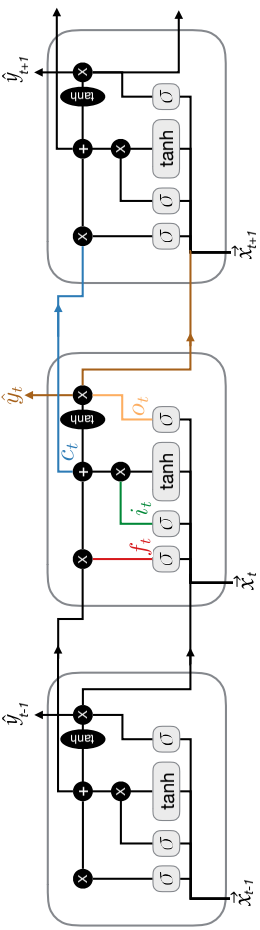

Figure 2.9: Internal view of an LSTM in three time steps.

2.4.1 Long Short-Term Memory

Previous work has attempted to solve this problem by adapting the optimisation method used (Bengio et al., 2013; Pascanu et al., 2013; Sutskever et al., 2014), however the more successful approach has been to modify the neural network architecture itself. Because of such efforts, there are several types of RNNs specifically designed to cope with this issue, essentially by enforcing a type of protection of the memory of the history of the input sequence, storing and maintaining important features, while neglecting and forgetting unimportant features. One such method, namely Long Short-Term Memory (LSTM), was described in Hochreiter and Schmidhuber (1997), and saw an explosion in popularity around 2014, following several influential papers (e.g. Sundermeyer et al. (2012); Sutskever et al. (2014); Dyer et al. (2015)). An LSTM is an extension of RNNs, with *memory cells*, engineered to cope with the issue of unstable gradients, and have been shown to be able to capture long-range dependencies (Hochreiter and Schmidhuber, 1997; Cho, 2015). For an overview of the many variations of LSTMs which appear in the literature, see Greff et al. (2017).

Whereas an RNN only has a single internal layer (Figure 2.8), typically with a *tanh* (hyperbolic tangent) activation, an LSTM is somewhat more complicated (Figure 2.9). An LSTM contains gates, denoted in the figure by the σ layers, which are used to modify the extent to which old information is remembered or forgotten. Part of the explanation for LSTMs involves the observation that they, on the surface, can be seen as a combination of Elman nets and Jordan nets, in that both the cell state and the hypothesis are passed between

states. In detail, an LSTM is implemented as follows

$$
\begin{aligned}
f_t &= \sigma(\mathbb{W}_f x_t + \mathbb{U}_f \hat{y}_{t-1} + b_f), \\
i_t &= \sigma(\mathbb{W}_i x_t + \mathbb{U}_i \hat{y}_{t-1} + b_i), \\
o_t &= \sigma(\mathbb{W}_o x_t + \mathbb{U}_o \hat{y}_{t-1} + b_o), \\
c_t &= f_t \circ c_{t-1} + i_t \circ \sigma_c(\mathbb{W}_c x_t + \mathbb{U}_c \hat{y}_{t-1} + b_c), \\
\hat{y}_t &= o_t \circ \sigma(c_t),
\end{aligned}
\tag{2.17}
$$

where f_t represents the output of the *forget gate*, i_t represents the output of the *input gate*, o_t represents the output of the *output gate*, c_t represents the *cell state*, $\hat{y}_t$ represents the output vector, $\mathbb{W}$ and $\mathbb{U}$ represent the weight matrices, x_t represents the current input, and b represents bias units. Each of these parts are covered in detail in the following sections.

Cell state

The cell state, c_t, is the line coded in blue in Figure 2.9. It is similar to the state vector, $\vec{s}$, in a simple RNN, but its content is maintained and protected by three gates. The forget gate's output f_t determines to which extent each feature in the cell state should be kept. This is done by observing the current input x_t, and the previous time step's output $\hat{y}_{t-1}$.

For instance, say we have a POS tagger which at time t observes a word which could be either a noun or a verb (e.g. *fly*). If the previous word was *to*, this would be useful to keep in mind in order to better predict the next tag. Following this time step, however, we might want to forget about this word, when predicting the next tag.

Adding information to the cell state happens in two steps. We first decide on which values in the cell state to update again observing x_t and $\hat{y}_{t-1}$, again deciding for each dimension the extent to which we will add information. This is denoted by the input gate i_t. The vector which is added to the cell state, is calculated by passing x_t and $\hat{y}_{t-1}$

through a non-linearity, marked by tanh in the figure. In the POS tagging example, we want to add the information regarding the preceding determiner. These two vectors are then summed, resulting in our new cell state c_t.

Output

Finally, we need to output a value from the current time step. This output is based on the cell state, c_t, which is first run through a non-linearity (usually tanh), and filtered again by o_t, which also observes x_t and $\hat{y}_{t-1}$, when deciding on which dimensions to keep, and to what extent.

Gated Recurrent Units

In addition to the many LSTM variants (Greff et al., 2017), Gated Recurrent Units (GRUs) represent a different variant of gated RNNs which was independently developed to LSTMs and similar in both purpose and implementation (Cho et al., 2014). The main difference between LSTMs and GRUs is the fact that GRUs do not have separate memory cells, and only include two gates – an update gate, and a reset gate (Chung et al., 2014). This in turn means that GRUs are computationally somewhat more efficient than LSTMs. In practice, both LSTMs and GRUs have been found to yield comparable results (Chung et al., 2014; Jozefowicz et al., 2015). On a general level, the performance of various gated RNN architectures is, at least in the case of large amounts of data, closely tied to the number of parameters (Collins et al., 2017; Melis et al., 2017).

Bi-directionality

Many properties of language depend on both preceding and proceeding contexts, so it is useful to have knowledge of both of these contexts simultaneously. This can be done by using a bi-directional RNN

variant, which makes both forward and backward passes over sequences, allowing it to use both contexts simultaneously for the task at hand (Schuster and Paliwal, 1997; Graves and Schmidhuber, 2005; Goldberg, 2015). Bi-directional GRUs and LSTMs have been shown to yield high performance on several NLP tasks, such as POS tagging, named entity tagging, and chunking (Wang et al., 2015; Yang et al., 2016; Plank et al., 2016).

2.4.2 Common use-cases of RNNs in NLP

In NLP, there are four general scenarios for producing some sort of analysis for a given text. Consider that we have the following sentence as input:

(2.18) *I'm not fussy.*[11]

We might want to analyse this unit as a whole, for instance in order to judge that the text is in English, and not in some other language, or to determine the native language of the person writing it, or the sentiment of the text itself. This can be referred to as a **many-to-one** scenario, since we have several smaller units (e.g. words or characters), which we want to translate into a single score or class, depending on the task at hand.

On the other hand, we might want to analyse the sentence word by word, by assigning, e.g., a part-of-speech (POS) tag or a semantic tag to each word in the sentence. This can be referred to as a **one-to-one** scenario, since every single unit in the text (e.g. each word token) has a direct correspondence to a single tag.[12]

[11] PMB 76/2032, Original source: Tatoeba

[12] The term 'one-to-one' is also used for simple classification cases where there is no sentential context available. We see this as simply being a special case in which the sequence length $= 1$. This is equivalent to the relation between FFNNs and RNNs, in which an FFNN can be seen as a special case of RNNs (or vice versa).

We might also want to carry out some task in which the sentence should be translated to some other form, for instance translating the sentence to German, or some other language. If the sentence was written in some non-standard form of English, we might want to produce a normalised version of the sentence, or in a different setting we might want to generate an inflected form of some word in the same language. This can be referred to as a **many-to-many** scenario, as there is no structural one-to-one correspondence between the input X and the output Y.

A final logically possible case, is the **one-to-many** scenario. This is a very uncommon scenario, as it is not generally the case that one tries to predict several things from an atomic unit. Although one could argue that some tasks fit this scenario, such as caption generation, this is not really a one-to-many scenario, as the image is not an atomic unit, but is read by the NN as a matrix of pixels.

A schematic overview of the three relevant scenarios is given in Figure 2.10. The versatility of these three scenarios is evident when observing the current NLP scene, in which common practise is to cast a problem to fit one of these scenarios, and to then throw a Bi-LSTM at the problem.

Many-to-one

Many NLP tasks deal with going from several smaller units to a single prediction. This essentially means that these units need to be compressed into a single vector, onto which a softmax layer can be applied in order to arrive at a probability distribution over the classes at hand (e.g. a set of languages to identify). For this type of problem, a number of possibilities exist, such as

1. Averaging the vectors representing each unit in the sentence (i.e. average pooling);

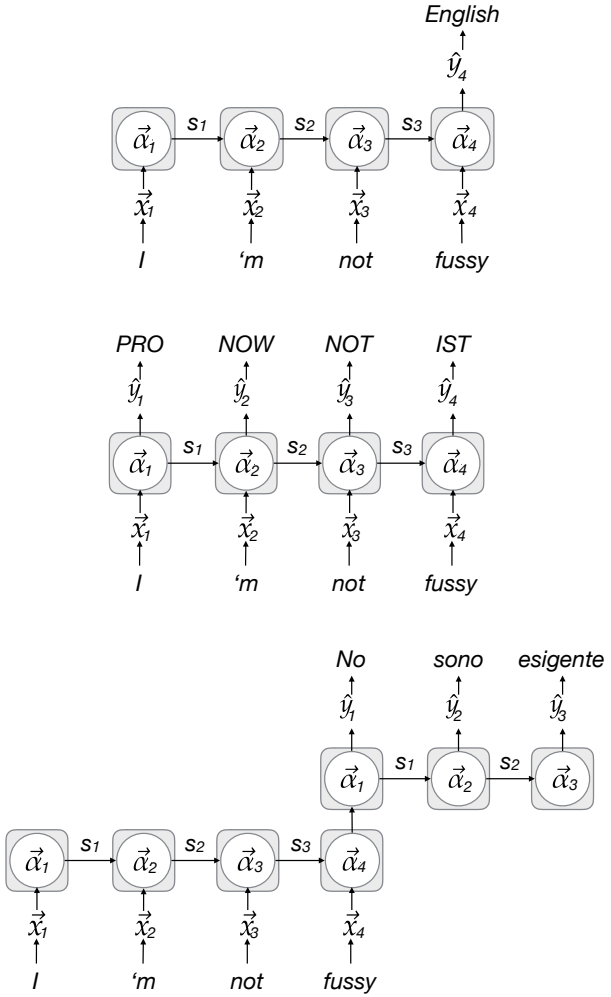

Figure 2.10: Common scenarios in which RNNs are applied in NLP. From top to bottom: many-to-one, one-to-one, and many-to-many.

2. Applying an RNN and using the final state output vector as a representation of the sentence (depicted in Figure 2.10);

3. Applying convolutions in order to arrive at a condensed representation.[13]

Approach 1) is the most simplistic of these, and has been successfully applied in previous work (e.g. Socher et al. (2013a); Zhang et al. (2015a)). The main advantage of this approach is indeed its simplicity, as calculating the mean of the vectorial representations of a sentence is both a very cheap operation, as well as an operation which allows for the application of a simple FFNN on top of this vector. With this approach, however, the structure inherent in the natural language signal is left unexploited.

Approaches 2) and 3) both offer more expressive power as compared to approach 1), as they both take advantage of the structure inherent in the input signal. This is not the case when summing the vectors, which is in a way analogous to a bag-of-words approach. Structure is naturally very important in natural language, so taking advantage of this is good. Although structure in natural language is generally hierarchical, even using the sequential structure is better than assuming no structure at all. There are, however, some recent architectures which do encode the hierarchical structure of language, such as tree LSTMs (Tai et al., 2015), and RNN-Grammars (Dyer et al., 2016).

These three approaches are meant to give an overview of some straight-forward manners of obtaining such representations. Other, more sophisticated approaches, include skip-thought vectors, in which sentence-level representations are learned with the objective of being able to predict surrounding sentences in a document (**?**). A systematic overview of such methods is given by **?**, who conclude

[13] Convolutional Neural Networks are described in Section 2.5.

that the best suited approach depends on the intended application of such representations.

Another use case for this approach is when building hierarchical models, in the sense that one, e.g., may want to have word representations which are aware of what is going on on a sub-word level. For this, one might apply a many-to-one RNN, and use the final state output vector as a word vector (Ballesteros et al., 2015; Plank et al., 2016). Alternatively, one can use convolutions to arrive at this type of word vector, as in dos Santos and Zadrozny (2014).

One-to-one

The one-to-one case is perhaps one of the most common scenarios in NLP. This covers tagging task scenarios, as well as simple classification scenarios. Many NLP tagging tasks have seen relatively large improvements when applying variants of RNNs similarly to what is depicted in Figure 2.10. Recently, gated variants such as LSTMs and GRUs, often in a bi-directional incarnation are applied (Wang et al., 2015; Huang et al., 2015; Yang et al., 2016; Plank et al., 2016). Such RNNs are highly suited for this type of task, as it is highly informative for, e.g., POS tagging to know which words occur both before and after the word at hand. Such dependencies might also have quite large spans, which both LSTMs and GRUs are able to capture well. To contrast with older feature-based models, it would require a fair bit of feature engineering to decide on what types of spans to include in feature representations, lest one wishes to suffer from the sparsity of simply using n-gram features with large values of n.

Many-to-many

This paradigm is also what is frequently referred to as an encoder-decoder architecture in the literature, or sequence-to-sequence learning problems (Bahdanau et al., 2014; Sutskever et al., 2014). A fre-

quent approach here, for instance in machine translation, is to apply an RNN from which one takes the final time step's output to be a representation of the entire sentence, as represented in Figure 2.10, which may seem like a bold thing to do.[14] It turns out that this is in fact often not sufficient, although one can obtain surprisingly good translations this way. However, results improve dramatically when going a step further by incorporating an attentional mechanisms. This is not focussed upon in this thesis, and will not be explained in full detail. Essentially, an attention mechanism can learn which parts of the source sentence to attend to, when producing the target sentence translation. For instance, such a mechanism might learn an implicit weighted word-alignment between the source and target sentences, thus facilitating translation.

Many NLP tasks can be solved with a many-to-many approach. Machine translation has already been mentioned, and has in no small degree been the driving force behind research in this direction. Apart from this, the approach has been applied to morphological inflection (Kann and Schütze, 2016; Cotterell et al., 2016; Östling and Bjerva, 2017; Cotterell et al., 2017), AMR parsing (Barzdins and Gosko, 2016; Konstas et al., 2017; van Noord and Bos, 2017b,a), language modelling (e.g. Vinyals et al., 2015), generation of Chinese poetry (Yi et al., 2016), historical text normalisation (Korchagina, 2017), and a whole host of other tasks.

2.5 Convolutional Neural Networks

Certain machine learning problems, such as image recognition, deal with input data in which spatial relationships are of utmost importance. While simpler image recognition problems, such as handwrit-

[14] *You can't cram the meaning of a whole sentence into a single vector!*
–Ray Mooney, as communicated by Kyunghyun Cho in his NoDaLiDa 2017 keynote (`https://play.gu.se/media/1_xt08m5je`)

ten digit recognition, can be carried out relatively successfully with simple FFNNs, this is often not sufficient. Recall that the input for an FFNN is simply a single vector $\vec{x}$, meaning that the network has no notion of adjacency between, e.g., two pixels. A Convolutional Neural Network (CNN) is a type of network explicitly designed to take advantage of the spatial structure of its input. The origins of CNNs go back to the 1970s, but the seminal paper for modern CNNs is considered to be LeCun et al. (1998a), although other work exists in the same direction (e.g., LeCun et al. (1989); Waibel et al. (1989)). CNNs have been used extensively in NLP, and can in many cases be used instead of an RNN (contrast, e.g., dos Santos and Zadrozny (2014) who use CNNs for character-based word representations, and Plank et al. (2016) who use RNNs for the same purpose).

Although NLP is the focus of this thesis, we will approach CNNs from an image recognition perspective, as this is somewhat more intuitive. This is in part due to the fact that image recognition was the intended application of CNNs upon their conception. On a general level, convolutions can be carried out on input of arbitrary dimensionality. As mentioned, two-dimensional input (e.g. images) were the original target for CNNs. More recent work has extended this to three-dimensional input (e.g. videos). In the case of NLP, it is often the case that one-dimensional input is used, for instance applying a CNN to a text string. There are three basic notions which CNNs rely upon: local receptive fields, weight sharing, and pooling.

2.5.1 Local receptive fields

In the case of image recognition, an image of $n \times n$ pixels can be coded as an input layer of $n \times n$ units.[15] In a CNN, this input is processed by sliding a window of size $m \times m$ across this image. This window, or patch, is known as a local receptive field. After passing this window

[15] This is assuming greyscale, i.e., one value per pixel.

over the input image, the following layer contains a representation based on $m \times m$ sized slices of the input image. Intuitively, this can be seen as blurring the input image somewhat, as the spatial dimensions of the image are generally reduced through this process.

The length with which this window moves is referred to as its *stride*, and is most often set to 1, meaning that the window simply shifts by one pixel at a time. Although stride lengths of 2 and 3 are encountered in the literature, it is fairly uncommon to see larger stride lengths than this. Figure 2.11 shows a convolution, where $n = 4$, $m = 2$, a stride of 2 is used, which yields a new layer with size 2×2.

2.5.2 Weight sharing

A key notion of CNNs is the fact that weights are shared between each such local receptive field, and the units to which they are attached. Hence, in our example, rather than having to learn $n \times i = 64$ weights (where $i = n \times n$ is the total number of units in the first hidden layer), as in an FFNN, only $m \times m = 4$ parameters need to be learned. Therefore, all units in the first hidden layer capture the same type of features from the input image in various locations. For instance, imagine you want to identify whether a picture contains cats (as in Figure 2.12). In this figure, units in the first hidden layer might encode some sort of *cat detector*. The fact that weights are shared in this manner, results in CNNs being robust to translation invariance. This means that a feature is free to occur in different regions in the input image. Intuitively, taking our cat detector as an example, it is naturally the case that a cat is a cat, regardless of where in the image it happens to hide.

Each such *cat detector*, is referred to as a *feature map*, or a *channel*.[16] For a CNN to be useful, normally more than one feature map

[16]The *channel* terminology makes sense when considering an input image in, e.g., RGB formatting, in which the intensities of each colour is represented in a separate colour channel.

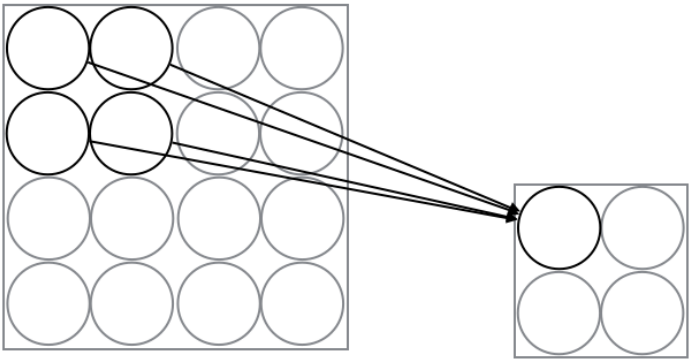

Figure 2.11: Illustration of a local receptive field of size $2x2$, which results in a new image of size $2x2$ due to the stride length being 2.

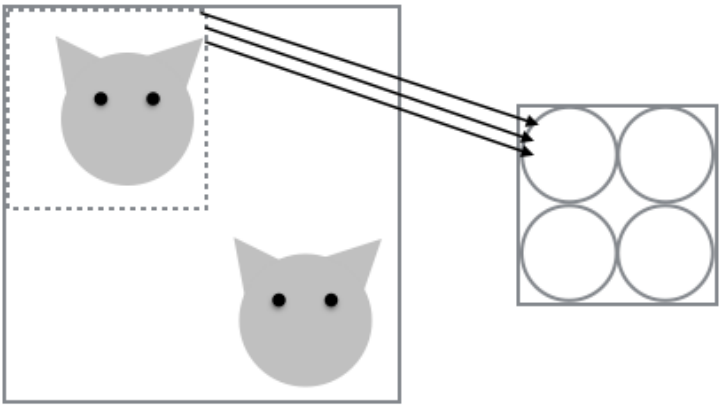

Figure 2.12: Weight sharing example with a cat. The dotted line represents the local receptive field used, the first square represents the entire input domain, and the second square represents the following convolutional layer.

is learnt. That is to say, while the figures shown so far only show a single feature map, an input image is normally mapped to several smaller images. As an example, another feature map in Figure 2.12 might learn a *dog detector*. A more realistic example can be found in facial recognition, in which one feature map might learn to detect eyes, while another learns to detect ears, and yet another learns to detect mouths. Local receptive fields generally speaking look at all feature maps of the previous layers, hence the combination of eyes, ears, and mouths might be used to learn a feature map representing an entire face.

The fact that we generally map to several feature maps means that, at each layer, the spatial size of the image shrinks (i.e. $m < n$), while the depth of the image increases, as shown in Figure 2.13. Finally, following a series of convolutional layers, it is common practise to attach an FFNN prior to outputting predictions.

In NLP, the situation is somewhat different, as we normally do not have an image as input, but rather some sort of textual representation. Commonly, this will either be a string of words, characters, or bytes. An intuition for how this works, is that something resembling an n-gram feature detector is learnt given a window. This type of approach has been applied successfully in various tasks, for instance to obtain word-level representations which take advantage of sub-word information (dos Santos and Zadrozny, 2014; Bjerva et al., 2016b), and for sentence classification (Kim, 2014).

2.5.3 Pooling

A pooling layer takes a feature map and condenses this into a smaller feature map. Each unit in a pooling layer summarises a region in the previous layer, generally using a simple arithmetic operation. Frequently, operations like maximum pooling (max pooling) or average pooling are used (Zhou and Chellappa, 1988). These operations take, e.g., the maximum of some region to be a representation of that

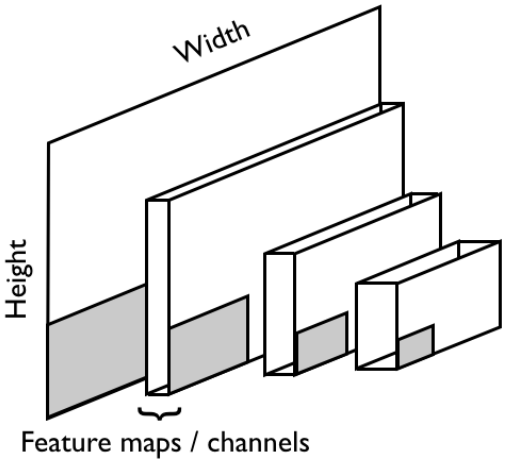

Figure 2.13: General CNN structure. Each layer shrinks the width and height of the input image, and increases the number of feature maps. The grey regions denote the sizes of the local receptive fields.

entire region, thus reducing dimensionality by essentially applying simple non-linear downsampling. Max pooling can thus be seen as a way for each max pooling unit to encode whether or not a feature from the previous layer was found anywhere in the region which the unit covers. The intuition is that the downsampled version of the feature map, which yields feature locations which are rough, rather than precise, is sufficient in combination with the relative location to other such downsampled features. Importantly, this operation reduces the dimensionality of feature maps, thus reducing the number of parameters needed in later layers.

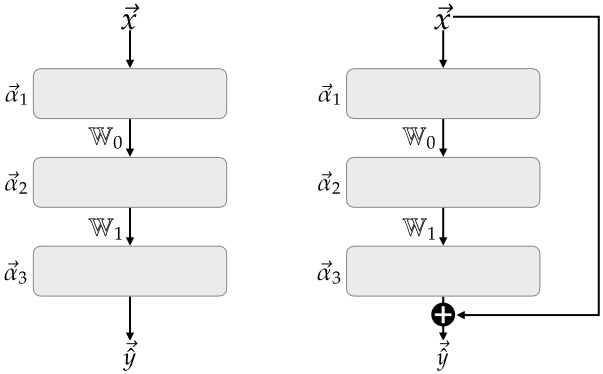

Figure 2.14: Illustration of a residual network (right) as compared to a standard network without skip connections (left). The skip connection here passes the input vector $\vec{x}$ and adds this to the activations $\vec{\alpha}_3$, thus providing the network with a shortcut. The squares represent an abstract block of weights, such as, e.g., a fully connected layer in an FFNN, a convolutional block in a CNN, or an LSTM cell.

2.6 Residual Networks

Residual Networks (ResNets) define a special class of networks with skip connections between layers, as depicted in Figure 2.14. This facilitates the training of deeper networks, as these skip connections ease the propagation of errors back to earlier layers in the network.

Such skip connections are referred to as residual connections, and can be expressed as

$$\begin{aligned} y_l &= h(x_l) + \mathcal{F}(x_l, \mathcal{W}_l), \\ x_{l+1} &= f(y_l), \end{aligned} \tag{2.19}$$

where x_l and x_{l+1} are the input and output of the l-th layer, $\mathcal{W}_l$ is the

weights for the l-th layer, and $\mathcal{F}$ is a residual function (He et al., 2016) such as the identity function (He et al., 2015a), which we also use in our experiments. Although ResNets were developed for the use in CNNs, the skip connections are currently being used, e.g., in LSTMs (Wu et al., 2016). ResNets can be intuitively understood by thinking of residual functions as paths through which information can propagate easily. This means that, in every layer, a ResNet learns more complex feature combinations, which it combines with the shallower representation from the previous layer. This architecture allows for the construction of much deeper networks. ResNets have recently been found to yield impressive performance in image recognition tasks, with networks as deep as 1001 layers (He et al., 2015a, 2016), and are thus an interesting and effective alternative to simply stacking layers. Another useful feature of ResNets is that they act as ensembles of relatively shallow networks, which may help to explain why they are relatively robust to overfitting in spite of their large number of parameters (Veit et al., 2016).

ResNets have recently been applied in NLP to morphological reinflection (Östling, 2016), language identification (Bjerva, 2016), sentiment analysis and text categorisation (Conneau et al., 2016), semantic tagging (Bjerva et al., 2016b), as well as machine translation (Wu et al., 2016). Recently proposed variants include wide residual networks, with relatively wide convolutional blocks, showing that resnets do not necessarily need to be as deep as the 1001 layers used in previous work (Zagoruyko and Komodakis, 2016).

2.7 Neural Networks and the Human Brain

While there are numerous reasons to use neural networks, there are camps which might argue for applying NNs because they are *'biologically motivated'*. While it can be tempting to use the conceptual metaphor (Lakoff and Johnson, 1980) of NEURAL NETWORKS ARE

THE BRAIN, this can be rather misleading. While usage of misleading metaphors can seem innocent, it has been shown that they do affect reasoning (Thibodeau and Boroditsky, 2013). Therefore, a misleading metaphor should certainly not be used as an argument for the usage of neural networks – there are plenty of other reasons for that.

Now, you might ask, is this metaphor really as misleading as it seems? At best, neural networks (as used in NLP) are a mere caricature of the human brain. On a physiological level, there is evidence that neurons do more than simply outputting some activation value based on a weighting of its inputs. For instance, recent research has shown that a single neuron can encode temporal response patterns, without relying on temporal information in input signals. Hence, the nature of how neurons work is quite different from what is encoded in a neural network, for instance in terms of information storage capacity (Jirenhed et al., 2017). There is in fact compelling evidence that memory is not coded in (sets of) synapses, but rather internally in neurons (cf. Gallistel and King, 2011, and Gallistel, 2016 and references therein, notably Johansson et al. (2014)). Additionally, backpropagation is not biologically plausible, although there is work on making biologically plausible neural networks (Bengio et al., 2015).

Convolutional Neural Networks and the Brain

Similarly to the ReLUs discussed in Section 2.3.2, CNNs are biologically inspired. When CNNs were invented (LeCun et al., 1998a), this was inspired by work which proposed an explanation for how the world is visually perceived by mammals (Hubel and Wiesel, 1959, 1962, 1968). The attempts to reverse-engineer a similar mechanism, as in CNNs, have proven fruitful, as CNNs are indeed highly suitable for image recognition. Furthermore, recent research has found some correlations between the representations used by a CNN, and those encoded in the brain in a study where the CNN could identify which image a human participant was looking at roughly 20% of the time

(Seeliger et al., 2017). However, it remains to be seen whether anything similar to this is even plausible for natural language.

2.8 Summary

In this chapter, an intuitive and theoretically supported overview of neural networks was given, including a practical overview NLP. We have seen that many common NLP problems can be classified into three categories: one-to-one, one-to-many, and many-to-many. Appropriate deep learning architectures suitable for each of these categories were suggested.

While this chapter is meant to be a sufficient introduction to neural networks to understand this thesis, it is by no means a complete account of the topic. For a more in-depth description of neural networks in general, I refer the readers to Goodfellow et al. (2016). For a primer which is more geared towards NLP, see Goldberg (2015).

CHAPTER 3

Multitask Learning and Multilingual Learning

Abstract|In this chapter, we build upon the background knowledge of neural networks presented in the previous chapter. We will focus on different, but highly related, paradigms – multitask learning, and multilingual model transfer. Multitask learning is first presented in a general context, and then in the context of neural networks, which is the primary focus of this thesis. We will then look at multilingual approaches in NLP, again first in a general context, and then in the context of model transfer with multilingual word representations, which is the secondary focus of this thesis. In this thesis, we consider the first setting in Part II, the second setting in Part III, and include an outlook for a combined multilingual/multitask paradigm in Part IV.

3.1 Multitask Learning

In Natural Language Processing (NLP), and machine learning (ML) in general, the focus is generally on solving a single task at a time. For instance, one might invest significant amounts of time in making a Part-of-Speech tagger or a parser. However, fact is that many tasks are related to one another. The aim of multitask learning (MTL) is to take advantage of this fact, by attempting to solve several tasks simultaneously, while taking advantage of the overlapping information in the training signals of related tasks (Caruana, 1993, 1997). When the tasks are related to each other, this approach can improve generalisation, partially since it provides a source of inductive bias, and since it allows for leveraging larger amounts of more diverse data.[1] Additionally, since related tasks can often make use of similar representations, this can lead to the tasks being learnt even better than when training on a single task in isolation.

The use of MTL is skyrocketing in NLP, and has been applied successfully to a wide range of tasks, for instance sequence labelling such as POS tagging (Collobert and Weston, 2008; Plank et al., 2016), semantic tagging (Bjerva et al., 2016b), as well as chunking and supertagging (Søgaard and Goldberg, 2016). In addition to this, it is the primary focus of this thesis, and having some background knowledge on this will be useful for the following chapters. The first part of this chapter is an attempt at providing an understanding of what MTL is and how it is applied. While some general MTL scenarios are covered, the focus will be on MTL in the context of neural networks, and in the context of NLP.

[1] Generally speaking, it is beneficial to have access to more data when training an ML model.

3.1.1 Non-neural Multitask Learning

Before going into MTL in neural networks (NNs), we first take a look at the usage of this paradigm in other frameworks. Generally speaking, we seek to exploit the fact that there are many tasks which are somehow related to one another (Caruana, 1993, 1997; Thrun and Pratt, 1998). For instance, MTL can have the role of being a distant supervision signal, in the sense that the tasks used might be fairly distantly related. Additionally, since MTL plays the role as a regulariser (see Chapter 2), and lowers the risk of overfitting (Baxter, 1997; Baxter et al., 2000), MTL often improves generalisation. This is in part because MTL reduces *Rademacher complexity* (Baxter et al., 2000; Maurer, 2006).[2] Furthermore, MTL will push the weights of a model towards representations which are useful for more than one task. Finally, MTL can be seen as a method of dataset augmentation, as it allows for using more data than when only considering a single task at a time.

A commonly made assumption in MTL is that only a handful of parameters or weights (see Chapter 2) ought to be shared between tasks, and conversely that most parameters should not be shared (Argyriou et al., 2007). This can intuitively be understood by considering that only a few features useful for a task t_1 might be useful for another task t_2. For instance, imagine that we are building a joint POS tagger and language identification system. A feature capturing capitalised words preceded by a determiner will both be a decent indicator of the language being, e.g., German, as well as that the capitalised word is a noun. Other features, on the other hand, such as one indicating that the language is likely to be Norwegian or Danish if the letter ø is encountered, is not likely to be beneficial for POS tagging at all. In other words, this type of *parameter sparsity* can be phrased

[2] A lower Rademacher complexity essentially indicates that a class of functions is easier to learn.

as that most parameters should not be shared, as many parameters are task specific. In this type of approach, all shared parameters are generally considered by all tasks involved. This puts the system at a relatively large risk of negative transfer, if one tries to combine this approach with tasks which are only slightly related. In NLP we are often interested in exploiting even relatively weak training signals, which makes this particularly problematic.

Another approach is to learn clusters of tasks, which allows for letting related tasks share certain parameters, and relatively unrelated tasks perhaps only a few. Such approaches have in common that they assume that the parameters which are beneficial for each other are geometrically close to one another in n-dimensional space (Evgeniou and Pontil, 2004; Kang et al., 2011). Other work has come up with other definitions of task similarities. For instance, Thrun and O'Sullivan (1995) consider two tasks to be similar simply if one improves performance on the other. While other approaches to MTL have been used in the past, such as Daumé III (2009) who approach MTL from a Bayesian perspective, and Toutanova et al. (2005) who train a joint classifier for semantic role labelling with automatically generated auxiliary tasks, the perhaps most popular approach in NLP is parameter sharing in NNs.

3.1.2 Neural Multitask Learning

We now turn to the main method used in this thesis, namely neural MTL. There are two main approaches to this, differing in the manner in which parameters are shared – *hard* and *soft* parameter sharing. Currently, the less popular variant of the two in NLP is soft parameter sharing, and will not be covered in detail. Briefly put, in this setting, parameters are constrained in a similar manner to the *parameter sparsity* approach. That is to say, the parameters between tasks are encouraged to be similar to one another, which allows for some transfer between tasks, or between languages (Duong et al.,

2015). However, as parameters are not explicitly shared between tasks, the risks of negative transfer are relatively low in this setting. This approach is not explored in this thesis as hard parameter sharing offers several advantages, including ease of implementation, and computational effectivity, as the amount of parameters is kept almost constant as compared to having a single task.

Hard parameter sharing is currently more common, perhaps mainly due to the ease with which a neural MTL system with several tasks can be created. This is the type of MTL discussed in the seminal works by Caruana (1993, 1997). In this thesis we consider research questions tied to this type of MTL in the context of NLP, partially due to the versatility of the paradigm. Apart from allowing for considering data from several tasks simultaneously, even corpora in different languages might be used in this approach, given some sort of unified input representations.[3] Then, if the output labels between tasks correlate with one another to some extent, it seems quite intuitive that this approach should be beneficial.

In NLP, MTL is generally approached from the perspective that there is some *main* task, i.e., the task in which we are interested, and some *auxiliary* task, which should improve the main task. It is important to note, however, that these labels are quite arbitrary.[4] There is not necessarily anything to distinguish a main task from an auxiliary task in an NN. One might lower the weighting of the auxiliary task (i.e. multiply the loss for each batch by some $\lambda < 1$), but this strategy appears to be relatively rare in the literature.

A common way of implementing hard parameter sharing, is to have a stack of layers for which weights are updated with respect to all tasks, with at least two output layers, each with task-specific weights (see Figure 3.1). Concretely, consider that we have t corpora

[3]This is covered further in the second half of this chapter.

[4]The exception being cases in which the performance on the auxiliary task is disregarded in favour of the main task performance.

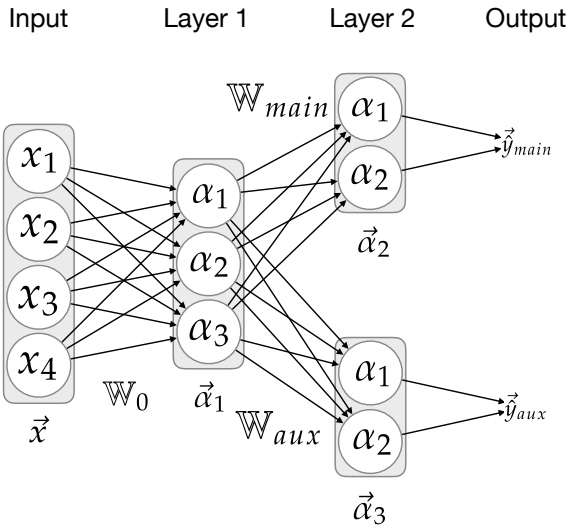

Figure 3.1: Common MTL architecture (bias units omitted for clarity).

with different annotations, each containing pairs of input and output sequences $(\vec{x}, \vec{y}^t)$ for a single task. While the inputs x will largely be part of the same vocabulary, and can be shared across tasks, the tag sets used, and therefore the labels $(y^0 \dots y^t)$ differ. Note that the vocabularies in the tasks at hand do not necessarily need to overlap, but when considering a single natural language, this tends to be the case. A common approach when training is to randomly sample such sequence pairs, predict a label distribution $\vec{y}^t$, and update model parameters as calculated by the loss relative to the true label distribution $\vec{y}^t$ with backpropagation (see Chapter 2 for an overview

of this). Each task t has a task-specific classifier ($f_{t=main}, f_{t=aux}$) with its own weight matrix ($\mathbb{W}_{main}, \mathbb{W}_{aux}$). The output of the task-specific layer is then calculated using the softmax function (cf. Section 2.3.2), such that

$$\vec{\hat{y}}_t = \text{softmax}(\mathbb{W}_t \vec{\alpha}_n + b), \tag{3.1}$$

where $\vec{\alpha}_n$ denotes the activations of the layer before the output layer.

This architecture is common in NLP, with weights typically shared between the *main* and *auxiliary* task at all layers, up to the task-specific classification layer (i.e. the output layer). A multitude of other possibilities do exist, as the task-specific output layers can be attached anywhere in the network. This can be advantageous, as Søgaard and Goldberg (2016) found that including the lower-level task supervision at lower levels in the network was useful, in the case of using the low-level task of POS tagging in combination with CCG supertagging (i.e. assigning CCG lexical categories). Most related work, including the experiments in this thesis, apply multitask learning akin to what is shown in Figure 3.1.

Neural Multitask Learning in Natural Language Processing

Hard parameter sharing in NNs is the target of considerable attention in the recent NLP literature. Practically speaking, there appear to be two main approaches to MTL in the NLP literature. Some work, such as Ando and Zhang (2005), Collobert and Weston (2008), Søgaard and Goldberg (2016), Plank et al. (2016), Bjerva et al. (2016b), and Augenstein and Søgaard (2017) take the approach of exploiting seemingly related NLP tasks, based on some linguistic annotation. Other work, e.g., Plank (2016), and Klerke et al. (2016), take the approach of exploiting data from non-linguistic sources (keystroke data and eye gaze data, respectively). While these approaches are both useful and interesting, the focus of this thesis is the first approach,

specifically in which an NLP sequence prediction task is used as an auxiliary task for some other NLP sequence prediction task. This is partially motivated by the fact that using a word-level input for all tasks, allows for a one-to-one mapping between labels in different tag sets (given a specific token in context), which in turn opens up for the information-theoretic approach considered in Chapter 5.

3.1.3 Effectivity of Multitask Learning

Plenty of studies demonstrate the success of MTL, such as in computer vision (Torralba et al., 2007; Loeff and Farhadi, 2008; Quattoni et al., 2008), genomics (Obozinski et al., 2010), and the aforementioned NLP studies. Apart from relatively straight-forward results showing that MTL is often beneficial, efforts have been put into experimentally investigating when and why MTL is advantageous in NLP. Martínez Alonso and Plank (2017) look at a collection of semantic main tasks while using morphosyntactic and frequency-based tasks as auxiliary tasks. They find that the success of an auxiliary tasks depends on the distribution of the auxiliary task labels, e.g., the distribution's entropy and kurtosis.[5] Bingel and Søgaard (2017) present a large systematic study of MTL in a collection of NLP tasks. They find that certain dataset characteristics are predictors of auxiliary task effectivity, corroborating the findings of Martínez Alonso and Plank (2017), and also show that MTL can help target tasks out of local minima in the optimisation process. In Bjerva (2017b), it is argued that entropy is not sufficient for explaining auxiliary task effectivity, and that measures which take the joint distribution between tasks into account offer more explanatory value (this is elaborated in Chapter 5).

In terms of data sizes Benton et al. (2017) suggest that MTL is effective given limited training data for the main task. Luong et al.

[5]The kurtosis of a distribution is essentially a measure of its tailedness.

(2015), however, highlight that the auxiliary task data should not outsize the main task data – this is contradicted by Augenstein and Søgaard (2017), who highlight the usefulness of an auxiliary task when abundant data is available for such a task, and little for the main task. Finally, Mou et al. (2016) investigate transferability of neural network parameters, by attempting to initialise a network for a main task with weights which are pre-trained on an auxiliary task, and highlight the importance of similarities between tasks in such a setting. Finally, a promising recent innovation is that of sluice networks, in which a NN learns which parts of hidden layers to share between tasks, and to what extent (Ruder et al., 2017).

3.1.4 When MTL fails

The cases in which MTL does not work are also deserving of attention. While up until now we have assumed that applying MTL is a piece of cake, there are times when one adds an auxiliary task, causing the system to collapse like a house of cards. This type of performance loss is referred to as *negative transfer*, and can occur when two unrelated tasks share parameters. This is generally something to avoid, as there are few, if any, advantages to worsening the generalisation ability of the network. However, such results are rarely shared in the community, in part due to the *file drawer problem* (Rosenthal, 1979). In short, the problem is that it is impossible to access, or even know of, studies which have been conducted and not published. In the case of MTL, this issue might be alleviated by publishing results on all auxiliary tasks experimented with, even if only one or two such tasks improved performance.

3.2 Multilingual Learning

In the second half of this chapter, we turn to multilingual approaches. Many languages are similar to each other in some respect,

and similarly to related tasks, this fact can also be exploited in order to improve model performance with respect to, e.g., a given language. While there are many approaches to multilingual NLP, with various use cases, the focus in this thesis is on model transfer, in which a single model is shared between languages. We will nonetheless begin with an overview of the most common approaches.

As an example, consider NLP tagging tasks, which can be summed up as learning to assign a sequence of tags from a tag set t to a sequence of tokens in language l. In cross-lingual multitask NLP settings, there are many l/t pairs which do not have any annotated data. For instance, there is (at the time of writing) no annotated data for Welsh in the Universal Dependencies (Nivre et al., 2017). However, many NLP systems require input data from specific tag sets. For instance, the Stanford Neural Network Dependency parser requires POS tags in its input (Chen and Manning, 2014), whereas the semantic parser Boxer requires semantic tags in its input (Bos, 2008; Abzianidze et al., 2017). Hence, for such tools to be applicable in multilingual settings, the tags they rely on need to be available for other languages as well, which highlights the importance of approaches which deal with this. There are three frequently used approaches to solving this problem:

1. human annotation;
2. annotation projection;
3. model transfer.

Although serious efforts have gone into furthering these approaches, they all have considerable drawbacks. In brief, human annotation is time consuming and expensive, annotation projection is only applicable to texts which are both translated and aligned, and model transfer is generally only used in mono-lingual or mono-task settings.

3.2.1 Human Annotation

Generally speaking, annotating data manually is a very expensive and time-consuming manner of, e.g., producing some sort of linguistic labels for a sentence. Although the process can be alleviated with gamification (Venhuizen et al., 2013; Chamberlain, 2014; Jurgens and Navigli, 2014; Bos and Nissim, 2015), considerable time and effort still needs to be invested into creating such crowd-sourcing systems.

3.2.2 Annotation Projection

Given an annotated sentence in a source language and a translation of that sentence in a target language, it is possible to transfer, or project, the annotation from the source language to the target language. This approach is known as annotation projection, and relies on having access to parallel text for which at least one source language is annotated (Yarowsky et al., 2001; Hwa et al., 2005). Usually, word alignments are used in order to project linguistic labels from source to target. The resulting annotations can then be used to train a new monolingual system for the target language(s). This approach has been applied successfully to various tasks, primarily syntactic parsing (Hwa et al., 2005; Tiedemann, 2014; Rasooli and Collins, 2015; Agić et al., 2016), POS tagging (Yarowsky et al., 2001), and recently also semantic parsing (Evang and Bos, 2016; Evang, 2016).

Annotation projection has two main drawbacks. Primarily, it is only applicable to texts which are both translated and aligned, whereas the majority of available texts are monolingual. Furthermore, this approach relies heavily on the quality of the automatic word alignments. Word-aligning parallel text is not always successful, for instance with very dissimilar languages, insufficient statistics, or bad translations (Östling, 2014, 2015). Another approach for annotation projection relies on automatic translation. This works by applying a machine translation (MT) system to generate a parallel

text for which source language annotation exists (Tiedemann et al., 2014). In other words, in addition to the difficulties of the annotation projection approach, this method places high requirements on availability of parallel texts for training an MT model. In addition to these prerequisites, the involvement of a fully-fledged MT system in an annotation pipeline, will in itself increase its complexity severely.

3.2.3 Model Transfer

Model transfer deals with learning a single model which is shared between several languages (Zeman and Resnik, 2008; McDonald et al., 2011a).[6] This type of approach has been explored extensively in previous work. Multilingual model transfer has been successfully applied to, e.g., POS tagging (Täckström et al., 2013), and syntactic parsing (Täckström, 2013; Ammar et al., 2016). This is commonly done by using delexicalised input representations, as in the case of parsing (Zeman and Resnik, 2008; McDonald et al., 2011a; Täckström et al., 2012, 2013). A related situation, is the case of exploiting language similarities in order to train models for low-resource languages (see, e.g., Georgi et al. (2010)).

In this thesis, model transfer is framed as a special case of MTL. That is to say, each language in the model can be seen analogously to a task. This means that we are also free to choose whether we want to code tag predictions jointly as a single output layer, or have one separate output layer per language. As with MTL with multiple tasks, we consider the same specific type of MTL across languages, namely hard parameter sharing in neural networks.

A common approach in parsing is to delexicalise the input representations in order to enforce uniformity across languages, by training a parser on sequences of PoS tags rather than sequences of words (Zeman and Resnik, 2008; McDonald et al., 2011a). However, as we

[6] Note the similarities to multitask learning with hard parameter sharing.

are looking at predicting such tags, we approach this by using input representations which are shared across languages. This allows for training a neural network for several languages simultaneously in a language-agnostic manner, while still taking lexical semantics into account. Apart from this advantage, implementing a system in this manner is straightforward. Additionally, this approach offers the possibility of out-of-the-box zero-shot learning, as simply adding input representations for a different language is sufficient to enable this.

Zero-shot learning is the problem of learning to predict labels y which have not been seen during training. This is especially relevant in cases such as MT, in which, e.g., many of the target forms which need to be produced for languages with rich morphology have not been seen. In recent years, zero-shot learning has become increasingly popular, for instance in image recognition (Palatucci et al., 2009; Socher et al., 2013b). Recently, it has also been applied to MT, resulting in a model which even allows for translation into unseen languages (Johnson et al., 2016). One way of enabling zero-shot learning, is to use shared input representations. For instance, in the case of character-based models, we can simply use the same alphabet in the inputs and outputs of each system, as in Östling and Bjerva (2017) and Bjerva (2017a). In the case of word level input representations, one can employ word embeddings living in the same space, regardless of language.

3.2.4 Model Transfer with Multilingual Input Representations

Looking further at the problem of model transfer across languages, consider the following example, of an English sentence and its translation, as two separate input sequences to a neural network, with their corresponding annotated output sequences.[7]

[7]PMB 01/3421. Original source: Tatoeba.

(3.2) *We must draw attention to the distribution of*
PRON VERB VERB NOUN ADP DET NOUN ADP
this form in those dialects .
DET NOUN ADP DET NOUN PUNCT

(3.3) *Wir müssen die Verbreitung dieser Form in diesen*
PRON VERB DET NOUN DET NOUN ADP DET
Dialekten beachten .
NOUN VERB PUNCT

Although the surface forms of these two sentences differ, as one is in English and one in German, multilingual word representations for the corresponding words in these two sentences ought to be close to one another. Hence, if the NN only sees the English sentence in training, and the German sentence during test time, it ought to be fairly successful in tagging this 'unseen' sentence with suitable tags.[8] However, one question is whether having access to the same sentence in a typologically more distance language such as Japanese also would be useful (this is approached in Chapter 5).

In order for such an approach to work, it is necessary that words with similar meanings in different languages are represented in a fairly similar way. How do we arrive at word representations with such properties? In the next few sections we will look at this, beginning at simple monolingual representations, and leading up to bilingual and multilingual representations.

[8]Considering that the semantic content of the two sentences ought to be highly similar, one could regard the translated sentence to be 'seen' if the original sentence was in the training data. This has the further implication that one, in this type of experiments, must take care not to allow corresponding sentences to occur in both training and evaluation data.

3.2.5 Continuous Space Word Representations

In many NLP problems, we are concerned with processing some word-like unit, in order to arrive at some linguistically motivated and appropriate label. In Section 2.3.1, we considered bag-of-words models for tasks such as this. To recap, in this type of model we assign an index to each unique word. Each word is then represented by a vector $\vec{x}$, with a dimensionality equal to the size of the vocabulary, since each word requires its own index. As an example, consider a vocabulary size of five words, with three of those words being *cat*, *dog*, and *coastal*, with their corresponding vector representations, such that

$$\begin{aligned} \vec{x}_{cat} &= [0, 0, 1, 0, 0], \\ \vec{x}_{dog} &= [0, 0, 0, 0, 1], \\ \vec{x}_{coastal} &= [0, 1, 0, 0, 0]. \end{aligned} \tag{3.4}$$

In NLP, we are often interested in comparing words with one another, either simply in order to have some measure of their similarity, or because we are interested in the fact that similar words tend to have similar properties in down-stream tasks. For instance, the words *cat* and *dog* are likely to have the same or similar linguistic analyses in many cases, such as both being tagged with the PoS tag `NOUN`. A commonly used similarity measure between vectors is the cosine distance. In this setting, a word representation as presented above is somewhat problematic, as the distances between the three words are equal, although we want a higher similarity between *cat* and *dog*.

The representation we have seen so far is known as a *sparse* feature representation, as each word is represented by a vector of zeroes with one element set to one, also known as a *one-hot vector*. Apart from the drawback of similarity, this type of input representation can run into other problems, such as the dimensionality of the representations becoming too large to handle as vocabulary size

grows. This can be remedied in many ways, for instance by applying dimensionality reduction algorithms. Commonly used algorithms include singular value decomposition (SVD), and random indexing (Kanerva et al., 2000; Sahlgren, 2005). This does not help with the problem of similarities, however.

It turns out that one can arrive at word representations with nice properties of similarity by taking advantage of the *distributional hypothesis*:

> *'Semantics is partly a function of the statistical distribution of words.'*
> –Harris (1954)

> *'You shall know a word by the company it keeps.'*
> –Firth (1957, p.11)

This means that the semantic content of a given word is related to other words occurring in similar contexts. Furthermore, Harris (1954) claims that the strength of this relation is proportional to the similarity between two words, such that if two words w_1 and w_2 are more similar in meaning than w_1 and w_3, then the relative distribution of the first pair will be more similar than that of the second pair. One way of implementing this type of distributional semantics is to count word co-occurrences in a large corpus. Let us now assign the two remaining indices in the five-dimensional representation used above to the words *pet* and *water*, such that

$$\begin{aligned} \vec{x}_{pet} &= [1, 0, 0, 0, 0], \\ \vec{x}_{water} &= [0, 0, 0, 1, 0]. \end{aligned} \tag{3.5}$$

These five vectors can then be used to generate new distributional vectors, $\vec{y}$, by representing each word by the sum of the vectors $\vec{x}$ of the words with which it co-occurs. If *cat* and *dog* frequently co-occur

with each other, and with *pet*, whereas *coastal* mainly co-occurs with *water*, the resulting representations may be similar to

$$\begin{aligned} \vec{y}_{cat} &= [25, 0, 20, 0, 10], \\ \vec{y}_{dog} &= [30, 0, 10, 0, 20], \\ \vec{y}_{coastal} &= [0, 10, 0, 100, 0]. \end{aligned} \tag{3.6}$$

In this representation, *cat* and *dog* are more similar to one another than they are to *coastal*, which is exactly what we want. A visualisation of such a word space is given in Figure 3.2.

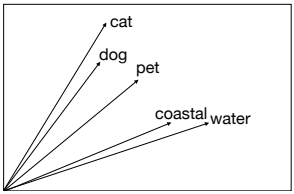

Figure 3.2: An example of a word space

The type of word representation discussed up until now is also known as a *count*-based representation, as opposed to a *prediction*-based representation (Baroni et al., 2014).[9] Whereas a count-based representation can be seen as counting the words in a given context, a prediction-based representation can be made by attempting to predict that context. Doing this with a neural network, the error obtained when attempting to make such predictions is used to update the representations, until a low error is obtained (see Chapter 2 for details on neural networks). With the entry of the deep learning tsunami on the NLP scene (Manning, 2015), this type of dense word representations has become increasingly popular. The availability

[9]Whereas Baroni et al. (2014) suggest that prediction-based methods outperform count-based ones, Levy and Goldberg (2014b) show that the underlying differences between the approaches are small.

of tools implementing such algorithms, such as *word2vec*, undoubtedly helped push the popularity of this approach further. This trend was introduced by Collobert and Weston (2008), Turian et al. (2010), and Collobert et al. (2011), and was further spearheaded by papers such as Mikolov et al. (2013c), which showed that a simple neural model would encapsulate linguistic regularities in its embedded vector space. The now infamous example of this property is shown in a figure in Mikolov et al. (2013c), replicated in Figure 3.3, where the following relation holds

$$\overrightarrow{king} + \overrightarrow{woman} - \overrightarrow{man} = \overrightarrow{queen}. \tag{3.7}$$

In other words, the distance between *man* and *woman* is similar to that between *king* and *queen*, so adding this difference to the vector of *king* results in a vector close to *queen*.

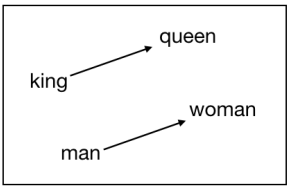

Figure 3.3: A word space in which adding the difference between *woman* and *man* to *king* results in *queen*.

In a prediction-based approach, the terminology used is that a word is *embedded* into n-dimensional space. Typically, this dimensionality is much lower (e.g. around 100) than what is generally used in count-based approaches (e.g. around 1000). In the case of neural networks, these embeddings can be trained together with the rest of the network. This results in a matrix of word vectors in which words with similar properties (under the task at hand) are close to one another.

Distributional vs. Distributed representations

An useful distinction to make is that of *distributed* vs. *distributional* representations. A distributional word representation is based on a co-occurrence matrix, taking advantage of the distributional hypothesis. Similarities between the resulting distributional word representations thus represent the extent to which they co-occur, and therefore also their semantic similarity. A distributed representation, on the other hand, is simply one that is continuous. That is to say, a word is represented by a dense, real-valued, and usually low-dimensional vector. Such representations are generally known as word embeddings, with each dimension representing some latent feature of the word at hand. One way to remember this is that such representations are *distributed* across some n-dimensional space. The first representations we saw were therefore distributional, but not distributed (Turian et al., 2010). The word embeddings, on the other hand, can be said to be both.

Bilingual and Multilingual Word Representations

Going from monolingual to bilingual word representations has been the subject of much attention in recent years. One of the first approaches to bilingual word representations was shown by Klementiev et al. (2012), followed by work such as Wolf et al. (2014), and Coulmance et al. (2015). Parallel to approaches which aim at making good multilingual embeddings, are attempts at producing better monolingual embeddings by exploiting bilingual contexts, as in Guo et al. (2014), Šuster et al. (2016), and Šuster (2016).

In essence, the approaches to building such representations can be divided up into several categories. Cross-lingual mapping can be done by first learning monolingual embeddings for separate languages, and then using a bilingual lexicon to map representations from one space to the other (Mikolov et al., 2013b). Another ap-

proach is to mix contexts from different languages, and training pre-existing systems, such as word2vec, on this mixed data (Gouws and Søgaard, 2015). The approach under consideration in this thesis is based on exploiting parallel texts, by jointly optimising a loss function when predicting multilingual contexts (Guo et al., 2016).

The true power of multilinguality is not unlocked until we can consider an arbitrary number of languages at a time. Whereas bilingual word representations only encode two languages, a multilingual word space contains representations from several languages in the same space. As before, we here also have the property that words with similar meanings are close to one another irrespective of the language (see Figure 3.4).

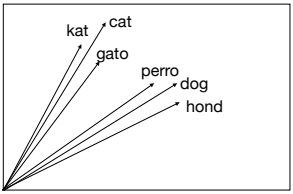

Figure 3.4: An example of a multilingual word space.

One such is the multilingual skip-gram model, as outlined by Guo et al. (2016).[10] As a variant of this model is used in Part III and Part IV, we will now cover this in more detail.

[10]The skip-gram method to create word embeddings, in which a neural network attempts to predict the context of a word, is not to be confused with skip-grams in the sense of n-grams which are not necessarily consecutive. In the second sense, we can define a k-skip-n-gram as a sequence of length n, in which words occur at distance k from each other. In this thesis, only the first sense is of importance.

Multilingual Skip-gram

The skip-gram model has become one of the most popular manners of learning word representations in NLP (Mikolov et al., 2013a). This is in part owed to its speed and simplicity, as well as the performance gains observed when incorporating the resulting word embeddings into almost any NLP system. The model takes a word w as its input, and predicts the surrounding context c. Formally, the probability distribution of c given w is defined as

$$p(c|w;\theta) = \frac{\exp(\vec{c}^T\vec{w})}{\Sigma_{c\in V}\exp(\vec{c}^T\vec{w})}, \tag{3.8}$$

where V is the vocabulary, and θ the parameters of word embeddings ($\vec{w}$) and context embeddings ($\vec{c}$). The parameters of this model can then be learned by maximising the log-likelihood over (w, c) pairs in the corpus C,

$$J(\theta) = \sum_{(w,c)\in D} \log p(c|w;\theta). \tag{3.9}$$

Guo et al. (2016) provide a multilingual extension for the skip-gram model, by requiring the model to not only learn to predict English contexts, but also multilingual ones. This can be seen as a simple adaptation of Firth (1957, p.11), i.e., you shall know a word by the *multilingual* company it keeps. Hence, the vectors for, e.g., *dog* and *perro* ought to be close to each other in such a model. This assumes access to multilingual parallel data, as word alignments are used in order to determine which words comprise the multilingual context of a word.

Formally, the learning objective in multilingual skip-gram is de-

fined in Guo et al. (2016) as

$$\begin{aligned} J &= \alpha \sum_{l \in L} J_{mono_l} + \beta \sum_{l \in L, \{EN\}} J_{bi_l, EN} \\ J_{mono_l} &= \sum_{(w,c) \in D_{l \leftrightarrow l}} \log p(c|w;\theta) \\ J_{bi_l, EN} &= \sum_{(w,c) \in D_{l \leftrightarrow EN}} \log p(c|w;\theta), \end{aligned} \tag{3.10}$$

where L denotes the set of all languages, and α and β are weight parameters for the monolingual and bilingual contexts, respectively. In our work, however, we do not rely on always using English as a pivot, and rather use all bilingual pairings to generate contexts. In other words, we also predict the French context based on the Spanish word, and vice versa, rather than only predicting from or to English. This is visualised in Figure 3.5, in which the dashed lines indicate the additional predictions made using the loss described here, and used in Bjerva and Östling (2017a).

Figure 3.5: Multilingual skip-gram utilising multilingual contexts. Dashed lines indicate the additions of our loss function, i.e., predictions between every language pair.

Formally, the joint objective function used here is defined as

$$\begin{aligned} J &= \alpha \sum_{l \in L} J_{mono_l} + \beta \sum_{l_1 \in L} \sum_{l_2 \neq l_1 \in L} J_{bi_{l_1}, bi_{l_2}} \\ J_{mono_l} &= \sum_{(w,c) \in D_{l \leftrightarrow l}} \log p(c|w;\theta) \\ J_{bi_{l_1}, bi_{l_2}} &= \sum_{(w,c) \in D_{l_1 \leftrightarrow l_2}} \log p(c|w;\theta). \end{aligned} \tag{3.11}$$

3.3 Outlook

In the first part of this chapter, we considered multitask learning, which is the focus of Part II of this thesis. We will first see a case study, in which a MTL paradigm is shown to improve performance on two sequence labelling tasks. Then we turn to a more theoretical investigation into why this is the case.

Following this, Part III also begins with a case study on multilinguality in a single NLP task. The subsequent chapter then includes an empirical study of multilinguality in several tasks, and looks at change in performance when multilinguality is employed.