Periodicals of Engineering and Natural Sciences ISSN 2303-4521
Vol. 8, No. 1, February 2020, pp.79-88

Design and implementation of image based object recognition

Farooq Safauldeen Omar[1], Sazeen Taha Abdulrazzaq[2], Muamar Almani Jasim[3]

[1,2,3] Kirkuk Technical Collage, Northern Technical University, Kirkuk, Iraq.
fkutalar@ntu.edu.iq[1*]; sazeentaha4@ntu.edu.iq[2]; muamar78@ ntu.edu.iq[3];

ABSTRACT

The aim of this paper is to design and implement image based object recognition. This represents more of a challenge when speaking of advance object recognition systems. A practical example of this issue is the recognition of objects in images. This is a task that humans can perform very well, but convolutional neural network systems don't struggle to perform. AlexNet pre-trained model was used for the training the dataset because of it trouble-free architecture on very large scale dataset "Cifar-10" using R2019a Matlab. The dataset was split with the ratio of 70% for training and 30% for the testing part. This has prompted convolutional neural network to start experimenting with networks architectures as well as new algorithms to train them. This research paper presents an approach to train networks such as to improve their robustness to the recognition of object images on R2019a Matlab. This training strategy is then evaluated for designed AlexNet network architecture. The result of the study was that the training algorithm could improve robustness to different image recognition at the expense of an increase in performance for the performance of images of objects (i.e. Dog, Frog, Deer, Automobile, Airplane etc.) with high accuracy of recognition. When the advantages of different types of architectures were evaluated, it was found that accuracy of all object recognition were around 98% based on the image. It followed the findings from classical object recognition that feed-forward neural networks could perform as well their high accuracy of recognition.

Keywords: Object recognition, Object identifying system, Pattern recognition

Corresponding Author:
Farooq Safauldeen Omar
Kirkuk Technical Collage
Northern Technical University
Kirkuk, Iraq
E-mail: fkutalar@ntu.edu.iq

1. Introduction

In this paper, a convolutional neural network (CNN) based digital image processing system for object recognition has been introduced that could identify the objects in less time with more accuracy. Today, object recognition and convolutional neural network are synonymous with artificial neural networks. This all started with the design of a neural network model by Kunihiko Fukushima in the 1980s [1]. For this he sequentially stacked, what are nowadays called convolutional neural network layers (Figure 1). This model was especially elegant, because it was efficient, while respecting the biological models of the time. First of all, it had hierarchical feed-forward 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 [3], 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 recognition could achieve good performances for very simple digit recognition tasks [4] by combining an unsupervised learning algorithm for the internal layers [5] and basic perceptron supervised learning for the last layer [6]. Once this type of convolutional architectures was combined with supervised learning techniques based on back-propagation, 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 [7] and dataset of images depicting different classes of objects [8].

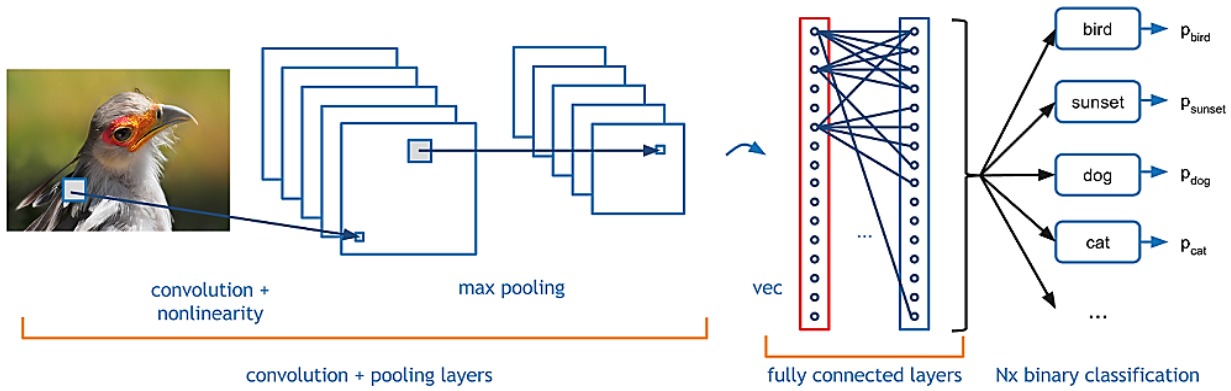

Figure 1. Convolutional architectures was combined with supervised learning techniques based on back-propagation for object recognition [8]

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 back-propagation. From this perspective, it makes sense that the next step would be to improve object recognition systems 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 field's studies in the inferior temporal cortex [9, 10]. It therefore made sense to target the design of rotation and scale invariant object recognition systems.

1.1. Contribution

This paper examines and contributes to some extent given by:

- The convolutional neural network (CNN) based adaptive object recognition system that can recognize different objects with more accuracy using advanced AlexNet pre-trained model on Cifar-10 dataset such that the network topologies are retained.
- Using training it is possible for convolutional neural network (CNN) based models to solve recognition problems of objects and adaptive recognition task.
- Adding the validation to the training process so, that the system doesn't overfits after the training and during the testing phase.
- The contributions are driven by two inquiries around designing and implementing of adaptive object recognition system based on convolutional neural network (CNN) with high accuracy using R2019a Matlab.

2. Background

In their study, Tang et al. [11] demonstrated different ways of recognizing objects starting from trained feedforward models. They could show that replacing the seventh neural network layer of ResNet 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 [12] and applying a machine learning using only the unoccluded images as input. This is a great feature from a neuro-scientific 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 [13]. 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 pre-trained 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 images and unoccluded images version 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.

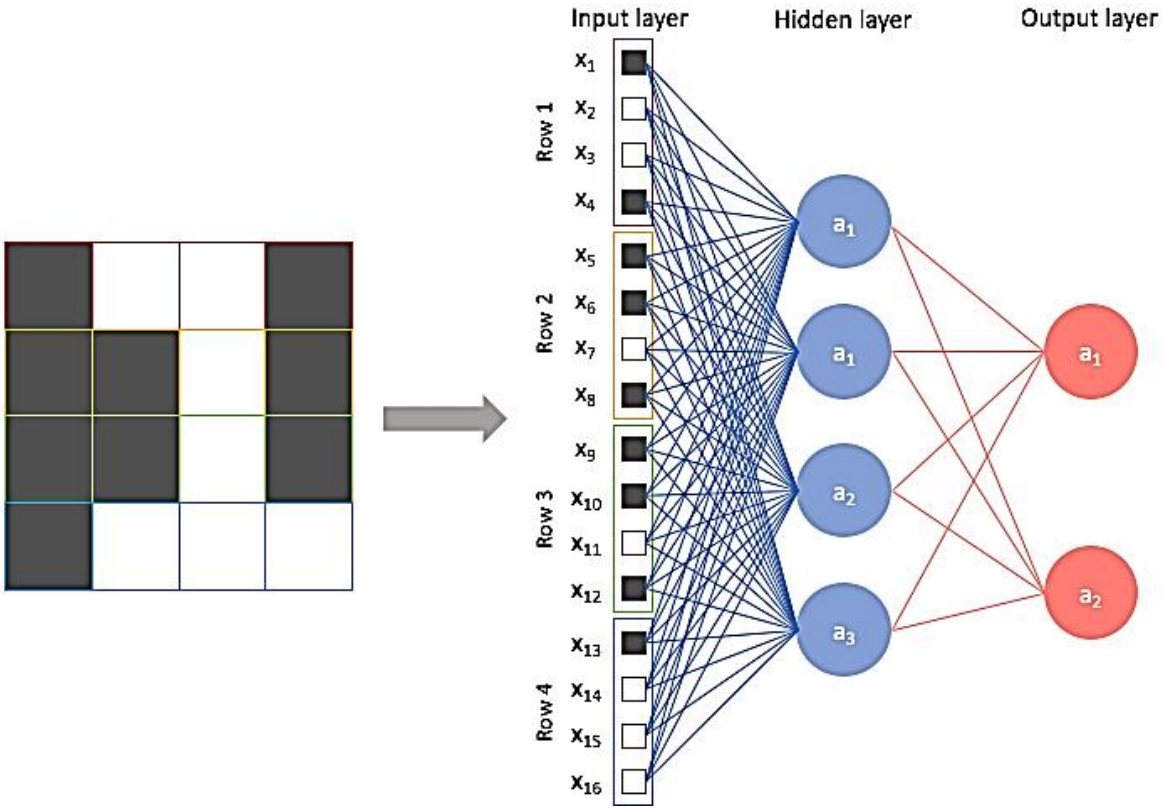

Figure 2.The convolutional neural networks with the complex feedforward architecture for the object recognition [13]

Object recognition 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 [14]. This structure, also known as the ventral stream became associated to the task of visual object recognition [14, 15] and was branded the 'what' pathway. 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 [15]. 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 led to important insights for the development of object recognition 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.

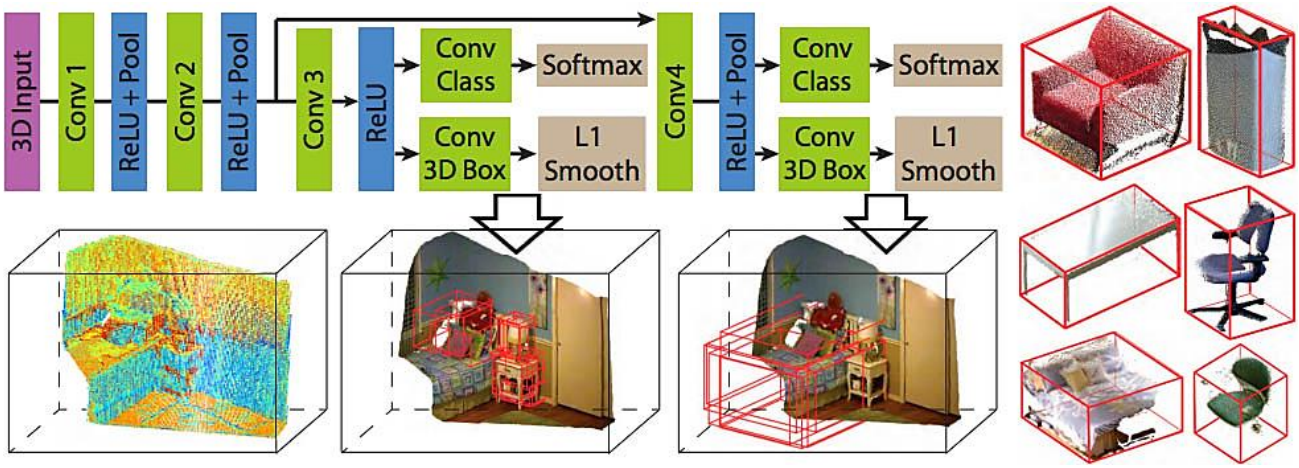

Figure 3. The process of detecting and recognizing the objects using CNN [13]

3. Methodology

3.1. Optimization of convolutional neural networks

For the convolutional neural networks, the choice for basic feedforward architecture was made in favor of AlexNet for this work. Besides of one network pretrained on Images and used as a control, for the purpose of object recognition. We designed and implemented a network architecture as the control, but were retrained. AlexNet had identical weights except for the 7th and 8th layer respectively. These trainable weights were the only ones modified during subsequent training. 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 for maximum optimization and increased accuracy. AlexNet 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 only the 7th layer was replaced by a recurrent layer and the 8th layer stayed feedforward as shown in the figure 4 below, but would also be part of the set of weights to be trained during optimization.

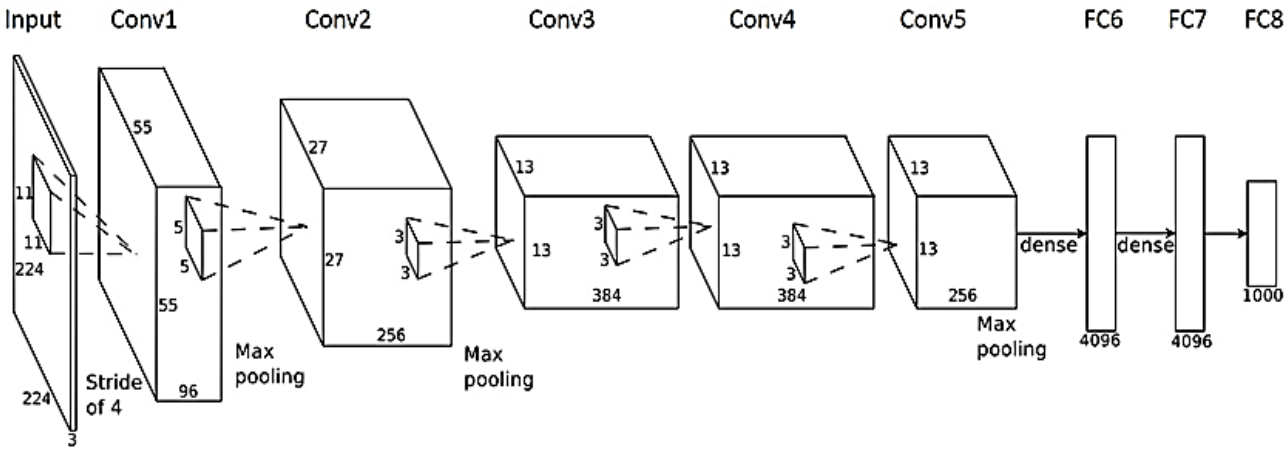

Figure 4. The optimized AlexNet architecture with max pooling for object recognition

A number of variations for sigmoidal activation functions exist such as the hyperbolic tangent and the rectified linear unit (Max Pooling, see Equation 1).

$$f(x) = \frac{0 \; if \; x>0}{x \; otherwise} \quad (1)$$

They are applied either in a feed-forward or recurrent (cyclic) manner, where the recurrent variant performs temporal transformations.

3.2. Object recognition architecture

It turns out that implementing ingenious object recognition network architecture is not the only way of achieving invariance towards a transformation. According to the universal approximation theorem for neural networks [12], 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. In theory there should therefore be no functional limitation to the power of standard neural network architectures for object recognition and image classification based on object localization as shown in figure 5. 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 Images from Cifar-10 dataset for object recognition.

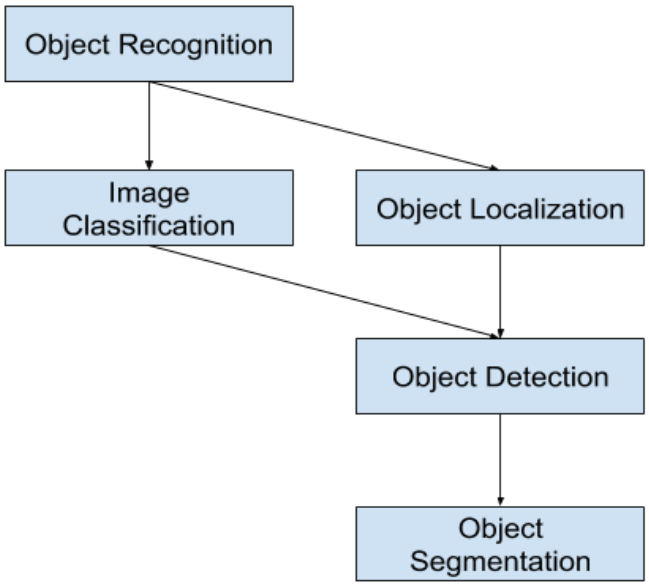

Figure 5. The approach being used for designing an optimized architecture for object recognition

3.3. Training of data

All of AlexNet network was trained using the Cifar-10 dataset consisting of images of different objects in dataset. Not the entire training set was used and only the images belonging to 40,000 out of the 10,000 possible images were used. This was because these results were originally planned 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 stochastic optimization method. Training aimed at minimizing the l2 distance between the activation patterns of the readout layer of the trained network with the activation pattern from the readout layer of the pre-trained AlexNet model in response to the same images, occluded or not. The set of weights trainable for each network correspond. AlexNet was trained to go through three recurrent loop time steps. Two different variations of AlexNet were trained, one with 2 time steps and one with 5 time steps of recurrence.

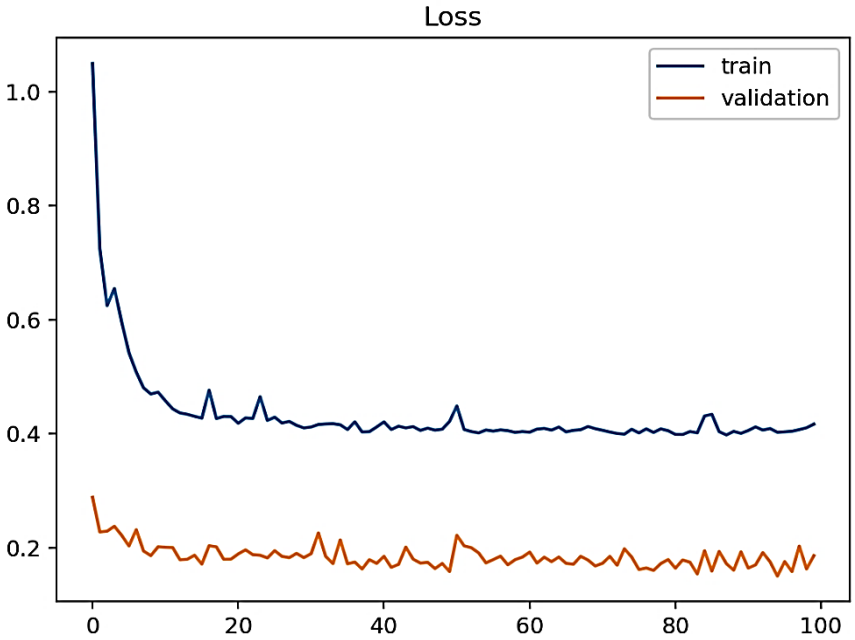

Figure 6. The training and validation loss during training process with split of dataset into 70% and 10% respectively

3.4. Dataset description

The Cifar-10 image dataset was chosen [2], because it had an appropriate size and difficulty, but also because many object recognition models are available that were pre-trained with Images. The Cifar-10 version-2019 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 Cifar-10 version-2019, because it is one of the few ones were the test labels have been released. It contains labeled images belonging to one of 10,000 different images.

All images were processed to size 227x227 resolution. An example image of 0%, 25%, 50% and 80% can be seen in Figure 7. As can be observed, the images 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.

Figure 7. Cifar-10 dataset for object recognition [2]

4. Evaluation of model

The performance of the models was evaluated on the Image 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 pre-trained AlexNet, it was necessary to restrict the model testing to images that belonged to one of the 40000 images 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. Taking a step back from this object recognition 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 evaluating method through back-propagation. 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 accuracy (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 images with 100% visibility. Finally, it is interesting to notice that although their performances is relatively similar, AlexNet is slightly better at recognizing images, but the contrary is true for images. In all cases the stable performance seems to be in the region of 90% visibility, in which all networks perform very similarly.

5. Results

For results, the most basic comparison to make is between networks that have been trained in the same way and are identical, but for the object detection of their connections. This contrast is visualized in both figures below. However, in object recognition network has a very slight edge on the feedforward one. In order to catch the differences between the different object recognition of the objects, their respective performance accuracy are depicted in Figures. There, one can notice that similarly to the retrained feedforward networks, they all became more robust to objects but paid a slight price in classification performance of objects. The main result that can be drawn from these object recognition figures are that all convolutional neural network using pre-trained model perform very extra-ordinary without the exception of error, which if very best for the object recognition.

Figure 8. The recognition of object as Frog with an accuracy of 99.74%

After the analysis of the object recognition for compliance with the requirements, the calculations and evaluation of development dynamics object detector using AlexNet were carried out. A model was developed for predicting objects using the methods for construction in order to improve the professional object detection and recognition.

In the course of object recognition, average values were obtained for each type of new object. Then the prediction of accuracy was made, where the object of measurement was one predicted, and the maximum value was 99.74%. The prediction of different objects can be seen in the figures below.

Figure 9. The recognition of object as Dog with an accuracy of 98.47%

Figure 10. The recognition of object as Deer with an accuracy of 98.78%

Figure 11. The recognition of object as Cat with an accuracy of 94.64%

6. 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 accommodate knowledge about how to classify occluded images into their weights. This therefore led to improvements in performance on images with imperfect visibility in comparison to the pre-trained AlexNet model that had never been subjected to explicitly occluded images. The increase in robustness to incomplete images lead to a tradeoff in performance for 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 led to the observed decrease in performance on objects. The second main observation that AlexNet is more affected by the training in the sense that its performance on occluded objects increased more and its performance on objects decreased. This can be explained by the bigger search space available to AlexNet during training. Indeed since the entire weight search space, it could be expected that if the training performed well, CNN should be able to move further away from the weight configuration of the pre-trained AlexNet model. This would imply that, if it was useful for the increase of overall performance to increase slightly the performance on objects, this is something AlexNet would be better capable to do in object recognition. Since this would make sense, as was described in the argument about giving up features that were not robust to images (here features are meant as neural activity patterns in layer 6 or 7), this explanation appears to hold so far.

7. Conclusion

This paper sat out to design and implement the object recognition system based on convolutional neural network relevance to the field of for AlexNet neural models was presented for object recognition on Cifar-10 dataset. Due to the large amount of completed experiments included in this report, there is no shortage object recognition material. This includes the actual transfer learning study on convolutional neural networks such as AlexNet with one or several feedforward layers replaced with recurrent ones. Besides of the variant of using a fraction of the 10,000 Images 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 Images for object recognition. Doing this could partially exclude the component of transfer learning due to the high similarity between some Images such as dogs, deer, frog etc. A rich set of images all different to Image is hard to come by, but a possibility was to use the openly available image dataset. It is a set of 2000 images for testing with 800 for validation, all depicted in a variety of styles. It is questionable whether such a dataset would sufficiently well sample the image space to adequately train the CNN layers without overfitting, but this research provides very satisfying results being successful.

References

[1] Possegger, H.; Mauthner, T.; Bischof, H. In defense of color-based model-free tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, pp. 7–12 June 2015

[2] Available Online: https://www.cs.toronto.edu/~kriz/cifar.html

[3] Bertinetto, L.; Valmadre, J.; Golodetz, S.; Miksik, O.; Torr, P.H.S. Staple: Complementary learners for real-time tracking. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1401–1409.

[4] Danelljan, M.; Robinson, A.; Khan, F.K.S.; Felsberg, M. Beyond correlation filters: Learning continuous convolution operators for visual tracking. In Proceedings of the European Conference on Computer Vision, Amsterdam, Netherlands, 8–16 October 2016; pp. 472–488.

[5] Held, D.; Thrun, S.; Savarese, S. Learning to track at 100 fps with deep regression net-works. In Proceedings of the European Conference on Computer Vision, Amsterdam, Netherlands, 8–16 October 2016; pp. 749–765.

[6] Tapu, R.; Mocanu, B.; Bursuc, A.; Zaharia, T. A Smartphone-Based Obstacle Detection and Classification System for Assisting Visually Impaired People. In Proceedings of the 2018 IEEE International Conference on Computer Vision Workshops, Sydney, Australia, 2–8 December 2018; pp. 444–451.

[7] Smeulders, A.; Chu, D.; Cucchiara, R.; Calderara, S.; Dehghan, A.; Shah, A. Visual Tracking: An Experimental Survey. IEEE Trans. Pattern Anal. Mach. Intell. 2018, 36, 1442–1468.

[8] Jia, Y. Caffe: An Open Source Convolutional Architecture for Fast Feature Embedding. Available online: http://caffe.berkeleyvision.org/.

[9] Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 2015, 115, 211–252.

[10] Zagoruyko, S.; Komodakis, N. Learning to compare image patches via convolutional neural networks. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4353–4361.

[11] Nie, Y.; Ma, K.K. Adaptive rood pattern search for fast block-matching motion estimation. IEEE Trans. Image Process. 2016, 11, pp. 1442–1449.

[12] A World Health Organization (WHO)—Visual Impairment and Blindness. Available online: http://www.who.int/mediacentre/factsheets/fs282/en/

[13] Rodríguez, A.; Yebes, J.J.; Alcantarilla, P.F.; Bergasa, L.M.; Almazán, J.; Cela, A. Assisting the Visually Impaired: Obstacle Detection and Warning System by Acoustic Feedback. Sensors 2017, 12, pp. 17476–17496.

[14] Tapu, R.; Mocanu, B.; Tapu, E. A survey on wearable devices used to assist the visual impaired user navigation in outdoor environments. In Proceedings of the 11th International Symposium on Electronics and Telecommunications (ISETC), Timisoara, Romania, 14–15 November 2017; pp. 1–4.

[15] Croce, D.; Giarré, L.; Rosa, F.G.L.; Montana, E.; Tinnirello, I. Enhancing tracking performance in a smartphone-based navigation system for visually impaired people. In Proceedings of the 24th Mediterranean Conference on Control and Automation (MED), Athens, Greece, 21–24 June 2016; pp. 1355–1360