ECOLE POLYTECHNIQUE FEDERALE DE LAUSANNE SCHOOL OF LIFE SCIENCES

Master project in Life Sciences and Technology

Neural circuits of visual pattern completion

Carried out in the Kreiman laboratory
at Harvard University
Under the supervision of Gabriel Kreiman, Ph.D.

Done by

Matthias Chinyen Tsai

Under the direction of
Prof. Wulfram Gerstner
In the laboratory of computational neuroscience

EPFL

Lausanne, August 17, 2018

Abstract

Being able to infer knowledge and take decisions in an environment of of which we possess only partial information is a necessity for humans. Luckily it is a necessity that we can meet quite naturally. This represents more of a challenge when speaking of artifically intelligent systems. A practical example of this issue is the recognition of objects in occluded images. This is a task that humans can perform very well, but computer vision systems struggle to perform. Recent neuroscientific evidence has suggested that humans explicitely require recurrent feedback connectivity for this task. This has prompted computer vision scientists to start experimenting with recurrent networks architectures as well as new algorithms to train them. This report presents an approach to train networks such as to improve their robustness to the recognition of occluded object images. This training strategy is then evaluated for different network architectures. The result of the study was that the training algorithm could improve robustness to occluded image recognition at the expense of a small decrease in performance for the performance of unoccluded images. When the advantages of different types of architectures were evaluated, it was found that recurrent connectivity didn't lead to any significant improvements in outcome of the training. In the contrary, it followed the findings from classical object recognition that feedforward neural networks could perform as well their recurrent counterparts.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my advisor Gabriel Kreiman for his enthusiasm and guidance throughout my project. I am very thankful for the great opportunity that he gave me by letting me join his lab at Boston Children's Hospital, Harvard University for a semester. I will miss our weekly meetings and scientific discussions around his office table. I would also like to thank the entire Kreiman group for the great atmoshpere to which they all contributed in the lab and I look forward to meet some of them again in the future.

I am also very grateful to my EPFL supervisor, Prof. Wulfram Gerstner, who accepted to oversee my thesis and let me have the chance to experience first hand the research taking place in Boston. I would also like to thank more broadly the Brain Mind Institute of EPFL for their financial support during my thesis and for the great environment they build to improve the life and research of their students. Finally, I would like to thank my friends and family for their love and support throughout my stay here in Boston.

Table of Contents

Page

List of Tables

Table **Page**

List of Figures

Figure **Page**

Chapter 1

Introduction

The broader goal of this thesis is to investigate how neural systems can stay robust to incomplete information. Humans are naturally gifted at extracting knowledge and taking decisions based on imperfect observations about the state of their environment. The evolutionary advantage of developing this ability is hardly questionable, however it remains to be clearly understood how exactly our brains can achieve this. Besides of being of great interest to improve our understanding of learning and information processing in neurobiology, this question will awaken interest in any engineer attempting to design more robust intelligent systems. In the context of this thesis, the focus will be kept on the second perspective by concentrating on the study of artificial neural networks. The specific problem of interest will be the recognition of objects in occluded images. This is a strategic choice for research on this topic, because classical object recognition is a well researched territory, both in the brain as well as in artificial systems. This is why it will be useful to review the current state of the field and introduce important theoretical notions before expanding on the core of the thesis.

1.1 Visual object recognition in the brain of primates

The visual system belongs to the most frequently researched and best understood parts of the brain. A natural preference for scientists to investigate the mechanisms of vision could be justified by the predominance of visual input among all sensory senses in humans. Another rationale for the appeal of studying the visual system is that vision is one of the easiest senses to manipulate in a controlled experiment. By correlating the neural activity of various regions with the exposition of a subject to different images, it is possible to functionally characterize the different parts of the visual system. A pioneer of this approach was Haldan Hartline. He was the first to associate the concept of receptive field with neurons in the 1930s after studying neural responses in the optic nerve to retinal illumination [12, 13]. After that first milestone, some major successes by Stephen Kuffler [23] as well as Hubel and Wiesel [17] followed in the 1950s and 1960s by characterizing the receptive fields of retinal ganglion cells and of neurons in the primary visual cortex. These results revealed that neurons in lower visual areas were specific to simple image characteristics such as shape, color and contrast. Later studies showed that neurons in higher visual areas exhibited selective responses to more complex patterns such as faces or hands [30]. One of the most extreme examples for this is the well popularized research of Quiroga et al. [32], who discovered neurons in the medial temporal lobe that were selective to images of distinct individuals, e.g. Jennifer Aniston.

Once sufficient neurons of different regions had been characterized, the mapping of their receptive fields uncovered a neural structure that sequentially extracted more and more complex features through a bottom-up hierarchy [21]. This structure, also known as the ventral stream became associated to the task of visual object recognition [8, 36] and was branded the ‘what’ pathway [43]. In this pathway, receptive fields are smallest for neurons in the lowest regions of the hierarchy, such as retinal cells and they increase in size for neurons in higher areas. This is further illustrated by the existence of neurons that are selective to complex patterns, such as faces, no matter where the pattern is located in the field of vision [42]. This property is more commonly referred to as the translational invariance of these neurons. Besides of developing a model for visual object recognition in the brain of primates, the study of receptive fields and of the ventral stream has lead to important insights for the development of computer vision systems. It is a notably famous field for applying biologically inspired architectures to artificial neural networks. Over the years this has considerably contributed to advancements that made computer vision the blooming field that it is today. For more details on this topic, the next section will be devoted to reviewing how the task of object recognition is tackled from the perspective of computer vision.

1.2 Object recognition in computers

Today, object recognition and computer vision are synonymous with artificial neural networks. This all started with the design of a neural network model by Kunihiko Fukushima in the 1980s [6]. This network model nicknamed the “neocognitron” was an attempt by Fukushima to emulate the visual information processing of the ventral stream. For this he sequentially stacked, what are nowadays called convolutional neural network layers (Figure 1.1). This model was especially elegant, because it was efficient, while respecting the biological models of the time. First of all, it had hierarchical feedforward architecture. Secondly, because convolutional layers implement a very local connectivity pattern between layers, it meant that neurons in the upper layers would have receptive fields of increasing sizes compared to neurons in lower layers. In addition, the use of convolutional layers solved the problem of translation invariance, which was not only biologically relevant [42], but also generally desired for an object detection network. Another benefit of the convolutional architecture was that neurons were not connected to all the other neurons from neighboring layers, as is the case in densely connected networks. This downsized connectivity and the weight sharing characteristic of convolutional networks produce a reduction in the size of the weight search space, which simplifies the network optimization procedure. Although training algorithms of the time were incompatible with deep multilayer networks, the neocognitron could achieve good performances for very simple digit recognition tasks [7] by combining an unsupervised learning algorithm for the internal layers [5] and basic perceptron supervised learning for the last layer [34]. Once this type of convolutional architectures was combined with supervised learning techniques based on backpropagation [24], the field of computer vision could start to get serious about object recognition. In the 1990s researchers began to successfully train systems that could recognize handwritten numbers [25] and dataset of images depicting different classes of objects [4].

Taking a step back from this historical perspective, it can be summarized by the advance in object recognition by designing a system that is translation invariant. Here the key features of the system were the convolutional neural network architecture as well as the training method through backpropagation. From this perspective, it makes sense that the next step would be to improve object recognition systems

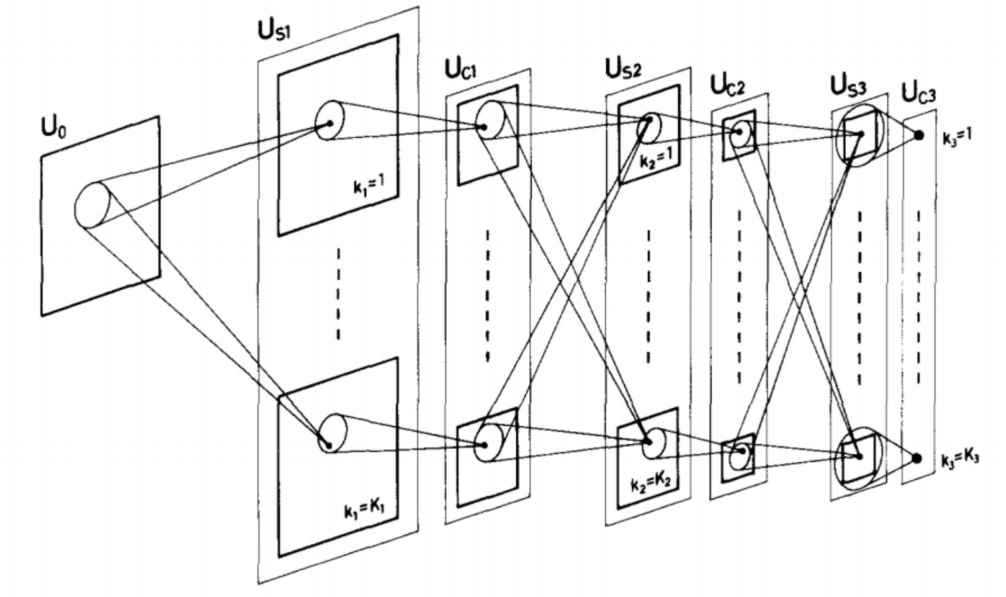


FIGURE 1.1. Schematic diagram illustrating the interconnections between layers of the neocognitron as well as the analogy between convolutional neural network layers and 2D convolutional filters. Figure reproduced from [7].

by making them scale and rotation invariant. Indeed, humans are capable of recognizing objects in images independently of their orientation and size as long as they remain reasonably visible. This is true on a functional level, but was also demonstrated on the neural level through receptive fields studies in the inferior temporal cortex [2, 18, 26]. It therefore made sense to target the design of rotation and scale invariant object recognition systems. More generally, the goal could be defined as improving robustness towards perceptual transformations. With this in mind, Perret and Oram [31] conjectured that a hierarchical pooling mechanism over neurons that were selective to different transformed variants of the same features would generate invariance to this transformation. A simplified example would be to pool over neurons that are selective to different rotated version of a square to get a rotation invariant representation of a square. Following this outline, Riesenhuber and Poggio pioneered a class of network models dubbed HMAX [33], which implements this pooling mechanism with a max pool operation over a window of neighboring neurons at different layers of the network.

It turns out that implementing ingenious network architecture is not the only way of achieving invariance towards a transformation. According to the universal approximation theorem for neural networks [16], standard multilayer feedforward neural networks are universal approximators. This means that given a sufficient number of neurons there exists a set of weights that will approximate any continuous function on a compact subset of R^n. In theory there should therefore be no functional limitation to the power of standard neural network architectures. In practice, the challenge of finding the right set of weights can be challenging. In the past, the main constraint was the computational power necessary to train big networks in order to converge to a good set of weights. With the increase in computation power, the limiting factor shifted towards the quality and quantity of data that was used to train the neural networks. It was during this period that the Imagenet Large Scale Visual Recognition Challenge [3] was launched as an annual competition. This was both in the response for the need of a new generation of datasets to train more complex networks as well as to encourage computer vision experts to beat a continuously growing dataset . The idea was that producing the most challenging

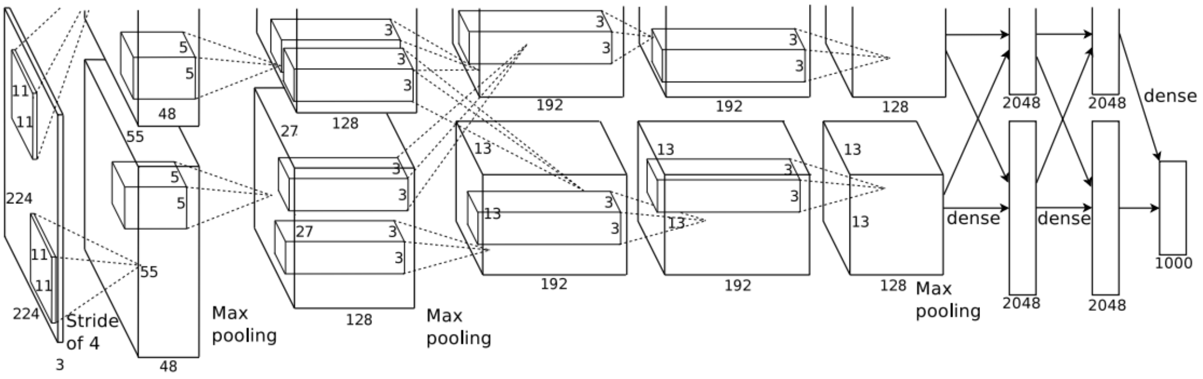


FIGURE 1.2. Architecture of Alexnet consisting eight neural network layers, three max pooling units and 62'378344 neurons. The first five layers are convolutional while the last three are dense neuron layers. Figure reproduced from [7].

image recognition dataset possible would automatically promote the right environment to push the field forward. The 2012 version of the database contained over one million images and 1000 image categories [35]. That year's competition was won by a landslide by AlexNet [22], which is a deep convolutional feedforward neural network with three max pooling units (Figure 1.2). Thanks to its success in the Imagenet challenge, Alexnet became a popular network for research purposes and quickly became a common network to use as a benchmark for studies. Its performance in the Imagenet has long been outdated by larger networks such as VGG16 [37], InceptionNet [38], ResNet [14] or Nasnet [46], but its simplicity and efficiency continue to make it a great network to work with for theoretical purposes.

As the room for improvement in classical object recognition tasks becomes slimmer, the focus of the field of computer vision has shifted towards topics such as object localization or image segmentation. Some of the community has moved to video data, while others have looked at more challenging aspects of object recognition as in occluded images. As will be presented later, it is conceivable that these two subjects might have more in common than meets the eye at first sight. The discussion will however be concentrated on occluded object recognition.

1.3 Occluded object recognition

After the great success of deep convolutional networks with classical object recognition, it seemed that this aspect of vision could be well replicated by feedforward network architectures. When such models were tested for robustness with respect to occlusion, it was found that the image recognition accuracy was very sensitive and quickly degraded [9, 19]. This deficiency can be resolved in several ways. The simplest is with data augmentation and adding occluded versions of the images to the training dataset [28]. Computer vision engineers have also conceived more elaborate modular systems, which explicitly combine different subsystems responsible for tasks such as segmentation, depth detection or other complex image representations [1, 11, 29, 44].

Another approach was to draw ideas from biology. Indeed, results from human studies had indicated that recognition of partially occluded images took more time than for unoccluded images [39]. Simple object recognition tasks elicited a selective electrophysiological response in the ventral stream around 100ms after an image was shown. Since this roughly corresponded to the time necessary for an input from the retina to arrive, it followed that this process should rely on a feedforward information processing cascade. When it was found that the same recognition selective response only arose around 200ms after stimulus, the conclusion was that additional recurrent processing was involved to handle occlusion in the brain. When computer vision scientists tried to train recurrent networks to perform object recognition of occluded images [27], they observed improvements in performance compared to feedforward networks [41]. The specific study by Tang et al. (2018) mentioned here is the main bedrock of this thesis. By playing with the implementation of recurrent models for occluded object recognition, they reached into the territory of training algorithms for recurrent neural networks and how these could improve robustness to incomplete information.

1.4 Neural networks and incomplete information

In their study, Tang et al. [41] demonstrated different ways of training recurrent neural networks starting from trained feedforward models. They could show that replacing the seventh neural network layer of alexnet 1.2 with an all-to-all connected recurrent layer lead to a significant increase in performance with occluded images. The main strength of the system was that the recurrent weights could be set by considering the recurrent layer as a Hopfield network [15] and applying a Hebbian learning using only the unoccluded images as input. This is a great feature from a neuroscientific perspective. Reproducing the one-shot learning capabilities of the brain is still an unsolved problem and any new strategy pushing towards this goal can be very valuable. In this case, the system did need many examples from the same class to learn, but at least the robustness to occlusion was not a result of adding occluded images to the training data. A second central aspect of learning that was touched upon by Tang et al. was the idea of transfer learning. The question was if by training the model with occluded images of certain classes, the robustness of the model towards occlusion could be improved for the recognition of objects of other classes not seen during training. They tested this approach on their dataset of 325 images each belonging to one out of five classes and found that indeed some transfer learning did occur. However due to their small dataset, the results remained fragile for practical purposes. This therefore opened the window to build a new project that aimed at applying the same transfer learning approach to a bigger dataset and investigate, if the transfer learning property would endure.

The roadmap was to start with a feedforward model that had been pretrained for object recognition with unoccluded images and add recurrent connections. Then the recurrent connections would be trained with the occluded images of a set of object classes. The specific method of training would be to extract the neuron activations of the feedforward network. Finally the network performance would be evaluated with occluded images from classes not used during training. The premise was simple and the outcome hopeful. However most of the results could not be completed by the end of the schedule. The only part finished was an experiment that trained different network architectures with an augmented dataset containing occluded and unoccluded versions of the training images. This experiment was meant as a benchmark and used the output of the readout layer as training target for supervised learning instead of the activations of the hidden units. As no transfer learning was expected in this setup the performance was evaluated on occluded images of objects of the same class as the images used to train the networks. Since only the results from this experiment will be presented and explained in the methods and results sections, this report will be restricted to a discussion about how data augmentation can improve robustness of object recognition to occlusion. A few different network architectures have been investigated and their different performances will therefore be compared.

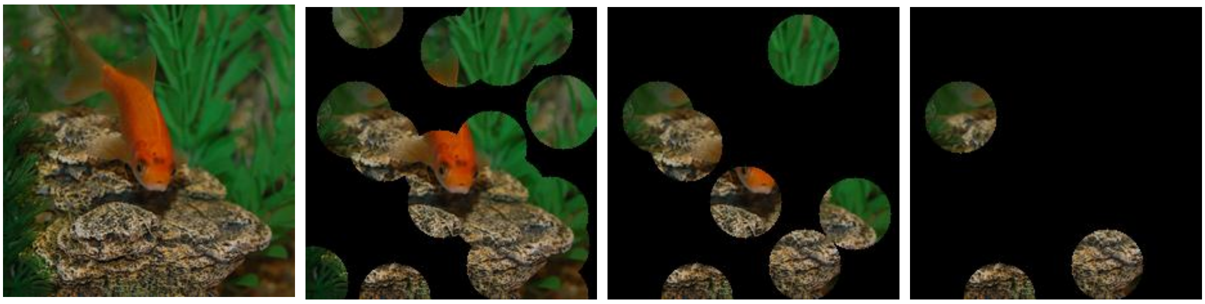
Chapter 2

Material and Methods

2.1 Dataset

The Imagenet dataset was chosen [35], because it had an appropriate size and difficulty, but also because many object recognition models are available that were pretrained with Imagnet. The ILVSRC2013 version of the training dataset was chosen, because it doesn't contain images with objects of different classes in the same image. This is a convenient feature for the purpose of isolating the performance with respect to the occlusion parameter and reducing uncontrolled sources of classification noise. The testing dataset was taken from the ILVSRC2012 version, because it is one of the few ones were the test labels have been released. It contains labeled images belonging to one of 1000 different classes.

All images were processed to size 227x227 and occluded with a bubble style. An example image of 0%, 25%, 50% and 80% occlusion can be seen in Figure 2.1. As can be observed, the occlusion of the object was only measured in terms of the entire image. This lead to some objects being effectively completely occluded, despite their label of partial occlusion. Such a case can be observed in the 80% occlusion example of Figure 2.1



Fully visible 25% occlusion 50% occlusion 80% occlusion

FIGURE 2.1. Example image of the goldfish class from the processed ILVSRC2013 training set. From left to right, there is an example of each: a fully visible, 25% occluded, 50% occluded and 80% occluded image with a bubble style.

2.2 Network architectures

For the neural networks, the choice for basic feedforward architecture was made in favor of Alexnet [22]. Besides of one original network pretrained on Imagenet [10] and used as a control, five other models were created (see Table 2.1). Two variants maintained exactly the same network architecture as the control, but were retrained. Fwdnet8 and Fwdnet78 had identical weights except for the 8th layer or the 7th and 8th layer respectively. These trainable weights were the only ones modified during subsequent training. Similarly, Recnet8 was identical to Alexnet up to the 7th layer and only replaced the 8th feedforward layer with a recurrent all-to-all connected layer with the same number of neurons. Recnet78 was equivalent with the only difference being that both the 7th and 8th layers were replaced by their recurrent counterparts. In the case of the final network Recnet7 only the 7th layer was replaced by a recurrent layer and the 8th layer stayed feedforward, but would also be part of the set of weights to be trained during optimization. A visual overview of these networks is rendered in Table 2.1.

Network	Layers 1-5	Layer 6	Layer 7	Layer 8
Alexnet	5 convolutional layers	Dense Layer of 4096 neurons	Dense Layer of 4096 neurons	Dense Layer of 1000 neurons
Fwdnet8			Dense Layer of 4096 neurons	Dense Layer of 1000 neurons
Fwdnet78			Dense Layer of 4096 neurons	Dense Layer of 1000 neurons
Recnet8			Dense Layer of 4096 neurons	Recurrent Layer of 1000 neurons
Recnet78			Recurrent Layer of 4096 neurons	Recurrent Layer of 1000 neurons
Recnet7			Recurrent Layer of 4096 neurons	Dense Layer of 1000 neurons

TABLE 2.1. Overview of the network designs used layer by layer depending on the network. Layers colored in blue were not trained and maintained the original weights of the pretrained Alexnet model [10]. The weights of the layers colored in yellow were all trainable.

2.3 Training with augmented data

All five variants of Alexnet were trained using the same dataset consisting of occluded and unoccluded images from Imagenet. Not the entire training set was used and only the images belonging to 400 out of the 1000 possible classes were used. This was because these results were originally planed to be benchmarks for the models trained to achieve transfer learning by only using a fraction of the available data.

Individual network training was performed using the Adam stochastic optimization method [20]. Training aimed at minimizing the l2 distance between the activation pattern of the readout layer of the trained network with the activation pattern from the readout layer of the pretrained Alexnet model in response to the same images, occluded or not. The set of weights trainable for each network correspond to the ones marked in yellow in Table 2.1. Recnet8 and Recnet78 were both trained to go through three recurrent loop time steps. Two different variations of Recnet7 were trained, one with 2 time steps and one with 5 time steps of recurrence. They will respectively be referred to as Recnet7.2 and Recnet 7.5 from this point.

2.4 Evaluation of the models

The performance of the models was evaluated on the Imagenet test set, which had never been seen by Alexnet nor any of the five other networks during training. In order to compute a performance comparable with the one of pretrained Alexnet, it was necessary to restrict the model testing to images that belonged to one of the 400 classes used for training. Two performance metrics were computed. The first was the accuracy of prediction, meaning the number of correct classifications over the total number of tested images. The second was the Top-5 Accuracy, which corresponds to the fraction of times that the network had the correct label among its five top picks. These performance measures were computed for a wide range of occlusion of the images. Finally, a last metric important to evaluate the optimization process was the loss at the end of training.

Using these measures, the performances between the networks were evaluated and compared between networks.

Chapter 3

Results

The results will be presented by splitting the experiment into two subjects. The first step will be to focus on the role of the training paradigm with the augmented dataset. This will be done by comparing exclusively the feedforward models that had the exact same architecture, but were trained differently . The second part will be devoted to the analysis of the potential of recurrent layers to improve the performance of this type of object recognition networks.

3.1 Training with an augmented dataset

In order to isolate the training component on the performance, the networks that have the same architecture will be compared. These are the pretrained Alexnet model, the network of which only the last layer was retrained (Fwdnet8) and the network of which both the 7th and the 8th layer were retrained (Fwdnet78). Their respective performances depending on the amount occlusion subjected to the images they had to classify are depicted in Figure 3.1. The first observation that can be made is that both the accuracy and the top-5 performance lead to qualitatively similar plots. The second obvious observation is that all performances decrease with increasing occlusion (respectively decreasing visibility). It can also be noted that the retrained networks perform better than the pretrained Alexnet control on occluded images, but suffer a slight decrease in performance for unoccluded images with 100% visibility. Finally, it is interesting to notice that although their performances is relatively similar, Fwdnet78 is slightly better at recognizing occluded images than Fwdnet8, but the contrary is true for unoccluded images. In all cases the stable performance seems to be in the region of 90% visibility, in which all networks perform very similarly.

3.2 The role of recurrent connections

For an analysis of the role of recurrence, the most basic comparison to make is between networks that have been trained in the same way and are identical, but for the recurrence of their connections. This contrast is visualized in Figure 3.2. In both plots the two networks show a very similar performance curve. However, in both cases the recurrent network has a very slight edge on the feedforward one.

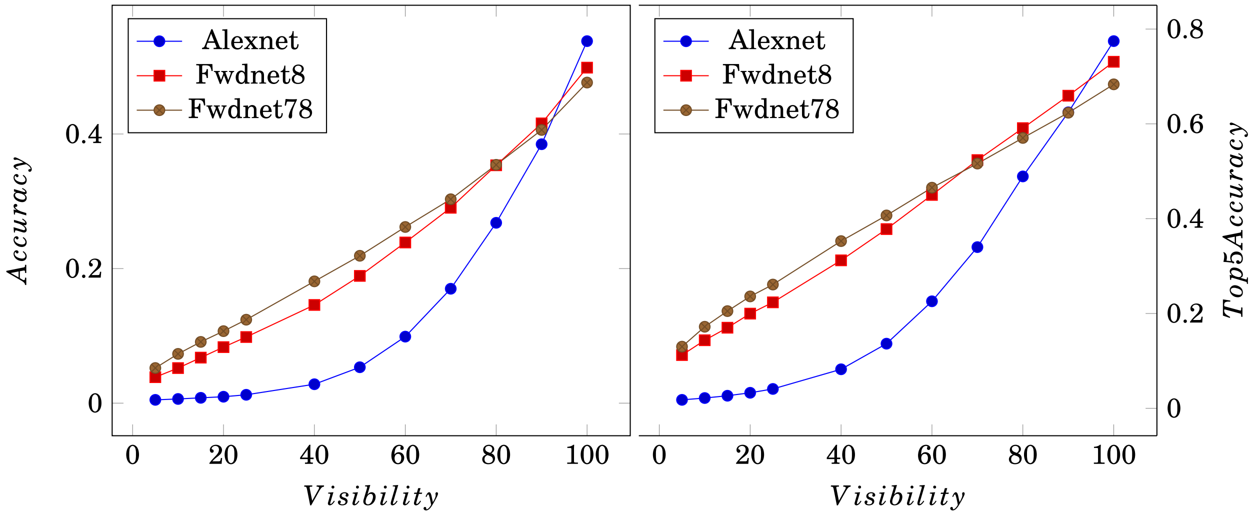


FIGURE 3.1. Performance comparison between the different feedforward networks with same architecture. On the left the Top-1 Accuracy is plotted and on the right the Top-5 Accuracy.

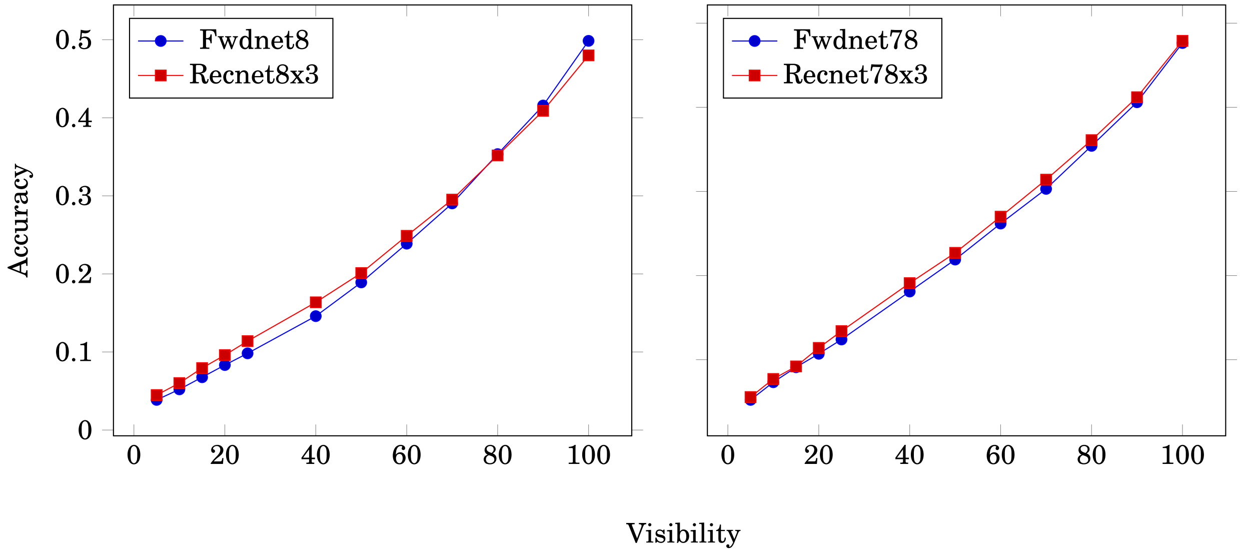


FIGURE 3.2. Comparison of the testing accuracy between feedforward networks and their recurrent counterparts. Left: comparison of Fwdnet8 and Recnet8. Right: comparison of Fwdnet78 and Recnet78 (see Table 2.1 for more details on their respective architectures).

The only exception being for the performances on unoccluded images of Fwdnet8 and Recnet8, where Fwdnet8 is slightly better.

In order to catch the differences between the different recurrent variants of the network, their respective performance curves are depicted in Figure 3.3. There, one can notice that similarly to the retrained feedforward networks (see Figure 3.1), they all became more robust to occlusion but paid a slight price in classification performance of unoccluded objects. The second main result that can be drawn from this figure is that all recurrent models perform very similarly with the exception of Recnet8, which if very slightly worse than the other three. These three others are almost indistinguishable however.

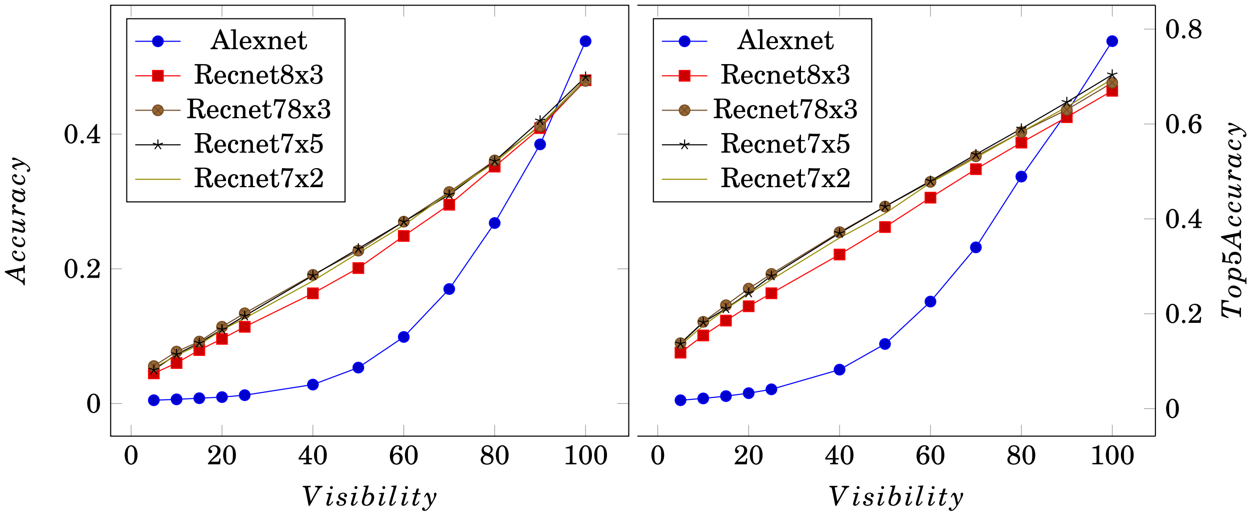


FIGURE 3.3. Comparison in performances of the different recurrent networks and of pretrained Alexnet. On the left the Top-1 Accuracy is plotted and on the right the Top-5 Accuracy.

Lastly, it can again be noted that the accuracy and the Top-5 performance are qualitatively extremely similar.

Finally, as an attempt to better evaluate how training differed, between the models and the role that this might have had on the observed outcome, the training loss was plotted beside of the performance of the networks on images around 40% visibility (see Figure 3.4. This specific amount of occlusion was chosen, because it was in the region in which the networks differed most in performance. The figure shows that all networks couldn't fit the augmented data equally well. It can be observed that the general tendency is for recurrent networks to better fit the data than feedforward ones. Secondly, the plot also shows that networks that had two layers retrained could reach lower training losses than the model in which a single layer was retrained. Now comparing the left red plot with the right blue one, it becomes immediatly obvious that networks that had smaller training loss could reach better performances in that particular visibility region.

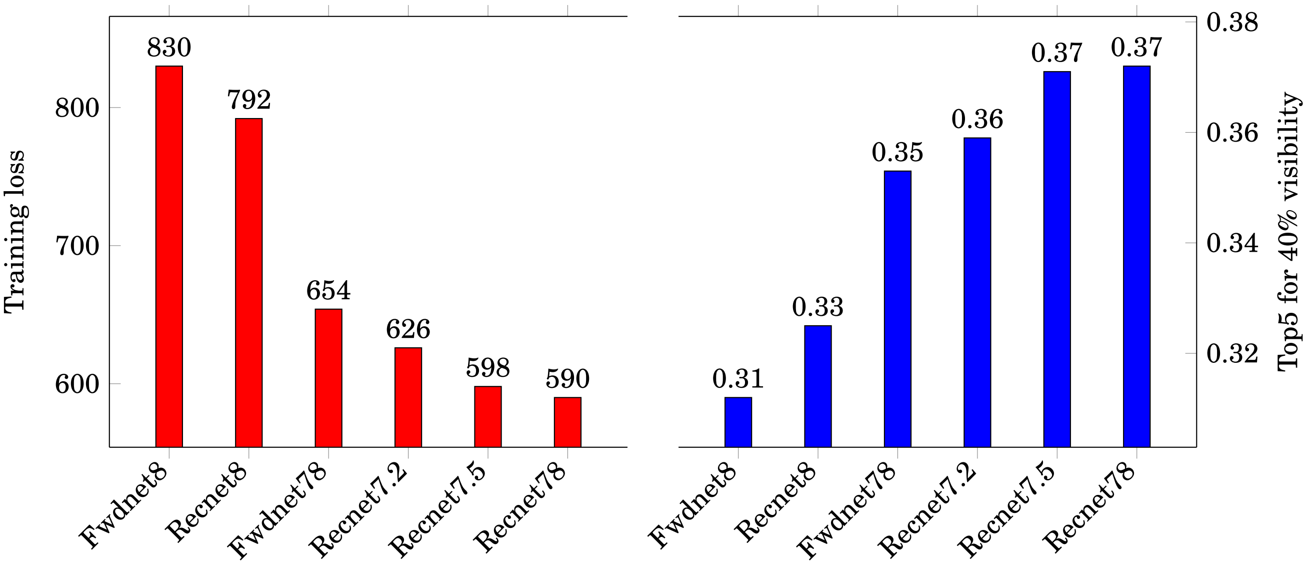


FIGURE 3.4. Comparison of the training loss of the models with their top-5 performances on images that had a visibility of 40%.

Chapter 4

Discussion

The discussion of the results will be done according to the same structure as the presentation of the results. It will therefore start with a discussion of the role of the training on performance, before heading into discussions about recurrence and finally come full circle to combine both of these components in the context of this study. Overall, the training paradigm achieved what would be expected from it. Thanks to the data augmentation the retrained models could try to accomodate knowledge about how to classify occluded images into their weights. This therefore lead to improvements in performance on images with imperfect visibility in comparison to the pretrained Alexnet model that had never been subjected to explicitly occluded images. This holds for the observation in both Figures 3.1 and 3.3. In both cases, the increase in robustness to incomplete images lead to a tradeoff in performance for unoccluded images. This could be interpreted as the networks getting forced by the training to rely less on patterns that are not robust to occlusion and therefore reducing false classifications in occluded images. Doing so would however have lead to the observed decrease in performance on unoccluded objects. The second main observation from Figure 3.1 was that Fwdnet78 was more affected by the training in the sense that its performance on occluded objects increased more and its performance on unoccluded objects decreased less than Fwdnet8's did. This can be explained by the bigger search space available to Fwdnet78 compared to Fwdnet8 during training. Indeed since the entire weight search space of Fwdnet8 was included in Fwdnet78's, it could be expected that if the training performed well, Fwdnet78 should be able to move further away from the weight configuration of the pretrained Alexnet model than Recnet8. This would imply that, if it was useful for the increase of overall performance to decrease slightly the performance on unoccluded objects, this is something Recnet78 would be better capable to do than Recnet8. Since this would make sense, as was described in the argument about giving up features that were not robust to occlusion (here features are meant as neural activity patterns in layer 6 or 7), this explanation appears to hold so far.

Now moving to discuss the behaviour of the recurrent model variants. The fact that they all performed very similarly prevents any strong hypothesis aside of the idea that both the number of recurrent iterative time steps used for training and the location of recurrent layers are not significant factors. This would be true but for the difference between all recurrent networks and Recnet8, which seems to be slightly out of the performance range of the other three recurrent networks. However the advantage of a recurrent layer 7 is highly questionable given strong similarity in performance between Fwdnet78 and Recnet78 (see Figure 3.2). The best explanation therefore seems to be that the single most important advantage for performance enhancement in this setup was the trainability of layer 7. This is not very surprising in itself, since as explained above, more trainable layers imply a bigger search space to converge to the best possible object recognition system. However, recurrent connections in addition of feedforward ones would also imply a bigger search space, yet these seem not to be very relevant. One explanation for this strange observation could be that the recurrent layers offer a too big search space, which allows the models to overfit and therefore lose any advantage that they might have from their bigger computational search spaces. This is where looking back at the training evaluation can be useful and indeed; the observation from Figure 3.4 that smaller training losses still correlated with improvement in testing performance do not permit any conclusion towards an overfitting hyptohesis. Untill new results are produced, it seems that the conclusion should be that recurrent layers didn't contribute a significant advantage to occlusion robustness at least with this training paradigm. This nonetheless takes nothing from the training strategy itself, which as assumed produces improvements in object recognition performance in occluded images.

Chapter 5

Future Work

Due to the small amount of completed experiments included in this report, there is no shortage material that can be mentioned in this section. The first experiments that should be effectuated next are evidently the ones that couldn't completed. This includes the actual transfer learning study on networks such as Alexnet with one or several feedforward layers replaced with recurrent ones. Besides of the variant of using a fraction of the 1000 Imagenet classes for training and the other fraction for evaluating performance, another planed experiment was to train the network with a set of images that had as little similarity as possible to the Imagenet classes. Doing this could partially exclude the component of transfer learning due to the high similarity between some Imagenet classes such as dog breeds. A rich set of images all different to Imagenet is hard to come by, but a possibility was to use the openly available pokemon image dataset [45]. It is a set of 4879 images with 800 different pokemons all depicted in a variety of styles (see Figure 5.1). It is questionable weather such a dataset would sufficiently well sample the image space to adequatly train the recurrent layers without overfitting, but it would be a very satisfying result, if found successful.

Aside of varying the data used, it could have been interesting to investigate in more depth how multiple recurrent layers could best be trained in this framework simultaneously. Assuming the goal would be to add recurrent connections to each layer of the feedforward network it is questionable how well the proposed training strategy would scale. Since it was designed only for a single layer, one could

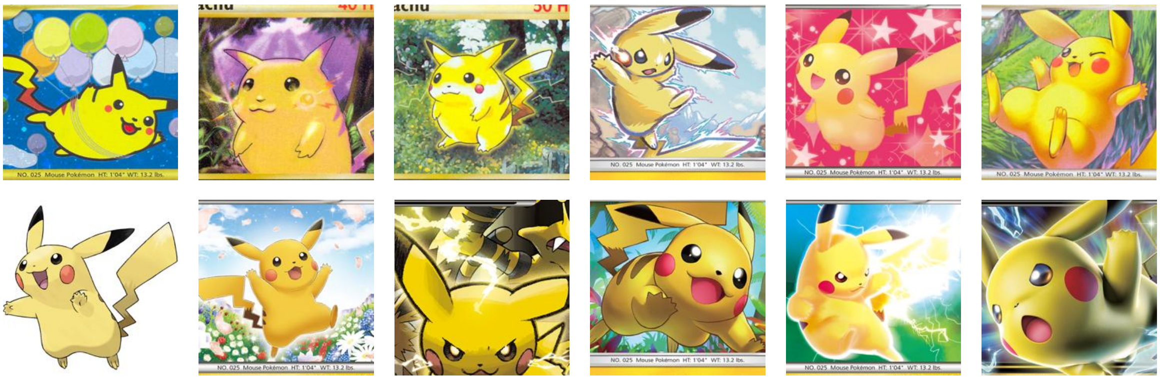


FIGURE 5.1. Examples out of the One-Shot-Pokemon Images dataset [45] illustrating the difference in styles and texture within a single class out of the 819 different classes.

imagine extrapolating the training of multiple consequent recurrent layers by backpropagating the error of the activations to the deepest layer. Another possibility would be to train layers individually with the classical method, but sequentially one after the other. For example by starting from the lower layers close to the input and optimizing the higher layers using the input processed by the optimized recurrent layers below. Finally, it would be useful to control, if the increase in robustness towards occlusion and the transfer learning property were characterstic to recurrent networks. As a control, it would be possible to use the same training methodology, but by either keeping single feedforward layer or by inserting multiple feedforward layers instead. The need for such a control is also enhanced by the preliminary results that were presented here, since they demonstrated that replacing a feedforward layer with a recurrent one didn't invariably lead to a significant improvement of the system. This observation could however be further tested, by taking deeper feedforward networks as basis to add recurrent connections. A second path of study would also be to push the study of this report and produce more variants of Alexnet in lower layers in order to perhapse find effects that couldn't be noticed in the uppermost layers.

Bibliography

[1] A. Aldoma, F. Tombari, L. Di Stefano, and M. Vincze, *A global hypothesis verification framework for 3d object recognition in clutter*, IEEE transactions on pattern analysis and machine intelligence, 38 (2016), pp. 1383–1396.

[2] M. Booth and E. T. Rolls, *View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex.*, Cerebral cortex (New York, NY: 1991), 8 (1998), pp. 510–523.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, *ImageNet: A Large-Scale Hierarchical Image Database*, in CVPR09, 2009.

[4] L. Fei-Fei, R. Fergus, and P. Perona, *Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories*, Computer vision and Image understanding, 106 (2007), pp. 59–70.

[5] K. Fukushima, *Cognitron: A self-organizing multilayered neural network*, Biological cybernetics, 20 (1975), pp. 121–136.

[6] K. Fukushima, *Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, biol cybem. 36 (1980) 193-202*, S. Shiotani et al./Neurocomputing 9 (1995) Ill-130, 130 (1980).

[7] K. Fukushima and S. Miyake, *Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position*, Pattern recognition, 15 (1982), pp. 455–469.

[8] M. A. Goodale and A. D. Milner, *Separate visual pathways for perception and action*, Trends in neurosciences, 15 (1992), pp. 20–25.

[9] K. Grm, V. Štruc, A. Artiges, M. Caron, and H. K. Ekenel, *Strengths and weaknesses of deep learning models for face recognition against image degradations*, IET Biometrics, 7 (2017), pp. 81–89.

[10] M. Guerzhoy and D. Frossard, *Alexnet model pretrained on imagenet*, 2016.
$http://www.cs.toronto.edu/\ guerzhoy/tf_alexnet/$.

[11] Y. Guo, F. Sohel, M. Bennamoun, J. Wan, and M. Lu, *A novel local surface feature for 3d object recognition under clutter and occlusion*, Information Sciences, 293 (2015), pp. 196–213.

[12] H. K. Hartline, *The response of single optic nerve fibers of the vertebrate eye to illumination of the retina*, American Journal of Physiology-Legacy Content, 121 (1938), pp. 400–415.

[13] H. K. Hartline, *The receptive fields of optic nerve fibers*, American Journal of Physiology-Legacy Content, 130 (1940), pp. 690–699.

[14] K. HE, X. ZHANG, S. REN, AND J. SUN, *Deep residual learning for image recognition*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[15] J. J. HOPFIELD, *Neural networks and physical systems with emergent collective computational abilities*, Proceedings of the national academy of sciences, 79 (1982), pp. 2554–2558.

[16] K. HORNIK, *Approximation capabilities of multilayer feedforward networks*, Neural networks, 4 (1991), pp. 251–257.

[17] D. H. HUBEL AND T. N. WIESEL, *Receptive fields of single neurones in the cat's striate cortex*, The Journal of physiology, 148 (1959), pp. 574–591.

[18] M. ITO, H. TAMURA, I. FUJITA, AND K. TANAKA, *Size and position invariance of neuronal responses in monkey inferotemporal cortex*, Journal of neurophysiology, 73 (1995), pp. 218–226.

[19] S. KARAHAN, M. K. YILDIRUM, K. KIRTAC, F. S. RENDE, G. BUTUN, AND H. K. EKENEL, *How image degradations affect deep cnn-based face recognition?*, in Biometrics Special Interest Group (BIOSIG), 2016 International Conference of the, IEEE, 2016, pp. 1–5.

[20] D. P. KINGMA AND J. BA, *Adam: A method for stochastic optimization*, arXiv preprint arXiv:1412.6980, (2014).

[21] E. KOBATAKE AND K. TANAKA, *Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex*, Journal of neurophysiology, 71 (1994), pp. 856–867.

[22] A. KRIZHEVSKY, I. SUTSKEVER, AND G. E. HINTON, *Imagenet classification with deep convolutional neural networks*, in Advances in neural information processing systems, 2012, pp. 1097–1105.

[23] S. W. KUFFLER, *Discharge patterns and functional organization of mammalian retina*, Journal of neurophysiology, 16 (1953), pp. 37–68.

[24] Y. LECUN, B. BOSER, J. S. DENKER, D. HENDERSON, R. E. HOWARD, W. HUBBARD, AND L. D. JACKEL, *Backpropagation applied to handwritten zip code recognition*, Neural computation, 1 (1989), pp. 541–551.

[25] Y. LECUN, L. BOTTOU, Y. BENGIO, AND P. HAFFNER, *Gradient-based learning applied to document recognition*, Proceedings of the IEEE, 86 (1998), pp. 2278–2324.

[26] N. K. LOGOTHETIS, J. PAULS, AND T. POGGIO, *Shape representation in the inferior temporal cortex of monkeys*, Current Biology, 5 (1995), pp. 552–563.

[27] R. C. O'REILLY, D. WYATTE, S. HERD, B. MINGUS, AND D. J. JILK, *Recurrent processing during object recognition*, Frontiers in psychology, 4 (2013), p. 124.

[28] E. OSHEROV AND M. LINDENBAUM, *Increasing cnn robustness to occlusions by reducing filter support*, in The IEEE International Conference on Computer Vision (ICCV), vol. 2, 2017.

[29] B. PEPIKJ, M. STARK, P. GEHLER, AND B. SCHIELE, *Occlusion patterns for object class detection*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3286–3293.

[30] D. PERRETT, *Temporal lobe cells of the monkey with visual responses selective for faces*, Neurosci. Lett. Suppl. 2, 340 (1979).

[31] D. I. PERRETT AND M. W. ORAM, *Neurophysiology of shape processing*, Image and Vision Computing, 11 (1993), pp. 317–333.

[32] R. Q. QUIROGA, L. REDDY, G. KREIMAN, C. KOCH, AND I. FRIED, *Invariant visual representation by single neurons in the human brain*, Nature, 435 (2005), p. 1102.

[33] M. RIESENHUBER AND T. POGGIO, *Hierarchical models of object recognition in cortex*, Nature neuroscience, 2 (1999), p. 1019.

[34] F. ROSENBLATT, *The perceptron: a probabilistic model for information storage and organization in the brain.*, Psychological review, 65 (1958), p. 386.

[35] O. RUSSAKOVSKY, J. DENG, H. SU, J. KRAUSE, S. SATHEESH, S. MA, Z. HUANG, A. KARPATHY, A. KHOSLA, M. BERNSTEIN, A. C. BERG, AND L. FEI-FEI, *ImageNet Large Scale Visual Recognition Challenge*, International Journal of Computer Vision (IJCV), 115 (2015), pp. 211–252.

[36] G. E. SCHNEIDER, *Two visual systems.*, Science, (1969).

[37] K. SIMONYAN AND A. ZISSERMAN, *Very deep convolutional networks for large-scale image recognition*, arXiv preprint arXiv:1409.1556, (2014).

[38] C. SZEGEDY, W. LIU, Y. JIA, P. SERMANET, S. REED, D. ANGUELOV, D. ERHAN, V. VANHOUCKE, AND A. RABINOVICH, *Going deeper with convolutions*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.

[39] H. TANG, C. BUIA, R. MADHAVAN, N. E. CRONE, J. R. MADSEN, W. S. ANDERSON, AND G. KREIMAN, *Spatiotemporal dynamics underlying object completion in human ventral visual cortex*, Neuron, 83 (2014), pp. 736–748.

[40] H. TANG, M. SCHRIMPF, B. LOTTER, C. MOERMAN, A. PAREDES, J. O. CARO, W. HARDESTY, D. COX, AND G. KREIMAN, *Recurrent computations for visual pattern completion*, arXiv preprint arXiv:1706.02240, (2017).

[41] H. TANG, M. SCHRIMPF, W. LOTTER, C. MOERMAN, A. PAREDES, J. O. CARO, W. HARDESTY, D. COX, AND G. KREIMAN, *Recurrent computations for visual pattern completion*, Proceedings of the National Academy of Sciences, (2018), p. 201719397.

[42] M. J. TOVEE, E. T. ROLLS, AND P. AZZOPARDI, *Translation invariance in the responses to faces of single neurons in the temporal visual cortical areas of the alert macaque*, Journal of neurophysiology, 72 (1994), pp. 1049–1060.

[43] L. G. UNGERLEIDER AND J. V. HAXBY, *'what'and 'where'in the human brain*, Current opinion in neurobiology, 4 (1994), pp. 157–165.

[44] S. YAN AND Q. LIU, *Inferring occluded features for fast object detection*, Signal Processing, 110 (2015), pp. 188–198.

[45] A. YIN, *One-shot-pokemon images*, 2018.
https://www.kaggle.com/aaronyin/oneshotpokemon/.

[46] B. ZOPH, V. VASUDEVAN, J. SHLENS, AND Q. V. LE, *Learning transferable architectures for scalable image recognition*, arXiv preprint arXiv:1707.07012, 2 (2017).