Abstract

Interpreting EEG measurements is of great relevance, both for developing underlying neuroscientific theory and improving existing applications. In this study, two networks with different approaches to time-frequency analysis and feature selection are compared on simulated and real data for semantic and emotional perception. The first network uses the Morlet wavelet transform to achieve adaptable feature selection. The second network uses a convolutional net to analyse reassigned spectrograms, in hope of improving component localisation. The result shows relatively good performance of the Morlet network with easily interpretable features, especially when combined with Grad-CAM, a method for visualising the gradients of the network to locate relevant data regions. The network using reassigned spectrograms performs less well, but comparisons with methods for ordinary spectrograms suggest that this is due to poor performance of the more traditional image-processing methods used, making it difficult to determine the effect of reassignment. Testing on novel data shows lower, but statistically significant, classification performance for emotional content, likely due both to methodological shortcomings and to the intrinsic difficulty of the problem. The study explores the use of transfer learning and finds promising results both in the accuracy on new subjects with models trained on data from others and in boosting training on single subjects by initialisation with transferred weights. Finally, the Morlet network is applied to analyse similarities between perception and memory retrieval, with significant results for networks trained on memory data and tested on perception data.

Acknowledgements

We would like to express our gratitude to Bo Bernhardsson at the Department of Automatic Control at LTH and Maria Sandsten at the Division for Mathematical Statistics for their valuable help and guidance during our thesis work. Also, thank you to Carolina Bergeling at the department for Automatic Control and Mikael Johansson at the Department of Psychology for their expert opinions and comments. Lastly, we would like to extend thanks to Sterre van de Langenberg and Linn Petersdotter for supplying us with data used in the thesis.

Contents

1

Introduction

The inner workings of the human brain have long been the subject of intense study, since a deeper understanding leads to progress in many crucial fields, from psychology [Steingrimsson et al., 2020] to prosthetics [Bright et al., 2016]. The interdisciplinary nature of the field has meant that a vast variety of different approaches have been employed to study the brain, its processes and their resulting actions. From more psychologically rooted experiments to biological modelling on the level of individual neurons these attempts have revealed much about the way the human mind functions, but a great deal remains unknown. Our entry-point into this field is, naturally, influenced by many of these previous approaches but takes as its starting point methods for large-scale data analysis recently developed within the fields of statistics and machine learning. Our work is less focused on mapping and modelling the actual processes of the brain and focuses instead on methods for interpreting the observable signals that are evoked. We use data relating to signals associated with image perception and the corresponding short-term memories that are created, but the methods used are applicable in many different areas of signal processing.

A diversity of methods is also present with regards to the measurements themselves. The method chosen here, the electroencephalogram (EEG), is well suited for our particular problem due to its high temporal resolution. Since a defining feature for many of the processes that we study is the frequency and positions in time, this allows for good identification and separation. The method is also useful due to the ease of measurement, meaning that more data can be gathered than would be possible with more elaborate measurement techniques, as well as the availability and relatively low cost compared to, for example, a Magnetic Resonance Imaging (MRI) scanner. This does not, however, mean that this is the optimal measuring method for studies of this kind and the EEG has its shortcomings compared to other techniques. For example, the spatial resolution of the EEG is very poor compared to that of the functional MRI (fMRI) [Lystad and Pollard, 2009], meaning that potentially important information about the origin of signals within the brain is lost. This is especially true for EEG readings made with a small number of electrodes. [Ferree et al., 2001]

Attempts to decode the contents of brain signals using more traditional existing psychological tools and models have relied on explicit knowledge of properties of the signals to find empirical patterns, for example by calculating the Event-Related Potential (ERP) [Luck, 2005]. This has in many ways been a successful approach, but risks missing unexpected patterns and fails to fully utilise the vast amount of data generated from many readings of brain signals. Building upon previous work, we hope to overcome some of these shortcomings by exploring different statistically based machine learning methods of data analysis to analyse and classify labelled signals without prior information.

EEG carries multiple types of information as previously explained: spectral, temporal, phase, channel correlation and spatial information. As EEG preserves certain information better than others our study will focus on some of the EEG features with higher information retention: temporal and spectral structures. These features are relatively easy to analyse.

Our approach also seeks to deliver a method that can quickly and with little data be used on previously unknown subjects. This would be a valuable advantage in many areas of application, since the circumstances for the use of methods like these do not always allow for detailed study of the subject a priori. A method working as fast as possible with as little data as possible will make the use of these methods easier, which in turn means that they can be applied in new areas in which they were previously of limited use. One good example is the interpretation of motor signals, as studied by [Zhao et al., 2019], the article that inspired one of our network architectures. In both mechanical prostheses and rehabilitation methods (one of the applications used as an example in [Zhao et al., 2019]) a fast method requiring little beforehand knowledge makes for easy to use applications that also require less on-board computing power. In psychological applications, which are more similar to the data we use in this study, a lower data requirement of these methods can mean many hours of time saved that would otherwise be used for calibration to the subject. In addition, long sessions of data collection is difficult since the participant gets tired after a limited time, resulting in both discomfort and decreased data quality. Overall, great improvements in the efficiency and applicability of EEG-based methods of psychological treatment would be possible if the desired improvements of the methods can be realised.

Aim

Our hope with this study is to lay the groundwork for the applications described above by examining and comparing two different methods of time-frequency analysis, the adaptive Morlet wavelet transform as proposed by [Zhao et al., 2019] and the scaled reassigned spectrogram presented in [Sandsten and Brynolfsson, 2015], in conjunction with a compact neural network. These methods and networks are evaluated by testing classification accuracy on simulated and real data, some of which derives from a neurological test we have helped design and implement. In addition, general machine learning tools frequently used to improve the performance or generalisability and to analyse the working processes of neural networks, such as data Grad-CAM and transfer learning, are applied to improve and learn more about the methods' functioning.

The methods are also used to draw conclusions from the real data and compare these to existing studies and relevant theory. Analogue to modelling and technical aims the thesis aims to correctly classify emotional context of images in a novel data set developed concurrently to this thesis work as well as to investigate eventual differences in this classification task compared to previous classification tasks for EEG data.

2

Background & Theory

In this section we describe the principles behind methods used in this thesis study. The thesis, as indicated by the title, focuses on classifying the contents of the mind both during active visual perception and during recall of previous perceptions. It is important to have a basic understanding of the neurological theory that is the basis for many of the assumptions and hypotheses used. This background will therefore focus (narrowly) on concepts that are directly applicable to the study at hand and not seek to give full explanations of the studied phenomena, although the foundations in sections 2.2 and 2.3 are explored more deeply.

2.1 Neuroscience, Memory and EEG

When we receive visual stimulus the corresponding neural centres of the brain activate, producing a mental representation of the image perceived. Information detailing this sequence of activations is then stored, which is what makes up our short-term memory. When a memory is triggered the same sensory information that was present during the initial perception is reactivated [Waldhauser et al., 2016], producing a mental representation of the previously seen image. This is what allows us to see past events in our "mind's eye". The experiments referred to in this study rely on this in combination with the fact that sensory information that relates to a previous experience can activate the stored memory [Tulving, 1983]. This is used by associating an arbitrary and previously unassociated word to the image that is to be stored and later activating the memory of the image by presenting the subject with the corresponding word.

EEG readings are widely used to analyse the processes taking place in the brain. The method basically consists of measuring the differences in electrical potential between different points on the scalp with the help of electrodes distributed across it. Extrapolating from the data points along the scalp allows for reconstruction of the spatial relations between the signals (to a certain extent), in addition to the time and frequency information collected by each individual electrode. In the context of analysing the resulting signals the electrodes are frequently referred to as *channels*, which will be used throughout this report.

The information correlating to events registered in the EEG is normally divided by frequency into five rough bands: the delta band (0.5 - 4 Hz), usually associated with brain activity during deep sleep, the theta band (4 - 8 Hz), activated during lighter states of sleep or when focusing, the alpha band (8 - 14 Hz), normally activated when the subject is awake but relaxed and with their eyes closed, the beta band (14 - 30 Hz), associated with most normal conscious activity and when the subject is concentrated, and the gamma band (30+ Hz), normally activated by sensory stimulus [Abo-Zahhad et al., 2015]. According to [Bazgir et al., 2018] data relating to emotional responses can be best classified by looking at intensities in the gamma band, although information is presented throughout lower bands as well.

Disturbances introduced by sources other than the activity of the brain are called *artefacts*. These can be detected and, if sufficient data is available, removed before the EEG is analysed.

The most common large artefact in our analysis is the eye blink, which can be removed by comparing the affected channels to the one placed directly below the eye (VEOG). Other common artefacts include movement by the subject and unconscious muscle activity like swallowing and heartbeat, which as far as possible is removed from the final EEG. Poorly attached electrodes and unforseen subject actions also create channels and epochs of largely unusable signal data. Despite the best efforts of the researcher, many artefacts may remain undetected in the final data. Taken together, this results in a large amount of noise relative to the signal content. Environmental artefacts may also disturb the EEG and should be minimised during measurements. Ordinary high-frequency disturbances such as power line noise also present a problem during analysis of high frequencies. [Tatum, 2014]

Bramao & Johansson 2018

The article by [Bramao and Johansson, 2018] is a cornerstone in the basis of our work. The authors investigate the possibilities of episodic memory decoding for semantic visual information. The report proves the possibility of using information during perception to broadly interpret later memories in certain cases.

The first real data set applied in this report and several others, denoted the semantic memory set in this study, originates from the experiments described in this paper. The experiment consists of two phases: a study phase and a retrieval phase. In the study phase participants were presented with an abstract word, an image and the same word and image as a pair. The participant was asked to remember the word-image pair. In the retrieval phase the participant was instead given one of two different tasks, a visual task and a verbal task. The visual task consisted of identifying a previously shown image in a pair with the image in addition to a mirrored version. In the verbal task participants were instead given the word and asked to respond with the paired image (or a description of it). The semantic classes of images shown were "landscape", "face" and "object".

The experiment was conducted on 36 subjects, half of which performed the visual task in the retrieval phase and the rest performed the verbal task (not used in this study). The experiment consisted of a number of blocks with breaks in between. Blocks consisted of a study phase and a retrieval phase, containing a total of around 180 word-image pairs shown to the subjects over the course of the experiment.

First the authors presented a classifier achieving reasonable accuracy (around 0.6) for three-way classifying image class information during perception in the experiment. [Bramao and Johansson, 2018] subsequently studied the possibility to extrude semantic information from EEG data taken at the time of image recollection by only using previous knowledge of the EEG data recorded during the study phase. To achieve this, authors cropped EEG data, used the wavelet transform, analysis of variance (ANOVA) feature reduction and trained a support vector machine (SVM) on image cue onset in the study phase to test on the visual or verbal task cue in the retrieval phase. The authors achieved a statistically significant accuracy indicating the possibility of generally recognising the brain activation during memory retrieval when comparing to brain activation when first perceiving the origin of the memory. The support vector machines achieved a higher accuracy during the verbal task when compared to the visual task.

2.2 Time-frequency Analysis

Wavelets & Morlet

One of the two methods of time-frequency analysis used in this study is a discrete Morlet wavelet convolution. A wavelet is in general an oscillating function used as a basis to decom-

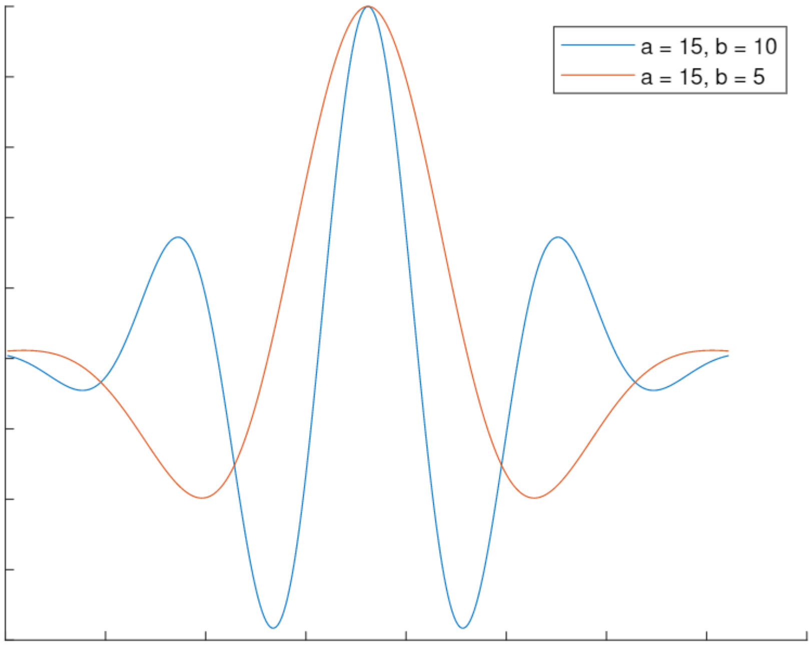


Figure 2.1: Two example wavelets initialised with the same bandwidth $\frac{1}{a}$ and different frequencies. The wavelets are depicted here in continuous time. When applied in the network each wavelet is sampled to discrete representation that depends on the sampling frequency of the data analysed.

pose complex functions into frequency components [Prakash, 2018]. In this study we restrict ourselves to applying the Morlet wavelet, but other design choices are possible. It is implemented in the network as a convolution of the signal with a window containing a real-valued Morlet wavelet, defined by the equation:

$$w(t) = e^{-\frac{a^2t^2}{2}} \cos\left(2\pi bt\right), \tag{2.1}$$

where the two parameters determine the bandwidth $\frac{1}{a}$ and central frequency b of the wavelet respectively [Zhao et al., 2019]. Figure 2.1 shows two example wavelets with parameters initialised within the range used in our layer design. When the wavelet is convolved with a given signal, it is these characteristics that determine what range of frequencies is captured by the convolution. If for example the wavelet with central frequency 5 Hz is used, the method captures the intensity of frequencies in a range centered around 5 Hz, the width of which depends on a. This method also retains phase information at the frequency given by b. [Zhao et al., 2019]

The Scaled Reassignment Spectrogram

The *spectrogram* is one of the most well-established methods for time-frequency analysis of stochastic processes. It displays the signal contents in an easily interpreted two-dimensional image showing the power of all component frequencies over time [Boashash, 2016]. A trade-off between resolution of individual peaks and spectral leakage at low-amplitude frequencies is determined by the choice of window function used to calculate the spectrogram. Due to general uncertainty there is an inherent trade-off between resolution of time and frequency depending on the chosen window length. When analysing noisy or complex signals, however, the spectrogram is less legible and has difficulty resolving the signal contents into well-defined separate components.

The reassignment method introduced in [Auger and Flandrin, 1995] and then further elaborated to scaled or matched reassignment method by [Sandsten and Brynolfsson, 2015] are both designed to improve the time and frequency resolution of single components compared to the ordinary spectrogram. Using a window and scaling factors designed to match the examined component the method can resolve a single, well-defined peak located at the "centre of mass" of the component. The *reassigned spectrogram* is defined as

$$RS_x^h(t,f) = \iint S_x^h(s,\xi)\delta(t-\hat{t}_x(s,\xi), f-\hat{f}_x(s,\xi))dsd\xi \tag{2.2}$$

where S_x^h is the spectrogram defined as

$$S_x^h(t,\omega) = |\int x(s)h^*(s-t)e^{-i2\pi fs}ds|^2$$

calculated using the short-time Fourier transform (STFT) with a chosen window function $h(t)$. Here t and f denote the time and frequency. The spectrogram values are relocated to $\hat{t}_x$ and $\hat{f}_x$, which are in turn calculated as

$$\hat{t}_x(t,f) = t + c_t \operatorname{Re}(\frac{F_x^{th}(t,f)}{F_x^h(t,f)}) \tag{2.3}$$

$$\hat{f}_x(t,f) = f - c_f \frac{1}{2\pi} \operatorname{Im}(\frac{F_x^{dh/dt}(t,f)}{F_x^h(t,f)}) \tag{2.4}$$

Here F_x^h, F_x^{th} and $F_x^{dh/dt}$ represent the STFT of the signal using $h(t), t*h(t)$ and $\frac{dh(t)}{dt}$ respectively as the window function. c_t and c_f are the scaling parameters and the special case $c_t = c_f = 1$ corresponds to the usual unscaled reassigned spectrogram. In this report we will exclusively be using the scaled reassigned spectrogram with specific values of the scaling parameters derived below. When we refer to the reassigned spectrogram from this point onward, this should be taken to mean the scaled reassigned spectrogram with this specific choice of parameters.

Using the example of a Gaussian windowed function

$$x(t) = e^{-\frac{(t-t_0)^2}{2\sigma^2}} e^{-i2\pi f_0 t} \tag{2.5}$$

[Sandsten and Brynolfsson, 2015] show that Equations 2.3 and 2.4 evaluated for this specific function give

$$\hat{t}_x(t,f) = t - c_t \frac{\lambda^2}{\lambda^2+\sigma^2} t \tag{2.6}$$

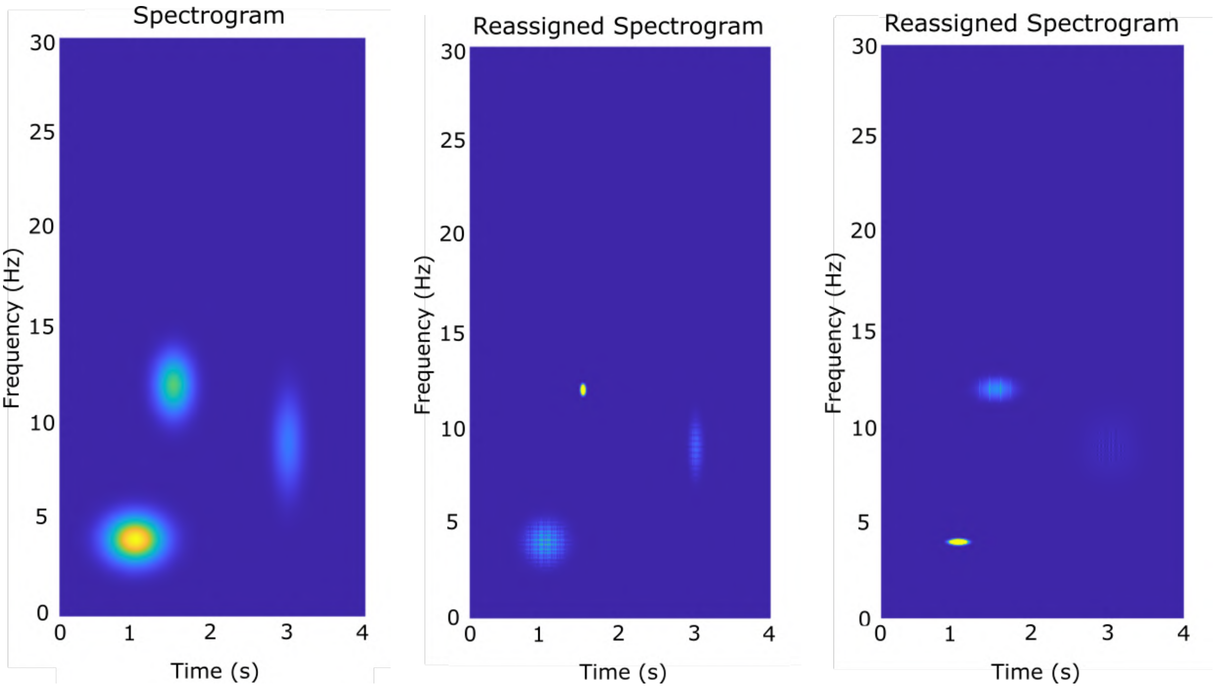
$$\hat{f}_x(t,f) = f - c_f \frac{\sigma^2}{\lambda^2+\sigma^2} f \tag{2.7}$$

where λ is the scaling parameter defining the shape of a unit energy Gaussian window $h(t)$ used. Furthermore, they show that in order for the spectrogram of this particular function to be entirely localised to the origin $\hat{t}$ and $\hat{f}$ should be zero. This gives the choice of the scaling parameters as

$$c_t = \frac{\lambda^2+\sigma^2}{\lambda^2} \tag{2.8}$$

$$c_f = \frac{\lambda^2 + \sigma^2}{\sigma^2} \tag{2.9}$$

which authors further specify by making a design choice of setting $\lambda = \sigma$, giving the *matched reassigned spectrogram*. This leads to $c_f = c_t = 2$ for perfect localisation of Gaussian signal components defined as equation 2.5. Figure 2.2 shows a comparison of the spectrogram and two different reassigned spectrograms for a simulated signal with three components. One can see that the method correctly reassigns different components in Figure 2.2b and 2.2c.



(a) (b) (c)

Figure 2.2: Reassignment localises specific components in the spectrogram by reassigning all nearby intensities to the "centre of mass" of components with selected length. In the figure we see the ordinary spectrogram, (a), compared to the reassigned spectrogram, (b) and (c), calculated with two different values for the parameter σ. The choice of sigma determines which of the components in (a) that the spectrogram intensity is reassigned to. In (b), the component at roughly (1.5 s, 12 Hz) is localised with a very large amplitude, while the other components are reduced in intensity. In (c), σ has been chosen to select the component at (1 s, 4 Hz) instead.

Sandsten & Brynolfsson, 2018

Aside from introducing the matched reassignment spectrogram [Sandsten and Brynolfsson, 2015] also conclude the possibility of the matched reassignment method to localise an unknown number of Gaussian components of equal lengths. In a subsequent paper, [Sandsten et al., 2018], by the same authors an extension is made that there is also localisation in the more general case when the window function $h(t)$ matches the transient envelope function of a component. In the same paper, [Sandsten et al., 2018] also conclude highly accurate time-frequency localisation by the matched reassignment method for components the method is not tuned against. They also showed accuracy in the presence of noise. Lastly, [Sandsten et al., 2018] bring up a possibility of using the method as a shape detector for unknown components, simply performing empirical search for parameters that reassign unknown spectrograms well. The authors present, beyond simple visual inspection of reassigned spectrogram, minimisation of the Rényi entropy [Baez, 2011] of a given reassigned spectrogram as a method used to find suitable spectrogram modifications, although this is a computation-heavy task.

2.3 Machine Learning & Neural Network

Multiclass classification is a problem posed in most natural sciences and computational models to perform such task are constantly being developed. One such model is the artificial neural network which can map any number of input data points to a class probability space. It does so by linearly sending information through layers of nodes (sometimes with non-linear activations in each layer) to a final layer of nodes each representing one class. To give a foundation for readers without a conceptual understanding of neural networks, a slightly sweeping foundation is presented. A mathematical representation of a neural network of depth L is for example given in [Caterini and Chang, 2018] as

$$F(x;\theta) = (f_L \circ \cdots \circ f_1)(x). \tag{2.10}$$

with a partial function f defined as a vector of artificial neuron model functions, given here for function number l,

$$\varphi\left(\sum_{i}^{n} \theta_i^l x_i\right) \tag{2.11}$$

where θ are weight parameters, φ is a activation function of a neuron and n is the number of inputs.

A softmax function,

$$\varphi_i(\mathbf{z}) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \qquad \mathbf{z} = [z_1 ... z_K] \tag{2.12}$$

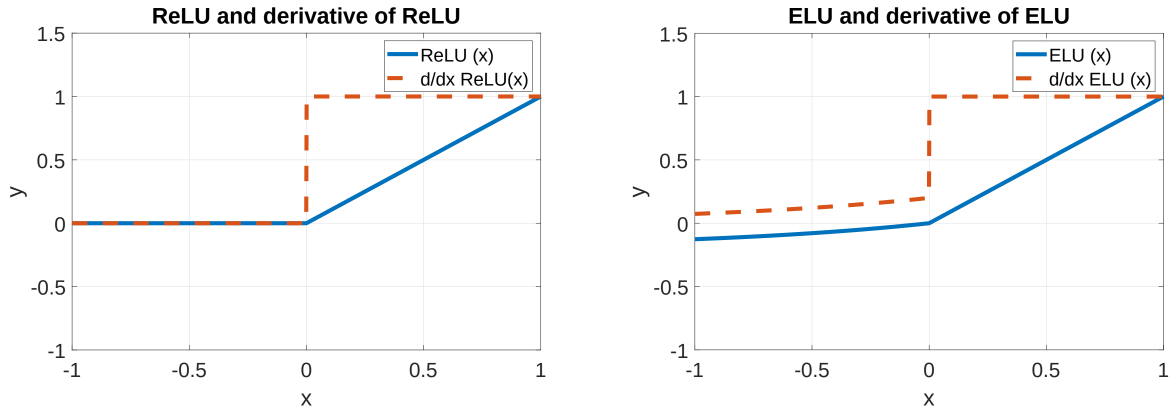
is finally applied to exponentially scale class node values z_i to the range [0,1], all values summing to 1 to give a probability representation. A classifier can then compare values and pick the class corresponding to the highest value. Each layer has a set of trainable parameters (also referred to as weights) that can be adjusted to change the predictions made by the network.

To train a neural network one must define a target for it to tune it self towards. For classification, a loss function is often supplied which the computer minimises in regard to given input data. In a multiple class classification context where outputs can be interpreted as class probabilities ($\hat{y}_i^k \in \mathbb{R} : 0 \leq \hat{y}_i^k \leq 1$ for sample i and class k) a normal loss is the categorical cross entropy defined, where k is specified to the correct label class, as

$$\text{Loss } = -\sum_{i=1} y_i^k \cdot \log \hat{y}_i^k \tag{2.13}$$

which, if outputs were probabilities would correspond to minimising cross entropy of predicted class distribution with actual class distribution. Minimising cross categorical entropy equates to minimising the Kullback-Liebler divergence [Kosheleva and Kreinovich, 2017] of the actual class distribution compared to the predicted.

Using the loss function as metric, a neural network can tune itself by auto differentiation, a method where the network utilises well defined mathematical functions and law of derivation to evaluate the derivative of output with regards to parameters of the model. The model can then



(a)

(b)

Figure 2.3: (a) Graph of ReLU activation function and its derivative. (b) Graph of ELU activation function and its derivative. The derivative of ELU is never zero (although it approaches zero asymptotically for extreme negative values).

iteratively loop an optimisation algorithm (for example gradient descent or *Adam*) to minimise the loss. The derivative can also be computed with regards to input for alternative conclusions.

The neural network will change weights during training to minimise the loss and improve the prediction of defined classes. The "deep" structure of neural networks will enable the networks to separate data with more complex patterns and structures.

Additionally, non-linear activation functions can help the model to solve non-linear classification problems. The output elements of layers are often sent through an activation function to augment the value in order to represent more complex connections. Examples of activations used in this study are ELU and ReLU.

ReLU An important problem to expanding the depth of neural networks is the vanishing gradient problem, where loss gradients are surpressed due to repeated sigmoid activations [Caterini and Chang, 2018]. A partial remedy to this problem is the ReLU activation

$$ReLU(x) = max(x, 0), \tag{2.14}$$

argued for in [Caterini and Chang, 2018], also shown in Figure 2.3a. The gradient of a ReLU activation is also easily calculated as 1 for $x > 0$ and 0 for $x < 0$. ReLU activation introduces instead the problem of shutting down some parts of the network due to the gradient being zero for negative input values. This is called the dead gradients problem.

ELU The activation ELU defined as

$$\mathrm{ELU}(x) = \begin{cases} x & \text{if } x \geq 0 \\ \alpha\left(e^x - 1\right) & \text{if } x < 0 \end{cases} \tag{2.15}$$

is very similar to ReLU but has a small but non-zero gradient for $x < 0$. The hyperparameter α defines the negative asymptotic limit of the activation function. The function is shown in Figure 2.3b. This means the ELU function doesn't suffer from either the vanishing gradient problem or dead gradients problem but is however slower to compute compared to ReLU.

Convolutional Neural Networks

Particularly relevant to this thesis study are convolutional neural neworks (CNNs) which are neural network models designed for images as input data. Layers of nodes form matrices representing the multiple channeled image sent into the model. The basic building blocks of a

CNN are convolution layers used to combine pixels/nodes with neigbouring pixels/nodes, creating new channels.[1] The trainable parameters of this layer is the element values of the matrix convolved with the image/data points. Thus a series of convolutions can send highly abstract information through the network while keeping the amount of parameters low compared to fully connected nodes. [Caterini and Chang, 2018]

Additional layers generally used in CNNs are dropout layers, normalisation layers, flatten layers and dense layers.

Dropout Layer Dropout is a layer which applies a probability to the connections of the layer to be removed. [Srivastava et al., 2014] presented the technique for regularisation. The layer prevents the model from learning co-adaptations of the entirety of the input data as during learning, the gradient is calculated only on a proportion of the data further upstream in the network compared to the dropout layer. Thus the layer increases regularisation during fitting.

Normalisation Layer A normalisation layer is usually added in the initial layers of a CNN. The purpose of the layer is to simply normalise the image data (sometimes over a large batch) which has been shown to increase performance of networks. The case of batch normalisation was presented in the paper [Ioffe and Szegedy, 2015] which was reconsidered to a layer normalisation presented [Ba et al., 2016].

Dense Layer Fully-connected dense layers treat matrix elements like nodes and connect input nodes to a determined number of nodes by multiplying each input with a uniquely tuned weight and applying a activation function to the output node. CNNs usually have one or multiple dense layers at the end of the neural structure. [Caterini and Chang, 2018]

Zhao et al. 2019

Previous work on learning feature selection and learning within a CNN in the context of decoding EEG data includes work outlined in [Zhao et al., 2019]. The authors present a novel network and layer in addition to a data augmentation method with the aim of addressing three main problems with EEG learning: the issue of feature selection, the issue of very large parameter spaces when learning features and the issue of training on small amounts of data. A schematic image of the network, named by the authors wavelet-spatial filters convolutional network (WaSF ConvNet), is shown in Figure 3.5. The first layer of WaSF ConvNet takes raw multi-channel EEG time series as input and applies 25 different wavelet transforms to each channel. The bandwidth and central frequency of each wavelet transform are defined as trainable parameters with random uniform initialisation over a range of frequency bands known to carry relevant information. The second layer spatially convolves all channels into a single representation, using 25 filters to enable multiple different weightings in the same forward propagation. The remaining layers are typical CNN layers used to reduce the number of dimensions to a single classification.

The method of data augmentation presented by the authors simply consists of cropping each data point into multiple windows and minimising loss over all new data points with a penalty towards differing labels for temporally adjacent data.

Authors of [Zhao et al., 2019] evaluated their network on three different data sets related to the motor imagery paradigm and the network compared well to other modern methods for the data sets in question.

[1] The concepts of nodes neighbouring each other is not present in neural network plainly built on dense connections.

Basic Knezevic & Heimerson 2018

A highly related, but separate problem formulation to ours was examined by [Basic Knezevic and Heimerson, 2018] in a master's thesis in 2018. In this thesis the authors evaluated the accuracy of feature selection methods and neural networks on the data set investigating semantic classification presented by [Bramao and Johansson, 2018]. The authors also made an effort to classify EEG during the retrieval phase. However, no models had any significant accuracy in this case.

The authors present multiple machine learning models with raw data as input as well as machine learning models requiring feature selection through time-frequency transformations. Models were applied to each subject who performed the visual task of the SM set (18 subjects) using the 10-fold validation average accuracy as measure of the accuracy of the models. As a final performance measurement an average over all subjects was presented for each model.

The highest performing network, when taking study phase accuracy into account, was a one dimensional convolutional neural network, denoted CNN1D, with around 90000 parameters. It was applied on raw data with no data augmentation and received an average accuracy of 0.82 in three-class classification.

In regard to feature selection significant to the work of this report, the authors [Basic Knezevic and Heimerson, 2018] present several time-frequency based feature selection methods. Of these methods the method of wavelet transforming input data and using a 2-D convolutional neural network, denoted CNN2D, performed the best, having a study accuracy of 0.62 for three-class classification.

3

Method

3.1 Data Sets

In this study, multiple data sets were used to evaluate methods and networks. Initially, simulated signals were used to establish a baseline validity of our networks. Secondly, the data set used by [Basic Knezevic and Heimerson, 2018] was used to compare the performance metrics with established neural networks for EEG-data as well as to reflect upon the experiments conducted in [Bramao and Johansson, 2018]. Lastly, and paramount to our context, our methods are assessed on a novel data set collected as part of a project at the Department of Psychology at Lunds University, which the authors also collaborated with and assisted. As a preface, deeper technical rundowns of each of the three data sets are given.

Simulated Set

Initially a simulated data set was used, consisting of single channelled data generated from 3 component Gaussian wavelets. Added one-over-f (pink) noise, alpha-noise and measurement (white) noise aimed to represent expected noise in real EEG data according to methodology presented in [Barzegaran et al., 2019]. The authors of [Barzegaran et al., 2019] combine the three noise components with weights estimated to resemble resting state EEG, i.e. EEG signals from brains not performing specific tasks. However, since we are not trying to create representations of EEG signals, but simply create data sets with similar inherent noise we settled for an equally weighted noise, in regards to power of the three components.

Semantic Memory Data Set (SM)

The Semantic Memory data Set (SM set) consists of task-related EEG-data from an experiment conducted in the report presented in [Bramao and Johansson, 2018], mentioned in section 2.1. As mentioned each experiment consisted of two separate phases, a study phase and a retrieval phase, shown in Figure 3.1. The study phase consisted of 180 instances of the subject being presented with a the word/image combination. The second phase of the visual task was comprised of the subject trying to recollect these 180 images again one by one.

Each phase resulted in a subset: the SM study set and the SM retrieval set respectively. Important to classification and learning is that each EEG-signal consists of 31 channels (one for each electrode) and was sampled with a frequency of 512 Hz. Each sample consisted of EEG data epoched from -1.5 to 2500 ms from post-image-stimulus onset. Additional data preparation consisted of manual eye and muscular activity removal, artefact removal through independent component analysis and manual inspection/cleaning of EEG signals. This was done by the authors of [Bramao and Johansson, 2018].

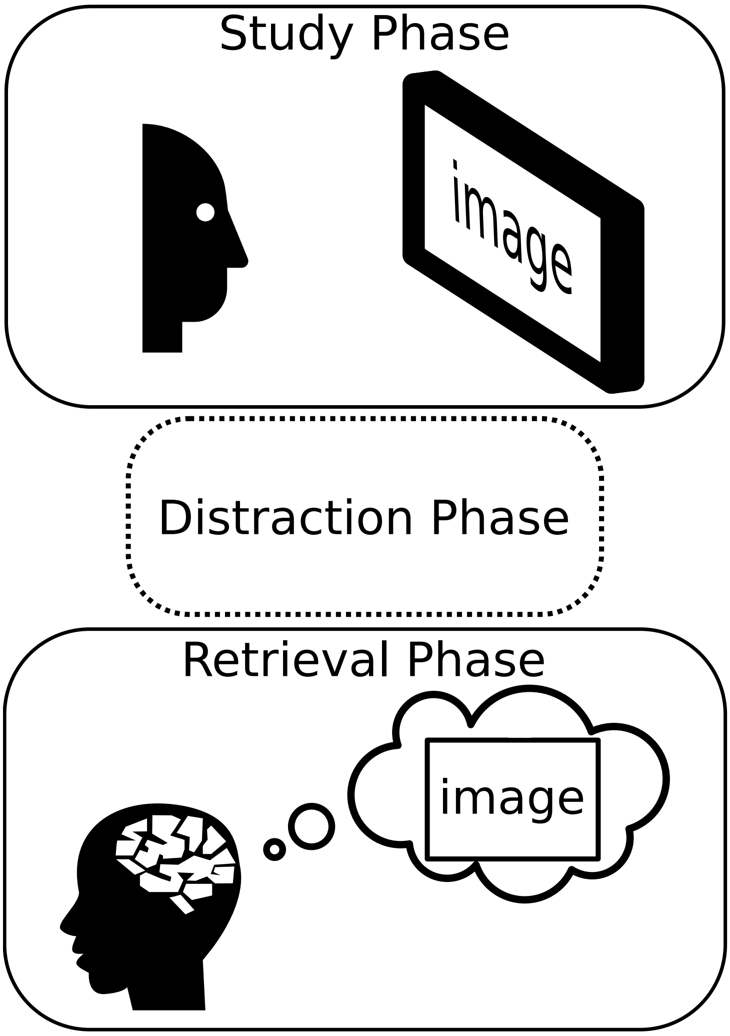


Figure 3.1: Outline of experiment structure used to generate the two non-simulated data sets. Both experiments were split into three phases: study phase, distraction phase and retrieval phase. Study comprised of seeing images of different classes and trying to remember them. Distraction consisted of arithmetic tasks to empty working memory. Lastly in retrieval phase memories of the same images were sparked for the participants. EEG readings were captured during all phases and two separate subsets of each set were saved, one for the study phase and one for the retrieval phase.

Emotional Memory Data Set (EM)

The novel emotional memory data set (EM set) was collected simultaneously to this thesis work and also resulted in two different data sets for each subject performing the experiment: the EM study set and the EM retrieval set keeping the same structure as the SM set.

The design of the experiment originated from a collaboration between Sterre Van de Langenberg, Mikael Johansson and the authors of this thesis. The experiment consisted of 8 blocks with each block containing a study phase and a retrieval phase. In the study phase the participant was run through 24 loops of being presented a word for 1.5 seconds, an image for 1.5 seconds and the word on top of the image for 2 seconds. In each loop, the participant was asked to memorise the word-image pair. The ensuing retrieval phase exhaustively presented the words to the participant again and subsequently the participant was asked three multiple choice questions about the content of the image. First two prompts in a random order: Choose emotional context of image (positive, negative, neutral) and choose semantic context of image (face, scene). These questions correspond to Q1 and Q2 in Figure 3.2. For each question subjects also had the possibility to enter "don't remember". The participant is lastly presented with a question to either correctly identify whether an image is mirrored (if the image was labelled "scene") or if the subject had seen the exact image before (if the image was labelled "face") given either the correct image or a different one of the same face. This is to check whether the subject has remembered the image in its entirety or merely associates the word presented with certain characteristics like for example emotion, colour or general composition. The distinction of two different last questions originates from the fact that it is harder to distinguish axial symmetry orientation in face images than scene images.

Between phases the participant is asked to perform an arithmetical counting task, continuously subtracting seven from a random integer for a period of time to empty the working memory of any information. This is important since we are interested only in the reactivation of brain regions due to the short-term memory, not in remaining impressions of the last few images that could still be active in the working memory.

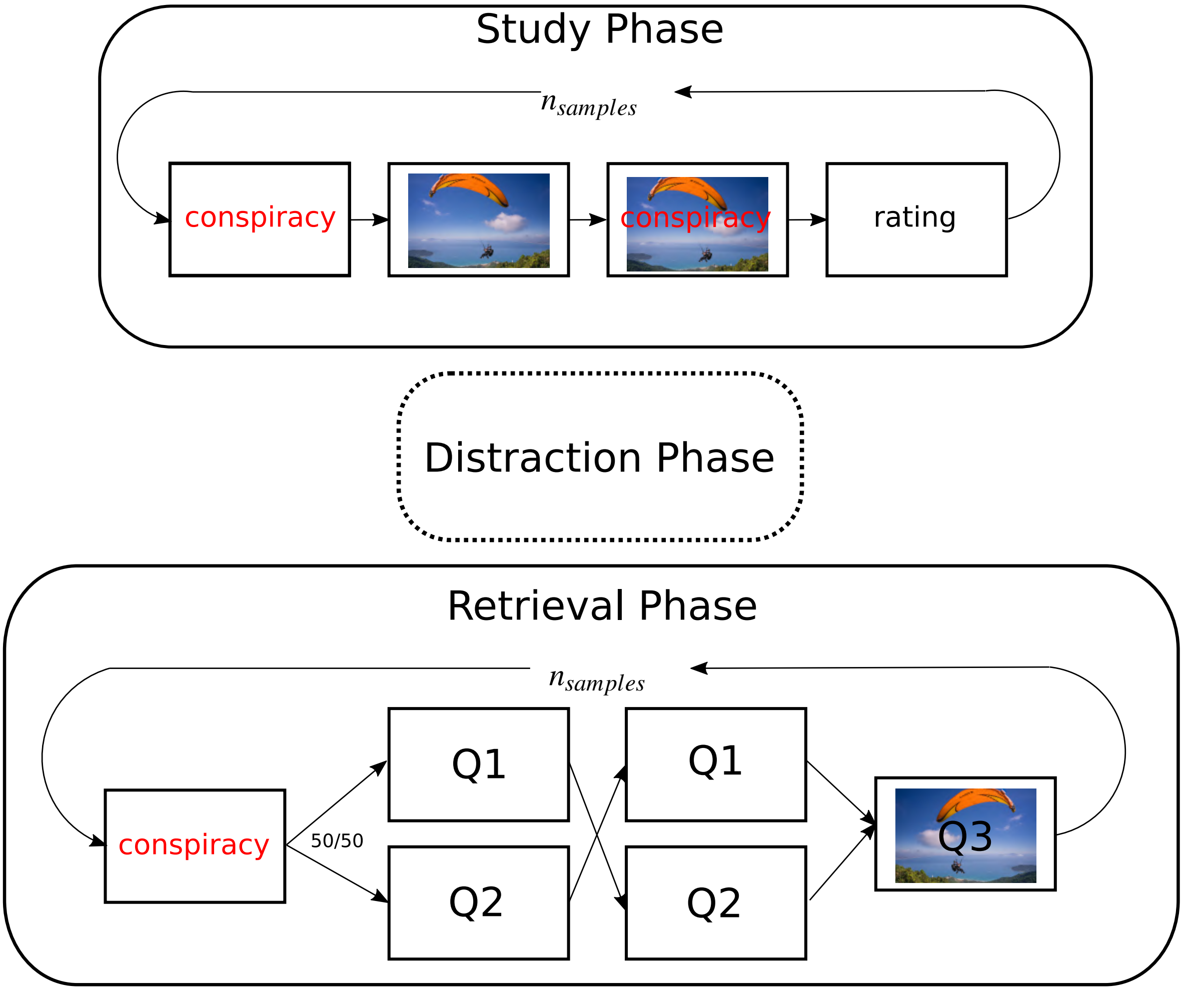


Figure 3.2: Design of one block of the novel experiment regarding emotions. A total of 8 blocks with 24 images each were recorded for each subject. The red word is related to the image shown in the study phase. During the retrieval phase only the word is shown and the subject is asked questions about the corresponding image. The questions about image content (here represented by Q1 and Q2) are presented in random order. Q3 prompts the subject to decide if the shown image coincides with the one shown during the study phase.

Each image bore two different types of labels: "face" / "scene" and "positive" / "neutral" / "negative". The image sets used were faces from The NimStim set of Facial Expressions by [Tottenham et al., 2009] and scenes from The International Affective Picture System (IAPS) by [Lang et al., 2008]. Faces were emotionally categorised by classifying their label according to existing ratings. Happy and surprise were classified as positive, calm and neutral as neutral. Fear, anger, disgust were classified as negative. Scenes were categorised through the IAPS rating system taking three metrics into account: valence, arousal and dominance. They were chosen so that the average valences were 3,5,7 for negative, neutral and positive class sets respectively as well as minimising the variance of the valence. Due to easier overreactions (connected to the arousal and dominance metrics) to negative valence images the image sets were manually sifted through to ensure there was an overall balance in the images.

In total, data from five subjects was used in the report to perform analysis. Experiments were conducted in a Faraday cage to reduce electrical background interference in the signal. 62-channel EEG time series, with a sample frequency of 1 kHz, were recorded continuously throughout the experiment using an active-electrode EasyCap and Neuroscan SynAmps RT 64-channel Amplifier. A vertical electrooculogram (VEOG) electrode was placed below the left eye to use for artefact extraction. Figure 3.3 shows one of the authors in the Faraday cage, wearing the measurement equipment. Data preparation for all tests unless otherwise is stated consisted of filtering data through a band-pass filter with lowest frequency 0.2 Hz and highest frequency 40 Hz, performing independent component analysis (ICA) and removing VEOG (and generally unreasonable) components. The data preprocessing and cleaning was conservative and as automatic as possible as a result of a lack of experience on our part. The data was then split into epochs from -100 to 1500 ms from image stimulus onset, with baselining from the first 100 ms of the epoch. Finally, the epoched data was downsampled to 512 Hz.

3.2 Software

The experiment used in the collection of the novel EM data was implemented in PsychoPy 3.0 [Peirce et al., 2019]. The subsequent artefact removal was done in MATLAB R2020a (the version used in all MATLAB implementations) using the FieldTrip toolbox for the SM set. For cleaning and epoching of the EM set the Python toolbox MNE was used [Gramfort et al., 2013]. Preprocessing of the data, which includes repackaging for all data and time-frequency analysis in the reassignment case, was also done in MATLAB. All neural networks were implemented in the TensorFlow 2.3 environment [Abadi et al., 2015] with Keras [Chollet et al., 2015] using Python 3.7. The Adam optimiser, as introduced by [Kingma and Ba, 2014], was used with default settings for all training unless otherwise is specified.

The TensorFlow environment had many advantages, including its accessibility (being open-source). Customisability was also a priority due to the unorthodox first layer of our Morlet transform network. This layer was implemented as a custom layer using only backend methods compatible with TensorFlow's automatic gradient calculation. The full implementation of the final version of this layer, which is the one used in all presented data for the network, can be found in Appendix A.

The MATLAB implementation of the function used to generate the reassigned spectrogram can be found in Appendix B. The function generates reassigned spectrograms separately for each channel of the EEG as well as ordinary spectrograms using the same window function for visual comparison. In this case the original signal is downsampled with a factor four using the MATLAB function `decimate` before computing the spectrograms. The final reassigned spectrogram is then cropped to contain only the 1.5 seconds directly after stimulus presentation. This

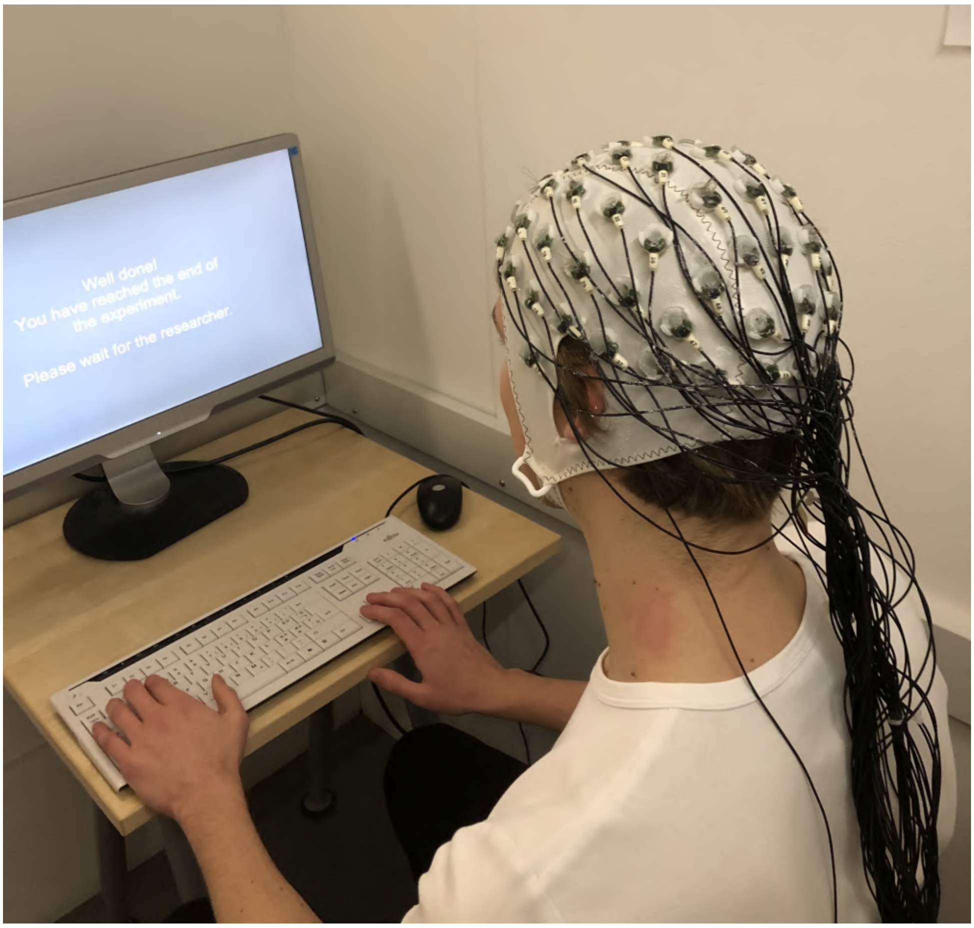

Well done!
You have reached the end of the experiment.
Please wait for the researcher.

Figure 3.3: Subject in the Faraday cage used for the experiment. Stimulus is presented on the computer screen and the subject responds using the keyboard.

is partly done to emulate the previous study by [Basic Knezevic and Heimerson, 2018], but is also justified by [Bramao and Johansson, 2018] showing that the majority of the predictive information is contained within this time interval.

For more details regarding implementation of methods, see `github.com/ohlindavid/ExjobbEEG`.

3.3 Morlet Network

Initial Layer

In the original paper by [Zhao et al., 2019] the WaSF ConvNet is proposed in which the first layer selects the features to be analysed by letting the parameters a_η and b_η in the Morlet wavelet equation

$$w_\eta(t) = e^{-\frac{a_\eta^2 t^2}{2}} \cos(2\pi b_\eta t) \qquad \eta = 1,2,...,25 \tag{3.1}$$

be the weights of the layer. This design does not correspond well to any preexisting layers found in the Keras/TensorFlow environment. Instead a custom layer was created to perform the operations described in the original paper. A set of 25 windows with differently shaped Morlet wavelets are constructed, each using a separate pair of weights (a_η and b_η). The b_η's were initialised by uniformly random distribution in the interval 2-30 Hz and a_η's were all set to an initial value of 10 Hz. Every window generated is then convolved along the time axis with all electrode channels separately as shown in Figure 3.4, generating a three-dimensional data structure with dimension:

$$(time - (L-1)/2) \times N_{chan} \times 25 \tag{3.2}$$

which is passed on to the following layer. N_{chan} is the number of electrode channels and L is the window length measured in sample points. This means that the dimension corresponding to the parameter η roughly represents 25 different selected frequencies and can be treated similarly to how the frequency dimension is treated in other time-frequency analyses. The main difference is the order, which is randomly decided by the network and can only be determined by looking at the value of each corresponding b_η. This is important to take into consideration when drawing conclusions from the features, since even a network where the parameters b_η are initialised in order may later be ordered differently as the parameters change somewhat independently of each other.

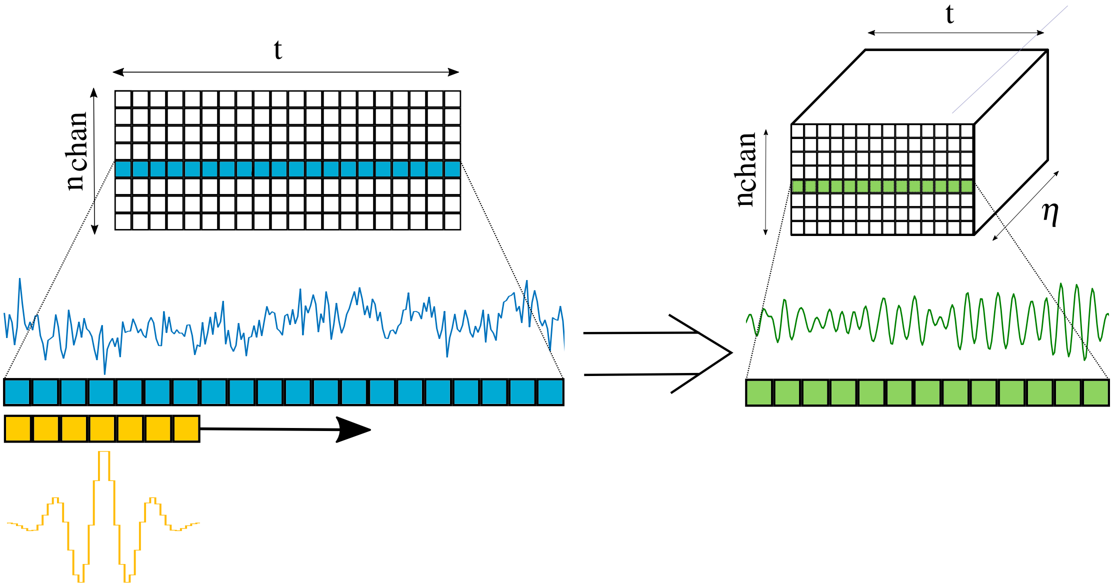


Figure 3.4: Schematic figure showing the operation of the Morlet layer. The signal of each electrode channel (blue) is convolved with 25 different Morlet wavelets (yellow) to distinguish the time varying power of different frequencies, one for each wavelet. One Morlet convolution returns a signal of one frequency (green) with contributions from surrounding frequencies depending on the bandwidth.This signal shows the presence (amplitude) of this frequency in the original signal over time.

By tuning both the central frequency and bandwidth of each Morlet wavelet, the idea is to initialise the weights so that wavelets are uniformly distributed in the desired frequency interval, 2-30 Hz for the semantic data. Much of the relevant information is thought to be located in this range (containing all bands except the gamma band) and the choice also allows us to safely reject high frequency noise in the data. This does not mean, however, that there is no relevant information outside this interval. Initial testing with this setting gave promising results and due to the large scope of examining all possible initialisations we have restricted our study to initially examine this setting. Some further attempts to use the interval 30-60 Hz (located in the gamma band) in combination with the lower bands were made to improve the performance on the EM data set, since these frequencies contain information relevant to emotional classification according to [Bazgir et al., 2018].

Given a sufficiently wide initial bandwidth the wavelets should, at least starting out, be able to register any event located in (and some distance outside) the interval. As the network trains, the intent is for the wavelets to then move towards those frequencies that contain predictive information, while at the same time narrowing the span of frequencies registered where advantageous to exclude noise. The position (in the frequency dimension) and bandwidth of the different wavelets correspond to the features that would have to be manually calculated and decided on before training the network in other non-adaptive applications. This also means that given a reasonable initialisation this network could be used for a variety of tasks with different features instead of having to tailor a network and the corresponding features to each problem. This is especially useful for processes where the current theory is not able to fully determine the optimal features.

Network Structure

The actual first layer of the model, which was not mentioned above since it basically amounts to preprocessing, is a LayerNormalization set to normalise across the channel dimension. This ensures that the data entering the network is on a suitable scale, but does not change the relative structure of the data.

The subsequent shape of the network can be seen in Figure 3.5 and is almost identical to the network structure presented by [Zhao et al., 2019]. The first layer, being the most novel part of the network was manually designed to perform according to the theory presented by [Zhao et al., 2019], as described above, using 25 different Morlet wavelets with a window length of 0.36 seconds.

The second layer performs a spatial 2D convolution of all previously generated images by convolving the electrode channel space using 25 filters with a width of a single data point in the temporal dimension. Since a 2D convolution is used the frequency dimension is treated as the colour channels of the image, which means that they are added together after the convolution. The number of filters in the convolution layer were selected to match the number of Morlet wavelets (η) to ensure that the output of the layer can represent all information previously contained in the frequency dimension. The convolution layer uses the ELU activation function. Subsequently a standard 2D pooling layer with a pool size of 71 data points taking strides of 15 data points averages the produced image along the time axis to reduce its size (although the parameters of this layer were changed to match the length of the signals). Following this, a dropout layer with 75% likelihood is used and the result is passed on to a final dense layer with either binary or three-way classification depending on the task at hand, using the sigmoid or softmax activation function respectively.

3.4 Reassignment Network

Preprocessing

The model that uses the reassigned spectrogram as implemented here lacks the ability to adapt the main parameters of the time-frequency analysis (λ and σ) due to the technical difficulty of implementing this in the TensorFlow environment. Initially we had hoped to design this automatic adaptation by implementing a custom layer. However, the method used to calculate the reassigned spectrograms required many functions that did not support automatic differentiation in TensorFlow. This meant that in order to implement the desired layer we would have to either radically change the way we calculate the reassigned spectrograms or find and use a different

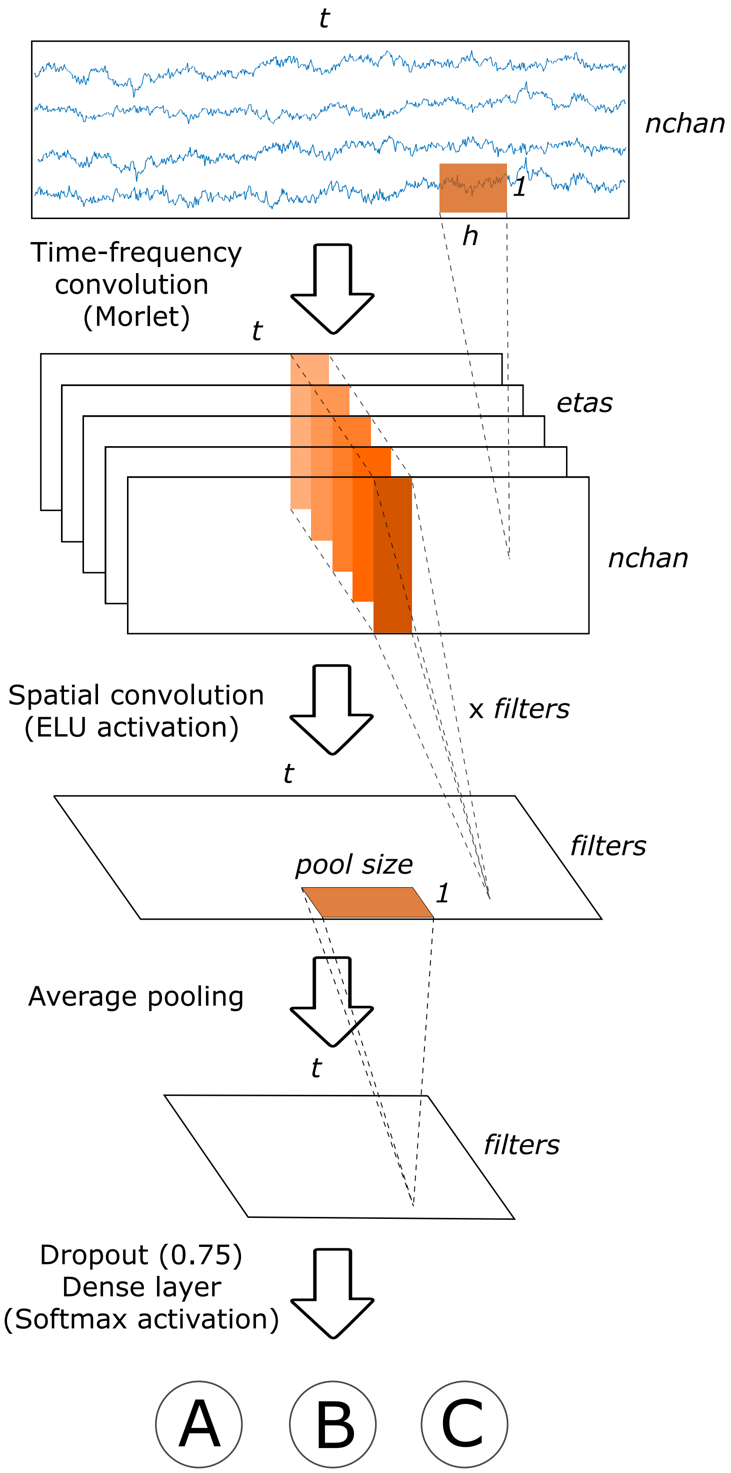


Figure 3.5: The architecture of the Morlet network, from normalised input to classification. Linear activation is used after all layers unless otherwise specified. The number of channels *nchan* varied between experiments, typically 31 (SM data set) or 62 (EM data set). The number of filters was set to 25 in all scenarios. The final dense layer had two or three nodes depending on the number of classes.

environtment with better tools for differentiation. This was judged to be far too time-consuming and we settled instead for a non-adaptive implementation.

The reassigned spectrogram can resolve most relevant frequencies using a single well-chosen parameter if the components analysed are somewhat close in frequency. The drawback to this approach is a deterioration of the resolution for certain components in the signal as a result of differing component temporal lengths and frequencies, but this was not enough to justify the time necessary for a manual implementation in a different environment. As a result, the reassignment model uses the reassigment of the spectrogram as pre-processing and a TensorFlow model takes these as input. The function used to generate the reassigned spectrogram of the input signal can be found in Appendix B. Due to the large size of the spectrograms calculated using the original sampling frequency of 512 Hz, the signal is downsampled to F_s = 128 Hz. This was done both to reduce the parameter count of the network and to reduce the size of the input data, which in turn shortened training times and reduced overfitting.

Following some of the conclusions by [Sandsten et al., 2018] the parameters λ and σ for unknown components in the EEG data can be empirically chosen. Thus we first considered the general scaled reassignment method. The estimation was performed by picking a few intuitive values of the window length $\lambda \times F_s$ (3, 5, 10, 12 and 15) and for each of these calculating the reassigned spectrogram for an interval between 1 and 50 for the parameter $\sigma \times F_s$. Visual inspection of the spectrograms show that at the ends of these spectra the signal is very distorted either in the time or frequency dimension and values outside of those tested are as a consequence very unlikely to give good results. The parameters giving the seemingly best resolution of most components in the spectrogram were then chosen as $\lambda \times F_s = \sigma \times F_s = 3$, and thus we settled for the resulting matched reassignment method. The window length parameter $\lambda = 3/F_s$ gives a window length of 0.234 s for F_s = 128 Hz. This length approximately matches the expected length of most signal components and roughly correponds to the one used by [Basic Knezevic and Heimerson, 2018], enabling a more direct comparisons of our results to theirs. The reassigned spectrogram is not sensitive to small changes in these parameter values, meaning that this very rough estimation of parameters should at worst only have a small negative impact on final performance compared to alternative noise in the processing of data. Additionally, the purpose of the proposed method is to correctly classify cases without prior knowledge of what specific frequencies are present in the different classes and extensive neurological research beforehand to better define these parameters would defeat this purpose.

Network Structure

The network using the reassigned spectrogram was designed to be identical to the one referred to as "CNN2D - v2" by [Basic Knezevic and Heimerson, 2018], in order to make valid comparisons to the previous results. The network is here referred to as "CNN2D - reassignment". This structure consisted of three convolutional layers using ELU activation, each followed by a dropout layer and an average pooling layer. This is then flattened and followed by two dense layers with 15 and 3 nodes respectively. A low learning rate was required for the network to converge. We used a learning rate of 0.0001, which is a factor 10 lower than the default used by TensorFlow's Adam optimiser, since it resulted in the best performance when testing on ordinary spectrograms.

Reconstruction

The results of our studies indicate that 2D-CNN may not be the optimal way to use the information in spectrograms to classify images in our data sets. Additional possible methods would be interesting to look in to. An alternate way to use information extracted in the reassignment method is to identify the time-frequency positions of the largest peaks in the reassigned spectrogram and use this information to reconstruct the signal again as input to the neural network. The idea is to effectively pick out these components and isolate them, enabling the recreation of noiseless signals retaining as much classifying information as be possible.

Our choice of method used the `imregionalmax.m` function in MATLAB to identify peaks in the reassigned spectrogram matrix. Progressing through all peaks in descending order of intensity removing other peaks located closer than a small distance to the current peak, in order to select as many different components as possible. This was especially necessary in the reassigned spectogram, as single components were usually reassigned to multiple peaks. We chose this distance empirically as a pixel length of 5 which removed peaks that we thought originated from the same component but kept close but distinct peaks. We also chose only to keep the 25 largest peaks that survived the purge of proximal peaks. The selected peaks for a given spectrogram and corresponding reassigned spectrogram can be seen in Figure 3.6. The choice of retaining only the largest peaks is not necessarily obvious as the noise components of EEG are of comparable magnitude to sought signal components. However, due to not having time to develop another more suitable option we settled for this simplification.

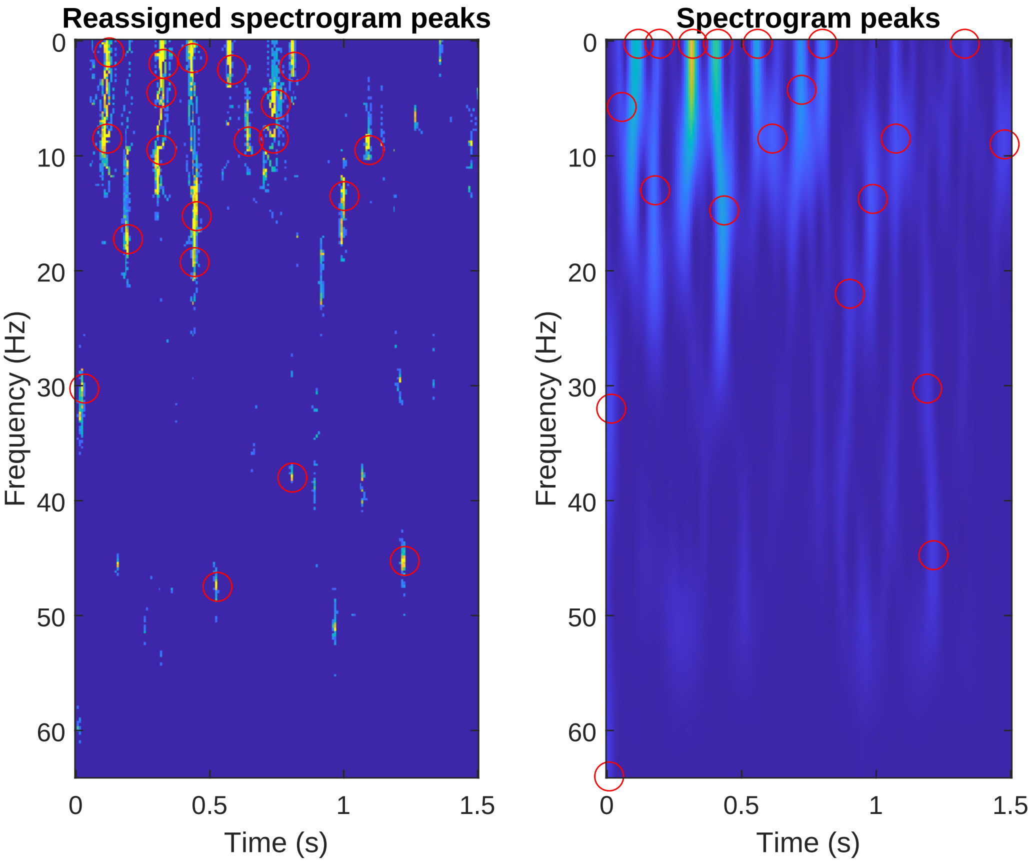


Figure 3.6: Selection of peaks in a spectrogram and the corresponding reassigned spectrogram. These peaks define the frequency and time of the components in the signal that is later reconstructed. Notice the difference in selected peaks in the two spectrograms shown.

During reconstruction, components were assumed to have time lengths of 0.1 seconds. Two different methods were used to estimate the amplitude and phase and recreate EEG signals. The first method simply used the relative powers at the time-frequency coordinates of the chosen peaks in the corresponding spectrogram (reassigned or untreated). Since the spectrogram contains no information about the phase of the signal it is not estimated and instead set to 0. The EEG signals are then estimated as a sum of Gaussian components with the calculated ampltitude, phase, temporal centre, frequency and temporal difussion. The signals generated in this way are termed *Spect* and *ReSpect* with amplitude originating in the spectrogram and reassigned spectrogram respectively.

The second method (Least Squares (LS) - estimation method) utilises the simple trigonometric relation

$$\alpha cos(x) + \beta sin(x) = \gamma cos(x + \phi) \tag{3.3}$$

which holds even though transient components are multiplied to each side of the equation. A LS-estimate of the two variables α_i,β_i for each component i in a n component estimation can be made through solving the system

$$X\theta = Y \quad \iff \quad \begin{bmatrix} \bar{v}_{cos} & \bar{v}_{sin} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} = Y \tag{3.4}$$

where $\bar{v}_{cos}$ and $\bar{v}_{sin}$ are vectors of generated transient periodic component vectors with each temporal centre, frequency and temporal diffusion calculated earlier and signal length corresponding to the EEG signal length. In addition, $\bar{v}_{cos}$ contains a simulated data vector created by a cosine function and $\bar{v}_{sin}$ is the respective data vector with sine function. Y is the raw EEG signal the reconstruction is aimed to reproduce. The reconstructed signal is calculated as $X\theta$. The signals generated in this way are termed $Spect_{LS}$ and $ReSpect_{LS}$.

3.5 Baseline Networks

CNN1D To act as a baseline for model performance comparison we have chosen to implement the best performing basic CNN1D network used by [Basic Knezevic and Heimerson, 2018] (referred to as "CNN1D"). This network uses only conventional one-dimensional convolutions of the raw signals, interspersed with dropout, pooling and normalisation layers. Since this is a very general model a higher accuracy when using the Morlet network would be required to consider the method beneficial for feature selection and classification.

CNN2D - Spectrogram In order to fairly judge the effect of using the reassignment method on spectrograms before training a second CNN2D network was trained on untreated spectrograms. The network was identical to the one used for the "CNN2D - reassignment" network and is referred to as "CNN2D - spectrogram".

3.6 Machine Learning Tools

Grad-CAM

To handle the unintuitive black box-like characteristics of neural networks several attempts to understand the information flow have been made. One particularly applicable to CNN is Grad-CAM (Gradient - Class Activation Mapping) presented in [Selvaraju et al., 2019]. The method visualises the gradient of class activations with regard to each pixel in the output of all convolution layers. Since previous convolution layers retain spatial information (in regards to the image) this method represents a coarse representation of the original image with gradient intensities as pixels. Pixels with high gradients have a large impact on the class prediction and are therefore judged to contain a large amount of predictive information. This method is developed for image analysis but has been applied and proven effective in finding sections of importance in EEG signals by [Jonas et al., 2019].

We have adapted this method of mapping gradients of top class predictions with regards to the outputs of our convolution layers in the models. One important difference is that we only have one dimension of input data that is structurally retained; the time dimension. The frequency dimension, as described above, is reconstructed by the convolutional layer in such a way that each vector in time is some linear combination of the original frequencies. This means that little intuition is gained by looking at the gradients in this dimension. The time dimension, however, is run through convolution and pooling but is not reconstructed in any other way. Thus our Grad-CAM-method is a one dimensional version of the method presented in [Selvaraju et al., 2019] with 25 feature channels.

If one views the output, A^k, from convolution layers of the network as 25 channels for a one dimensional image (with length T and increments i) and y^c as the non softmaxed class outputs, then the Grad-CAM heat map is calculated through

$$L^c_{\text{Grad}-\text{CAM}} = \text{Re}LU \underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}} \tag{3.5}$$

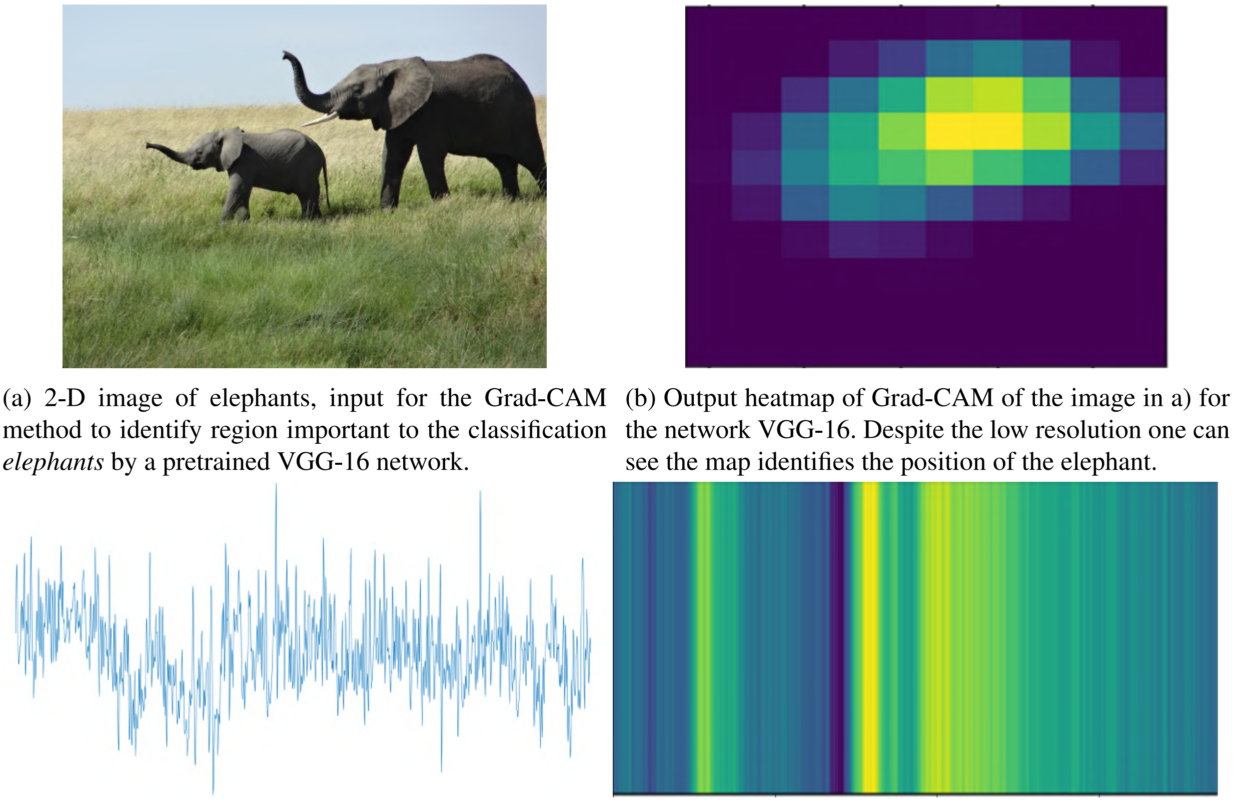
where

$$\alpha_k^c = \frac{1}{T}\sum_i \underbrace{\frac{\partial y^c}{\partial A_i^k}}_{\text{gradients via backprop}} \tag{3.6}$$

Transfer Learning

One of the main difficulties in the area of application for our networks is the lack of subject-specific training data. The approach we have chosen tries to minimise the amount of training data required to reach acceptable results by reducing the number of network parameters and initialising the network in a way that gives the model a decent start without loss of generalisation. While making a noticeable difference, these methods are limited in effect. They also require unnecessarily long training time to reach acceptable accuracy and risk overfitting on the available data. These drawbacks could impact many of the possible applications of the method negatively.

In order to alleviate these issues *transfer learning* is employed. This means that the network is initialised with layer weights learned from training on a large set of similar data from different subjects. The idea behind this is for the network to learn those features that are generalisable between subjects before training on the subject data to learn the subject specific features. This should in theory result in a network that takes less time to train on the specific subject, since the weights of the network are already somewhat well tuned and also prevent overfitting, since the network has already learned more generalised features instead of data specific patterns that only apply to this subject. We have implemented a very simple method of transfer learning by saving a previously trained network and initialising the network intended for a specific subject with the layer weights from the first network. As shown in the paper by [Zhao et al., 2019], more sophisticated methods of transfer learning also exist and have proven effective with similar data and models.



(a) 2-D image of elephants, input for the Grad-CAM method to identify region important to the classification *elephants* by a pretrained VGG-16 network.

(b) Output heatmap of Grad-CAM of the image in a) for the network VGG-16. Despite the low resolution one can see the map identifies the position of the elephant.

(c) 1-D input for Grad-CAM in this thesis study. The network in study is the Morlet network.

(d) Corresponding 1-D heatmap of input in c) for classifying the signal. As one can see the resolution in time gives us a possibility to analyse the networks reasoning about valuable time data points.

Figure 3.7: Set of figures comparing a 2-D case and a 1-D of Grad-CAM.

It is not implausible that many of the processes measured by the EEG are highly individual. This means that although some rough general features can be learned by the transferred weights for some tasks, the effect is limited or sometimes even negligible. In order to determine for which networks and data sets transfer learning would be effective, we used the accuracy achieved by the network trained on multiple subjects on a validation set taken from all subjects. If this accuracy is not significantly above chance, we judge the network unable to extract any useful features that generalise across subjects and do not present any separate results for the transferred network.

4

Results & Discussion

This section presents the results of all tests that we have designed to compare the performance of the models and analyse the novel data collected. The results are presented in three sections, one for each data set. Generally, the accuracies presented for experimental data are calculated using 10-fold cross-validation, which is suggested to be an optimal compromise between negative bias (due to small training sets) and variance (due to small validation sets) according to [Raschka, 2018]. These accuracies are then averaged over all 18 subjects for the SM data set and five subjects for the EM data set unless otherwise is stated. The low number of test subjects in the EM study analysed is unfortunate, since it likely induces a larger variance in these results. This is a consequence of the experiment being severely delayed by the ongoing pandemic. No test set separation was done because of limited intra-subject data sizes for both SM set and EM set. For the test on simulated data, the exact accuracy is not important and only a basic method to reduce the variance induced by random initialisation is used.

The chapter is concluded with a more thorough discussion of interdisciplinary insights and observations mainly from the EM set.

4.1 Insights from Simulated Data

In order to know more about how the models work we initially designed test cases using simulated input data. This is intended both to test the methods in a controlled environment where parameters can be independently tuned and to compare the performance of the methods.

The first test was intended to compare the noise sensitivity of the two time-frequency analysis methods. Three signals (shown in Figure 4.1) of length 1 s, each with a single Gaussian component of identical amplitude, were generated such that two of the signals had their components occur at the same time (0.4 s) but with separate frequencies (8 and 12 Hz) while the third occurred at a later time (0.7 s) with the same frequency as the first signal (8 Hz). The thought behind this was to ensure that the networks could distinguish between events separated only in time or frequency. From these three signals evenly divided training (totalling 180 data vectors) and testing (totalling 60 data vectors) sets were constructed by adding simulated EEG noise as described in section 3.1 (under Simulated Set) with different average power relative to the signal. In Figure 4.2 the results of the two models are compared for sets generated with different amounts of noise. Each model was trained for 25 epochs and then evaluated on the testing set. This was repeated with five separate initialisations to give a more stable result. The models used were the same as those used for similar signal lengths (1.5 s) in the rest of the study.

From these results it is clear that both models perform their basic function correctly and are easily able to separate the three signals when little noise is present. The test also indicates that the Morlet network is less sensitive to noise since it is able to perform well even for very noisy signals, with a classification accuracy above 98 % when the power of the noise is up to ten times larger than that of the signal (SNR = 0.1). It should be noted that given the simplicity of the task

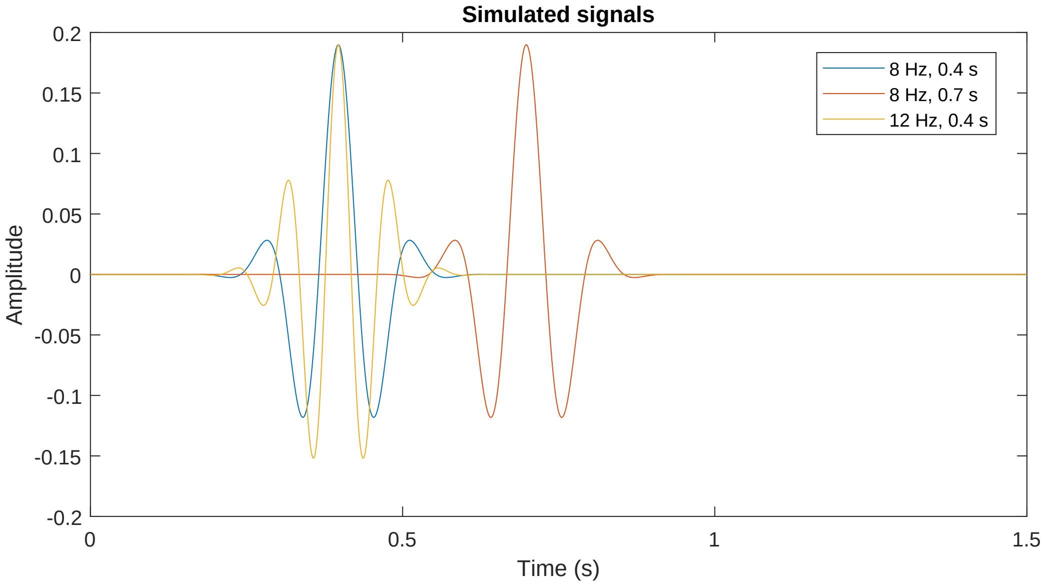


Figure 4.1: The three simulated signals used to test the noise sensitivity of the Morlet and reassignment networks, presented here in the same figure without added noise.

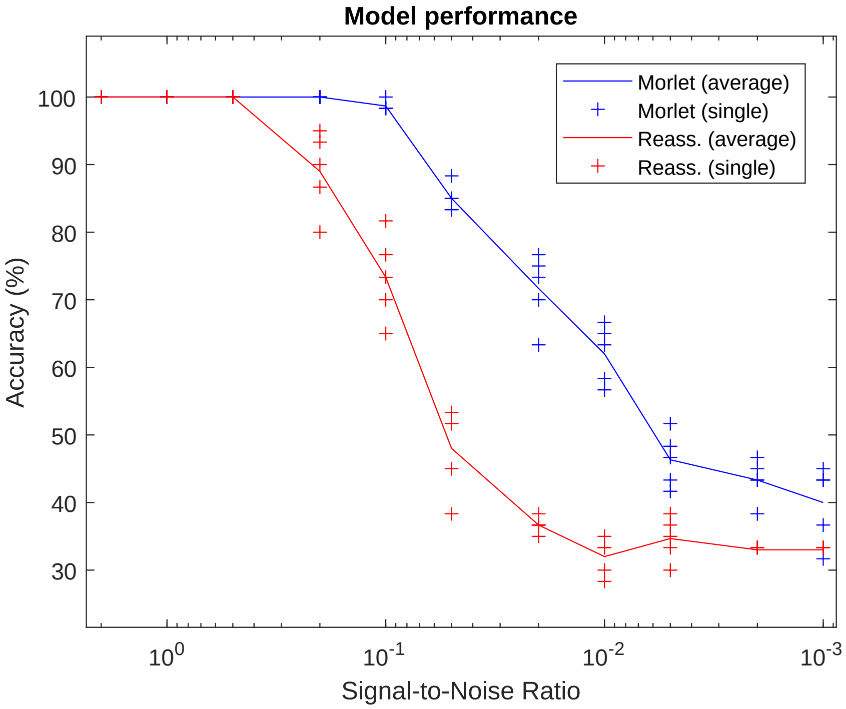


Figure 4.2: Performance of the Morlet network (blue) and the CNN2D - reassignment (red) trained for 25 epochs on three classes of simulated signals with different levels of noise. The Signal-to-Noise Ratio (SNR) specifies the power of the signal relative to the noise in the generated data sets. Each test was performed five times to account for the random initialisation of the model and the resulting accuracies were averaged. The accuracies are calculated as the fraction of correct predictions on an unseen set of 60 signals with added noise.

this performance will likely be worse in a real scenario for both models, but the comparison between the two models is still a good indication of which is more noise sensitive. It should also be noted that for signal lengths of 1.5 s the Morlet network is significantly smaller than the reassignment network (4753 parameters for Morlet compared to 8219 for reassignment). Additionally, the training time was longer for the reassignment network, even when excluding the preprocessing of the data.

4.2 Results from SM Data Set

Model accuracies and comparisons

To start the results and discussion section of the SM set, the overall performance of investigated networks are presented in table 4.1. As initial discussion, the table shows clearly significant performance for three-way classification for each new model. New models are also compared to existing counterparts presented in [Basic Knezevic and Heimerson, 2018], named CNN1D and CNN2D networks which were designed by the cited authors. [Basic Knezevic and Heimerson, 2018] also applied their two networks to the SM set, and achieved a higher accuracy. However, as the accuracies were not replicable the new inferior accuracies resulting from our tests of these networks are included for a fair comparison to network models presented in this report.

Table 4.1: Table showing predictive capabilities of our networks applied to classify semantics (faces/landmarks/objects) in the SM set. Presented values are the average 10-fold cross-validation accuracies, averaged over all subjects. The Morlet network performs best and compares well to the previously developed methods (CNN1D) and the CNN2D networks perform similarly to similar methods applied on ordinary spectrograms. The classification is between the three semantic classes, meaning that random chance would give an accuracy of 0.33.

Network \ Data Set	SM Study set
CNN1D	0.66
Morlet	0.74
CNN2D - Spectrogram	0.45
CNN2D - Reassigned Spectrogram	0.44

In table 4.1, the one dimensional models are in general better compared to the time-frequency convolutional neural networks. One partial explanation for this could be the higher amount of parameters needed in two dimensional CNNs as well as the fact that two dimensional CNNs are the best models for pattern recognition in images. Separating identities in time-frequency representations are locations and overall distributions of power in the spectrogram, not localised patterns. As seen in Figure 4.3, reassignment increases contrast in the time frequency image and introduces a property of sparsity to the image. This is a positive and negative characteristic since it is easier to work with while also creating very small gradient in many parts of the spectrogram. One should also note that the highest peaks in both spectrograms are not necessarily the most informative as EEG has high magnitude noise. It is also noteworthy that the reassignment network performs slightly worse than the CNN2D using ordinary spectrograms. This could be due to the CNN being even worse at classifying the more sparse reassigned spectrograms or suggest that reassignment, when used for signals containing components with different properties, needs to be more precisely tuned.

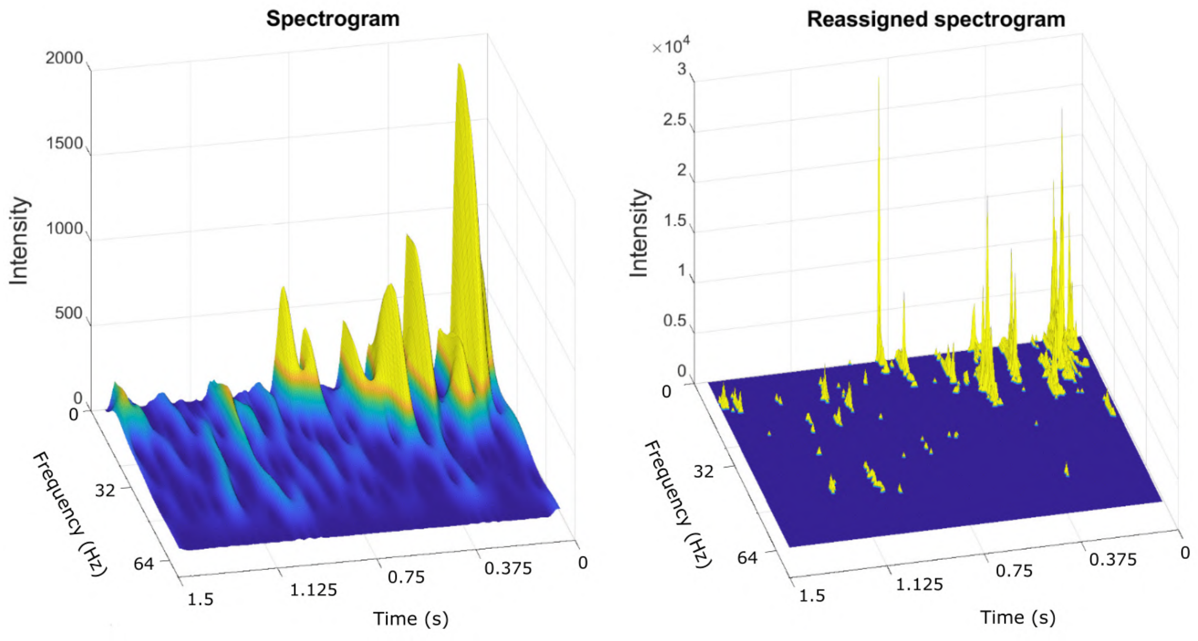


Figure 4.3: Three-dimensional plot showing the spectrogram and reassigned spectrogram of one trial for one EEG channel. Notice the large difference in intensity scale between the two subfigures and the much greater localisation of most peaks in the reassigned spectrogram.

Evaluation of Reconstructed Signals from Reassigned Spectrogram

Exploring alternative ways to process information from the reassigned spectrograms, different methods of reconstructing signals from spectrograms, as outlined under *Reconstruction* in section 3.4, were tested. The reconstructed signals were then classified using the Morlet network. As explained in the Method section, multiple different ways of estimating the amplitude and phase of components in the reconstruction were tried. The average cross-validation accuracy of all subjects of the semantic data set for multiple ways of reconstruction are presented in table 4.2.

Table 4.2: Average 10-fold cross-validation accuracies of Morlet network over all 18 subjects using reconstructed signals of the SM set as input data. Ordinary and reassigned spectrograms are denoted as $Spect$ and $ReSpect$, with LS signifying that least-squares estimation was used to match the phase and amplitude from the original signal. As one can see the reassignment and subsequent reconstruction slightly impairs class predictability compared to using the ordinary spectrogram. However, a lot of class separating information is retained despite the very sparse reconstructed representation, as can be seen in the accuracies of reconstructed signals. The classification is between the three semantic classes, meaning that random chance would give an accuracy of 0.33.

Signal	Acc.
Raw	0.74
$Spect$	0.44
$Spect_{LS}$	0.62
$ReSpect$	0.43
$ReSpect_{LS}$	0.59

An interesting result here is that no matter which reconstruction method was used there is a significant level of predictability. Due to time constraint the methods were not optimised over multiple small hyper-parameters such as discriminant distance between peaks, number of included peaks and actual reassignment parameters. Thus one could probably expect a slightly higher possible accuracy. Here we again find that the results for reassigned spectrograms are nearly identical and even slightly worse than those for ordinary spectrograms. This further implies that the reassignment does not contribute significantly when not properly adapted to the signal content.

Transfer Learning & Epoch Cropping

Transfer learning was implemented by training a Morlet network on study data from all subjects except one. This network was evaluated, trained and again evaluated separately on the excluded subject. Due to large training times for evaluation, the evaluation was repeated for six random subjects to give an average accuracy over epochs. In Figure 4.4 we can see that model transferring results in both faster training and a higher final accuracy. This is a comforting result, that data analysis of the SM set (and most certainly the EM set) can rely on the common performance improvements of transfer learning. As EEG data is limited, using similar larger data sets to increase learning speeds and validation generalisability even in unseen data is greatly appreciated.

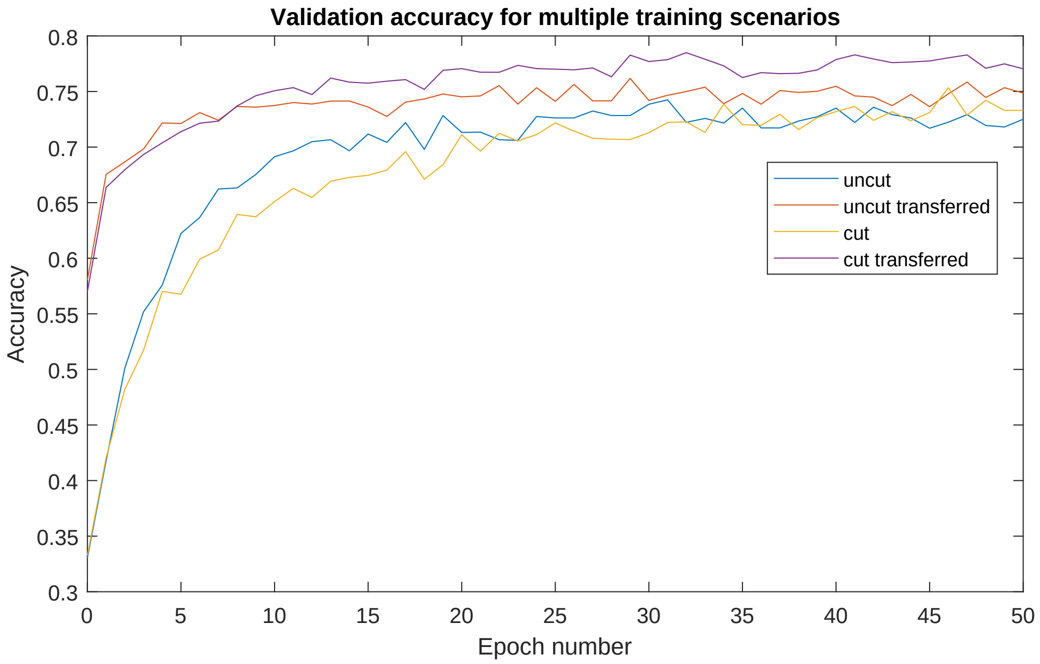


Figure 4.4: Performance of the Morlet network averaged over 6 different subjects. The model was trained in four differing scenarios. Two scenarios learning on whole epoch data and two scenarios learning on cut data from image onset to 1.5 second after image onset. Each of the pairs of scenarios had one scenario running on a freshly initiated network and the other starting out with a model transferred from a training scheme on data from 17 other subjects. One can see a clear increase in initial accuracy for transferred networks compared to newly initiated Morlet networks. Transferred networks also performed better after many epochs as well.

Furthermore, the initial validation accuracy (epoch 0 in Figure 4.4) reached an average of over 0.55 for the six subjects for the transferred models compared to the freshly initiated model predict accuracy of 0.33. From this we can see that there is indeed generalisability in the data, meaning that the processes classified are similar between subjects. For subjects with inherently high accuracy probabilities, the newly transferred model tested well on the subject. For example, the transferred Morlet network had, on subject 3, an accuracy of 0.83 before starting to train on the subject. This could be interpreted as that the classification model actually generalises fairly well over subjects overall, but has problems predicting strongly deviating signals from previously seen data.

Heat Maps of Important Time-Frequency Stamps

As an investigative method to find time intervals that have a significant effect on classification of the SM set, grad-CAM was used. The method was only applied to the SM set because of the large amount of total trials over subjects, which was required for creating reasonable non-overfitting models.

Figure 4.5 shows a grad-CAM heatmap for subject 14 after a pretraining algorithm has fitted a model towards 17 other subjects. The grad-CAM is averaged over multiple trials of different classes to get a reasonable grad-CAM, because grad-CAMs from individual signals were very noisy and hard to make conclusive observations on. The signal shown in the figure is not cropped, but has gone through convolutions (each shortening the signal by the length of one window) and thus image onset lies at approximately 1.25 seconds into the signal. Thus one can see that the grad-CAM method successfully identifies an interesting period from 150-600 ms after onset. Stimulus information is processed in this time range according to previous works on this data set as well as previous studies [Bramao and Johansson, 2018]. Thus grad-CAM may prove useful for investigation of unknown EEG processes, and perhaps even other time series, when using convolutional neural networks and when specific expertise is unavailable. In addition to these conclusions, the grad-CAM also highlights a problem with the model. This is the fact that pre-stimulus-onset ($t < 0$) the heatmap is not zero. Logically, all these pixels should have zero gradient and zero input into the classification since the subject has not been shown any image at that point in time. However, these pixels have a non zero gradient and thus make a difference in classification, which causes the model to be less accurate generally, since any gradient in this area is the product of overfitting. Thus, grad-CAM shows the strength for properly epoched and preprocessed signals for the performance of the Morlet network.

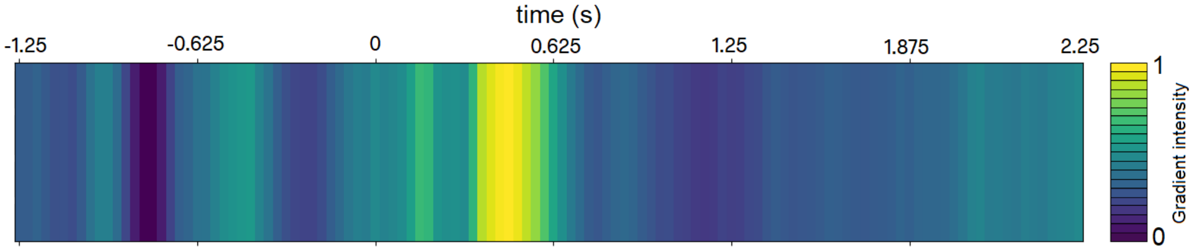


Figure 4.5: An averaged grad-CAM heatmap from multiple trials from subject 14 after pretraining a Morlet network on a pool of trials from multiple separate subjects. Gradient of the grad-CAM is normalised to the range 0 to 1. The input signal here is epoched from -1.5 to 2.5 second from stimulus onset in the study phase of the SM set. Although noise and variance in the gradient from the small sample size, the grad-CAM has arguably identified areas of interest in the signal as there is a highlighted time period soon after stimulus onset.

Retrieval Performance

Some tests were conducted on data collected during the retrieval phase in order to use this easily available data to compare model performance on a different task and draw some conclusions regarding the difference between the SM data and the novel EM data. Analysis of this data was, however, not the main focus of this thesis and as a result of time constraints and lacking data (for certain subjects) these tests are not performed for all subjects. Several tests were performed to test the hypothesised connection between signals during the study phase and the retrieval phase, with varying degrees of success. The accuracies presented in Table 4.3 were calculated averages from 10-fold cross-validation on three subjects using our best performing network, the Morlet network, on the SM data set. The three subjects (1, 3 and 8) were randomly selected from those that have a sufficient number of retrieval trials, at least around 90 trials, to give accuracy when using retrieval data as training set.

Table 4.3: 10-fold cross-validation accuracies of the Morlet network using data from different phases of the SM experiment as training and testing sets averaged over three test subjects selected to have a sufficient number of retrieval trials. Significant accuracies are achieved for networks trained on retrieval data both when testing on retrieval and study data, which was not the case in previous studies. The classification is between the three semantic classes, meaning that random chance would give an accuracy of 0.33.

Testing: / Training:	Study phase	Retrieval phase
Study phase	0.74	0.34
Retrieval phase	0.39	0.46

In the top left corner of Table 4.3 the accuracy when training and testing on data from the study phase (for all subjects, already presented above) is included for comparison. The first and most intuitive test on retrieval data (identified as training: study, testing: retrieval) is similar to those performed by [Bramao and Johansson, 2018] and [Basic Knezevic and Heimerson, 2018]. The test simply trains a network on the data collected during the study phase, hoping that the features extracted from this data can be generalised to give predictive accuracy on the retrieval data. This would support the theory that similar processes in the brain occur for observation and recall of an image. No significant accuracy (0.34) was found in this test, which was somewhat expected due to similar or only marginally better results from the previous studies on the same task. To analyse further, in order to examine if there is any predictive information whatsoever in the retrieval data that allows classification into the three semantic classes, a second test used retrieval data both as training and validation set. This resulted in an accuracy (0.46) that suggests that there is indeed information present in the trials that can be used to classify into the desired classes. A similar test performed by [Basic Knezevic and Heimerson, 2018] found no significant accuracy, but we can not definitively say if this is due to differences in the test itself or in model performance.

These results leave two possibilities for the relation between the two data sets. One is the possibility that the signals in the two sets are indeed not similar and originate from different processes. This would mean that there is no underlying connection that can be found by the model. We are hesitant to accept this explanation, since results indicating that there is indeed such a link have been presented by [Bramao and Johansson, 2018]. The second possibility would be that there is a connection, but that the method for some reason fails to find it. This could in turn have several causes, for example that the experiment fails to capture the signal at the moment of visualisation (this is discussed further below). Another possible reason for the lack of accuracy when training on study data could be that the network extracts features that are not useful for classifying retrieval data. Under the working assumption that there is some connection we devised a third test.

The third test on retrieval used the retrieval data as training set and the study data as testing set; a simple inversion of the original test, also similar to tests performed by [Basic Knezevic and Heimerson, 2018]. The idea behind the test is that a network trained on study data could be using features based on information from visual stimulus that is not present in the retrieval data. If the signal originating in information related only to direct observation is significantly stronger than signals that originate in processes that are common between observation and recall (the ones we are trying to find) this could (hypothetically) cause the network to rely entirely on features related only to direct observation. In reversing the problem, the hope is that the features extracted from retrieval data could be the ones that are common between the two sets. The test result from ten separately initialised networks trained on all retrieval data for ten epochs from three subjects and tested on all available study data (for each subject) gives an average accuracy of 0.39. Given the large testing set (in this case 185 trials) and the consistency of the average over ten separate initialisations (with outliers reaching accuracies above 0.51) and averaged over three subjects this seems to indicate a relationship beyond random chance.

An issue affecting the quality of the retrieval data in both studies is the time gap between the end of measurements and the questions in the retrieval phase. Data is only collected while the stimulus word is shown, after which the subject is given five seconds to reply to each question. This means that it is very possible that the processes we wish to measure do not occur until after the measurements have stopped. The result is that the subject could answer both questions correctly while the associated signal is devoid of any information about the relevant brain process. Additionally, the number of trials available for each subject is limited by the number of correct responses in the retrieval phase, since an incorrect response indicates that the correct mental image is not present. This means that some subjects have very few trials and cannot be reliably trained on the retrieval data.

4.3 Results from EM Data Set

Comparison of Semantic and Emotional Classification

Before any analysis on the performance and success of predictions of EEG in the context of emotional images can be done one should recognise a fairly large impediment. This is the plain fact that the authors of this thesis preprocessed the EM set. Due to a lack of experience within EEG preprocessing and artefact cleaning a more automatised and general application of preprocessing and data cleaning was chosen. Although unknown, a substantially lower theoretical accuracy ceiling exists for this set for several reasons discussed below. Therefore cannot reliably expect accuracies to be nearly as good as those obtained from the SM data set.

The Morlet network achieved an average cross validation accuracy of 0.41 (cohen's $\kappa = 0.12$) for classification into the three emotional classes after training for 30 epochs, averaged over all four subjects from the EM set. Initially this could be seen as quite low. During the training the cross validation accuracy lies above 0.33 which suggests an identification of some sort of emotion classification separability. A two-way semantic classification was also performed for this data set as well (since all images also had face/scene labelling). This test yielded a respective accuracy of 0.75 (cohen's $\kappa = 0.5$). We know this level of classification accuracy should be possible, as high accuracy has been redundantly shown for the SM set in this very report. This is not as good as semantic classification in the SM set, but this can reasonably be explained by poor prepossessing and cleaning of data. If one compares the respective Cohen's κ, as a simple

Table 4.4: Confusion matrix for validation predictions for 10-fold cross-validation, averaged over all subjects in the EM set. One can compare the prediction accuracies for each class as well as inspect the total accuracies. The classification is between the three emotional classes, meaning that random chance would give an accuracy of 0.33.

	Pred Pos	Pred Neu	Pred Neg
True Pos	0.43	0.32	0.26
True Neu	0.31	0.38	0.31
True Neg	0.30	0.30	0.40

way to compare model performance in the two cases, the two metrics do not coincide. This suggests that the model doesn't under-perform due only to the quality of the data, but instead that there is some problem with classifying emotions in the EM set.

To see if the model can predict certain classes more easily than other, a confusion matrix was calculated for each fold and averaged over folds and over subjects. This matrix is presented in table 4.4. Note that each row in the matrix sums to 1. One can see that the previously mentioned 0.41 accuracy is shown in the confusion matrix as well. If one more closely studies the matrix, one can note that when the model misclassifies positive images the model is more likely to classify the image as neutral. In addition to this, when negative images are misclassified there is an equal likelihood for a predicted positive or neutral label. This would suggest that the model can classify the negative against either positive or neutral image, but has a harder time to separate positive and neutral images from each other. However, a contradiction exists in the fact that there is an equal prediction chance for positive and negative images when misclassifying neutral images, according to the confusion matrix.

Multiple two-way classifications between two out of three emotion images were performed as a response to these observations from the confusion matrix, as well as 2-versus-1 classifications, however these performed equally to three-way classification, and thus we concluded that no certain emotional class separations were easier or clearer than another, at least in regards to our results.

According to previous reports, emotional content of images are coded in the gamma band which we had excluded during testing on the SM set [Mohammadi et al., 2015] [Bazgir et al., 2018]. Thus we conducted multiple model evaluations with b_η initialisation up to 60 Hz. During these tests we also reprocessed the raw data to filter with a low-pass filter frequency of 100 Hz to avoid filtering away any important components. This had no distinguishable effect on the performance of the model. Since this was implemented by simply spreading the same number of Morlet wavelets over a larger interval it is possible that relevant information between the wavelets was missed to a higher extent in these tests. An intuitive improvement would be to increase the number of wavelets proportionally to the increase in interval length, but this runs into the problem of giving the network far too many parameters. In combination with the very small data sets available we found that this only resulted in overfitting. However, given a sufficient amount of data this could be an effective option.

Uncertainty in the Emotional Memory Data

There are several aspects of the novel EM experiment that affect the quality and validity of the resulting data. One of the largest factors that potentially ruins parts of the data is mislabelling of the stimuli. This could conceivably occur for a variety of reasons, depending both on the content of the image and - in the case of emotional data - the subjective experience of the test subject. Some of the images in the stimulus set do not neatly fall into the semantic categories of "faces" and "scenes". For example, many "scenes" contain prominent faces in the image and it is not impossible that this could cause the subject to react to the image as if it belonged to the "faces" category. This would mean that the trial contained information that to the models looks like the reaction to a face but is labelled as a "scene". In the previous SM data set, the semantic categories were more separate ("faces", "landmarks" and "objects") although it is still theoretically possible that an image categorised as "landmark" could contain an object that the test subject reacted to. Although this kind of mislabelling is technically possible, we do not judge it to be a large source of error, in part due to our own experience as test subjects as we found little difficulty in classifying the semantic content of the images.

We believe, however, that the most significant contribution to mislabelling in the EM data is the subjective experience of the test subject compared to the emotional label. When partaking in the experiment ourselves we felt that a significant number of images were labelled differently than our emotional response. For example, images of skydiving and parachuting were labelled as "positive", but the reaction to these images for a test subject that is afraid of heights would likely be negative and the reaction of a test subject that simply does not enjoy skydiving could be neutral. Another illustrative example is the explicit erotic material that is categorised as "positive". The erotic material consists mostly of images of women and the reaction of the test subject can be assumed to vary greatly depending on whether or not they are attracted to women.

Since almost all of the images are to some extent subjective in their emotional effect we can expect a significant number of mislabelled trials for every subject. This causes two problems in the training phase. Firstly, the sample size of correct training data is reduced, meaning that we can expect lower overall accuracies and quicker overfitting to the training data. Secondly, we introduce false data points that mislead the network, changing the weights in ways that destroy the structure for correct classification into the three classes. The first problem is unavoidable and can only be compensated by collecting more data. The second problem could be avoided if the experiment included a subjective labelling phase, where the subject after having completed the experiment could label the seen images according to their reaction. This data could then be used to remove incorrectly labelled trials individually for each subject.

Another source of uncertainty in the EM data is the difficulty in making the test subject actually experience the emotion depicted in the image presented. Since the experiment is made in a very safe controlled environment and the subject is focused on completing the assigned task this is far from certain. This could lead to further disparity between the expected and actual content of the EEG and in the extension means that conclusions drawn from this test are not necessarily true in a scenario where the subject experiences more intense "real" emotional responses.

In summary, the effects presented above likely cause significant deterioration of the EM data when compared to the SM data. This can serve to explain at least part of the disparity in accuracy between the emotional and semantic classifications.

5

Conclusion

The main comparison of this thesis study, that between the Morlet network and the reassignment network, resulted in superior performance for the Morlet network on both simulated and real data. A clear conclusion from the tests is also, however, that the conventional 2D convolutional net used here to analyse the reassigned spectrograms is not well suited to the task. The traditional convolution focuses too much on kernel-level geometry, does not in an efficient way use the difference between temporal and frequential information and is a very inefficient way to analyse the sparse matrices generated by reassignment. It would therefore be unwise to discard the method of reassignment as a whole due to a poorly adapted network.

From the analysis of the EM data set we can conclude that the task of emotional classification is more difficult than semantic classification. Even when taking the less sophisticated preprocessing of the EM data into account by comparing semantic and emotional classification on the same set we get significantly lower performance on the latter task. This could depend on both poor adaptation of the networks to the problem of emotional classification, since they were mainly tested using the SM data set, and on the difference between emotional and visual EEG data.

The use of GradCAM in conjuncture with the easily interpretable frequency features of the Morlet network has resulted in an intuitively understandable model. This is important if further neurological conclusions are to be drawn from the results of applying the network and simplifies the design and optimisation process considerably. Given scientifically relevant tasks the model is capable of contributing to a deeper understanding both of the learning process itself and what specific features in time and frequency are used to draw conclusions.

The tests of transfer learning with the Morlet network show surprisingly good results, considering that validation is carried out on previously unseen subjects. This strongly implies that the network finds features that are very similar for equivalent visual processes across subjects. In addition, the method is shown to significantly increase performance after subject specific training for the previously unseen subject is performed. This effect is especially pronounced if only a small amount of data or a short training time is available for the new subject.

Using the Morlet network for reverse classification of the retrieval data (training on retrieval data and validating on study data) gave significant accuracies across multiple subjects and folds where previous tests on the same data had yielded no significant results. The same was true for training and validating on retrieval data. This supports previous theories claiming similarities between the perception and memory processes and suggests that the feature selection of the Morlet network is indeed selecting relevant high-level features that correspond to the underlying processes, as opposed to only finding surface-level patterns in the data.

6

Continuation

The results of this study reveal several interesting possibilities for further study. Looking at the specific problem examined in the context of the novel emotional experimental data, it is clear that the task of classifying emotional content is not equivalent to classifying semantic content. The accuracies of our models further suggest that the emotional problem is more difficult than the semantic one, at least using these methods. Further study into what methods perform best at this task and why, possibly with a more robust basis in neurological theory to aid model design, would be valuable since it would allow for better analysis of for example the link between observed and remembered emotional data.

The area of spatial analysis was left mostly unexplored in this study due to the limited time and scope. Integrating the existing information about the location of electrodes into the networks tested here (for example by weighing the spatial convolution in the Morlet network based on electrode distance) could enhance performance and, by applying analysis methods like the GradCAM, illustrate the spatial relations between signal components.

In order to further examine the reassigned spectrogram for EEG signals our results seem to indicate that a better method for analysing the resulting spectrograms is required. Conventional image classification networks like 2D convolutional nets display many weaknesses when applied to this kind of data. It is also clear that the reassignment network as implemented here is more sensitive to differences in the given task compared to the Morlet network. Increased performance could likely be achieved if the reassignment network was more precisely tailored to the task at hand or, alternatively, made more adaptable by exploring a range of different parameters (similar to the Morlet network). If, for certain problems, the reassignment method has desirable properties one possibility would be to use the two models presented here in tandem, using the Morlet network to identify features and using the corresponding parameters for reassigning the spectrograms. Making the reassignment method learnable, unsuccessfully tried during our work, might still be possible.

As expected of a study with limited scope and time, many parts of our methods could be further optimised and more thoroughly tested, likely resulting in somewhat improved performance and more statistically grounded conclusions.

Bibliography

Abadi, M., A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng (2015). "Tensorflow: large-scale machine learning on heterogeneous distributed systems". URL: http://download.tensorflow.org/paper/whitepaper2015.pdf.

Abo-Zahhad, M., S. Ahmed, and S. N. Seha (2015). "A new EEG acquisition protocol for biometric identification using eye blinking signals". *International Journal of Intelligent Systems and Applications (IJISA)* **07**, pp. 48–54. DOI: 10.5815/ijisa.2015.06.05.

Auger, F. and P. Flandrin (1995). "Improving the readability of time-frequency and time-scale representations by the reassignment method". *IEEE Transactions on Signal Processing* **43**:5, pp. 1068–1089. DOI: 10.1109/78.382394.

Ba, J. L., J. R. Kiros, and G. E. Hinton (2016). *Layer normalization*. arXiv: 1607.06450 [stat.ML].

Baez, J. (2011). "Renyi entropy and free energy". *Centre for Quantum TechnologiesNational University of Singapore*.

Barzegaran, E., S. Bosse, P. J. Kohler, and A. M. Norcia (2019). "EEGsourcesim: a framework for realistic simulation of EEG scalp data using MRI-based forward models and biologically plausible signals and noise". *Journal of Neuroscience Methods* **328**, p. 108377. ISSN: 0165-0270. DOI: https://doi.org/10.1016/j.jneumeth.2019.108377.

Basic Knezevic, D. and A. Heimerson (2018). "Statistical and machine learning methods for classification of episodic memory". *Faculty of Engineering, Centre of Mathematical Sciences, Mathematical Statistics*.

Bazgir, O., Z. Mohammadi, and S. A. H. Habibi (2018). "Emotion recognition with machine learning using EEG signals". In: *2018 25th National and 3rd International Iranian Conference on Biomedical Engineering (ICBME)*, pp. 1–5. DOI: 10.1109/ICBME.2018.8703559.

Boashash, B. (2016). "Chapter 2 - heuristic formulation of time-frequency distributions". In: Boashash, B. (Ed.). *Time-Frequency Signal Analysis and Processing (Second Edition)*. Second Edition. Academic Press, Oxford, pp. 65–102. ISBN: 978-0-12-398499-9. DOI: https://doi.org/10.1016/B978-0-12-398499-9.00002-9.

Bramao, I. and M. Johansson (2018). "Neural pattern classification tracks transfer-appropriate processing in episodic memory". *eNeuro* **5**:4. DOI: 10.1523/ENEURO.0251-18.2018.

Bright, D., A. Nair, D. Salvekar, and S. Bhisikar (2016). "EEG-based brain controlled prosthetic arm". In: *2016 Conference on Advances in Signal Processing (CASP)*, pp. 479–483. DOI: 10.1109/CASP.2016.7746219.

Caterini, A. L. and D. E. Chang (2018). *Deep Neural Networks in a Mathematical Framework*. 1st. Springer Publishing Company, Incorporated. ISBN: 3319753037.

Chollet, F. et al. (2015). *Keras*. URL: `https://github.com/fchollet/keras`.

Ferree, T., M. Clay, and D. Tucker (2001). "The spatial resolution of scalp EEG". *Neurocomputing* **38-40**. Computational Neuroscience: Trends in Research 2001, pp. 1209–1216. ISSN: 0925-2312. DOI: `https://doi.org/10.1016/S0925-2312(01)00568-9`.

Gramfort, A., M. Luessi, E. Larson, D. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, and M. Hämäläinen (2013). "Meg and EEG data analysis with mne-python". en. Frontiers in NeuroScience; Volume 7, pp. 1–13. ISSN: 1662-4548. DOI: `10.3389/fnins.2013.00267`. URL: `http://urn.fi/URN:NBN:fi:aalto-201705114073`.

Ioffe, S. and C. Szegedy (2015). *Batch normalization: accelerating deep network training by reducing internal covariate shift*. arXiv: `1502.03167 [cs.LG]`.

Jonas, S., A. Rossetti, M. Oddo, S. Jenni, P. Favaro, and F. Zubler (2019). "EEG-based outcome prediction after cardiac arrest with convolutional neural networks: performance and visualization of discriminative features". *Human Brain Mapping* **40**. DOI: `10.1002/hbm.24724`.

Kingma, D. and J. Ba (2014). "Adam: a method for stochastic optimization". *International Conference on Learning Representations*.

Kosheleva, O. and V. Kreinovich (2017). "Why deep learning methods use KL divergence instead of least squares: a possible pedagogical explanation". *Departmental Technical Reports (CS), UTEP* **1192**.

Lang, P. J., M. M. Bradley, and B. N. Cuthbert (2008). "International affective picture system (iaps): affective ratings of pictures and instruction manual. technical report a-8".

Luck, S. J. (2005). *An Introduction to the Event-Related Potential Technique*. Cambridge.

Lystad, R. and H. Pollard (2009). "Functional neuroimaging: a brief overview and feasibility for use in chiropractic research." *The Journal of the Canadian Chiropractic Association* **53 1**, pp. 59–72.

Mohammadi, Z., J. Frounchi, and M. Amiri (2015). "Wavelet-based emotion recognition system using EEG signal". *Neural Computing and Applications* **28**, pp. 1985–1990.

Peirce, J., J. Gray, S. Simpson, M. MacAskill, R. Höchenberger, H. Sogo, E. Kastman, and J. Lindeløv (2019). "Psychopy2: experiments in behavior made easy". *Behavior Research Methods* **51**. DOI: `10.3758/s13428-018-01193-y`.

Prakash, A. (2018). "Wavelet and its applications". *International Journal of Scientific Research in Computer Science, Engineering and Information Technology* **3**, pp. 95–104. DOI: `10.32628/CSEIT183820`.

Raschka, S. (2018). "Model evaluation, model selection, and algorithm selection in machine learning". *CoRR* **abs/1811.12808**. arXiv: `1811.12808`.

Sandsten, M. and J. Brynolfsson (2015). "The scaled reassigned spectrogram with perfect localization for estimation of Gaussian functions". eng. IEEE Signal Processing Letters **22**:1, pp. 100–104. ISSN: 1070-9908. DOI: `10.1109/LSP.2014.2350030`.

Sandsten, M., J. Brynolfsson, and I. Reinhold (2018). "The matched window reassignment". English. In: *26th European Signal Processing Conference, EUSIPCO 2018*. European Association for Signal Processing (EURASIP), pp. 2340–2344. ISBN: 978-908279701-5. DOI: `10.23919/EUSIPCO.2018.8553204`.

Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra (2019). “Grad-CAM: visual explanations from deep networks via gradient-based localization”. *International Journal of Computer Vision* **128**:2, pp. 336–359. ISSN: 1573-1405. DOI: `10.1007/s11263-019-01228-7`. URL: `http://dx.doi.org/10.1007/s11263-019-01228-7`.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov (2014). “Dropout: a simple way to prevent neural networks from overfitting”. *Journal of Machine Learning Research* **15**:56, pp. 1929–1958.

Steingrimsson, S., G. Bilonic, A.-C. Ekelund, T. Larson, I. Stadig, M. Svensson, I. S. Vukovic, C. Wartenberg, O. Wrede, S. Bernhardsson, and et al. (2020). “Electroencephalography-based neurofeedback as treatment for post-traumatic stress disorder: a systematic review and meta-analysis”. *European Psychiatry* **63**:1, e7. DOI: `10.1192/j.eurpsy.2019.7`.

Tatum, W. O. (2014). *Handbook of EEG Interpretation*. 2nd. Demos Medical. ISBN: 1620700166.

Tottenham, N., J. Tanaka, A. Leon, T. Mccarry, M. Nurse, T. Hare, D. Marcus, A. Westerlund, B. Casey, and C. Nelson (2009). “The nimstim set of facial expressions: judgments from untrained research participants”. *Psychiatry research* **168**, pp. 242–9. DOI: `10.1016/j.psychres.2008.05.006`.

Tulving, E. (1983). “Elements of episodic memory.” *Oxford University Press*.

Waldhauser, G. T., V. Braun, and S. Hanslmayr (2016). “Episodic memory retrieval functionally relies on very rapid reactivation of sensory information”. *Journal of Neuroscience* **36**:1, pp. 251–260. ISSN: 0270-6474. DOI: `10.1523/JNEUROSCI.2101-15.2016`.

Zhao, D., F. Tang, B. Si, and X. Feng (2019). “Learning joint space–time–frequency features for EEG decoding on small labeled data”. *Neural Networks* **114**, pp. 67–77. ISSN: 0893-6080. DOI: `https://doi.org/10.1016/j.neunet.2019.02.009`.

A

Morlet Layer

```
class MorletConvRaw(keras.layers.Layer):

def __init__(self, input_dim, Fs, input_shape=[75,31,1],etas = 25,
        wtime = 0.36):
    super(MorletConvRaw, self).__init__()
    self.nchan = input_dim[1] #Antal kanaler
    self.ttot = input_dim[0] #Tiden per trial
    self.etas = etas #Antal fönster
    self.wtime = wtime #Fönsterbredd i tid
    self.wlen = int(self.wtime*Fs)
    self.a = self.add_weight(name=’a’, shape=(self.etas,1),
    initializer=keras.initializers.Constant(value=a_init),
        trainable=train_a)
    self.b = self.add_weight(name=’b’, shape=(self.etas,1),
        initializer=keras.initializers.RandomUniform(minval=b_init_min, maxval=b_init_max, seed=1),
        trainable=train_b)

def call(self, inputs):
    #Create a Morlet window tensor.
    win = tf.convert_to_tensor(np.linspace(-self.wtime/2,self.wtime/2,self.wlen,dtype=’float32’))
    win = tf.raw_ops.MatMul(
        a = tf.raw_ops.Diag(diagonal=win),
        b = tf.constant(np.ones((self.wlen,self.etas),
        dtype=’float32’)))

    aterm = tf.raw_ops.Transpose(x = tf.raw_ops.MatMul(
        a = tf.raw_ops.Diag(diagonal = tf.raw_ops.Mul(x = self.a,y =self.a/2)[:,0]),
        b = tf.constant(np.ones((self.etas,self.wlen),dtype=’float32’))),perm=[1,0])

    mwin = tf.raw_ops.Exp(x = -tf.raw_ops.Mul(x = tf.raw_ops.Mul(x=win,y=win),y = aterm))
    costerm = tf.raw_ops.Transpose(x = tf.raw_ops.Cos(x = tf.constant(2*math.pi)*tf.raw_ops.MatMul(
        a = tf.raw_ops.Diag(diagonal= self.b[:,0]),
        b = win,transpose_b=True)),perm=[1,0])

    mwin = tf.raw_ops.Mul(x= costerm,y = mwin)

    # Expand
    tinput = tf.raw_ops.ExpandDims(input = inputs,axis = -1)
    mwin = tf.raw_ops.ExpandDims(input = mwin, axis=1)
    mwin = tf.raw_ops.ExpandDims(input = mwin, axis=1)
    # Convolve.
    output = tf.raw_ops.Conv2D(input = tinput,filter = mwin,strides = [1,1,1,1], padding=’VALID’)

    return output
```

B

Reassignment

```
function [SS,MSS,TI,FI,H]=screassignspectrogram1(X,lambda,candsig,NFFT,NSTEP,Fs,e);

% SCREASSIGNSPECTROGRAM [SS,MSS,TI,FI,H]=screassignspectrogram(X,lambda,candsig,NFFT,NSTEP,Fs,e);
% computes and plots the windowed spectrogram and the scaled reassigned spectrogram.
%
% Output data
%
% SS:  The Gaussian windowed spectrogram
% MSS: The scaled reassigned windowed spectrogram
% TI:  Time vector for the time-frequency plots
% FI:  Frequency vector for the time-frequency plots
% H:   The Gaussian window
%
% Input data
%
% X:    Data sequence
% lambda:    Parameter of Gaussian window.
% candsig: Candidate sigma, the assumed scaling factor of the signal
% NFFT: The number of FFT-samples, default NFFT=2048.
% NSTEP:The time-step between to spectrum calculations, default NSTEP=1.
% Fs:   Sample frequency, default Fs=1
% e:    Smaller spectrum values than this number are not reassigned, default e=0.
%

% Gaussian window calculation

Hl=10*lambda; %Long enough window
H=exp(-0.5*([-Hl/2:Hl/2-1]'/lambda).^2);

% TH and DH needed for the reassignment

Tvect=[-Hl/2+1:Hl/2]';
TH=Tvect.*H;
DHd=diff(H);
DHd2=interp(DHd,2);
DH=[0;DHd2(2:2:end)];

data=X;
data=data(:);

% Spectrogram calculation

%mvect=[0:NFFT-1];
data=[zeros(fix(Hl/2),1);data;zeros(fix(Hl/2),1)];
datal=length(data(:,1));

timevect=[0:NSTEP:datal-Hl-1];
TI=[];
FF=[];
TFF=[];
DFF=[];
MSS=zeros(NFFT/2,length(timevect));
nmat0=zeros(NFFT,length(timevect));
mmat0=zeros(NFFT,length(timevect));
nmat=zeros(NFFT,length(timevect));
mmat=zeros(NFFT,length(timevect));
for i=0:NSTEP:datal-Hl-1
   testdata=data(i+1:i+Hl);
   testdata=testdata-mean(testdata); % Mean value reduction!
   F=fft(H.*testdata,NFFT);
   TF=fft(TH.*testdata,NFFT);
   DF=fft(DH.*testdata,NFFT);
   FF=[FF F(1:NFFT/2)];
   TFF=[TFF TF(1:NFFT/2)];
   DFF=[DFF DF(1:NFFT/2)];
   TI=[TI i];
end
SS=abs(FF).^2;
% SS = SS + 0.1; % Add epsilon to check effect on final spectrogram
TI=TI/Fs;
FI=[0:NFFT/2-1]'/NFFT*Fs;
e = 0.02*max(max(SS)); %Mixtra med senare!
% Scaling factors for the scaled Gaussian reassignment

fact=(lambda^2+candsig^2)/(lambda^2);
fact2=(lambda^2+candsig^2)/(candsig^2);

% Scaled reassignment calculation

% imaginary = max(max(imag(FF)))

for n=1:length(TI)
    for m=1:NFFT/2
        if SS(m,n)>e
            nmat0(m,n)=fact/NSTEP*(real(TFF(m,n).*conj(FF(m,n))./SS(m,n)));
            mmat0(m,n)=NFFT/2/pi*fact2*(imag(DFF(m,n).*conj(FF(m,n))./SS(m,n)));
            nmat(m,n)=n+round(nmat0(m,n));
            mmat(m,n)=m-round(mmat0(m,n));
            if mmat(m,n)>0 & mmat(m,n)<=NFFT/2 & nmat(m,n)>0 & nmat(m,n)<=length(TI)
                MSS(mmat(m,n),nmat(m,n))=MSS(mmat(m,n),nmat(m,n))+SS(m,n);
            else
                mmat(m,n)=0;
                nmat(m,n)=0;
            end
        end
     end
end
```