Advanced Cognitive Neuroscience

Portfolio Exam – Fall 2019

Line Kruse - 201608877

Study Group 6

Content

All applied code can be found in the appendix.

An applied example of neural networks

Line Kruse

MA Cognitive Science, Aarhus University

Code in appendix.

Neural networks are a class of algorithms characterized by small, interconnected computing units organized in multiple layers. This architecture has turned out to be highly efficient for learning from training, and neural networks have thus become a prominent method of machine learning. The current example constructed two types of neural networks with two types of learning algorithms and aimed to compare their performance on a classification task of handwritten digits.

Neural units

Neural units are inspired by the biological neuron and resemble these in that they convert a complex pattern of inputs into a single output. The output is an activation value, conceptually corresponding to the action potential of the biological neuron. The basic computation of a neural unit has three steps. It takes a weighted sum of its input, adds a bias term and computes a new weighted sum (Nielsen, 2015). The weighted sum is thus a vector composed of a weight vector, a scalar bias, and an input vector. The bias acts as a threshold, modulating the likelihood of the unit to be activated. However, if the output of the neural unit is simply determined by whether it exceeds the threshold or not, i.e. a step function (1 or 0), then very small changes in the weights and biases can cause large changes in the output, making learning problematic (Nielsen, 2015). Rather, the most commonly used neural units apply a non-linear activation function to the weighted sum. The current example employed two types of activation functions. First, the Rectified Linear Unit (ReLu) function, which outputs the linear result if it is positive, and 0 if it is negative. Because it is almost linear, it readily deals with high input values enabling efficient learning, in contrast to functions as Sigmoid. Second, the Softmax activation function turns output values into probabilities that sum to 1, such that the output is a probability distribution over all potential outcomes. Hence, it can profitably be applied in the output layer, particularly in the case of non-binary outcomes (Kriegeskorte, 2015). In general, the application of activation functions facilitates learning as they allow exploration of the most optimal weights and biases for accurate performance.

The architecture of neural networks

The two networks employed in the current example had an identical number of neural units organized in four layers, and only differed in their organizational architecture; one composed a feedforward neural network (FNN) and the other a convolutional neural network (CNN). Both networks constitute deep neural networks, i.e., they have hidden layers. Hidden layers function to decompose the task into simple and specific sub-problems (Nielsen, 2015) allowing more detailed and abstract representations and facilitating more advanced learning. This is necessary to discriminate between complex categories in the input that are not linearly separable. In feedforward neural networks, the output from each layer is passed directly as input to the next layer establishing a hierarchy of increasingly complex feature representations (Jurafsky and Martin, 2018). However, feedforward networks are computationally demanding. Since each neural unit has a weight parameter and a bias parameter to be estimated, even a simple two-layer network readily has to estimate many thousands of parameters. Further, this carries a high risk of overfitting.

The core idea of CNNs is to explore spatial features of the input by applying small filters. These filters, termed feature maps or kernels, each search the input image for a specific feature. Each neuron in a layer searches its own local receptive field of NxN pixels. Thus, the map convolves over the entire image analysing a small field at a time. Each layer has one or more feature maps and each neuron analyses its receptive field for every feature map. This procedure implies that spatial specificity decreases as we go deeper into the network, since each output of the neural units is a "sum" of the entire receptive field. Thus, representations become more abstract and less sensitive of space as we move through the layers. CNNs resemble processing in the visual cortex in the sense that the feature maps can be compared to for instance edge- and bar-detectors, which passes their output to higher-level visual areas with increasingly abstract feature detection. This architecture is highly efficient for learning, as each feature map applies the same weight and bias to each receptive field, which greatly reduces the number of parameters. Hence, CNNs often enable more rapid learning and are less prone to overfitting.

Building, training and evaluating the neural networks

The task employed in this example was based on a subset of the MNIST dataset from the package Keras (Chollet, 2015) and consisted of 6.000 28x28 pixels grayscale images of 10 different digits, each labelled with the correct class. 5.000 of the images were used as training set, while the last

1.000 images constituted the test set. The neural networks were built using the Keras package. Both neural nets had an input layer of 784 neural units (28x28 pixels), and an output layer of 10 neural units (10 possible digits) using the softmax function. The networks additionally included three fully connected hidden layers of 32, 64, and 32 neural units, respectively, each using the ReLU activation function. However, in the case of the CNN the first two hidden layers were convolutional layers, while the last was a fully connected layer. The two convolutional layers applied a kernel of size 3x3 pixels. The fully connected layer in CNNs takes the results of the convolutional layers and generates an outcome prediction. Further, CNNs necessitate a pooling layer and a flattening layer. A pooling layer is typically applied after the convolutional layer, and functions to simplify the information in the output of the convolutional layer, reducing the number of parameters and noise. Each unit in the pooling layer makes a summary of a region of units in the convolutional layer for each feature map. In the current example, max pooling was implemented, i.e., each unit in the pooling layer output the maximum activation of the analysed region. A region of 2x2 neural units was used. The flattening layer is applied to transform the 2D matrix of previous layers into a vector that can be passed to the fully connected output layer.

Both networks employed supervised learning and were trained with the categorical cross entropy loss function and the Stochastic Gradient Descent (SGD) learning algorithm, both from the Keras package. Categorical cross entropy compares the distribution of the predictions to the true distribution, assuming that each input belongs to only one class (Peltarion, 2019). Gradient descent seeks to find the set of weights and biases that makes the cost as small as possible, i.e., the global minimum of the cost function. This is obtained through the computation of the derivative of the cost function. Since neural nets have many parameters, the partial derivative is computed with respect to each parameter, and hence reflects the local shape of the cost function. The partial derivative then represents how a small change to a particular weight (or bias) affects performance. The algorithm initialize with a random set of weights and biases, then iteratively computes the partial derivatives and changes the weights and biases according to the effect on the error, until it reaches the global minimum of the cost function (Kriegeskorte, 2015). A learning rate parameter determines the size of steps to be taken down the slope of the cost function. In CNNs the gradient is calculated on each local receptive field. Gradient descent relies on error back-propagation, which works backwards through the connections of all intermediate layers to calculate the partial derivatives (Kriegeskorte and Golan, 2019). *Stochastic* gradient descent increases the speed of learning by calculating the

gradient on a random mini-batch of the training input and averages over these samples to approximate the true gradient. Learning is divided into epochs, in which different mini-batches are used until the training inputs are exhausted. The current example networks were trained on mini-batches of 128 inputs in 5 epochs.

Performance and optimization

The FNN had a performance accuracy of 0.46 %, and hence classified the handwritten digits worse than expected by chance. The CNN had a performance accuracy of 0.54 %, performing slightly above chance. However, both networks exhibited rather poor performance. To improve classification accuracy an alternative learning algorithm was employed; the Adadelta from the Keras package. Adadelta is an alternative optimizer of the Stochastic Gradient Descent algorithm, and differs in that it uses a dynamic learning rate calculated for each parameter at each time step. The parameter is dependent on the average step size of the previous iteration and the current gradient, such that the learning rate will be larger for infrequent parameters and smaller for frequent parameters (Zeiler, 2012). Thus, it removes the need to manually tune the learning rate, and copes with a problem inherent in the default SGD; that the learning rate continuously decreases as training progresses. The Adadelta optimizer significantly improved performance in both types of networks. The FNN now had an accuracy of 0.83 %. In comparison, the CNN had a performance accuracy of 0.96 %, Hence, the dynamic learning rate of Adadelta improved the ability of the networks to learn efficiently.

As expected, the CNN outperformed the FNN in classification of handwritten digits, although both networks had identical number of neural units distributed equally in four layers. Thus, the current example illustrates the efficiency of feature learning performed by CNNs. Further, the reduction of dimensions in CNNs not only increases learning efficiency but also saves memory of the network, which may further explain the improved performance. More memory capacity allows the CNN to capture more details and hence build a more complex feature representation.

References

Chollet, F. (2015). Keras, Github repository: https://github.com/fchollet/keras

Jurafsky, D., & Martin, J. H. (2018). Neural Networks and Neural Language Models. In Speech and Language Processing (3rd ed. draft ed.): https://web.stanford.edu/~jurafsky/slp3/7.pdf.

Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. Annual Review of Vision Science, 1, 417-446.

Kriegeskorte, N., & Golan, T. (2019). Neural network models and deep learning. Current Biology, 29, R231-R236.

Nielsen, M. A. (2015). *Neural networks and deep learning* (Vol. 25). San Francisco, CA, USA:: Determination press.

Peltarion (2019). Categorical crossentropy . Retrieved from: https://peltarion.com/knowledge-center/documentation/modeling-view/build-an-ai-model/loss-functions/categorical-crossentropy

Zeiler, M. D. (2012). ADADELTA: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Temporal differences in the processing of fearful compared to neutral faces: a MEG study

Authors: Line Kruse, Martin Ito, Christoffer Olesen and Simon Hansen

Abstract (Together)

This paper is the first part of a dual fMRI/MEG study investigating the spatio-temporal dynamics in the processing of fearful and neutral faces. This first study investigated temporal differences in the processing of fearful and neutral faces using MEG. Earlier literature suggests differences in the processing of faces displaying a range of emotions and the current study aimed to investigate the temporal components of such differences. To analyse the data, three different classifiers were employed; a Naive Bayes, a linear Support Vector Machine and a k-Nearest Neighbours classifier. Although the accuracy across time followed very similar patterns for the three classifiers, the linear Support Vector Machine performed the best and had a maximum accuracy of 0.94 at 108 ms. after stimulus onset. Maximum accuracy seemed to co-occur with a spike in activity, suggesting the first main spike in activity after stimulus onset is a good predictor of the emotion of the target face. A number of limitations in the methodology of the study should be acknowledged. First, there was only one participant, limiting the generalizability of the results. Second, both magnetometers and gradiometers were included in the analysis, which complicates the interpretation of the results.

Introduction (Line**)**

One of the most crucial sources of information in human social interactions is that conveyed by facial expressions. In particular, emotional expressions allow interlocutors to infer the intentions, feelings, and state of mind of the other person, which is critical for successful interpersonal interaction and cooperation (Morel et al., 2009). It is well established in the literature on visual perception, that the perceptual processing of faces exhibits unique and prioritized processing mechanisms (Kanwisher, 2000). The Fusiform Face Area, located in the Inferior Temporal Cortex, has been identified as one of the structures that is selectively activated by the presentation of a face. Further, this structure appears to be part of a distributed cortical network including the Superior Temporal Sulcus, associated with theory of mind, and limbic circuits encompassing the Hippocampus, Amygdala, and Orbitofrontal Cortex (Ishai et al., 2005). This suggests a strong interdependence of emotions and face processing, in which information from one is highly important for the other. In fact, preferential processing of emotionally salient stimuli has been

shown in both visual perception of words, objects, and faces. Milders and colleagues (2006) reported more frequent detection of emotionally expressive faces compared to neutral faces, and a particularly strong processing facilitation of fearful facial expressions.

Influence of emotional expression of perceptual face processing (Line**)**

While much is established regarding visual processing of faces, it still remains unclear if and how emotional expressions influence various stages of the processing. Traditionally, it has been assumed that initial perceptual analysis of faces could be dissociated in time from the processing of facial emotions (Morel et al., 2009). Generally, the face selective N170 component in electroencephalogram (EEG), and the magnetic concomitant M170 in magnetoencephalography (MEG), has been associated with the initial stage of face processing. It has been commonly assumed that emotional expressions solely influence visual processing of faces beyond this time range of initial perceptual analysis.

However, the saliency of emotional expressions on cognition and its criticality for social perception and interaction has led scholars to suspect that emotions may influence face processing at very early stages, potentially influencing subsequent processes as memory encoding and attentional mechanisms. One study investigated the temporal dynamics of emotional influence on face perception using both EEG and MEG and reported differential responses to emotional expressions as early as 40-50 ms after stimulus onset (Morel et al., 2009). Furthermore, the N170/M170 component also exhibited modulation dependent on emotional vs. neutral expressions. Lastly, a selective effect for fearful faces was found in late latencies (280-320 ms) in which the M300 component decreased with repetition of the same fearful face. This may reflect the higher arousing value of fearful stimuli that rapidly habituates with repetition. Thus, it appears that the visual processing of emotional expressions can be differentiated already at the earliest stages of processing, and that fearful faces exhibit particularly differentiated temporal effects in processing. Such findings raise the question of the function of the diverse influence of emotional expressions on face perception. Righi and colleagues (2012) investigated the interaction between emotional facial expressions and memory and showed that the early components P100 and N170 were modulated by the emotional expressions, which influenced memory encoding and retrieval. While the N170 is associated with initial face detection, the P100 is a visually evoked potential component associated with the initial detection of visual stimuli in primary visual cortices. Hence, the very initial stages of face processing appear to be subject to emotional modulation. Righi further reported an enhanced

positivity around 150 ms selectively for fearful faces, even when these were processed unconsciously. Such findings suggest that fearful faces, in contrast to neutral faces, might elicit enhanced memory encoding. Thus, it appears that face recognition can be influenced by top-down processes of emotional memory, particularly for threatening stimuli. These findings are consistent with the notion that potential threat signals are prioritized in sensory processing (Williams et al., 2006).

In line with such findings, Maratos and colleagues (2011) reported differential effects of emotional expressions on attention processes. They showed that fearful faces elicited rapid attentional prioritization, which subsequently enhances processing of these faces. Additionally, it has been suggested that the rapid processing of fearful faces is mainly facilitated by efficient processing in the Amygdala. One study used intracranial field potential recordings (Sato et al., 2010) and reported greater Amygdala gamma oscillations as early as 50-150 ms. after stimulus onset in response to fearful faces. Thus, it appears that amygdalar processing of fearful faces is faster than the initial face-specific activity in visual cortices (N170/M170) and may subsequently facilitate visual processing in neocortex and drive the commonly observed faster temporal dynamics of this emotion. However, there are serious challenges associated with using MEG to detect signals from such deep structures as the Amygdala. Hence, the relationship between the spatial and temporal components of differential emotional face processing remains theoretical.

As outlined above, it is generally established that fearful emotional expressions exert differential influence on face processing from both neutral faces and faces displaying other emotions. However, the specific spatio-temporal scales of this influence are continuously debated. The current paper is the first part of a two-part study, aiming to assess the interaction between spatial and temporal dynamics of fearful face processing compared to the processing of neutral faces. Subjects were presented with either a fearful or a neutral face and were to respond with a keypress which face they had seen. In addition to the traditional paradigm, subjects were presented with an emotionally valenced (or neutral) word prior to the presentation of the target faces. These words were predictive of the emotional valence of the subsequent face. Thus, due to priming effects we expected face processing to be slightly faster than normally observed. Additionally, this setup allowed us to investigate whether this priming had differential effects on processing in the two conditions. If fearful faces do have prioritized attentional effects, as suggested in the literature, we

would expect that this early activation is either faster or stronger in this condition compared to the neutral condition.

In this study, one subject performed the task while MEG signals were obtained, allowing assessment of the temporal dynamics in the two conditions. In the second study, fMRI derived BOLD-responses were measured for the same paradigm, enabling analysis of the spatial components activated in the two conditions. This procedure facilitated comparison of the temporal and spatial components differentiating processing of fearful emotional expressions from neutral faces. In the current paper, classification analysis was employed to assess where the largest differences between fearful and neutral conditions could be observed in the MEG signal. Three distinct classification algorithms were used, in order to assess the robustness of conclusions across classifiers.

MEG (Christoffer**)**

The temporal components of face processing in the two conditions were assessed using MEG. MEG is a method for measuring electrical current in the brain, and thereby neuronal activity (Buzsáki, 2006). Any electrical current creates a magnetic field perpendicular to the direction of the current. MEG measures the magnetic fields outside the scalp and for this reason it is not necessary to attach sensors directly on the scalp as it is when using the closely related method EEG. The type of sensor used in MEG is a superconducting quantum interference device or a SQUID for short. A SQUID consists of a small electrical circuit with a temperature of -270°C, making it superconductive i.e. without any electrical resistance (Hämäläinen, 1993). The magnetic field originating in the brain interferes with the electrical circuit in the SQUID and it is this interference, which is measured. However, due to very low magnitude of this magnetic field it is of extreme importance to have the MEG equipment located within a magnetically sealed room. Otherwise, anything in its vicinity with magnetic properties, such as electronic devices or metal objects, could interfere with the SQUID to such a degree that measuring the brain would be impossible. The SQUIDS are positioned in a helmet-like shape in the MEG apparatus in which the head of the human subject is placed. On each position, there are three SQUIDS; one magnetometer and two gradiometers. The magnetometer measures the strength of the magnetic field at the given position. The gradiometers are positioned relative to each other in such a way that they can measure the angle of the magnetic field (Hämäläinen, 1993).

Because MEG measures magnetic fields it is only able to pick up currents that run somewhat tangential to the scalp (Buzsáki, 2006), as the magnetic field of other currents does not reach outside the surface of the head, where the SQUIDs are located. This is a shortcoming in the MEG measure one ought to keep in mind when analysing MEG data. As with EEG there are various sources of noise such as eye movements or blinks, which needs to be accounted for. However, in contrast to EEG, measuring magnetic fields are not subject to the pitfall that the scalp conducts and thereby scatters the measured current across the surface of the brain (Buzsáki, 2006). This makes MEG superior to EEG in spatial resolution. However, it is important to note that MEG still has a lower spatial resolution than functional magnetic resonance imaging (fMRI), partly because MEG mostly measures the activity in the surface area of the cortex (Buzsáki, 2006). However, MEG have much higher temporal resolution and for this reason this method is preferable when the timing of neuronal activity is of essence. Hence, with MEG it is possible to get high temporal resolution without losing too much spatial information.

Classification algorithms (Christoffer**)**

The current analysis employed classification algorithms to investigate where in the MEG signals the two conditions exhibited the largest differences. In data science a classifier is in essence an algorithm that sorts data into given classes, according to some patterns in that data. The overall purpose of the classifier is to be able to correctly assign new data to one of the given classes. The data that the classifier uses to establish these patterns are called the training data and the new data that it classifies are called the test data. The proportion of correct classifications of the test data is called the accuracy of the classifier.

There exists a bunch of different algorithms developed to accomplish this task. Many of these use a technique called supervised learning, which is just to say that the data is pre-labelled with class labels and the algorithm uses this information to learn about which data points fits the given classes (Guerra et al., 2011). Popular examples of such classifiers are Naïve Bayes (NB) (Hand & Yu, 2001), Support Vector Machines (SVM) (Cortes & Vapnik, 1995) and k-Nearest Neighbor (k-NN) (Altman, 1992), although the latter is not strictly a learning algorithm. NB is a probabilistic method based on Bayes theorem, which in simplistic terms defines the probability of an event given some evidence, as the likelihood of the event multiplied with a numerical representation of prior evidence (normally just called the prior) (Hand & Yu, 2001). NB uses the

training data to inform the prior and is "naïve" because it assumes that the different variables in the data are independent (which they rarely are). SVM is a geometrical method of classification and does not assume independence of variables. Thus, SMVs are usually preferred to NB in more complex tasks were interactions in the data are suspected. SVMs map the data onto a hyperplane and divides the space into areas corresponding to the different classes (Cortes & Vapnik, 1995). The algorithm attempts to find a way to divide the hyperplane such that as little data points as possible falls within the wrong class. The simplest version of SVM uses a linear vector through the hyperplane and are therefore given the name linear SVM. k-NN works on the assumption that different data points assigned to the same class appears in close proximity of each other. The algorithm simply just assigns new data to a class based on what class is most represented in an area around the data point (Altman, 1992). This area is defined by the value k. Hence, the accuracy of this method is hugely depended on its basic assumption of proximity but is less sensitive to data with a high degree of noise (for simple explanations see Gahukar, 2018).

Although classifier algorithms have existed for decades and has been used and implemented in a variety of contexts, only recently have they been discovered to prove useful in the brain sciences (Lemm et al., 2011). Due to the complexity of brain imaging data, it has always been a challenge to find appropriate statistical methods for analysis in these fields. Using classifiers to investigate the neuronal activity represented in such data, has the potential of answering new questions about cognition and the brain. The simplest way to use a classifier as a means of analysis of brain data, is to train the classifier on a subset of brain signals observed in different classes of stimuli. If the classifier can assign the rest of the data, i.e. the test data, to the correct classes of stimulus with a high accuracy, then the scientist is able to infer at the very least that there is a difference in the neuronal processing of the different classes of stimulus. In this way the classification method has its strength in its ability to predict. However, it is at the cost of detail and insight into informative aspects of the data, such as signal strength. With that said, a notable advantage of this method is that it makes comparison between different neuroimaging techniques more straight forward, as it is possible to apply the same type of classifier to different data structures and get results in the same format. Hence, one can benefit from both the advantages of fMRI and MEG, given that the same participants were tested on the same experimental paradigm using first one and then the other of these techniques (for examples see Cetin et al., 2016; Kaiser et al., 2016).

Method (Martin**)**

Preparation of participant

The study had one female participant aged 24. In preparation for the experiment several electrodes were attached to the participant. Six ECoG electrodes were attached to the participant's face, one above and one below each eye to detect blinking and one on each temple to detect vertical eye movement. Additionally, four head position indicator (HPI) coils were attached to the participants head, one on each side of the forehead and one behind each ear. This was done to track head movement. Two electrodes were used to measure heart rate, one on the left collarbone and one on the right hip. The reference and ground electrodes were attached to the right elbow and the right wrist. The three cardinal points of the head were registered using a Polhemus FASTRAK. Additional points along the scalp were also registered.

Experiment

Each session of the experiment consisted of 60 trials evenly split between the two conditions, i.e. the fearful and the neutral condition. A trial started with a fixation cross being displayed at the centre of the screen. After the fixation cross a word was shown, followed by another fixation cross and then a face that was either fearful or neutral (figure 1). The order of fearful and neutral trials was randomized. The behavioural task of the experiment was to press one of two buttons depending on which face one saw. The framerate of the monitor was 120 Hz. The fixation cross was displayed for either 180 frames or 336. The mean number of frames the fixation crosses were displayed was 258 frames. The words and faces were shown for 84 frames. At 120 Hz., this means that the stimulus was shown for 700 ms. In total, four sessions of 60 trials were run with the participant.

Figure 1: The experiment stimuli. Left: fearful face, Right: neutral face.

All the words were obtained from a corpus by Binder et al. (2016), where each word has been scored on different dimensions (e.g. colour, large, small). Using a principal component analysis based on the four categories pleasant, unpleasant, happy and sad a sentiment score was calculated. Additional sentiment scores were retrieved from a corpus by Warriner, Kuperman and Brysbaert (2013). Based on the first component of the PCA analysis and the sentiment score by Warriner et al. (2013) the words were sorted into three categories: Positive, negative and neutral. In the experiment, negative words were followed by a fearful face, positive words were followed by a neutral face and neutral words could be followed by either a fearful or a neutral face. This was part of the behavioural task and designed to make the experiment more engaging.

Experimental setup

The experiment was conducted using an Elektra Neuromag Triux MEG with 102 magnetometers and 204 planar gradiometers. The participant was seated below the MEG detector and the seat was raised to get the participant's head as close to the helmet as possible for better recordings. The MEG recordings were made in a magnetically shielded room. The experiment stimuli were projected onto a screen from a neighbouring room. To make sure the stimuli were indeed time locked with the MEG recordings, a small white dot was also projected onto the lower right hand corner of the screen when the words and faces were presented, and the white dot was registered by a sensor to record stimulus onset time.

Analysis (Simon)

Preprocessing

In order to analyse the data, different preprocessing steps were carried out. First, the data was max filtered. This step has multiple functions. It removes noise, it detects bad channels, realigns activation data based on movement measure from the HPIs, account for the eye movements and blinks, and move the data into a standardised space making comparison across subjects possible.

The data was high pass filtered at 70 Hz. No low pass filter was applied. An event was defined as 500 ms. before the main stimulus (the face) onset and 1000 ms. after stimulus onset. For this analysis, only the face stimuli were analysed, while the data of the behavioural task, i.e. pressing a button, were ignored. The epochs were calculated using the epochs function in the MNE package in Python (Gramfort et al., 2013). Based on the epochs the average evoked responses were

calculated. These were used for plotting the evoked responses. The epochs from each trial were converted into NumPy files in order to be able to perform classification analysis on them.

Classification analysis

In order to investigate which components were predictive of the two stimulus classes (neutral vs. fearful) a classification was performed. The scikit-learn package was applied for this purpose (Pedregosa et al., 2011). First, the activation data was scaled. Then the classifier looped through each sample (timepoint) training the classifier using the method of cross-validation. The data was split into 10 folds and accuracy was calculated on the test partition of each fold. This accuracy score was then averaged across folds and 95 % confidence interval was calculated in order to estimate uncertainty. To assess the stability of the findings different classifiers were applied. The following classifiers were used: Gaussian NB, Linear SVM (C=1) and k-NN (K=3). The accuracy for each timepoint was plotted in order to identify predictive components together with the 95 % confidence interval.

Global Field Potential

While the classification analysis of the MEG data allows us to infer whether there are differences in the signal it is also interesting to consider what type of stimulus elicit the strongest activation. To compare the strength of activation across the two stimulus classes we plotted the Global Field Potential for all the magnetometers. This allowed us to assess the general level of brain activation measured by the strength of the magnetic field.

Results (Together)

Inspecting the Raw Data

The raw measurements (figure 2 and 3) indicated an early component roughly 110 ms. after stimulus onset in both conditions.

Neutral stimulus

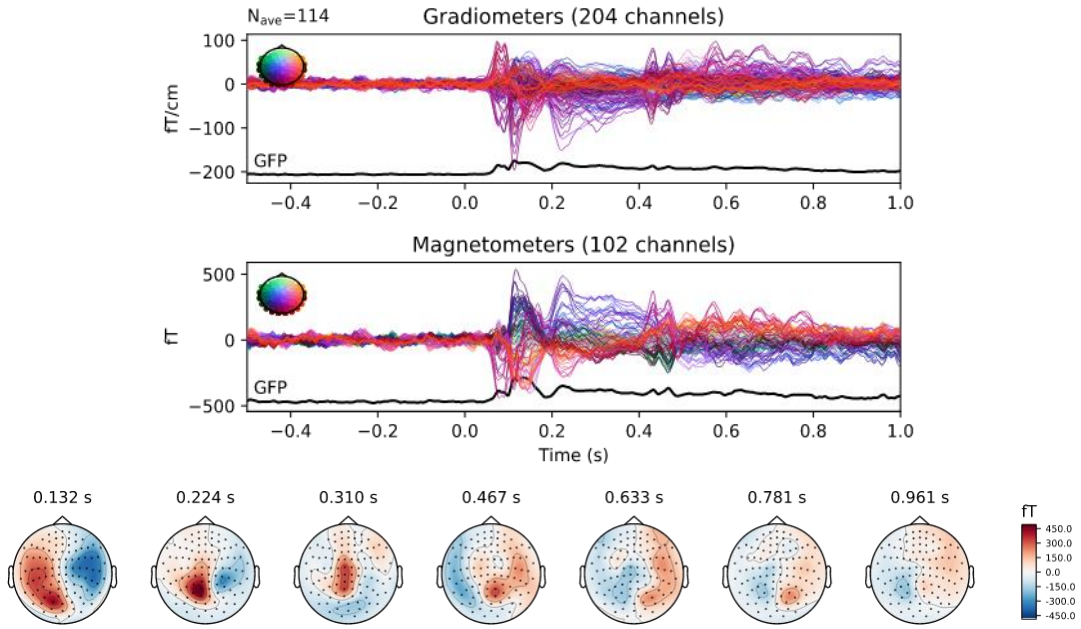


Figure 2: TOP: Butterfly plots of activation for the neutral faces in the specified time window (-500 ms to 1000 ms). The colours indicate the position of the sensors. BOTTOM: The lower plot shows activity maps for different peak activity windows identified by the MNE python package

Fearful stimulus

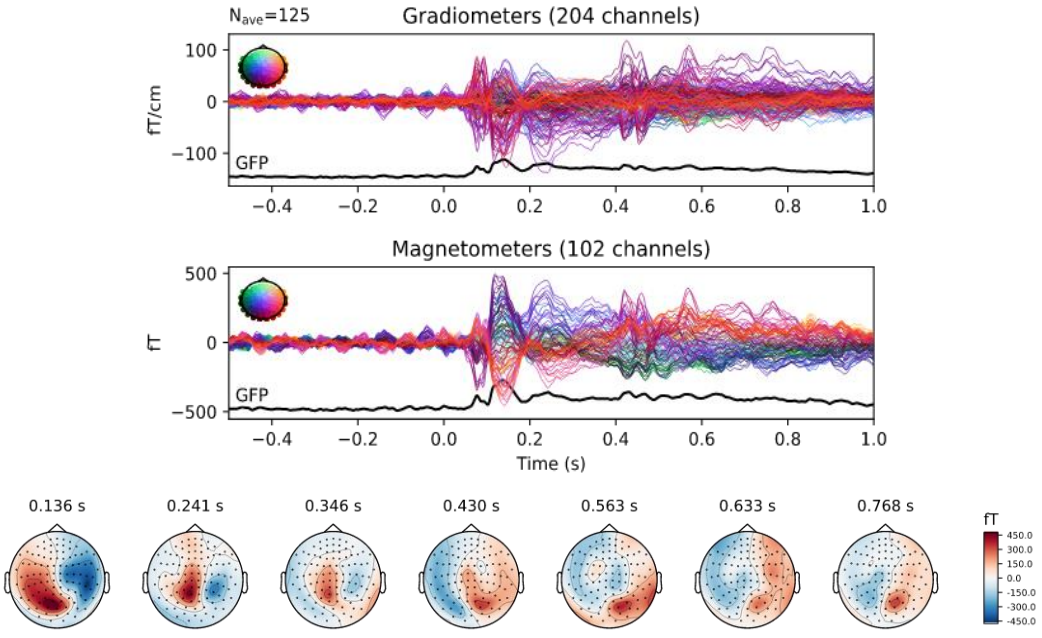


Figure 3: TOP: Butterfly plots of activation for the fearful faces in the specified time window (-500 ms to 1000 ms). The colours indicate the position of the sensors. BOTTOM: The lower plot shows activity maps for different peak activity windows identified by the MNE python package.

Assessing Classifier Accuracy

The accuracy of the three classifiers across time were calculated and plotted (figure 4 to 6). Table 1 summarizes the maximum accuracy of each classifier and when maximum accuracy was reached.

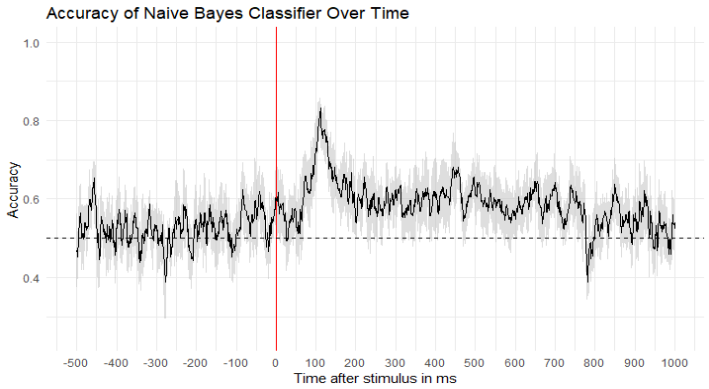


Figure 4: Accuracy of a NB classifier across time samples. The 95 % confidence intervals are shown in grey. The red, vertical line signifies stimulus onset and the dashed, horizontal line shows chance level. Highest accuracy 111 ms after stimulus.

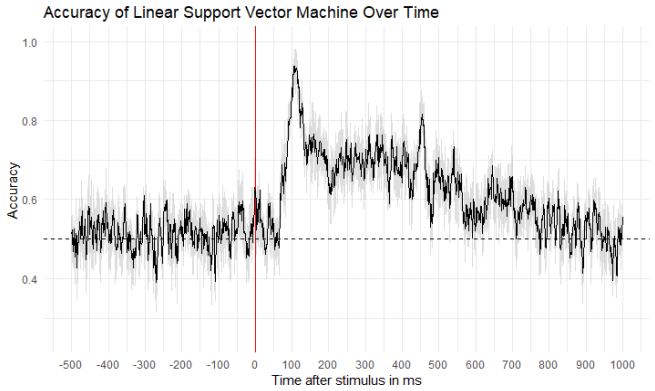


Figure 5: Accuracy of a linear SVM across time samples. The 95 % confidence intervals are shown in grey. The red, vertical line signifies stimulus onset and the dashed, horizontal line shows chance level. Highest accuracy 108 ms after stimulus.

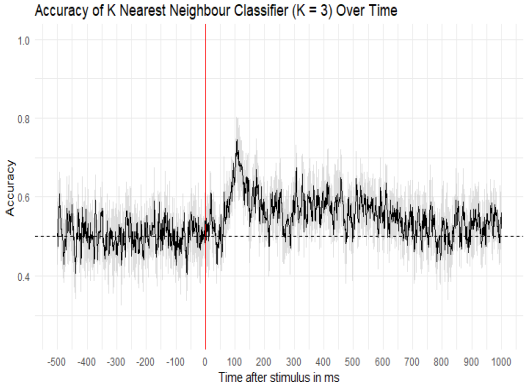


Figure 6: Accuracy of a k-NN classifier, where k = 3, across time samples. The 95 % confidence intervals are shown in grey. The red, vertical line signifies stimulus onset and the dashed, horizontal line shows chance level. Highest accuracy 107 ms after stimulus.

Classifier	Max Accuracy	Lower 95 % CI	Upper 95 % CI	Time of max accuracy
Naïve Bayes	0.83	0.79	0.88	111 ms
Linear Support Vector Machine	0.94	0.89	0.98	108 ms
K Nearest Neighbours, K=3	0.74	0.68	0.81	107 ms

Table 1: Summary of classifier performance

Investigating the Grand Field Potential

The grand field potential across time for all the magnetometers were plotted to investigate which components improved predictive power (figure 7).

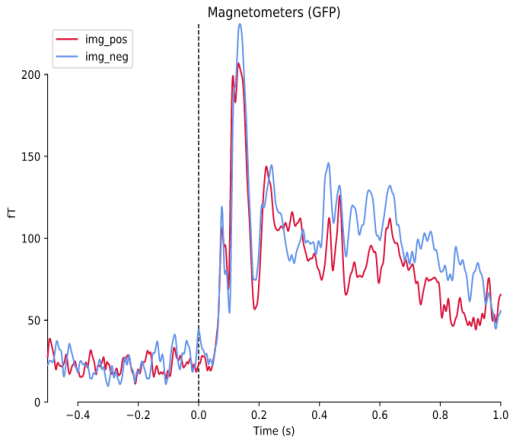


Figure 7: Global Field Power of the magnetometers. The plot shows a stronger activation to fearful faces compared to neutral faces at different time intervals. This difference is especially clear at around 150 ms, but there also seems to be an increased activation to fearful faces between approx. 350 ms to 900 ms.

Discussion

Results and interpretation (Martin)

The activation in the two conditions were very similar, although the activation in the fearful condition seemed stronger than for the neutral condition (figure 7). Looking at figure 2 and 3 we observed that before stimulus onset and roughly 100 ms after stimulus onset, there was mostly noisy activation. Approximately 100 ms after the stimulus onset there was a smaller spike and then at 170 ms after stimulus onset a large spike in activity occurred and from this point onwards, activation levels were higher for the rest of the epoch. The activation in the larger spike seemed to be driven by a spike in activation in the left occipital lobe and right parietal lobe. This study will not dive further into source localization of the activation, this will instead be investigated in the fMRI study. The timing of these two spikes, approximately 100 and 170 ms after stimulus onset suggests that they reflect the visual detection component P100 and the face detection component M170.

The three classifiers all had very similar patterns of accuracy across time (figure 4 to 6) and also peak in accuracy around 110 ms after stimulus onset, roughly coinciding with a peak in activation which could be the P100. Before this main spike in accuracy, the accuracy fluctuated around chance level. This indicates that the main component differentiating the two conditions occurs before the participant becomes aware that the stimulus is a face. After the spike in accuracy, the accuracy slowly decayed towards chance level but was slightly higher than before the spike. Roughly 900 ms. after stimulus onset, the conditions were almost indistinguishable again. The face stimuli lasted 700 ms., so one would expect that in this period the conditions would be distinguishable. The fact that conditions are distinguishable for longer than 700 ms suggests that processing continues after the stimulus has disappeared. This might simply just reflect processing time in the brain. However, the defining activity occurs shortly after stimulus onset. That all three classifiers followed the same pattern, which was also reflected in the data suggests a sound result. Overall, the linear SVM performed the best and had a peak accuracy of 0.94 at 108 ms after stimulus onset (table 1). NB had a peak accuracy of 0.83 at 111 ms after stimulus onset and k-NN had a peak accuracy of 0.74 at 108 ms after stimulus onset. Thus, the results suggest that fearful and neutral faces exhibit important differences in processing on the temporal scale, and that the most significant difference is in the earliest stages of processing. This could either indicate that the amplitude of activity is greater in one condition than in the other around this time point, or that the activity occurs faster in one compared to the other. Figure 7 suggests the first possibility to be more likely. However, the specific nature of this difference at early latencies cannot be inferred based on the current analysis. As the accuracy peaked roughly at 110 ms. after stimulus onset, this result could not have been found using fMRI, as fMRI's temporal resolution is simply not high enough.

Limitations and implications (Simon)

The study presented in this paper has some limitations. Firstly, only one participant was analysed. This has important implications on which inferences we can make based on the results. The findings should therefore not be generalized beyond this subject but could provide researchers with directions for further studies. The linear support vector machine achieves a very high accuracy which peaks at 0.94 (CI: 0.89-98). This high accuracy might reflect some degree of overfitting to the participant. Including more participants in the analysis would shed some light on whether this is the case.

Another concern of the study is the stimuli presented. The stimuli consisted of two types of faces. The first being fearful and the second being neutral. While the pictures were designed to mimic human facial expressions, we don't know whether this has been successful, as the stimuli were quite smiley-like. However, we do seem to observe the M170 face detection component suggesting that the smileys were perceived as faces. The fearful smiley is characterised by dilated eyes and an open mouth which is supposed to resemble a fearful response. However, one might argue that this facial expression could also resemble surprise, which according to some studies have similar morphological structures (Neta et al., 2017). Further studies should therefore investigate whether the effects found in this paper corresponds to real human faces.

A final problem of the study is associated with the analysis. The MEG scanner consists of different sensor types: gradiometers and magnetometers. The former measures and calculate the derivative of the strength of the magnetic field with respect to the position of the magnetic field. The later measures the strength. This means that the values obtained from each sensor are on different scales. This fact was not included into the analysis and therefore this might have impaired the classifiers. This also complicates the interpretation of what the classifiers used to differentiate between the two conditions. One could argue that since the derivatives are systematically related to the actual strength of the magnetic field, the overall pattern of activation would not necessarily change in a very drastic way if we had excluded one type of sensor from the analysis. Another problem is that the NB classifier assumes independence of the different sensors, which is not true given that activation of one sensor is going to be affected by the activity of other sensors. In addition, brain activity in one part of the brain cannot be said to be independent to the activity in other areas of the brain. However, as the NB classifier produce accuracy scores similar to the other classifiers, we can argue that the results remain valid.

Setting aside all the problems and concerns mentioned, the findings suggest that fearful faces are processed differently than neutral faces and that the difference is evident around 50-100 ms. after stimulus onset. This is consistent with previous evidence suggesting modulation of face perception in the earliest stages processing. Further, the fact that the activity difference occurs before the M170 component generally associated with initial face detection in cortical areas, suggest the possibility that emotional expressions are processed in subcortical structures, such as the amygdala, prior to cortical visual processing. This is particularly likely in the current study, given

that participants were presented with words prior to the face stimuli that predicted the emotional expression of the target face. Hence, priming may have activated emotional processing of the words prior to presentation of the faces, influencing the visual processing in an expectation-based manner and causing the early modulation. Alternatively, the early signal differences might reflect the P100 component, and indicate that low-level visual features specific to different emotional expressions are processed in the initial stage of visual processing prior to the detection of a face. Using more spatially sensitive measures such as fMRI, could potentially provide clarification on the specific nature of this early difference. In either case, the findings suggest that humans are really efficient in distinguishing different emotional facial expressions from one another. Further, the slightly stronger activation elicited by fearful faces in the very early stages of processing, indicates that this emotion may elicit attentional priority compared to neutral faces. That is, priming may have caused stronger recruitment of attention in the fearful compared to the neutral condition. Being social creatures, this makes intuitive sense as quickly identifying fearful responses in another human might prompt an appropriate fear response resulting in behaviour that is optimal in terms of increasing one's chances of survival. This should of course be further investigated.

Conclusion (Together**)**

The present study investigated the temporal difference in the processing of neutral and fearful faces using classification methods to investigate data recorded with MEG. Visual inspection of activation patterns revealed similar activation pattern across the two conditions. However, the different classifiers applied were able to discriminate between the two patterns of activity at different time windows suggesting that there are indeed processing differences between fearful and neutral faces. The best performing classifier proved to be the linear support vector machine, which maxed out at an accuracy above 0.9 approximately 110 ms. after stimulus onset. Multiple concerns are associated with the study, the major ones being only including one participant in the analysis and the ambiguous emotional expression of the fearful stimuli. The results should therefore be interpreted with extreme caution.

References

Altman, N. S. (1992). An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. The American Statistician, 46(3), 175-185. doi:10.2307/2685209

Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., & Desai, R. H. (2016). Toward a brain-based componential semantic representation. *Cognitive neuropsychology*, *33*(3-4), 130-174.

Buzsáki, G. (2006). Rhythms of the brain. Oxford: Oxford University Press.

Cetin, M. S., Houck, J. M., Rashid, B., Agacoglu, O., Stephen, J. M., Sui, J., ... & Calhoun, V. D. (2016). Multimodal classification of schizophrenia patients with MEG and fMRI data using static and dynamic connectivity measures. *Frontiers in neuroscience*, *10*, 466.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297. doi:10.1007/bf00994018

Gahukar G. (2018, Nov 8) *Classification Algorithms in Machine Learning…*. Retrieved from https://medium.com/datadriveninvestor/classification-algorithms-in-machine-learning-85c0ab65ff4

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., ... & Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in neuroscience*, *7*, 267.

Guerra, L., McGarry, L. M., Robles, V., Bielza, C., Larrañaga, P., & Yuste, R. (2011). Comparison between supervised and unsupervised classifications of neuronal cell types: A case study. Developmental Neurobiology, 71(1), 71-82. doi:10.1002/dneu.20809

Hämäläinen, M., Hari, R., Ilmoniemi, R. J., Knuutila, J., & Lounasmaa, O. V. (1993). Magnetoencephalography---theory, instrumentation, and applications to noninvasive studies of the working human brain. Reviews of Modern Physics, 65(2), 413-497. doi:10.1103/RevModPhys.65.413

Hand, D. J., & Yu, K. (2001). Idiot's Bayes: Not So Stupid after All? International Statistical Review / Revue Internationale de Statistique, 69(3), 385-398. doi:10.2307/1403452

Ishai, A., Schmidt, C. F., & Boesiger, P. (2005). Face perception is mediated by a distributed cortical network. *Brain research bulletin*, *67*(1-2), 87-93.

Kanwisher, N. (2000). Domain specificity in face perception. *Nature neuroscience*, *3*(8), 759.

Kaiser, D., Azzalini, D. C., & Peelen, M. V. (2016). Shape-independent object category responses revealed by MEG and fMRI decoding. *Journal of neurophysiology*, *115*(4), 2246-2250.

Krolak-Salmon, P., Fischer, C., Vighetto, A., & Mauguiere, F. (2001). Processing of facial emotional expression: Spatio-temporal data as assessed by scalp event-related potentials. *European Journal of Neuroscience*, *13*(5), 987-994.

Lemm, S., Blankertz, B., Dickhaus, T., & Müller, K.-R. (2011). Introduction to machine learning for brain imaging. *NeuroImage*, 56(2), 387-399.

Maratos, F. A. (2011). Temporal processing of emotional stimuli: The capture and release of attention by angry faces. *Emotion*, *11*(5), 1242.

Milders, M., Sahraie, A., Logan, S., & Donnellon, N. (2006). Awareness of faces is modulated by their emotional meaning. *Emotion*, *6*(1), 10.

Morel, S., Ponz, A., Mercier, M., Vuilleumier, P., & George, N. (2009). EEG-MEG evidence for early differential repetition effects for fearful, happy and neutral faces. *Brain Research*, *1254*, 84-98.

Neta, M., Tong, T. T., Rosen, M. L., Enersen, A., Kim, M. J., & Dodd, M. D. (2017). All in the first glance: first fixation predicts individual differences in valence bias. *Cognition and Emotion*, *31*(4), 772-780.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, *12*(Oct), 2825-2830.

Righi, S., Marzi, T., Toscani, M., Baldassi, S., Ottonello, S., & Viggiano, M. P. (2012). Fearful expressions enhance recognition memory: electrophysiological evidence. *Acta psychologica*, *139*(1), 7-18.

Sato, W., Kochiyama, T., Uono, S., Matsuda, K., Usui, K., Inoue, Y., & Toichi, M. (2011). Rapid amygdala gamma oscillations in response to fearful facial expressions. *Neuropsychologia*, *49*(4), 612-617.

Warriner, A.B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*, 1191-1207.

Williams, L. M., Palmer, D., Liddell, B. J., Song, L., & Gordon, E. (2006). The 'when'and 'where'of perceiving signals of threat versus non-threat. *Neuroimage*, *31*(1), 458-467.

Spatial differences in the processing of fearful compared to neutral faces: an fMRI study

Authors: Line Kruse, Martin Ito, Christoffer Olesen and Simon Hansen

Abstract (Together**)**
Although visual processing of faces is largely thought to be dependent on selective activity of the Fusiform Face Area, evidence suggests that this area is part of a distributed network including the primary visual cortex, the insula and the amygdala. These areas are thought to be critically involved in emotional processing, and it has subsequently been suggested that the processing of faces is modulated by the emotional expression of the face. The current study is the second part of a two-part study of emotional face processing. The experimental paradigm consisted of a predictive face detection task of neutral and fearful faces. The target faces were preceded by the presentation of emotionally valenced words predictive of the following facial expression. Using fMRI and multivariate classification analysis the current study aimed to investigate spatial activation differences in the processing of the two types of emotional faces. Results indicated that the largest differences between the two conditions were found in the primary visual cortex, suggesting that differences between the two conditions primarily is due to low-level features of the emotional expressions, such as shape, size, and spatial frequency. A limitation of these results is that activation differences in deeper subcortical structures, such as the amygdala, may not have been detectable with fMRI. Thus, a potential modulatory effect of emotional content cannot be excluded.

Introduction (Together)
This fMRI study is the second half of a two-part study on the processing of emotional faces, the other half being a MEG study (see portfolio assignment on MEG). As mentioned in the MEG study, there seems to be a processing stream specialized for faces, which is intertwined with emotional processing streams. This is important to investigate, as it plays a crucial role in successful social interaction. Whereas the MEG study focused on the temporal components of neural activity in response to fearful and neutral faces, this study focused on the spatial components of this neural activity.

Face processing (Martin)
One area of interest for face processing in the brain is the Fusiform Face Area located in the Inferior Temporal Cortex. In an fMRI study, Kanwisher et al. (1997) found that the Fusiform Face Area responded significantly more to pictures of faces than scrambled faces, objects,

houses and hands. This suggests that the Fusiform Face Area is specialized for unscrambled faces specifically. However, the Fusiform Face Area might not inherently be specialized for face processing. Instead, it might be involved in processing familiar visual stimuli and thus reflect learned expertise in recognizing certain stimuli rather than a specific face detection area. In an fMRI study by Gauthier and colleagues (1999) activation was detected in the right Fusiform Face Area when people were looking at “greebles” (figure 1). This activation was stronger in participants who had been familiarized with the greebles and thus become greeble experts and the activation was attenuated by inverting the greebles’ features. Thus, viewing greebles can elicit the same activation as viewing faces. Similar results have later been found for cars and birds as well (Gauthier et al., 2000). This highlights the importance of face processing, because the association suggests that most people are face experts.

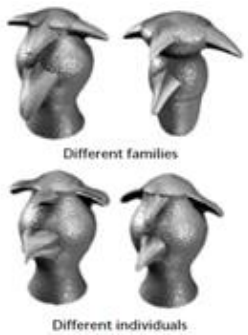


Figure 1: Greebles (Gauthier et al., 1999).

A distributed network of emotional face processing

Ishai, Schmidt and Boesiger (2005) have found evidence that the Fusiform Face Area is part of a greater distributed cortical network processing both faces and the emotions they display. This network is posited to have two systems; a core and an extended system. The core system can be split into a ventral and a dorsal part. The more ventral region of the core system encompasses the Inferior Occipital Gyrus and the Fusiform Gyrus including the Fusiform Face Area. This area appears to be involved in face-based recognition of individuals. The more dorsal area of the core system, the Superior Temporal Sulcus, seems to be involved in the processing of social cues in faces, e.g. gaze following and lip movements. The extended system consists of the Amygdala, Insula, the Nucleus Accumbens, Prefrontal Cortex and Hippocampus. The study showed bilateral activation throughout both the core and the extended system when viewing faces, but the activation was stronger and spanned a slightly larger area in the right hemisphere. When comparing line drawings and black and white photographs of unfamiliar faces, the activation pattern

was not significantly different. This suggests that faces do not have to be photorealistic to activate this system. Emotional faces elicited a stronger bilateral response in the core system compared to neutral faces. These results indicate that the roles of the different parts of the network are not always clear cut, as one would have expected the extended rather than the core system to react more strongly to emotional faces. Noesselt et al. (2005) compared neural activation when viewing fearful and neutral faces. They found that participants exhibited stronger activation in the Amygdala when viewing fearful faces compared to neutral faces. In addition, the activation was stronger in the right hemisphere. This finding suggests that the extended system is involved in emotional processing as expected. However, although the Amygdala seems to be an obvious area of interest, there are a number of problems associated with doing fMRI detection of signals in this area. Boubela et al. (2015) found that in addition to inhomogeneities in the magnetic field, caused by air filled cavities in the skull near the Amygdala, the signal from the Amygdala is confounded by blood flow from nearby veins. Consequently, we decided not to investigate the Amygdala further. The other area of the extended system that is posited to process emotion is the Insula. The Insula has bidirectional connections with many different areas of the brain and seems to be involved, among other things, in interoception. Based on this, the Insula has been posited to play a crucial role in generating emotions from interoception. Earlier it was believed that the Insula was specialized in disgust, but it seems to be involved in the processing of several emotions suggesting a general function rather than a specific one (Gasquoine, 2014).

Face-like stimuli

How face-like stimuli need to be in order to activate the face processing system is subject to debate. As mentioned earlier, Ishai et al. (2005) found that this face processing system was activated when viewing line drawings of faces, indicating that stimuli do not need to be photorealistic to elicit a response. The stimuli of the current experiment were simpler than line drawings however, and more smiley-like. In a study by Yuasa, Saito and Mukawa (2006) investigating activation in photographs of emotional faces and Japanese-style text smileys, e.g. a happy smiley depicted as (^_^), they found that both emotional faces and smileys had activation in the right middle frontal gyrus, interpreted as emotional processing. However, the photographs elicited activation in the Fusiform Gyrus, whereas the smileys did not, suggesting that the emotional content of the smileys were processed whereas the processing of the “face” content had been abstracted away. In a study on pareidolia (the tendency to see faces in non-face stimuli), Akdeniz, Toker and Atli (2018) found that

compared to scrambled faces, photographs of faces and pareidolia elicited almost identical patterns of activation in V1 and V2 of the occipital lobe, the Fusiform Face Area and Prefrontal Cortex. This suggests that non-face facelike stimuli can elicit activation in areas associated with face processing. This might have been caused by the participants expecting to see a face given the context of the experiment. This context is different from the experiment of Yuasa et al (2006), where the participants might have been more focused on the emotion communicated by the smiley rather than seeing a face. In general, drawing a line between what is regarded as a face and what is not appear to be a complex task. We did observe a component in the MEG part of the study, which could reflect the M170 component, suggesting face detection did occur, at least in the MEG study. Although using smiley-like stimuli comes with some challenges, it offers better experimental control compared to photos of faces, as real faces would be judged on multiple personal factors such as gender, age and race. Therefore, smiley-like stimuli were chosen for better experimental control.

As processing of emotions and faces rely on multiple areas of the brain, fMRI was chosen as a method, as it offers good spatial resolution and also detects differences in strength of activation, which might differentiate types of processing within the same area.

fMRI (Simon)

Functional Magnetic Resonance Imaging (fMRI) has been a popular method for combining structural and functional imaging of the brain. There are multiple reasons for this. Firstly, the method allows for good spatial resolution and also with new fMRI equipment a decent temporal resolution. Another important point is that it is non-invasive, which means that you do not need to open up people’s skull and insert electrodes, which is required for multiunit and single-unit recording, or inject radioactive material into the participant, which is done in PET studies. Due to these advantages fMRI has become a popular choice for localizing different functions of the brain (see figure 2 for a comparison of the spatial and temporal resolution of different neuroscience methods).

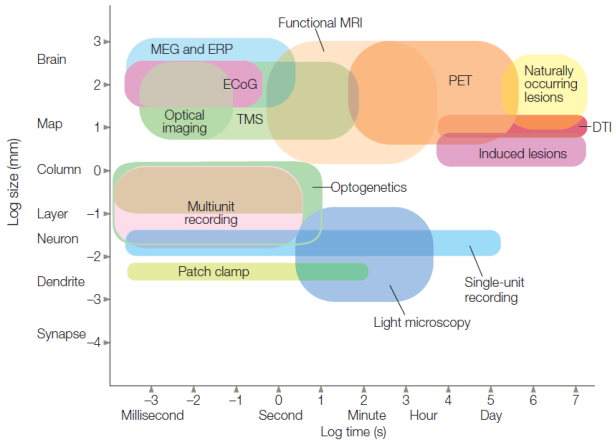


Figure 2: fMRI compared to other neuroscientific methods. The x-axis represents the temporal resolution and the y-axis represents the spatial resolution. From the plot it is evident that fMRI has a good spatial resolution for a non-invasive method and decent temporal resolution although not as fast as MEG and EEG. (Gazzaniga, 2013)

An fMRI scanner basically works by detecting changes in the magnetic field to allow for identification of brain structure (MRI) and function (fMRI). First, a structural scan is acquired by the fMRI scanner. When the scanner is activated it creates a strong magnetic field that aligns the spin of the hydrogen protons in the brain to the magnetic field of the scanner. The scanner then emits radio frequency pulse which alters the orientation of the protons. When realigning to the magnetic field of the scanner, also known as relaxation, the protons emit a radio pulse, which is picked up by the scanner (Buxton, 2013). Two characteristics makes it possible to distinguish different kinds of tissue from another. First, the hydrogen density varies across regions, resulting in differences in magnetic strength. Second, the relaxation time varies across tissue (Sprawls, 2000). This is used to compute the structural image of the brain.

The functional component of fMRI measures the ratio of deoxygenated blood to oxygenated blood. The idea being that if an area of the brain show increased activation more oxygenated blood will flow to that area in order to supply the neurons with energy. Studies have found that increased activity in the brain is correlated with a higher oxygenated to deoxygenated hemoglobin ratio (Gusnard & Raichle, 2001; Buxton, 2009). As they have different magnetic properties (deoxygenated blood is weakly magnetic and oxygenated blood is not), we can use differences in the signal emitted from the protons to infer activation of different brain regions. This measurement is called the Blood-Oxygen-Level-Dependent (BOLD) signal. However, it should be clear that the BOLD signal is an indirect measure of activity. Research on the coupling of neural activity and the BOLD signal has proven the relationship to be non-trivial. The BOLD signal is a complex function that is affected by both cerebral blood flow, cerebral metabolic rate of oxygen and cerebral blood volume (Murta et al., 2015). Research

into the BOLD signal has found that the signal is not correlated with neuronal spiking activation, but rather local field potential. This has led to questioning of what type of information processing we actually model with the BOLD signal (Kayser & Logothetis, 2013).

An important element in modeling the relationship between neural activity and the signals picked up by the scanner is the hemodynamic response function. This is a model of how the BOLD signal develops following brain activation (elicited by for example presentation of visual stimuli). The canonical hemodynamic functions states that the BOLD signal first has an initial dip, followed by a sharp rise. After this it drops to below baseline level and then stabilises around baseline level. Using the canonical hemodynamic response function we assume that the BOLD signal has specific shape and temporal properties. This assumption has been questioned by different studies. Firstly, the BOLD signal has been found to vary across different age groups (West, 2019) and have slightly different temporal characteristics across different areas of the cortex (Taylor et al., 2018).

In general, two different types of study designs are used in fMRI studies. These are called block and event-related study design. In a block design the stimuli of one condition is presented in one block. Then there is a break and after this, a block of the other condition is initiated. The block design method allows for detection of very subtle effects and is considered to have better power than the event-related design (Chee et al., 2003). The other method is event-related where different stimuli classes are presented in a mixed order and pause duration might be varied. Using an event-related study design is thought to reduce predictability (Chee et al., 2003).

One of two methods of analysis is typically applied to fMRI data. The univariate approach (e.g. Flandin & Novak, 2013) and the multivariate approach (e.g. Haynes, 2015). The first approach is the more traditional one. It takes the activation of voxels in different conditions and tests whether there is a significant difference in signal strength between conditions in the same voxel using the General Linear Model. While this approach has proven to be quite effective for localisation purposes it also has some serious problems. One very important one is the multiple comparison problem. This problem happens because the univariate analysis violate the assumption of independence because activity across voxels is not independent from activity in neighbouring voxels. Thus, running multiple independent test on every voxels increases our chance of getting a false-positive by chance even though there in fact is no statistical relationship. This shortcoming can be overcome by applying the method of multivariate analysis, where patterns of voxel activation is used to determine what stimuli has been presented to the participant. This allows researchers to infer what areas are

important for processing of different stimulus classes. Given that the analysis is more complex interpretation of activation levels is difficult. The current study used a combination of univariate and multivariate analysis to address differences in the spatial activation patterns when observing a neutral compared to a fearful face. The argument for applying both analyses is two-fold; 1) We want to be able to compare results across analyses to assess the robustness of the findings, 2) the methods can compensate for some of the shortcomings of the other method (e.g. the multiple comparison problem in univariate analysis).

Methods (Christoffer)

The study was the second part of a two-part study, in which our previous MEG study was the first. An important difference between the two is that another participant was tested in the present study than in the previous study. This participant was a 23 year old male. The same event-related experimental paradigm was used with some practical adjustments due to the distinct nature of the method. As mentioned, fMRI measures the BOLD signal, which has a temporal delay and for this reason the periods in between stimuli was doubled in duration by dividing the frame rate by 2. In this study we ran 6 sessions of the experimental paradigm (in contrast to the 4 sessions in the MEG study). The participant had a short break outside the scanner, between the 4th and the 5th session. Otherwise the experimental setup was the same across the two studies (for details see portfolio assignment on MEG). In short, one trial consisted of a predictive word stimulus presented to the participants followed by a fixation cross and then the target facial stimulus. There were two classes of face stimuli; neutral and fearful. The predictive words where either positive (always followed by neutral faces), negative (always followed by fearful faces) or neutral (followed by either of the face stimuli classes). Before the experimental paradigm was initiated, structural scans of the participants brain was done. The scanner used was a Siemens Prisma and it were set to a TR value of 1000ms and a TE value of 29.6ms. The participant was told that he could leave the experiment at any time.

Analysis (Line**)**

Preprocessing

The data was preprocessed in SPM12 (Penny et al., 2011) and included the following four steps. First, realignment was applied to correct for effects of subject movement. Realignment serves to determine the rigid body transformation that most effectively represent all scans obtained in a common space (Flandin & Novak, 2013). The rigid body transformation has six parameters, three translations and three rotations, and were estimated by maximizing the mean squared difference between each scan and reference scan. The reference scan consisted of the mean image. The resulting movement parameters were subsequently used

in the statistical model. Second, the structural and functional images of each subject were co-registered. This procedure allows the estimations from the structural image to be applied to the functional image, and thus increases the precision of later spatial normalisation (Flandin & Novak, 2013). Co-registration was performed by estimation of the six parameters of the rigid body transformation based on maximization of the mutual information shared between the two images (in terms of image intensities). Third, segmentation was performed such that each voxel was categorized depending on the tissue type. This facilitated the fourth step, in which images were spatially normalised, rendering them in a common space. This allows for comparison of signal patterns across subjects and was performed using the Montreal Neurological Institute (MNI) space. The data was retained unsmoothed for the analysis, since searchlight decoding was performed which takes into account the activity in neighbouring voxels for classification (see below).

Univariate statistical modelling

The statistical analysis relied upon a mass-univariate approach, in which the general linear model was applied to estimate regression coefficients at each voxel separately, resulting in a statistical parametric map (SPM) representing the statistical effect at each voxel (Flandin & Novak, 2013). The general linear model constituted a design matrix with the following regressors. First, it included stimulus functions modelling the stimuli pattern, as we expect voxel activity to alternate with the presence or absence of stimuli. The current study employed an event-related design, and stimulus was thus modelled as a delta function (an event) at the trial onset. This method assumes that the duration of the trials were zero. The 60 trials of the data were divided into chunks of five events. Thus, the model included 12 regression coefficients (60/5) each modelling five trials of either neutral or fearful faces. Further, the model included the haemodynamic response function (HRF), to account for the fact that the BOLD-signal represents the haemodynamic change mediated by neuronal activity, rather than direct neuronal activity. Third, the signal was expected to have a non-zero mean, which was modelled by a regressor constant in time. Fourth, the model included a combination of cosine waves at several frequencies, to account for expected slow fluctuations in the data due to either scanner drift or cardiac and respiratory cycles. Fifth, the movement parameters obtained in the realignment procedure were modelled as six regressors. Lastly, the model included an error-term. To account for multiple comparisons, model estimation was performed with family-wise error correction. This procedure ensures that the null hypothesis takes into account the whole search volume and thus adjust the p-values to control the false-positive rate. In total, the univariate model fits resulted in 72 beta images containing the estimated beta coefficients for each voxel in the brain for either

neutral or fearful faces. These images were used as input in subsequent classification analysis.

Contrasts

Two contrasts were computed allowing statistical inferences about the difference of activation in the two conditions on an individual voxel level. Using the t-statistics we computed two contrasts; first, activation in fearful condition subtracted from that in the neutral condition (pos-neg), and second, activation in neutral condition subtracted from that in the fearful condition (neg-pos). The t-statistic tests the null hypothesis that the difference in activation between the two conditions is zero. Contrasts were tested with a family-wise error corrected p-value below 0.05. This procedure resulted in a statistic image depicting which voxels exhibited a statistical effect indicating neural activation differences in the two conditions.

Multivariate classification analysis

In order to investigate the extent to which patterns of activation in the brain differ between the two conditions, we performed a multivariate analysis. Using multivariate pattern classification, brain activity was analysed as patterns of activation across multiple voxels and hence allow inferences about content-based processing of the brain. Further, multivariate classification accounts for the fact that the marginal distributions of activation in each experimental condition often overlap among voxels (Haynes, 2015).

We implemented supervised classification in order to assess where in the signal the largest differences in the processing of neutral versus fearful faces could be observed. The input of the classifier was the series of univariate models fit to five face trials. The data was split into training and test set based on the condition labels. The test consisted of 20% of the data (n=15). To investigate the predictive accuracy of particular areas in the brain, we employed searchlight decoding from the NiLearn package (Pedregosa et al., 2011). Searchlight decoding iteratively runs a classifier on a sphere of selected voxels, and give a score to the voxel in the centre of the sphere based on the ability of this area to predict condition class. The searchlight method assumes that information is contained in local clusters of voxels, which enables us to perform the analysis on unsmoothed data. In the current analysis, searchlight was conducted with a whole-brain mask of the training data, ensuring that voxels outside the brain are discarded. In order to assess the stability of the results, three types of classifiers were utilized. First, the spherical searchlight used a Gaussian Naïve Bayes classifier (NB), second, a Linear Support Vector Machine (SVM) was implemented and third, k-Nearest Neighbor (k-NN) was used. This allowed us to compare the extent to which the

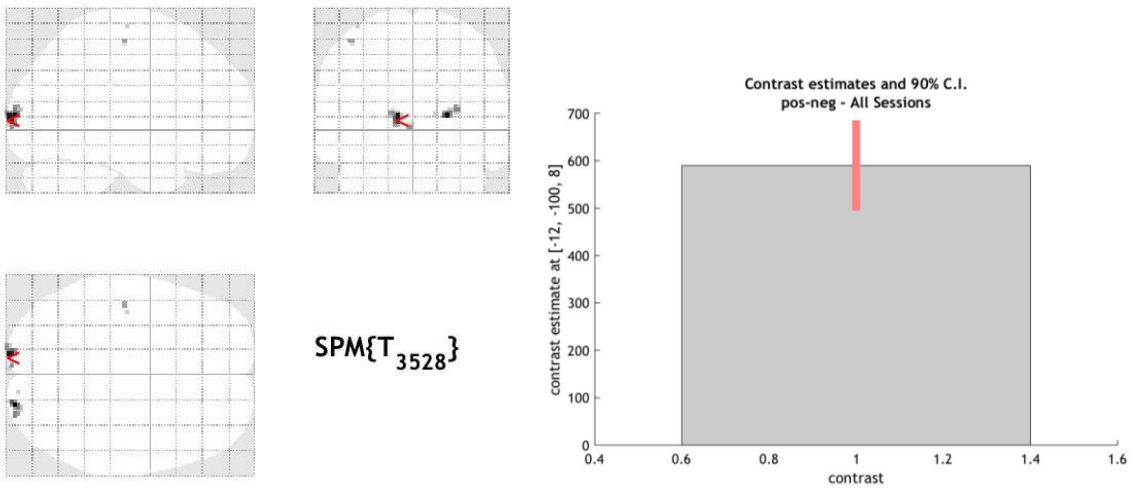
most accurate voxels were detected in similar areas of the brain. All searchlights had a radius of 5 voxels and performed 8-fold cross-validation.

From the searchlight scores, a new mask for each classifier was created based on the 500 voxels with highest classification accuracy. Using the NiLearn package based on scikit-learn tools (Pedregosa et al., 2011), we then performed classification on the test set masked with the 500 best voxels. This was done for each mask using classification methods that matched the searchlight algorithms which each mask were derived from (i.e. NB, SVM and k-NN). All classifiers employed a 4-fold cross validation. Lastly, permutation tests were carried out to address the quality of the three classification analyses. The classifications were run again with the same parameter settings, but now with 100 permutations, i.e., 100 rounds of randomized labels. The result is a p-value indicating the likelihood of obtaining the original classification results with randomized labels.

Results (Christoffer)

Univariate analysis

The univariate analysis showed most activation around the primary and secondary visual cortex. The neutral-fearful contrast showed mostly activation in the primary visual cortex. Two peak values were identified at the coordinates [-12, -100, 8] and [22,-96,8]. (figure 3)



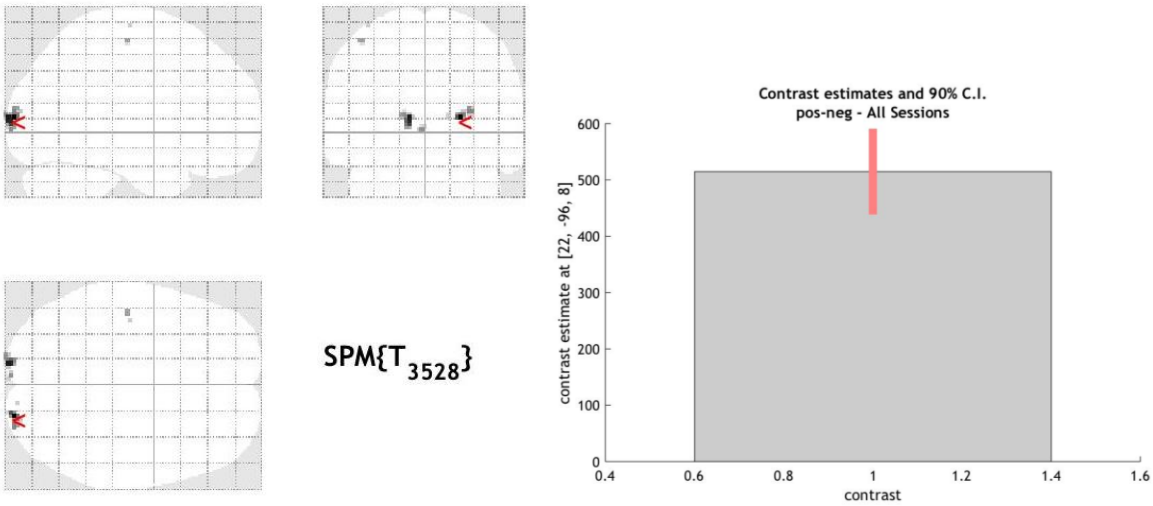


Figure 3: Positions and contrast estimates of the two peak values in the neutral-fearful contrast. Above: coordinates [-12, -100, 8]. Below: coordinates [22,-96,8].

The fearful-neutral contrast showed activation in the primary and secondary visual cortex. Two peak values were identified at the coordinates [-12, -100, 8] and [22,-96,8] (figure 4).

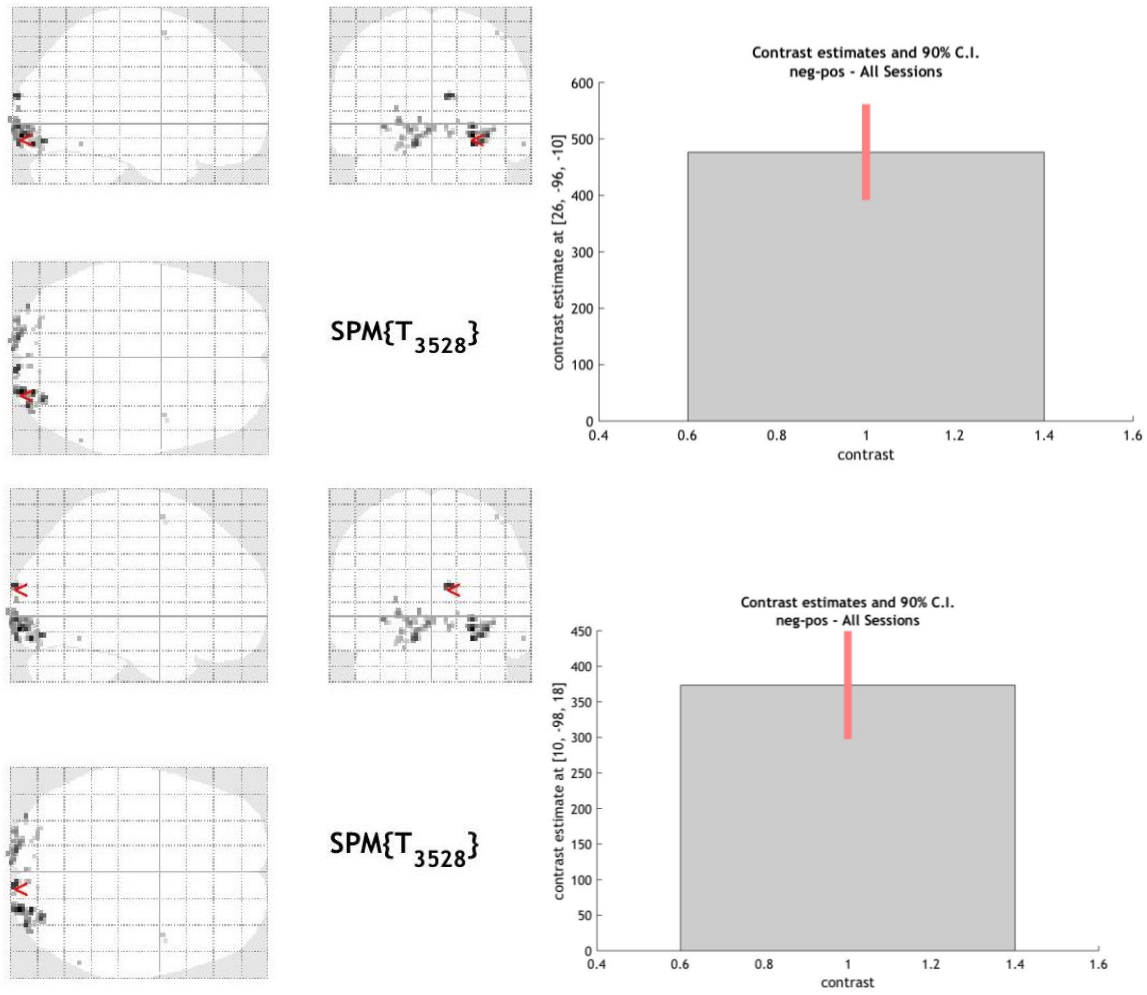


Figure 4: Positions and contrast estimates of the two peak values in the fearful-neutral contrast. Above: coordinates [26, -96, -10]. Below: coordinates [10,-98,18].

Multivariate analysis

The 500 best predicting voxels in all searchlight analyses where mostly located in the primary and the secondary visual cortex (figure 5).

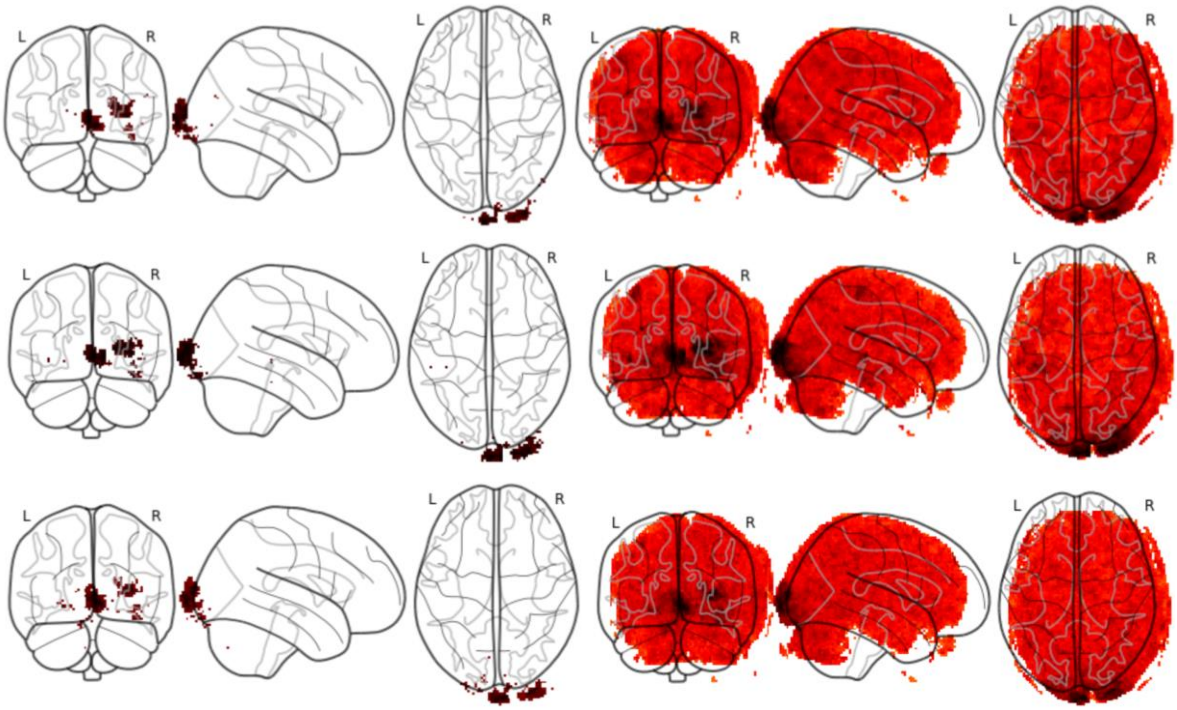


Figure 5: The dark red areas represent the predictive effect of voxels using NB (first row), SVM (middle row) and k-NN (last row). The darker the more predictive. The three first brains of each row shows only the 500 best predicting voxels and the last three shows the predictability across the whole brain. Note that there is an offset between the brain illustration and the activation image which is imposed on top of it. However, it gives a rough idea of where these voxels are in the brain and makes the 3 classifications comparable on this matter. The last three brains of each row gives an idea of the extent of this off set.

The NB classification gave a mean accuracy of 0.73. The permutation test gave a significant result ($p < .05$) (figure 6.A). The SVM classification gave a mean accuracy of 1. The permutation test gave a significant result ($p < .05$) (figure 6.B). The k-NN classification gave a mean accuracy of 0.85. The permutation test gave a significant result ($p < .05$) (figure 6.C).

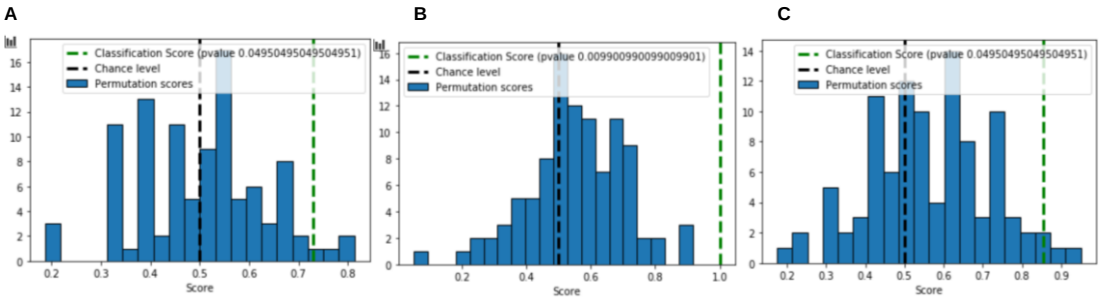


Figure 6 - Plot showing the permutation scores and the classification score of each classifier. A: NB. B: SVM. C: k-NN.

Discussion

Interpretation of results (Christoffer)

The current paper analysed fMRI data from one participant using two different paradigms of analysis; univariate and multivariate analysis. In the univariate analysis we observed that the primary difference in activation was in the primary visual cortex. In the neutral-fearful contrast, it was observed that the activation was a bit higher up in the visual cortex compared to the fearful-neutral contrast. In addition, a second peak activation with a smaller spatial extent was found in the fearful-neutral contrast (figure 4), located slightly above the main visual activation in a medial position.

Turning to the multivariate analysis we found that the voxels that are most capable of distinguishing the two stimulus classes from each other were also located in the visual cortex, in line with the results produced by the univariate analysis. This further supports the effect. The areas consisting of the 500 best predicting voxels were roughly the same across the three different searchlight analyses. We found two prominent areas, one in the right hemisphere and one in the left. The former was located a bit lower than the latter. The classifiers all yielded relatively high accuracies, where the highest was the SVM classification, with an extraordinary accuracy of 1, and the lowest was NB, with an accuracy of 0.73. The high accuracy from the classifiers, especially the SVM accuracy of 1, hints towards a pitfall of the current study, namely that only one participant was tested. First, this makes it highly likely that the classification is overfitted to this participant and second, it shows a limit in the amount of data, in which the number of images that were in the test set (n = 15) inflated the probability of getting all correct by chance. Taken together, these two points makes it probable to get a mean accuracy of 1 by chance, despite the fact that the permutation test shows otherwise, simply for the reason that the model is likely overfitted to begin with.

Comparing the results of the univariate and multivariate analyses, the second peak activation identified in the neg-pos contrast in the univariate analysis was not found in the multivariate analysis. This suggests that even though it is significantly more active when the participant is presented with fearful faces, it is not predictive of the stimulus classes when taking the spatial patterns of activation into account. Both analyses showed a height difference in activation. In the univariate analysis, the height differences was between contrast, in the multivariate difference the height difference was between the hemispheres. This could signify a difference in emotional processing, but looking at the areas where the activation occurred, this height difference probably reflects a difference in the processing of visual features. Both analyses found activation in primary visual cortex and surrounding areas, which are all involved in early stages of visual processing. The primary visual cortex of macaques has been found to be sensitive to orientation, spatial frequency and color of stimuli (Mazer et al., 2002; Johnson et al., 2008). Our findings suggest that the activation extends beyond just V1 and into V2. V2 receives projections from V1 and further projects to V3. V2 also projects back to V1, and this is part of a larger system of feedback connections, in which the later visual areas project back to earlier visual areas (Hupé et al., 1998). V2 has been implicated in the processing of complex shapes like circles and arcs, which is what our face stimuli consisted of (Hegdé & Van Essen, 2000).

Face processing system and MEG (Together)

There are several possible explanations for why we do not see significant activation in the face processing system described earlier. One possible explanation is that the stimuli weren't sufficiently realistic and thus perceived more like smileys. However, we also see no activation in areas associated with emotional processing of smileys. The focus on classification rather than emotions or faces in the behavioral task might explain why there seemed to be no activation in either the core or extended system despite pareidolia stimuli being able to elicit activation. Additionally, the stimuli probably would not elicit a strong emotional response, even if the participant paid attention to the emotional content of them. Another reason why we see no activation related specifically to processing faces is that the contrasts might not afford it properly, as the contrasts subtract the activation occurring in one condition from the other, without having a scrambled face baseline. We would then expect the results of the contrasts to be the difference in the emotional content of the two faces. We observed no differences in areas related to emotional processing. As the insula is posited to play a role in multiple emotions, the only way to differentiate insula activation in different emotions using fMRI would be the differences in strength and spatial extent of the activation. As mentioned above, the stimuli did not elicit strong emotions, rendering differences in

strength of activation unreliable. When the strength of the signal is weak, the differences in spatial extent of activation would also be masked by noise. Combined, this could explain why we see no insula activation. Lastly, the differences between emotions might be temporal, making an fMRI study unsuited to detect the differences. If differences in emotion could not be detected reliably, one would expect that lower level visual areas would be the areas that predicted condition the best, as more basic features such as shape would consistently stay different between conditions and contrast.

Comparing the results of this study with the temporal results of the MEG part of this study, they support each other. The finding that the best classification accuracy was around 110 ms which is an early component, fits very well with the fMRI analysis suggesting that discrimination of the two stimuli classes happens at a very early stage of visual processing in V1 and V2. This is further supported by findings that the P100 dipole has been suggested to originate from the lateral bottom of the calcarine fissure roughly corresponding to somewhere within the primary and secondary visual areas (Seki et al., 1996). Additionally, the highest activation peak in the signal was observed around the M170 component, which reflects initial face detection. Thus, MEG results suggests that emotional classification occurred prior to actual face-detection. Further, the MEG results indicate that the stimulus was actually perceived as face like stimulus. The results from the fMRI study seem to support the interpretation that the early differences in processing of the two emotional expressions are primarily driven by processing of low-level features in the visual cortex. The very early classification accuracy observed in the MEG results, may have been facilitated by the predictive effect of emotional content words. If this is the case, it appears that the predictive effects primarily facilitated early low-level visual processing. However, given the difficulties in picking up signals from the limbic structures, such as the amygdala, using fMRI, there might have been some modulation by this area, which we are simply not capturing with the method used.

Conclusion (Together)

The human brain is specialized in detection and processing of faces, and particularly emotional expressions of faces provide crucial information for social interactions. Combining MEG and fMRI analyses, the current paper used a predictive face detection paradigm to investigate the extent to which emotional expressions modulate the processing of faces, as well as the specific nature of such effects. The fMRI results suggested significant spatial activation differences in the processing of fearful versus neutral faces. In particular, the two conditions elicited activation in slightly different areas of the visual cortex. This suggests that the processing difference between emotional expressions is driven by low-level visual

features of the facial expression. Additionally, the MEG results indicated that this difference occurred in the earliest stage of processing. This may be due to the predictive words preceding the target faces causing very early processing differences between the two conditions. If this is the case, it appears that the priming effect primarily occurred in the visual cortex. However, further analysis contrasting predictive and non-predictive words is needed to establish whether this is the case. The finding that the largest differences occurred in processing of low-level visual features is supported by the fact that no differences were detected in the insula. However, the current study was not able to detect signals from the amygdala, hence the possibility that this difference is, at least partly, modulated by early emotional processing remains.

References

Akdeniz, G., Toker, S., & Atli, I. (2018). Neural mechanisms underlying visual pareidolia processing: An fMRI study. *Pakistan journal of medical sciences*, *34*(6), 1560.

Boubela, R. N., Kalcher, K., Huf, W., Seidel, E. M., Derntl, B., Pezawas, L., ... & Moser, E. (2015). fMRI measurements of amygdala activation are confounded by stimulus correlated signal fluctuation in nearby veins draining distant brain regions. *Scientific reports*, *5*, 10499.

Buxton, R. B. (2009). Introduction to Functional Magnetic Resonance Imaging (pp. 20-21). Cambridge University Press

Buxton, R. B. (2013). The physics of functional magnetic resonance imaging (fMRI). *Reports on Progress in Physics*, *76*(9)

Chee, M. W., Venkatraman, V., Westphal, C., & Siong, S. C. (2003). Comparison of block and event-related fMRI designs in evaluating the word-frequency effect. *Human brain mapping*, *18*(3), 186-193.

Flandin, G., & Novak, M. J. U. (2013). fMRI Data Analysis Using SPM. In S. Ulmer & O. Jansen
(Eds.), *fMRI. Basics and Clinical Applications* (2nd edition ed., pp. 51-76). Berlin: Springer.

Gasquoine, P. G. (2014). Contributions of the insula to cognition and emotion. *Neuropsychology review*, *24*(2), 77-87.

Gauthier, I., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature neuroscience*, *3*(2), 191.

Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P., & Gore, J. C. (1999). Activation of the middle fusiform'face area'increases with expertise in recognizing novel objects. *Nature neuroscience*, *2*(6), 568.

Gazzaniga, M., Ivry, R. B., & Mangun, G. R. (2013). Chapter 3: Methods of Cognitive Neuroscience. *Cognitive Neuroscience: The Biology of the Mind* (4th edition, pp 70-119). WW Norton.

Gusnard, D. A., & Raichle, M. E. (2001). Searching for a baseline: functional imaging and the resting human brain. *Nature Reviews Neuroscience*, *2*(10), 685.

Haynes, J. D. (2015). A Primer on Pattern-Based Approaches to fMRI: Principles, Pitfalls, and
Perspectives. *Neuron*, 87, 257-270

Hegdé, J., & Van Essen, D. C. (2000). Selectivity for complex shapes in primate visual area V2. *Journal of Neuroscience, 20*(5), RC61-RC61.

Hupé, J. M., James, A. C., Payne, B. R., Lomber, S. G., Girard, P., & Bullier, J. (1998). Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons. *Nature, 394*(6695), 784.

Johnson, E. N., Hawken, M. J., & Shapley, R. (2008). The orientation selectivity of color-responsive neurons in macaque V1. *Journal of Neuroscience, 28*(32), 8096-8106.

Ishai, A., Schmidt, C. F., & Boesiger, P. (2005). Face perception is mediated by a distributed cortical network. *Brain research bulletin*, *67*(1-2), 87-93.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience, 17*(11), 4302-4311.

Kayser C & Logothetis NK (2013). The Electrophysiological Background of
the fMRI Signal. In: Ulmer S & Jansen O (eds.) fMRI. Basics and Clinical Applications. 2nd edition ed. Berlin: Springer.

Mazer, J. A., Vinje, W. E., McDermott, J., Schiller, P. H., & Gallant, J. L. (2002). Spatial frequency and orientation tuning dynamics in area V1. *Proceedings of the National Academy of Sciences, 99*(3), 1645-1650.

Murta, T., Leite, M., Carmichael, D. W., Figueiredo, P., & Lemieux, L. (2015). Electrophysiological correlates of the BOLD signal for EEG-informed fMRI. *Human brain mapping*, *36*(1), 391-414.

Noesselt, T., Driver, J., Heinze, H. J., & Dolan, R. (2005). Asymmetrical activation in the human brain during processing of fearful faces. *Current Biology*, *15*(5), 424-429.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, *12*(Oct), 2825-2830.

Penny, W. D., Friston, K. J., Ashburner, J. T., Kiebel, S. J., & Nichols, T. E. (Eds.). (2011). *Statistical parametric mapping: the analysis of functional brain images*. Elsevier.

Seki, K., Nakasato, N., Fujita, S., Hatanaka, K., Kawamura, T., Kanno, A., & Yoshimoto, T. (1996). Neuromagnetic evidence that the P100 component of the pattern reversal visual evoked response originates in the bottom of the calcarine fissure. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, *100*(5), 436-442.

Sprawls, P. (2000). Chapter 1: Magnetic Resonance Image Characteristics. *Magnetic resonance imaging: principles, methods, and techniques*. Retrieved from http://www.sprawls.org/mripmt/MRI01/index.html

Taylor, A. J., Kim, J. H., & Ress, D. (2018). Characterization of the hemodynamic response function across the majority of human cerebral cortex. *NeuroImage*, *173*, 322-331.

West, K. L., Zuppichini, M. D., Turner, M. P., Sivakolundu, D. K., Zhao, Y., Abdelkarim, D., ... & Rypma, B. (2019). BOLD hemodynamic response function changes significantly with healthy aging. *Neuroimage*, *188*, 198-207.

Yuasa, M., Saito, K., & Mukawa, N. (2006, April). Emoticons convey emotions without cognition of faces: an fMRI study. In *CHI'06 extended abstracts on Human factors in computing systems* (pp. 1565-1570). ACM.