Bachelor thesis
Computing Science

Radboud University

Cognitive Psychology For Black Box Models

On Holistic And Featural Face Processing in Neural Networks

Author:
Arne Diehl
S4451392

First supervisor/assessor:
Prof. Dr. Ir. A.P. de Vries
arjen@acm.org

Second assessor:
MSc T.M. van Laarhoven
t.vanlaarhoven@cs.ru.nl

April 4, 2020

Abstract

A known problem with neural networks is the lack of interpretability they offer. This lack hinders the application of the technology to fields where comprehensibility is critical, such as the judicial system. It also undermines trust in the decisions made by neural network based systems. This paper explores the application of knowledge gathered in the cognitive sciences to the field of neural network interpretability. It seems that well established experiments from psychology can also help us understand neural networks behaviour and decisions. The hypothesis that I will test in this paper is the claim that neural networks are simple featural classifiers.

Contents

Chapter 1

Introduction

1.1 Context

In the last few years the role of artificial neural networks (ANN) has increased dramatically within the fields of artificial intelligence and computer science, as well as in the industry, with companies like Google and Facebook investing heavily in creating frameworks and technologies surrounding ANNs. This trend kicked of with Krizhevsky et al. publishing AlexNet [13] in 2012, a convolutional neural network (CNN) which outperformed the competition for the ImageNet Large Scale Visual Recognition Challenge significantly. Following this the interest in so called deep learning architectures increased, as they could be trained for many different tasks and achieve human like accuracy in many cases.

1.2 Explainability is an unsolved problem

Interestingly however our understanding of how exactly these networks reach their conclusions has remained an open question. Neural network models routinely have millions of parameters that are being learned, which makes it hard to reason about them as a whole. There have been attempts to understand trained models based on ANN algorithms by fitting a simpler model to a part of the neural network, such as LIME [25] or FullGrad [32]. However these approaches obviously fall short of the goal of increasing the general understanding of models created with deep learning architectures. Other approaches such as models producing an explanation in addition to their normal output, have an obvious shortcoming: the explanation itself is created by a process we don't understand, which makes the explanation just as trustworthy as the regular output.

1.3 Inspiration from Cognitive Psychology

Yet another approach is to apply methods from cognitive psychology to trained models in order to explain them the same way we try to explain the mental processes of humans. Ritter et al. [26] used results from infant research and applied them successfully to the domain of ANNs. The team hypothesised that the shape bias found in human categorisation should also exist in one-shot neural network models trained on object categorisation tasks. They were able to show that indeed differently seeded ANN architectures were displaying different levels of shape bias. This lead to further research into the shape bias as a property of a model, which in turn lead to a study by Geirhos et al. [7], in which they showed that stronger shape bias leads to higher accuracy and robustness of models and also showed how to achieve that. This illustrates how promising this kind of research is and why it helps us understand neural networks.

1.4 Neural Network Face Detection

In this research we want to apply cognitive psychology (CP) to face detection in neural networks. Especially to the question as to whether neural networks process faces one feature at a time (featurally) or as a whole (holistically). Hinton et al. [27] proposed a new kind of neural network based on capsules, that is directly inspired by theories from cognitive psychology. They are trying to process objects as a whole, rather that just features. Wieland et al [3] on the other hand have done research that shows how close a "bag of local features" model can come to state of the art CNN performance on image data sets. This would imply that CNNs take a featural approach, as the bag of local features does not take pose and location into account. By using a cognitive psychology inspired approach, we can simultaneously test the claims from both Hinton and Wieland. In order to do this, I will measure the performance of networks with differing architectures on some face processing tasks and compare them to human performance on the same tasks. This allows us to use theories and concepts from human face detection research for analysis and interpretation of NNs.

1.5 Human Face Detection

Similar to how we know about the shape bias in humans, we also know about different properties of face detection in infants [23]. Face detection in humans is a big research topic within the field of cognitive psychology. Researchers are trying to create a complete model of how humans process faces. To that end, one of the main questions that is being researched is whether humans process faces by individual parts (featurally) or as a whole

(holistic). Multiple effects have been found during research that shape the proposed models that are being discussed. Some of the most important effects are "Face Inversion" [35], "Thatcher Effect" [16] and "Perceptual Closure" [22].

1.6 Research Question

The main interest of this paper is exploratory in nature, as I want to understand how feasible the cognitive psychology approach is for neural network explainability research. So the main question is:

- Can cognitive psychology help us understand neural networks?

To that end, I will apply theories from cognitive psychology to the domain of face detection, therefore fitting sub questions are:

- How close is NN performance on face detection tasks to human performance on the same tasks?
- Do NNs process faces featurally as implied by Wieland?
- Do capsule networks process faces differently from regular CNNs?

Chapter 2

Preliminaries

2.1 Artificial Neural Networks

The following is a high level description of artificial neural networks. Further information can be found in the book "Deep Learning with Python" by Chollet [4]. Artificial neural networks (ANNs) are a family of brain inspired computing systems. The brain is an accumulation of interconnected neuronal cells, also called neurons. Neurons can receive an electric potential from other neurons, and also transmit a electric potential, when the received potential(s) is big enough. This principle is imitated by ANNs. The artificial neurons get input in the form of a number, for example the brightness value of a pixel in an image, perform some calculation and pass the result to one or more artificial neurons in the next layer. When an artificial neuron receives more than one input, it performs a computation that reduces it to one output. This calculation can take on different forms but most importantly it has to weigh the importance of the different incoming information against each other. This weight is what is learned in an ANN.

2.2 The Architecture Zoo

The simplicity of the underlying building blocks of neural networks make it possible to easily connect them in new ways. This in turn allows for a multitude of different architectures that all vary in what they can learn and which tasks they will excel at.

2.2.1 Densely connected

Densely connected neural networks (DenseNets) are a rather simple family of architectures. The structure of the input is not important for DenseNets. Thus a black and white image could be represented as an array of pixel values. These values would be separately put into the input layer neurons. The

input layer of neurons is followed by the first layer of hidden units. That is just another layer of neurons. The name suggests that these neurons are neither in the input layer nor in the output layer. Every input layer neuron is connected to every neuron in the first layer of hidden units. They then apply a calculation on the entire input array with an individual weight for every value. These values can be set randomly or by hand before training. Since every neuron connects to all following neurons, and every input value has it's own weight, this class of neural networks is called densely connected. Once the output layer neurons are reached a number of things can happen. In a multiclass classification setup such as in the experiments done for this paper, we will have as many neurons in the output layer as we have classes. The value they output has to sum to one, such that their outputs are a probability distribution over the sets of classes. During training we calculate the difference between the real label for the input and the given probability distribution. The size of that discrepancy determines how strongly the weights need to be adjusted in order to reach the correct label. This is repeated many times over with possibly many thousand input label pairs.

2.2.2 Convolutional

A convolutional Neural Network (CNN) works much like a densely connected one. The main differences are that the structure of the input plays an important role and that convolutional layers don't have a simple one-to-all correspondence. A convolutional input layer receives the input as a whole, thus for images, it receives a two dimensional array of pixel values. It then performs a computation known as convolution: a small consecutive part for example 2 by 2 pixels of the input image is taken at a time and some computation is performed on them that reduces the partial image to a single value. We then move that 2 by 2 window on the input (also known as a kernel) by a pixel or more to the right and perform another convolution. The result of that convolution is then written to the right of the first result. This way we basically create a new two dimensional matrix from the input matrix. The only weights learned in this CNN are weights attached to the 4 places in the 2 by 2 kernel. Usually we create multiple matrices from the input layer. In the literature these are often referred to as feature maps. Every feature map can then again be convolved. In the last stage there usually is a densely connected neural network which takes the output of the last layer of convolution operations as input. The densely network works just like described above.

2.2.3 Capsule

The capsule neural network by Geoffrey Hinton is yet another architecture for image classification tasks. This network is inspired by an observation

from psychology: It seems that humans have internal prototypes of objects, containing reference frames. This means that when we see a car which is turned upside down, then we only know that this car is upside down because of the transformation needed to get from the standard frame to the current frame. Thus the idea is to change the architecture of neural networks such that the network incorporates this information in it's decisions. Therefore the first few layers are convolutional layers. After that there are first a few convolutional capsule layers (ConvCaps) and at the end one final ConvCaps layer. The network begins exactly like a convolutional neural network creating feature maps. Then these feature maps are fed into the first capsule layer, say PrimaryCaps1. That layer takes in all feature maps and performs convolutions such that the output is an array of 8 dimensional vectors. Each vector in this case represents a node in the network. Each of these vectors is a "capsule". If the next layer is also a capsule layer, a special routing algorithm is calculated that decides which capsule contributes information to which capsule in the next layer. When the final capsule layer is reached, the input capsules will be transformed into as many capsules as there are classes in the classification task with 16 dimensions. The output of the network is then generated by calculating the length each output capsule just like we calculate vector-length, which reduces it to one dimension. In Hinton's implementation, the length of these vectors represents the probability of the object being present. The main idea of this capsule approach is that we want the internal representations of objects to contain information about the pose (rotation, shear, distortions), instead of only containing a probability of the object existing. Figure 2.1 illustrates how a simple convolutional network processes objects.

2.3 Neural Network Evaluation

Neural networks for classification tasks are often evaluated in four metrics: Accuracy, Positive Predictive Value, True Positive Rate and F1 Score. Accuracy is the true positives plus divided by the sample size for the class in question. Positive Predictive Value is the number of true positives divided by the number of true positives plus false positives. This value is also called precision. True Positive Rate is the number of true positives divided by true positives plus false negatives. This value is also known as Recall. The F1 Score is calculated like this: 2*((precision*recall)/(precision+recall)). This score goes to 0 when either precision or recall go to zero but is 1 when both precision and recall are 1. Thus the F1 score can enable us to spot a badly trained network directly, while precision and recall enable us to see the weaknesses and the strengths of a network. We usually calculate these values per class and compute a weighted average corresponding to the ratio of the class to the entire sample size.

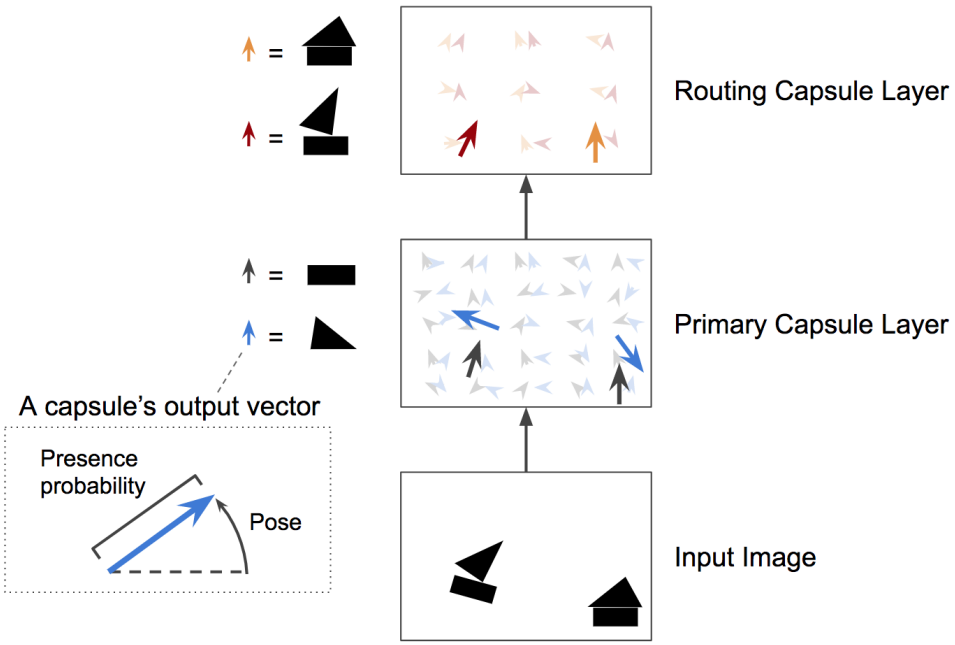

Figure 2.1: A two-layer CapsNet. In this example, the primary capsule layer has two maps of 5 by 5 capsules, while the second capsule layer has two maps of 3 by 3 capsules. Each capsule outputs a vector. Each arrow represents the output of a different capsule. Blue arrows represent the output of a capsule that tries to detect triangles, black arrows represent the output of a capsule that tries to detect rectangles, and so on. Caption and image by Aurélien Géron[2]

2.4 Explaining Neural Network Behaviour

Arrieta et al. [1] have defined 9 goals of explaining machine learning models: Trustworthiness, Causality, Transferability, Informativeness, Confidence, Fairness, Accessibility, Interactivity and Privacy Awareness. A short definition of these terms can be found in appendix A in table A.1. Our research will directly contribute towards Causality, Transferability, Informativeness and Confidence. Also this indirectly heightens our ability to enable Trustworthiness, Fairness, Accessibility, Interactivity and Privacy Awareness. This research will help domain experts to judge and interpret models more easily which in turn allows them to enhance the aforementioned goals.

Arrieta et al. also distinguish between three kinds of machine learning transparency: Simulatability, Decomposability and Algorithmic Transparency. Simulatability refers to the ability of humans to run the machine learning process in their head, or strictly think about it. Decomposability describes the property of a machine learning model to be explained in parts. Algorithmic transparency means that a machine learning model can be fully explored by mathematical analysis and methods. Neural network models as described in this paper do not fulfil any of the above criteria for transparency. The simulatability is not fulfilled since state of the art neural network models all contain thousands of parameters, which makes simulatability impossible due to complexity. Decomposability is not given since neural network models only works as a whole, and therefore can't be analysed in parts. Algorithmic transparency is likewise not given, since we lack a mathematical understanding of many parts making up neural networks as noted by Sun [33].

Thus neural networks are opaque systems and the only way we can explain their behaviour is through post hoc analysis, which Arrieta et al distinguishes into six categories: Text explanations, Visual explanation, Local explanations, Explanations by example, Explanations By Simplification and Feature Relevance Explanation. See Appendix A table A.2 for definitions. This research falls under the categories of explanation by example, local explanation and feature relevance, as we reproduce experiments with prototypical local input, to infer the internal processes of a neural network by analysing feature relevance.

2.5 Cognitive Psychology

Cognitive Psychology is explaining human behaviour by crafting theories explaining capacities, which are bounded by effects. A capacity is thought

to be an ability that a human possesses. It does not need to be discovered but can be observed easily, for example the ability to speak or walk. An effect on the other hand is some specific measured fact that is exhibited by humans under certain conditions which have to be discovered. An example of this is the McGurk effect reported by McGurk et al. [21] which shows that when the syllable "ba" was played along a video of someone saying "ga", a subject would report that he heard "da". Thus any model that aims to explain the language capacity of humans must account for the McGurk effect. A model that explains the capacities while incorporating the effects can then be tested by comparing it to the data gathered in experiments.
This process is comparable to the post hoc explanation approaches of the machine learning community, with the human as the black box. This similarity is what prompted the research question for this paper. That is in how far can we use the cognitive psychology approach to understand neural networks.

2.6 Face detection effects

Research by Yin [35] has revealed that humans can recognise upright faces better than other upright objects after memorising upright examples. However, that effect disappeared once the subjects had to recognise the stimuli upside down after memorising them upright. This effect was found in adults as well as in infants, since a study on 4-month olds [34] revealed that the time it took them to recognise a face was significantly longer when the face was inverted, with respect to the version they memorised. This is called the face inversion effect. In a similar vain face preference research has shown that children preferred top heavy non face like stimuli to bottom heavy stimuli [31]. This means that even without facial features present, the preference for upright configurations was present. Similar research was done with scrambled pictures of normal faces [30]. The research in this field is not conclusive towards the preference of one of these over the other, as Otsuka [23] notes. It seems that there is a preference for top heaviness but it is unclear whether there is a preference for normal faces over scrambled faces. On the other hand, Sekuler et al. [29] have shown that adults do in fact recognise scrambled faces.
Yet another important study on human adults [22] by Craig M. Mooney investigated the closure effect for faces that only had global information, with all local information removed. The study found that the participants were able to recognise faces in the Mooney pictures, but not when the images had been inverted.
There exist different theories about the reason why the effects above exist, but the most popular of them is that faces are processed within different mechanisms that have to evolve over time. Infants may rely more heavily on

holistic mechanisms, while adults may use both holistic and featural mechanisms. However, it may also be the case that these findings are not unique to faces, but objects in general. Thus it could be that faces are inherently better suited for holistic processing than a car or an airplane.

Chapter 3

Research

3.1 Methods

3.1.1 Architectures

Four different neural network architectures have been chosen to be investigated: NatCapsNet [17], IBMCapsNet [8], Dense Baseline (DenseBase) and Convolutional Baseline (ConvBase). NatCapsNet and IBMCapsNet are capsule networks based on Geoffrey Hintons idea of capsules in convolutional networks. NatCapsNet is based on the network described in the paper "Dynamic Routing between capsules" [27] while IBMCapsNet is based on the network described in "Matrix capsules with EM routing" [9]. DenseBase and ConvBase are both baselines of standard implementations of respectively densely connected and convolutional neural networks. The reason these architectures were chosen is because they either relate to the topic of human like perception (NatCapsNet, IBMCapsNet) or are often used in many different tasks and are therefore relevant architectures (DenseBase, ConvBase). It should be noted that the architectures produce a probability distribution over three classes, which is a closed set calculation. Humans however are performing an open set calculation by having the option of not knowing an answer. However since the human experiments used as reference for this paper were forced choice tasks, this should not be an issue.

3.1.2 Training Data

The dataset used for training consists for one third of images of faces, another third of cars and the last third of cats and dogs as a stand in for a "other" class. The images of faces come mainly from the CelebA dataset [18] from Liu et al. The car data comes from the MIO-TCD Dataset by Luo et al. [19] The cat and dog images come from the kaggle Cats vs Dogs dataset [11]. The dataset contains 75000 28x28x1 images. The values are integers between 0 and 255. Since the images come from wildly different sources it

is almost certain that some images have been preprocessed in various ways. However since they come from random sources there is no expectation of a bias towards one or the other class in that regard. After collecting the images, they were resized using the scikit-image python library and converted to greyscale using gleam [12]. The images of faces are at most rotated by 90 degrees while most are upright. This seems to be in line with the faces that humans are seeing [10] during infancy. The car images are upright as well. The random object images are upright images of natural scenes including animals. The evaluation set consists of 15000 pictures with the same characteristics and sources as the dataset.

3.1.3 Training

The task of the networks was to distinguish between three classes: face, car, random scene. The networks were trained and tested for the following metrics: Accuracy, Positive Predictive Value, True Positive Rate and F1 Score with weighted averages. The results can be found in table 3.1.

Table 3.1: metrics for the networks trained on three classes

	Acc	PPV	TPR	F1
NatCapsNet.	0.98	0.98	0.98	0.98
IBMCapsNet	0.97	0.97	0.97	0.97
ConvBase	0.98	0.98	0.98	0.98
DenseBase	0.87	0.87	0.87	0.87

3.1.4 Clarification of the terminology

The facial information is described with certain words, for which the meaning is not always clear. For a review of these terms in the literature, please refer to Piepers et al. [24] For this paper, I will use the term local information for relatively small features such as the eyes in a face or a door handle on a car. First order relational information is qualitative information such as the information that the eyes are above the mouth. Second order relational information is quantitative information such as the exact distance between the eyes or the head lights on a car. Configurational information is used to describe all information that emerges as a result of feature placement in an object, thus first and second order relational information.

3.1.5 Experiments

The experiments described in this section were chosen to investigate the importance of local vs first order vs second order relational information. I created a data set respectively for the 4 experiments, which will be classified

by the models. Afterwards I compared the accuracy during experiments with the values from table 3.1.

Face Inversion Effect

To investigate the face inversion effect (FIE) the experiments of one of the classic FIE papers were used as inspiration. The paper in question is "Looking at upside down faces" [35] by Robert K. Yin. In the paper Yin examines the ability of subjects to recognise images they have previously seen during the training phase. The photos used belong to four classes: faces, houses, airplanes and "Men in action". The latter category contains stickfigure drawings of certain activities. Experiment 1 investigates upright recall after upright training and inverted recall after inverted training. Experiment 2 investigates inverted recall after upright training and upright recall after inverted training. Experiment 3 investigates whether an artists drawings of faces without any shadows still shows the same characteristics as photographs of faces during previous experiments.
Yins experiments can be seen as classification tasks where the important variable change is the inversion of the classified images, which led to remarkable intercategory change but only a small cross category change. I created a similar classification task where the same variable is changed and compare the results. Thus the inversion effect is tested by measuring the accuracy for the 5000 images of faces and cars respectively and then repeating the experiment with the inverted versions. The results were then compared to those of Yin. The images stem from our evaluation set. If there is a face inversion effect, faces should be recognised better than other objects when upright and they should be correctly recognised just as often as objects when upside down.

Mooney Face Test

The Mooney face test [22] developed by Craig M. Mooney is a well established experiment which reveals the human ability for perceptual closure. That is the ability to form a rich mental image from very little visible information. To investigate this ability Mooney blurred grayscale images of faces and then increased the contrast resulting in a black and white image. These images were then displayed to subjects which described what they saw. A newer version of this test is created by Schwiedrzik et al. [28] in which subjects were shown newly created Mooney like images and also similar images that didn't contain faces. Subjects were then asked, whether or not they detected a face in the image. This experiment can be used directly as a classification task for neural networks, making use of the data set provided by Schwiedrzik et al. I will measure the accuracy of the face classification on their dataset.

Figure 3.1: From left to right: Normal, Inverted, Mooney Face

Scrambled Face

In order to understand the importance of low level vs relational information, psychologists have often used experiments involving scrambled faces [23]. Low level or local information would be the shape or orientation of the eye or that of the mouth, while relational information of the first order includes attributes visible in all normal human faces, such as containing two eyes, or that the mouth is under the nose. Relational information of the second order are specific measurements between the eyes for example. A local disruption can show the importance of single facial features, while relational disruption, such as moving the pair of eyes shows the significance of relational information. In order to review the ratio of importance of local information vs relational information, images of faces were cut into pieces and rearranged. This could be done automatically, so all 5000 face images of the evaluation set were prepared this way.
To review the influence of first order relational information, 10 face images were prepared such that the eyes were switched with the mouth. The distance between the eyes was kept the same, in order to preserve the second order relational information as much as possible. Thus for this experiment 10 images were prepared with first order relational perturbations and 5000 with first and second order relational perturbations. The results were then compared to the paper "Face processing stages: Impact of difficulty and the separation of effects" [14] by Latinus et al. The images for this stem from our evaluation set.

Negative Face

Photographic negative images are an easy way to detect the importance of local information vs relational information. In a negative, the pixel values are altered while all relational information and some of the local information is kept intact. The results of this experiment will be compared with the experiment of Galper et al. [6]

Figure 3.2: From left to right: First Order Perturbed, Scrambled, Photographic Negative

3.2 Results

3.2.1 Face Inversion Effect

Table 3.2: Accuracy for face and car classification with upright and inverted stimulus

	Upright	Inverted
NatCapsNet	Face: 0.99 Object: 0.97	Face: 0.31 Object: 0.53
IBMCapsNet	Face: 0.97 Object: 0.98	Face: 0.29 Object: 0.77
ConvBase	Face: 0.99 Object: 0.97	Face: 0.24 Object: 0.63
DenseBase	Face: 0.90 Object: 0.86	Face: 0.19 Object: 0.12
Random	Face: 0.33 Object: 0.33	Face: 0.33 Object: 0.33

3.2.2 Mooney Face Test

Table 3.3: Accuracy for Mooney face detection on the three class trained networks

	Normal	Mooney
NatCapsNet	0.99	0.42
IBMCapsNet	0.97	0.15
ConvBase	0.99	0.03
DenseBase	0.90	0.39
Random	0.33	0.33

3.2.3 Scrambled Face

Table 3.4: Accuracy for normal, first order perturbed and scrambled faces

	Normal	FO	S
NatCapsNet	0.99	0.80	0.20
IBMCapsNet	0.97	0.40	0.15
ConvBase	0.99	0.60	0.13
DenseBase	0.90	0.40	0.18
Random	0.33	0.33	0.33

3.2.4 Negative Face

Table 3.5: Accuracy for normal and photographic negative face images

	Normal	Negative
NatCapsNet	0.99	0.01
IBMCapsNet	0.97	0.05
ConvBase	0.99	0.04
DenseBase	0.90	0.00
Random	0.33	0.33

3.3 Discussion

3.3.1 Face Inversion Effect

A human decision task done by Yin [35] showed a stronger ability to recognise faces vs. a weaker ability to recognise objects from several categories (outcomes averaged), when the stimulus was upright. When the stimulus was inverted however, the classification ability of both objects and faces was worse, but at the same level, see table 3.6. In our experiments all convolutional and capsule networks have similar accuracy for faces (Mean=0.96, SD=0.04), objects (Mean=0.95, SD=0.06) and inverted faces (Mean=0.26, SD=0.05), but not for inverted objects (Mean=0.56, SD=0.19). The mean accuracy for inverted objects is actually higher than that of inverted faces for the convolutional networks. The dense network does have a higher accuracy for faces than for objects, but the accuracy for the inverted stimuli are not close to each other. Therefore none of the investigated neural networks display a face inversion effect as found in humans. However, the convolutional group has an inverted effect, since faces suffer more from inversion

than cars do. The large standard deviation suggests that the difference between different architectures is enormous. It exceeds the standard deviation found in human groups by more than tenfold.

Table 3.6: Average accuracy for face and object classification with upright and inverted stimulus

	Upright	Inverted
Human	Face: 0.98 Object: 0.93	Face: 0.92 Object: 0.92
Random	Face: 0.50 Object: 0.50	Face: 0.50 Object: 0.50

3.3.2 Mooney Face Test

The Mooney face test leaves first and second order relational information in faces untouched. But the local information is deleted. This reveals the importance of this kind of information for humans, since human accuracy is 93 percent. For convolutional neural networks it seems that local information is paramount, with two of them performing worse than the random classifier (Mean=20, SD=20). The dense neural network seems to be better at detecting global information, as it outperforms the random classifier by around 6 percent. It is interesting that the best performing neural network however is a convolutional neural network, since my intuition would have suggested that convolutional neural networks with kernel sizes of 3 or 2, are mostly good at capturing and combining small features instead of recognising global features.

Table 3.7: Human accuracy for Mooney face detection

	Accuracy
Human	0.93
Random	0.50

3.3.3 Scrambled Face

Neural networks performed poorly on first order perturbed images (Mean=0.55, SD=0.19) and worse on scrambled faces (Mean=0.16, SD=0.03). The best network in both categories (IBMCapsNet) however performed with 80 percent accuracy on first order perturbed images, and below the random classifier on scrambled images. However it should be noted that the sample size for the first order perturbed images is rather small. The human performance for scrambled images is more than three times higher than the best

performing network. The human results for scrambled images can be found in table 3.8.

Table 3.8: Human accuracy for scrambled face detection

	Normal	Scrambled
Human	0.99	0.96
Random	0.50	0.50

3.3.4 Negative Face

All networks performed worse than the random classifier, with all networks having an accuracy under 6 percent. These results are somewhat surprising as with this change the first and second order relational information is preserved as well as most of the local information. The only thing that isn't preserved is the intensity of the pixels. The relative intensity of the pixels hasn't changed however. This result is telling us that the investigated networks didn't learn an abstract pattern, but rather a pattern in a concrete intensity, or in colour images, a certain colour.

Table 3.9: Average accuracy for normal and photographic negative face images

	Normal	Negative
Human	0.98	0.73
Random	0.50	0.50

3.3.5 How close is NN performance on face detection tasks to human performance on the same tasks?

The experiments above have demonstrated that neural networks as a whole do not show any of the specific effects found in humans. However they do follow general trends of the face detection capabilities, e.g. they perform worse when humans perform worse and they perform well when humans perform well. On the other hand, they are far less robust than human object recognition, as their accuracy decreases substantially for inverted faces, where human accuracy only decreases by a few percentage points.

3.3.6 Do NNs process faces featurally as implied by Wieland?

The data has shown that both featural and relational information is important for the neural networks. This implies that neural networks are not just bag-of-features algorithms, but somewhere in between fully featural and

fully holistic processing. Thus this research question can be answered with a clear no.

3.3.7 Do capsule networks process faces differently from regular CNNs?

There is no clear trend that capsule networks behave fundamentally different from convolutional networks. The accuracies in all experiments are very much in line with the convolutional baseline network. We can only see a clear difference in accuracies between the dense network and the rest. Therefore we know that capsule based networks in their current state are not necessarily better at representing human object recognition.

3.3.8 Face Processing in Neural Networks

With the experiments above I have been able to show that neural networks process faces neither purely holistically or purely featurally. In all cases both features and composition seemed to play a role in deciding whether the picture contained a face or not, since disruptions of both local and configurational information had an impact on the networks accuracy.
It is also the case that some architectures display more or less configurational processing, as for example NatCapsNet seemed to be less affected by the first order perturbed images in the scrambled faces Experiment. This suggests that the local features are weighed as more important than the relational information. On the other end of the spectrum is the dense neural network that was able to recognise some faces in Mooney face images, which can only draw on the strength of the configurational processing. Thus it seems that the architecture of neural networks dictates how local information is weighed compared to relational information. And it seems that all neural network architectures lie on a spectrum between purely relational and purely featural. Being on the featurally processing end of the spectrum implies that classifier in question acts more like a bag-of-features model as described by Wieland [3]. Using this knowledge, we might be able to create adversarial attacks more easily by exploiting the fact that some networks rely mostly on featural information.

3.3.9 Face Processing in Humans

It is still an open question whether the face inversion effect reveals image processing mechanisms in humans specific to face recognition or whether the effect for faces is circumstantial. Circumstantial means that we are just better at recognising faces compared to objects, because we spend much more time with faces, compared to other objects. Since we saw that faces suffered more from an inversion than objects, it seems like the effect might be inherent to the face stimuli. This means that this research supports the

expert theory [29] [20], which stipulates circumstantial reasons for the face inversion effect. However, it also does not disprove the inherent mechanism theory [15] which stipulates that humans have an inherent mechanism that processes faces, which is separate from the mechanisms for recognising objects. This is because we saw that the trained models lack the effects that humans display, which implies that they are not a representation of human face processing.

Chapter 4

Related Work

The related work is research that tries to explain ANN behaviour post hoc.

4.1 Cognitive Psychology

Ritter et al. [26] have used the knowledge about the shape bias in children to study CNN behaviour. They found that while showing similar accuracy, the shape bias was very diverse among differently seeded networks and between different architectures. This was the first paper that specifically tackled the neural network explainability problem from the perspective of cognitive psychology. Afterwards Feinman [5] et. al have done research on the results of Ritter et al. and tried to relate the results in neural networks to the shape bias data from experiments in children. In my research the focus is however on the domain of face detection, for which it seems that this paper is the first to use the cognitive psychology approach. Furthermore I am using multiple different network architectures, while Ritter et al. is researching one architecture with different seeding.

4.2 Decision Tree Simplification

Zhang et al [36] have trained decision trees on the contribution of feature maps in convolutional neural networks. In their research they used NN models that have disentangled filters (feature maps that encode features that no other feature maps encode). Given that disentangled feature maps are usually representing one "concept", they were able to label them accordingly, which means that the decision tree learned on these feature map activations is able to "explain" a classification decision by showing the most important contributing part in the image and also print multiple important concepts. For example, for the image of a bird we might get a list of answers that reads: white head, black wing, yellow feet. A drawback of this approach is that it only works for a specific type of neural network namely convolutional

neural networks with disentangled feature maps. The research in my paper however can be applied to all kinds of networks, if there is a corresponding part of research in cognitive science to draw upon.

4.3 Salience Maps

Salience map producing methods such as LIME [25] and FullGrad [32] are widely used as explanations for CNN decisions. In all cases the salience map is a perturbed version of the input image, with a colour coded highlighting of the important pixels vs the less important pixels. This approach is easy to use and does allow for quick verification that a network didn't learn an unwanted bias in the training data. However, it falls short of actually explaining neural network behaviour, as it only shows the explanation for one example at a time. It also needs to be inspected manually in every case. And furthermore this kind of technique is most useful for image processing CNNs, but has as of late also been applied to text processing networks.

Chapter 5

Conclusions

I was able to answer all research questions with this cognitive psychology inspired approach to understanding neural networks:

- Neural networks do not show the same effects as humans on comparable tasks, but they follow the same trends
- Neural networks use both holistic and featural information in face recognition
- Capsule networks did not process faces different than CNNs

These findings underline the strength of the cognitive psychology approach, which helps to answer the overarching question of whether cognitive psychology can help us understand neural network behaviour. As it turns out, the rich history of psychological theories and experiments lends itself perfectly for neural network research. Especially psychology experiments that only require input from the visual domain, can easily be converted to a machine learning task, since the setup can often remain the same, just like the experiments I performed for this study. However one must not forget that we are far away from understanding the human brain and the behaviour it facilitates. Even in the case of holistic processing as discussed in this paper, the research is still out on the question how configural and featural face processing mechanisms work together in humans and how they develop over time.

It might also be interesting to pair this approach with powerful tools like FullGrad salience maps [32], when we have comparable human eye tracking research at hand. All in all, this direction promises to be a fruitful endeavour, which can enhance model evaluation and our understanding of how neural networks arrive at their classifications.

Bibliography

[1] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, 2019.

[2] Aurélien Géron. Angle estimation as performed by capsule networks. `https://www.oreilly.com/content/introducing-capsule-networks`, 2018. [oreilly; accessed April 04, 2020].

[3] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet, 2019.

[4] Francois Chollet. *Deep Learning with Python*. Manning Publications Co., USA, 1st edition, 2017.

[5] Reuben Feinman and Brenden M. Lake. Learning inductive biases with simple neural networks, 2018.

[6] Ruth Ellen Galper. Recognition of faces in photographic negative. *Psychonomic Science*, 19(4):207–208, Oct 1970.

[7] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.

[8] Ashley Daniel Gritzman. Avoiding implementation pitfalls of "matrix capsules with EM routing" by hinton et al. *CoRR*, abs/1907.00652, 2019.

[9] Geoffrey E Hinton, Sara Sabour, and Nicholas Frosst. Matrix capsules with EM routing. In *International Conference on Learning Representations*, 2018.

[10] Swapnaa Jayaraman, Caitlin M. Fausey, and Linda B. Smith. The faces in infant-perspective scenes change over the first year of life. *PLOS ONE*, 10(5):1–12, 05 2015.

[11] Kaggle. Cats vs dogs dataset. `https://www.kaggle.com/c/dogs-vs-cats/data`, 2014.

[12] Christopher Kanan and Garrison W. Cottrell. Color-to-grayscale: Does the method matter in image recognition? *PLOS ONE*, 7(1):1–7, 01 2012.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. Imagenet classification with deep convolutional neural networks. *Neural Information Processing Systems*, 25, 01 2012.

[14] Marianne Latinus and Margot J. Taylor. Face processing stages: Impact of difficulty and the separation of effects. *Brain Research*, 1123(1):179 – 187, 2006.

[15] Helmut Leder, Juergen Goller, Michael Forster, Lena Schlageter, and Matthew A. Paul. Face inversion increases attractiveness. *Acta Psychologica*, 178:25 – 31, 2017.

[16] Michael B Lewis and Robert A Johnston. The thatcher illusion as a test of configural disruption. *Perception*, 26(2):225–227, 1997. PMID: 9274755.

[17] Huadong Liao. Capsnet tensorflow. `https://github.com/naturomics/CapsNet-Tensorflow`, 2018.

[18] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[19] Z. Luo, F. Branchaud-Charron, C. Lemaire, J. Konrad, S. Li, A. Mishra, A. Achkar, J. Eichel, and P. Jodoin. Mio-tcd: A new benchmark dataset for vehicle classification and localization. *IEEE Transactions on Image Processing*, 27(10):5129–5141, Oct 2018.

[20] Daphne Maurer, Richard Le Grand, and Catherine J. Mondloch. The many faces of configural processing. *Trends in Cognitive Sciences*, 6(6):255 – 260, 2002.

[21] MACDONALD J. MCGURK, H. Hearing lips and seeing voices. *Nature*, 264:746–748, 1976.

[22] CM Mooney. Age in the development of closure ability in children. *Canadian Journal of Psychology*, 11:219–226, 12 1957.

[23] Yumiko Otsuka. Face recognition in infants: A review of behavioral and near-infrared spectroscopic studies. *Japanese Psychological Research*, 56, 01 2014.

[24] Daniel Piepers and Rachel Robbins. A review and clarification of the terms "holistic," "configural," and "relational" in the face perception literature. *Frontiers in Psychology*, 3:559, 2012.

[25] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144, New York, NY, USA, 2016. ACM.

[26] Samuel Ritter, David G. T. Barrett, Adam Santoro, and Matthew M Botvinick. Cognitive psychology for deep neural networks: A shape bias case study. In *ICML*, 2017.

[27] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules, 2017.

[28] Caspar M. Schwiedrzik, Lucia Melloni, and Aaron Schurger. Mooney face stimuli for visual perception research. *PLOS ONE*, 13(7):1–11, 07 2018.

[29] Allison B Sekuler, Carl M Gaspar, Jason M Gold, and Patrick J Bennett. Inversion leads to quantitative, not qualitative, changes in face processing. *Current Biology*, 14(5):391 – 396, 2004.

[30] Francesca Simion and Elisa Di Giorgio. Face perception and processing in early infancy: Inborn predispositions and developmental changes. *Frontiers in Psychology*, 6, 07 2015.

[31] Francesca Simion, Eloisa Valenza, Viola Macchi Cassia, Chiara Turati, and Carlo Umiltà. Newborns' preference for up–down asymmetrical configurations. *Developmental Science*, 5(4):427–434, 2002.

[32] Suraj Srinivas and François Fleuret. Full-jacobian representation of neural networks. *CoRR*, abs/1905.00780, 2019.

[33] Ruoyu Sun. Optimization for deep learning: theory and algorithms, 2019.

[34] Chiara Turati, Sandy Sangrigoli, Josette Ruely, and Scania de Schonen. Evidence of the face inversion effect in 4-month-old infants. *Infancy*, 6(2):275–297, 2004.

[35] Robert K. Yin. Looking at upside-down faces. *Journal of Experimental Psychology*, 81(1):141, 1969.

[36] Quanshi Zhang, Yu Yang, Haotian Ma, and Ying Nian Wu. Interpreting cnns via decision trees, 2018.

Appendix A

Appendix

Table A.1: XAI Goals according to Wieland et al. [1]

Trustworthiness	A trustworthy model behaves as expected under known circumstances.
Causality	The underlying causal relationships in data may not be broken in AI systems.
Transferability	A model should be able to transfer it's knowledge to new problems never seen before.
Informativeness	AI systems should show information about the problem they are solving and about how they derived at their solution.
Confidence	A generalisation of robustness and stability of a model.
Fairness	A measure of how ethical a network is in respect to human morale and expectation.
Accessibility	Ease of use and wide range of potential users.
Interactivity	A systems potential to be tweaked and fine tuned by the end user for a specific task.
Privacy Awareness	Knowledge about and power to change the privacy implications of AI systems for users.

Table A.2: XAI post hoc explanations according to Wieland et al. [1]

Text	Textual explanations, such as rule based explanations and every symbol generating explanation approach.
Visual	Every explanation that uses a visualisation of some sort. Examples are saliency maps and sensitivity analysis.
Local	Explanations that elicit only a part of the solution space.
Example	Explanations that are based on grasping a representative example from the input.
Simplification	All approaches that create a simpler model that is nearly as efficient belong to this category of explanations.
Feature Relevance	Approaches that rank individual features from the input space according to their influence on the output. For image data this usually includes some image segmentation method as well.