LAPPEENRANTA UNIVERSITY OF TECHNOLOGY
School of Engineering Science
Computational Engineering and Technical Physics

REPORT

Denis Sedov, Sinan Kaplan

Fish detection and classification using Convolution Neural Networks

Examiners: Professor Arto Kaarna

Supervisor: Professor Arto Kaarna

ABSTRACT

Lappeenranta University of Technology
School of Engineering Science
Computational Engineering and Technical Physics

Denis Sedov, Sinan Kaplan

Fish detection and classification using Convolutional Neural Networks

2017

27 pages

Examiners: Professor Arto Kaarna

Keywords: fish detection, fish recognition, convolutional neural networks

Automated fish detection and classification is considered to increase the performance and diminish human error significantly in comparison with manual processing. In this report, different methods for fish detection and classification are described. State-of-the-art convolutional neural networks is one way how to approach the given problem. There are two CNN architectures considered. The first one is based on traditional LeNet and AlexNet architectures. The second one uses advantage of separation on high and low resolution frames and combining the information from both channels for the classification purpose. Pretests show that the resolution does not have a significant effect on the performance, therefore only first architecture is implemented. Experimental part includes the analysis of the activation functions, filter sizes and data augmentation. The overall performance of the model is 71%.

CONTENTS

1 INTRODUCTION

1.1 Problem description

The rising level of environmental issues across the world leads researchers to understand the core of the problems and develop solutions for those issues. Extinction of some fish species is one of those important issues. One way to solve this problem for a particular area could be by simply observing fishes in this area.

Saimaa channel/canal is a kind of bridge for species in the Lake Saimaa between Russia and Finland. In order to analyze the biological diversity of species, which travel across the channel, we decided to detect and recognize fishes passing through the channel. This task is focused on analyzing surveillance videos using deep convolutional neural networks [1].

1.2 Structure of the report

The paper is structured as follows: Section 2 gives the short history of convolutional neural networks, describes different methods, which were developed for the research problem as well as application of CNN to the research problem. Section 3 states the research question and hypothesis. Section 4 considers the proposed approach for solving the problem stated in the research question. Section 5 evaluates the time which was needed in order to complete the research. Section 6 describes the experiments and obtained results, and the conclusions are made in Section 7.

2 CONCEPTS AND RELATED WORK

2.1 Deep Convolutional Neural Networks

Convolutional Networks are first introduced in 1990 by Yann LeCun [9]. LeNet architecture (Figure 1) is one of the best and earliest examples of the concept developed in the paper. The main application of this architecture is to read digits and zip codes.

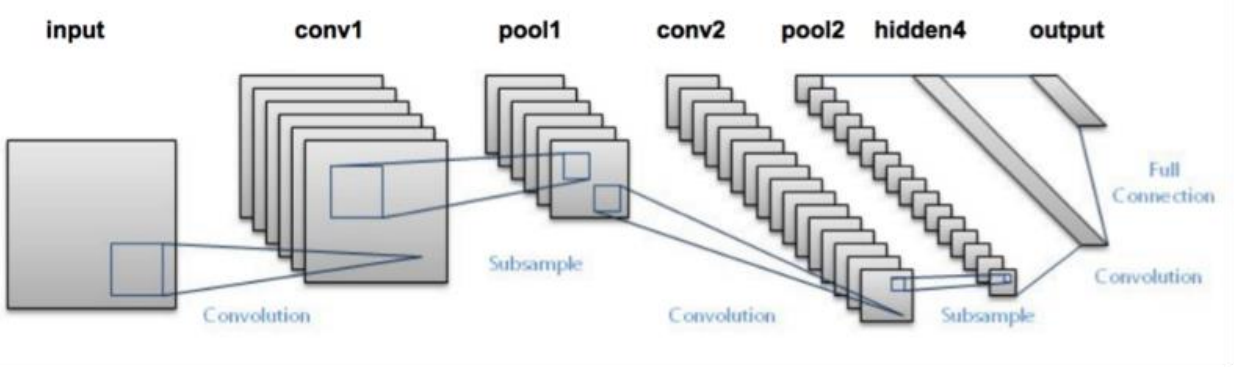


Figure 1. LeNet Architecture consists of two convolutional layer [9].

The popularity of the convolutional networks in computer vision has begun with Alexnet [10] in 2012. AlexNet is particularly developed for the challenge of ImageNet Classification [11]. The architecture of AlexNet (Figure 2) is quite similar to LeNet. However, the main difference in the architecture is to have more layers or, in other words, deeper architecture.

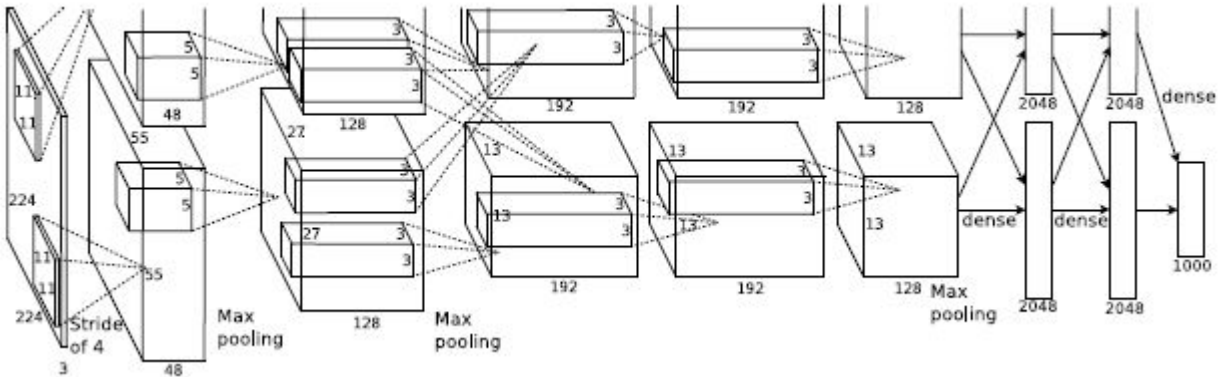
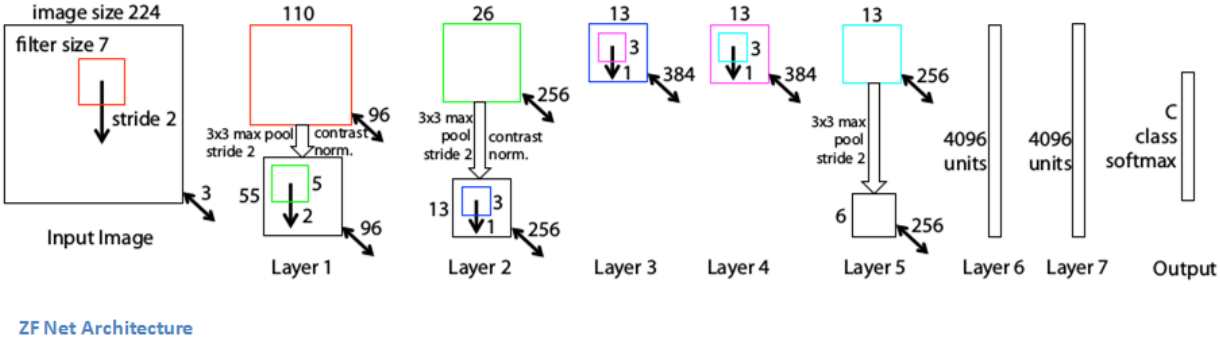


Figure 2. AlexNet Architecture consists of eight layers: five convolutional and three fully connected [10].

ZF Net [12] is another example of convolutional networks implemented for the same purpose as AlexNet. The main contribution of ZF Net architecture (Figure 3) in the field is that it presented the different type of convolutional networks by simply changing parameters and number of filters in each layer.



ZF Net Architecture

Figure 3. ZF Net Architecture [12].

After the success of ZF Net, Google introduced new convolutional network called GoogleNet [13] for ImageNet competition. In this network, total number of parameters is decreased by the use of inception layer. The architecture is shown on the Figure 4 and we can already see the increasing complexity of the deep neural networks.

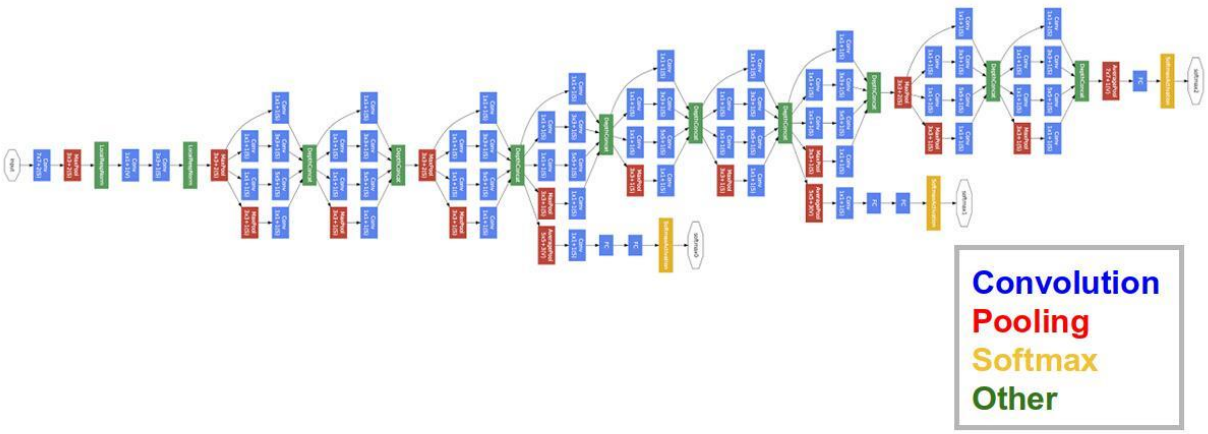


Figure 4. GoogleNet Architecture [13].

VGGNet [14] was developed in order to reveal the importance of the deep architecture of convolutional networks. VGGNet architecture (Figure 5) shows that the depth of the network

is a crucial design pattern for the convolutional networks. Increasing level of depth of the networks, it is possible to increase the accuracy of predictions. However, the downside of making the architecture deeper is that it takes long time to train the whole network.

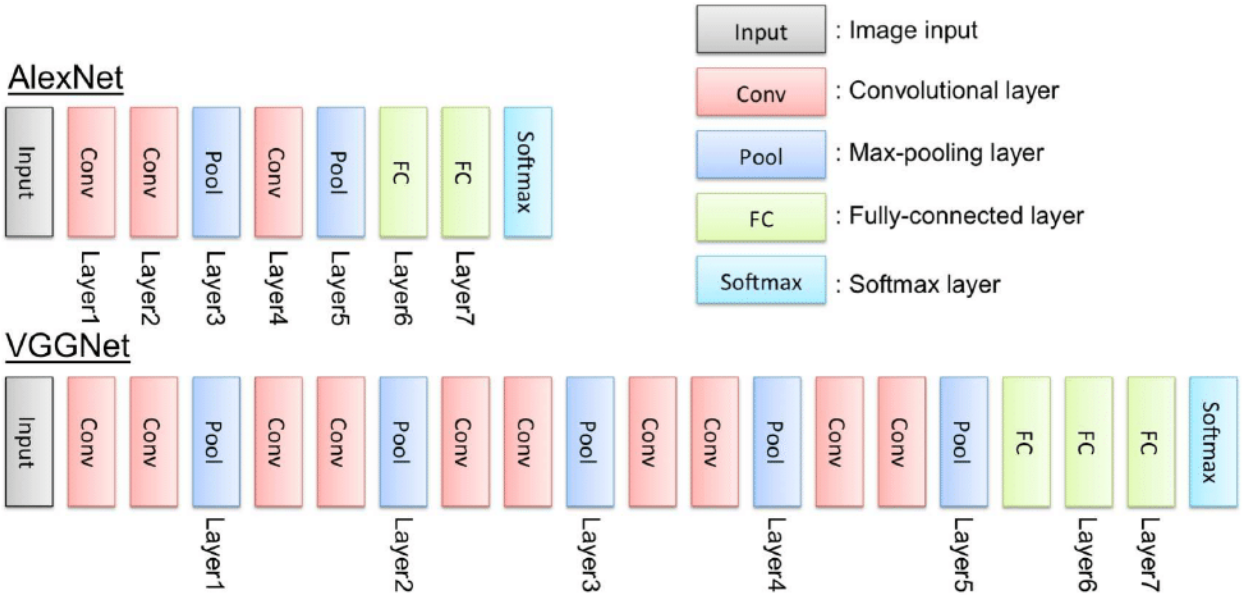


Figure 5. VGGNet [14] and ALexNet [10] architecture comparison.

The state-of-the art network is ResNet [15], which has been developed recently (Figure 6). The architecture of ResNet differentiates from above mentioned networks by using batch normalization densely and introducing skip connection within the architecture.

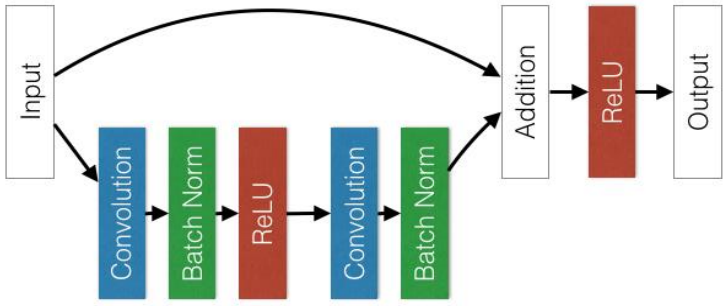


Figure 6. A simple building block from ResNet [15].

2.2 Related work

2.2.1 Fish Detection and Recognition Methods

The problem of fish detection and recognition has been studied mostly for environmental concerns [2, 3]. The main motivation behind the studies in this field is to automate the fish classification/recognition process in terms of time efficiency. There are several papers, which are based on this motivation.

In study [4], the authors summarize the research efforts with their specific methods to provide a guideline in this field. After giving literature review, they compared Principal Component Analysis (PCA) with Scale-invariant Feature Transform (SIFT). By using a synthetic dataset from NOAA SWFSC Benthic Resources Group, they find out that the large training set provides better classification accuracy. This study can be seen as a small guideline, however, it does not provide detailed knowledge in the field and the compared methods are only limited to PCA and SIFT.

In study [2], the authors proposed a method, which takes advantage of Adaptive Gaussian Mixture Model and Adaptive Mean Shift Algorithm for detection, tracking and counting of fishes using low quality videos from Ecogrid in an uncontrolled environment. They succeeded to develop the model, which provides reasonable detection and classification accuracy even in varying environmental conditions. The system is capable of analyzing videos to give some relevant information, such as the number of fish present, quality of the video (clear, murky, smoothed, etc.), dominant color of the video. The main drawback of the system is that it is not still robust enough to drastically varying conditions, such as e.g. murky water, moving plants and unknown objects. In addition to those shortcomings, low contrast and low frame rate can be considered as main limitations of this study.

The paper [5] studied both fish classification and clustering in order to automate the fish classification system, which enables also fish tracking. In that system, they tested combination

of spatial Gabor filtering, co-occurrence matrix, Curvature Scale Space transform and histogram of Fourier descriptors using Ecogrid data set. As a result, they manage to combine texture features with shape descriptors for better classification. However, their methods are not robust enough to low quality video streams and rapidly changing water environment, since it is assumed to be a fixed underwater habitat.

In recent study [6], a framework is developed to classify and cluster fishes based on extracted features. This framework is a new type of systems in this field, which introduces different type of methods to integrate into system. At the feature extraction step, PCA, SIFT and vector of locally aggregated descriptors are used, whereas Artificial immune network algorithm (aiNet), Adaptive radius immune algorithm (ARIA) and k-Means are applied for clustering. The classification task is done by using k-Nearest neighbor classifier (k-NN), SIFT, and k-means. Different combinations of those methods are tested to decide, which methods provide better performance than other for the final implementation. The combination of PCA, ARIA and k-NN superiors to the rest of combinations.

The authors in [7] studied Support Vector Machines for fish classification task compared to Artificial Neural Networks (ANN), k-NN and k-means. After relevant experiments on the texture datasets from Fishery Departments of the Federal University of Technology, Akure (FUTA), Nigeria and Adekunle Ajasin University, Akungba-Akoko (AAUA), Nigeria, they found that SVM clearly enables more accurate results that other methods. On the other hand, this study suffers from the small sized dataset.

So far, the idea in aforementioned studies for fish detection and classification follows three steps: removing background from foreground, removing noise (or applying morphological operations) and object classification. The same procedure is applied in [25] with different algorithms in each stage. While Gaussian mixture model is used for first step, in the second

stage morphological operations are applied. The final step is done with different classification algorithms. However, it is shown that random forest superiors the other used methods.

Instead of keeping camera stable, the researchers in [26] move the video camera and try to detect and classify fishes with this particular setting. The primary goal of this study is to count and classify fishes with moving camera. For counting fishes, the authors applied canny Edge-detection, blob-counting methods while they used Zernike moments for classification. The findings from this study are that the performance of fish count is highly dependent on goodness of background estimation and Zernike moments seem to work quite well for classification.

Recently, Jäger, Jonas et al. [8], have used convolutional neural networks (CNN) for detection and classification of fishes in low-resolution videos as part of SeaCLEF 2016 challenge. Their study revealed that object proposal classification (OPC) provides more accurate results than mostly used traditional technique background subtraction for fish detection and classification. The method used for classification part of CNN is binary SVM classifier. The negative feature of this study is usage of relatively small CNN architecture when compared to state-of-art architectures [1].

2.2.2 Neural Network Approaches to Fish Detection and Recognition

With the recent rise of neural networks and its extension to computer vision related tasks, Convolutional Neural Networks are used to solve classification and recognition tasks as well. In this section, we will go through recently conducted researches about CNN, in order to give the main intuition for what kind of applications or problems CNN can be used.

AlexNet [10] is one the earliest designs of CNN, which is used for classification of large-scale images during the ImageNet Challenge. In this study, rectified linear units (ReLu) is introduced as part of CNN with the purpose of enabling robust regularization over the

network. ReLu removes the overfitting, which is the main drawback of neural network based approaches. AlexNet was selected as the best method in ImageNet challenge and superior to all the developed methods so far for this particular contest. The performance of CNN on image classification tasks has lead the researchers around the world to focus on this area to advance the studies with different structures of networks and different optimization methods. For instance, in the case of aforementioned VGGNet [14], the effect of adding more layers to CNN has been studied to reveal whether there is an increase in performance or not. Even though this approach increases the computational burden, it is proven that adding more layers increases the classification accuracy. After this point, the importance of adding more layers, the concept of deep convolutional neural networks (DCNN) is merged and new architectures are proposed, which follow this trend. In the context of image classification from video streams, it is quite significant to discriminate low-resolution frames from high-resolution frames in order to get better explanation of the data [18]. However, training the network without low-resolution frames might cause loss of some relevant information. By taking this fact into consideration, the authors in [17] proposed a new architecture in which the low-resolution frames and the high-resolution frames are trained differently with two parallel DCNN, however their fully connected layers are combined into one for classification purposes. Respectively, the experiments on Sports-1M dataset show that this approach learns powerful features for the classification. The same concept with different settings is applied in [20]. The main characteristic of this study is to divide the networks into two streams, one is called spatial stream ConvNet and the other one is called temporal stream ConvNet. While the spatial ConvNet is trained with a single frame at a time, the temporal stream ConvNet is trained on multiple-frame dense optical flow. Subsequently, the results obtained from experiments is shown the training of the ConvNet on optical flow in fact gives better performance compared to raw stacked frames.

While above mentioned studies have proven that the mining of all the frames, particularly low resolution and high resolution, from video sequence play an important role on the performance

of classification tasks, it is still necessary to detect and track an object of interest on the frames precisely. In this case, the localization of the object from one frame to another one is considerably significant that can enable the system having better tracking and detection performance. This problem motivated the authors in [21] and they studied the combination of Recurrent Neural Networks with Long Short-Term Memory (LSTM) [1, 18] and CNN together. The main characteristic of RNN here is to memorize the location of features extracted from CNN. In this way, the incorporating information across models leads the better classification results. The indicated idea here is recently applied in [23] as a temporal CNN (TCNN). The main characteristic of this study is that it combines still-image object detection with generic object tracking for tubelet proposal. In other words, the localization of an object of interest is studied for efficient tracking and detection. This approach also tested for low quality videos by the combination of RNN and CNN [24].

Aforementioned studies focus on implementation of CNN for particular tasks in their domain, such as ImageNet challenge [10-15]. However, when the amount of time and computational resources needed for training of CNN are considered, it is not efficient just to train CNN for a single task and use learned features for this task. Instead of that, it could be more beneficial to reuse those learned features for other tasks. This idea is studied as a transfer learning in [22]. This study is inspired from the point of view to use learned features across the different domains. In that way, it is possible to solve some particular problems, which have small data sets by enabling usage of the learned features from other domains. For instance, the study uses the features from ImageNet challenge to solve Pascal VOC challenge, which notably has small amount of data. One should bear in mind that the main ingredients of CNN is large-scale data set [1,18,21,22].

As the main goal here is to study CNN on the videos, the paper [16] is a good example of that, in which the authors developed a system for human action recognition by utilizing CNN. The study focuses on analysis and recognition of certain actions with 3D CNN. The main

contribution of this paper is that it uses 3 dimensional kernels/filters instead of traditional 2D kernels/filter in the architecture of CNN. By approaching the problem in that way, the authors show that 3D kernels extract more reliable and robust features for the given task.

3 OBJECTIVES

By taking into account both our problem and literature review, we formed our research questions as follow:

- Can one classify the fishes passing a fish pass?
- Does CNN work for fish detection and recognition tasks with low quality videos?

Above mentioned research questions lead us to define our hypothesis by considering both our motivation and literature review. The hypothesis in this case is that:

- Deep Neural Network enable reliable fish classification.
- It is possible to use CNN for fish detection on videos and for further tracking/counting - recognition of fishes.

4 APPROACH

Current approach to solve fish detection task is to consider 2 different CNN architectures, which seem to be promising according to the literature review on the topic. Subsequently, it would be possible to measure the performance on both architectures and choose the best one. The first architecture (Figure 7) is more traditional one and similar to LeNet or AlexNet.

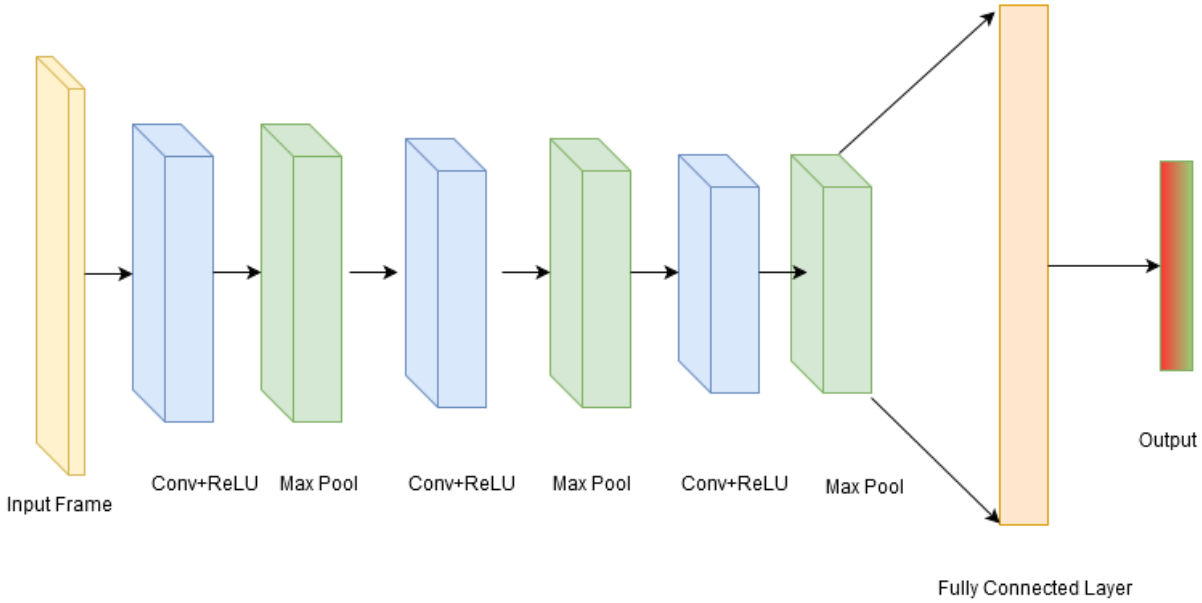


Figure 7. First CNN architecture, similar to LeNet or AlexNet.

The second CNN architecture (Figure 8) is based on the paper [17], where high and low resolution frames are separated into two different channels and afterwards the information from both channels is combined into fully connected layer.

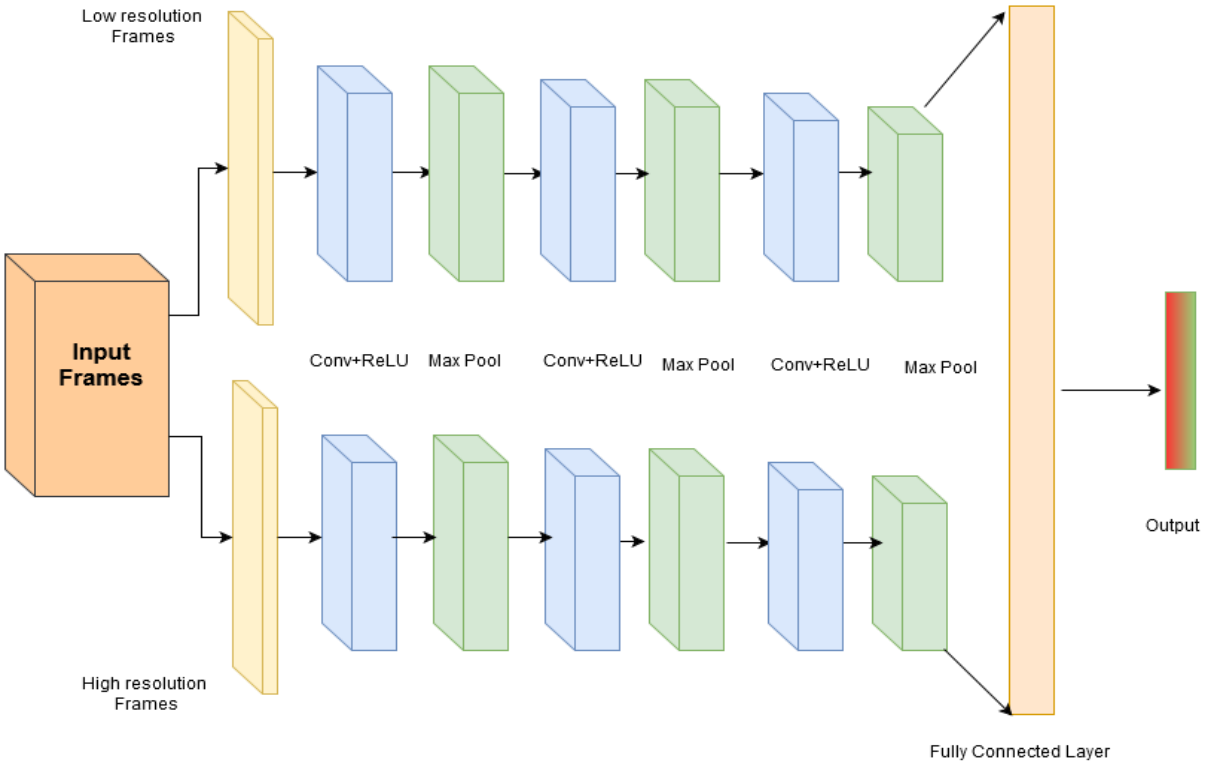


Figure 8. Second CNN architecture, which separates high and low resolution frames.

Some parameters of the network, such as the size of the convolutional filter, padding and the type of rectifying function in pooling layer, will be chosen during implementation, depending on the classification performance of the network, since there is no relevant information on the best parameters and they highly depend on the application. In addition, in the current approach each video frame is considered separately, i.e. temporal information is not used at all, since there will not be enough data to train CNN.

5 PROJECT WORKLOAD

This research was conducted in 4 months. The estimated workload for Denis is 84 hours (3 ECTS credits), for Sinan is 168 hours (6 ECTS credits). The research process consists of the following steps: literature review and research plan, implementation of algorithm (including setting up the environment) and testing, final report. Each student contributed to every step of

the research, since it helped to get deeper knowledge of the area by reading greater amount of papers, eliminate errors in the code and practice pair programming, thus producing significantly more valuable results. The large portion of implementation step is devoted to studying Python and corresponding libraries for deep learning (Lasagne and Theano). The Table 1 describes the workload of the group members according to the related tasks for the project.

Table 1. Project working hours.

	Denis (h)	Sinan (h)
Literature review (Nov - Dec)	20	50
Research plan (Dec)	20	30
Implementation of algorithm (Jan - Feb)	20	35
Testing (Feb)	10	25
Final report (Feb - Mar)	14	28

6 EXPERIMENTAL ANALYSIS

6.1 Description of the dataset

In this study, three different underwater fish surveillance videos are used to train and test the proposed CNN architectures. However, as an optimizer, Adam optimizer [28] is used instead of traditional gradient descent based optimizer. There are three videos in total and the length of all videos together is 25 seconds. After extracting all the frames, we get 876 frames in total and we keep 493 frames from video one as a test set and rest of the frames from both video 2 and video 3 as a training set. All the frames are down-sampled to the size of 40x40 pixels in order to decrease the computational burden.

6.2 Experiments

First, the architecture presented on the Figure 7 was implemented. Several pretests were done with different resolutions of the video frames. The results revealed that the resolution does not have a significant effect on the classification accuracy, which means that the frames can be down-sampled without the loss in the performance. Since the results were similar both for high and low-resolution frames, their combination in the second CNN architecture (presented on the Figure 8) will not produce any advantage for the classification performance and therefore we decided to concentrate only on the first architecture.

To reveal the potential of CNN for detecting the images, we have conducted several experiments. Each experiment focuses on the particular property of CNN in terms of classification accuracy. Therefore, the experiments are divided into three subcategories, which are as follows:

1. Activation Function Analysis
2. Filter size Analysis
3. Data Augmentation Analysis

6.3 Activation function analysis

Three different activation functions are analyzed by providing learning curves of training loss and training accuracy for each epoch. The experiments are done by using ReLU, Sigmoid and Tangent as activation functions.

While the number of epoch increases, one can see from Figure 9 that the loss decreases for each activation function and tangent superiors the other activation functions. The main reason behind this situation is that tangent provides stronger gradients and in our case, as we have a low-quality data, this is an important factor which increases the overall quality of the model. The more detailed information about the activation functions is described in the study [27].

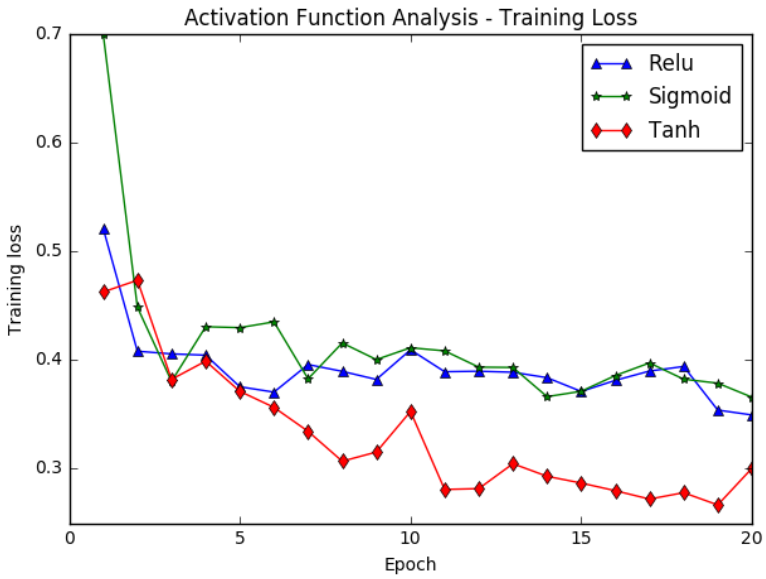


Figure 9. Activation Function Analysis - Learning curve of loss.

During epochs, we have also measured the training accuracy for each activation function. Figure 10 shows the results obtained from this analysis. One can easily see from the figure that there is a correlation between the performance of activation function on the loss value and the training accuracy.

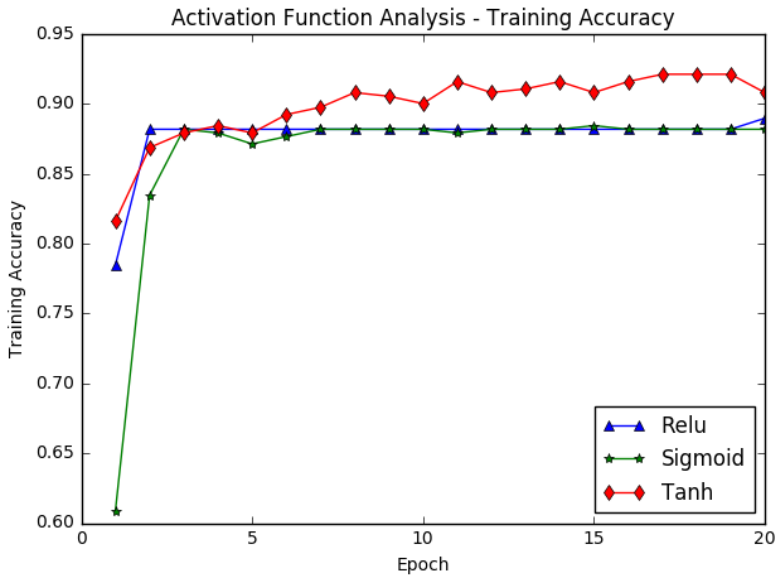


Figure 10. Activation Function Analysis - Learning curve of training accuracy.

The performance of each activation function is also examined by applying them on the test data. Table 2 shows the test accuracy of each activation function. As the tangent provides best accuracy on training data, it also superiors the other activation functions on the test data. Since the dataset does not contain variable context for training and testing data, this result is not something unexpected.

Table 2. Activation Function Analysis - Test Accuracy.

Activation functions	**Test Accuracy (%)**
ReLU	69.476082004555806
Sigmoid	67.425968109339408
Tanh	72.665148063781331

6.4 Filter size analysis

In this part, the impact of the filter size on the overall performance of the network is examined by using both training and testing data. The experiments are done by fixing activation function to ReLU. The aforementioned steps are also followed in this part. The small filter size does not capture the local features well in the frames and that is why the loss value doesn't change that much during each epoch. However, both 5x5 and 4x4 filter size have almost the same performance (Figure 11).

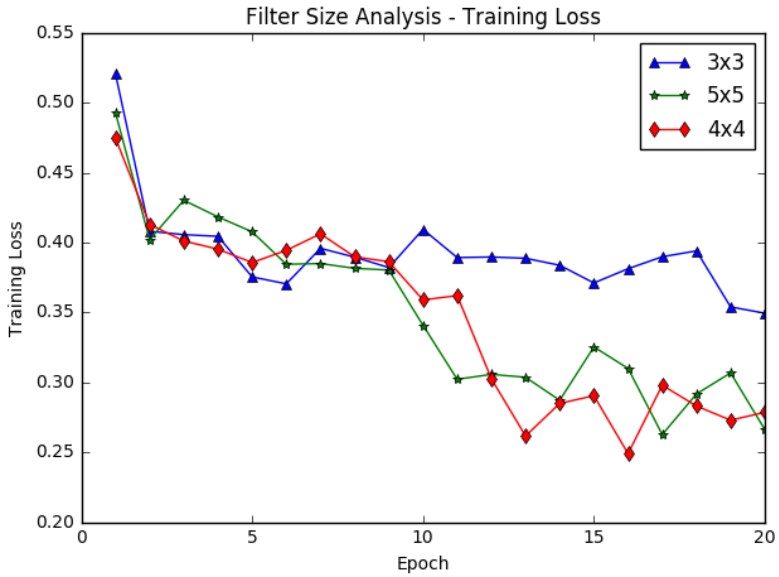


Figure 11. Filter Size Analysis - Learning curve of loss.

The training accuracy in this context is also analyzed and the results are given in Figure 12. The same situation explained above is also valid in this part.

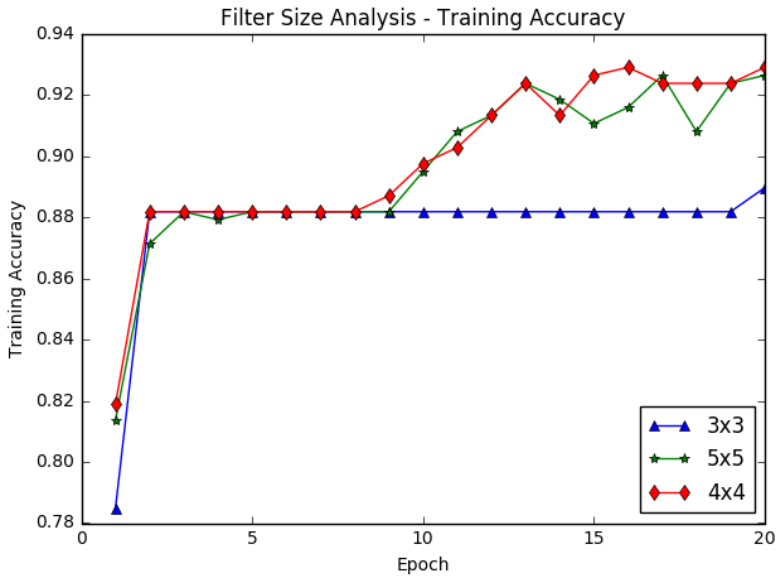


Figure 12. Filter Size Analysis - Learning curve of training accuracy.

Furthermore, we have studied performance of filter size on the test set and results are given in Table 3. As it can be seen, the cases with 4x4 and 5x5 filter sizes show approximately the same performance. However, the test accuracy is not that big, which can be explained by the low quality of the videos and very small training dataset.

Table 3. Filter Size Analysis - Test Accuracy.

Filter Size	**Test Accuracy (%)**
3x3	69.476082004555806
4x4	71.070615034168554
5x5	71.526195899772205

6.5 Data augmentation analysis

Since there is such a small dataset available for the training, it is possible to extend it by augmentation of the data. The augmentation is performed using 4 operations: horizontal flip, vertical flip, width shift (20% of the image width to both directions) and height shift (20% of the image height to both directions). The results of the training process are presented on the Figure 13 and Figure 14. As it can be seen the overall training loss and training accuracy are almost the same at the final epoch.

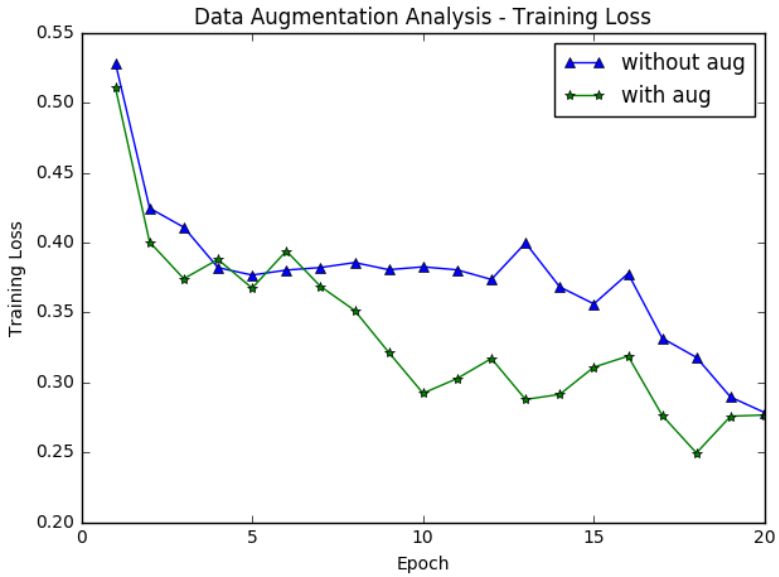


Figure 13. Data Augmentation Analysis - Learning curve of loss.

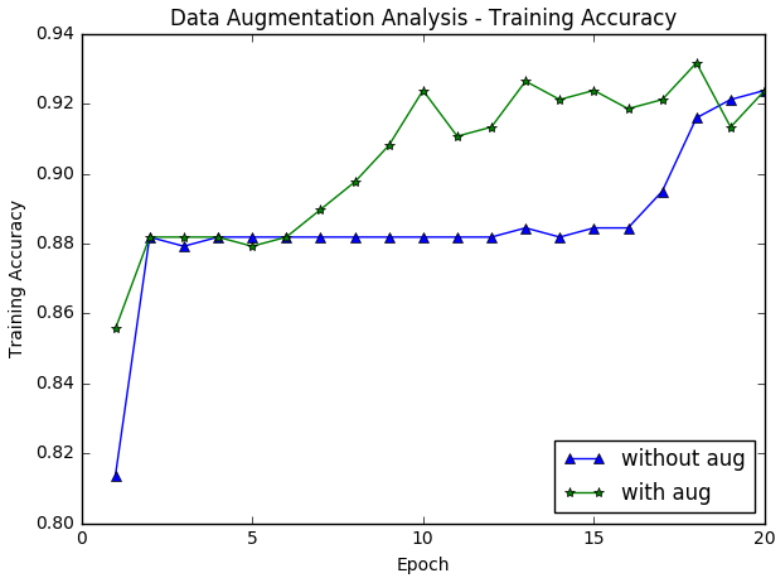


Figure 14. Data Augmentation Analysis - Learning curve of training accuracy.

However, according to the Table 4, the test accuracy for the CNN trained on the augmented dataset is higher rather than on the unaugmented one. It means that for this particular problem, augmentation helps to obtain better generalization for the model.

Table 4. Data Augmentation Analysis - Test Accuracy

Data Augmentation	**Test Accuracy**
without augmentation	67.425968109339408
with augmentation	71.75398633257403

7 CONCLUSION

Since convolutional neural networks show outstanding performance for different computer vision and pattern recognition tasks, this paper proposes an approach for fish detection and classification task using deep CNNs. By applying LeNet and AlexNet based CNN architecture, it is possible to detect and classify the fishes from low-quality underwater videos. The experiments are conducted by studying different aspects of CNN architecture. These aspects are the effect of activation function, filter size and data augmentation on the overall performance of the detection system. We have shown that tangent activation function superiors both ReLU and sigmoid both on the training and testing accuracy. In addition to that, it is revealed that the filter size does not play important role on the overall performance of the system. Finally, based on the conducted experiments, it can be said that data augmentation is an important preprocessing step to consider in terms of increasing the overall accuracy of the system. Although the system provides reasonable performance on the task of detection and classification of the fishes from low-quality video streams, the system suffers from the low amount of the data both for training and testing. From this perspective, it can be said that it is necessary to have large amount of data for deep learning applications to get better accuracy and performance.

As a future work, by considering recent development and performance of CNN on computer vision tasks, one can think of extension of this work to classify detected fish species and track them by applying deep learning methods.

REFERENCES

[1] Gu, Jiuxiang, et al. "Recent Advances in Convolutional Neural Networks."*arXiv preprint arXiv:1512.07108* (2015).

[2] Spampinato, Concetto, et al. "Detecting, Tracking and Counting Fish in Low Quality Unconstrained Underwater Videos." *VISAPP (2)* 2008 (2008): 514-519.

[3] White, D. J., C. Svellingen, and N. J. C. Strachan. "Automated measurement of species and length of fish by computer vision." *Fisheries Research* 80.2 (2006): 203-210.

[4] Matai, J., et al. "Automated techniques for detection and recognition of fishes using computer vision algorithms." *NOAA Technical Memorandum NMFS-F/SPO-121, Report of the National Marine Fisheries Service Automated Image Processing Workshop, Williams K., Rooper C., Harms J., Eds., Seattle, Washington (September 4–7 2010).* 2010.

[5] Spampinato, Concetto, et al. "Automatic fish classification for underwater species behavior understanding." *Proceedings of the first ACM international workshop on Analysis and retrieval of tracked events and motion in imagery streams*. ACM, 2010.

[6] Rodrigues, Marco TA, et al. "Evaluating cluster detection algorithms and feature extraction techniques in automatic classification of fish species." *Pattern Analysis and Applications*" 18.4 (2015): 783-797.

[7] Ogunlana, S. O., et al. "Fish Classification Using Support Vector Machine." *African Journal of Computing & ICT* 8.2 (2015): 75-82.

[8] Jäger, Jonas et al. "SeaCLEF 2016: Object Proposal Classification for Fish Detection in Underwater Videos." *CLEF* (2016).

[9] LeCun, Yann, et al. "Gradient-based learning applied to document recognition." *Proceedings of the IEEE* 86.11 (1998): 2278-2324.

[10] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*. 2012.

[11] S. V. L., "ImageNet large scale visual recognition competition 2012 (ILSVRC2012)," 2012. [Online]. Available: http://www.image-net.org/challenges/LSVRC/2012/. Accessed: Nov. 10, 2016.

[12] Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." *European Conference on Computer Vision*. Springer International Publishing, 2014.

[13] Szegedy, Christian, et al. "Going deeper with convolutions." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

[14] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).

[15] He, Kaiming, et al. "Deep residual learning for image recognition." *arXiv preprint arXiv:1512.03385* (2015).

[16] Ji, Shuiwang, et al. "3D convolutional neural networks for human action recognition." *IEEE transactions on pattern analysis and machine intelligence*35.1 (2013): 221-231.

[17] Karpathy, Andrej, et al. "Large-scale video classification with convolutional neural networks." *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2014.

[18] Schmidhuber, Jürgen. "Deep learning in neural networks: An overview." *Neural Networks* 61 (2015): 85-117.

[19] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. Large-scale video classification with Convolutional neural networks (CVPR 2014). Retrieved November 17, 2016, from http://cs.stanford.edu/people/karpathy/deepvideo/

[20] Simonyan, Karen, and Andrew Zisserman. "Two-stream convolutional networks for action recognition in videos." *Advances in Neural Information Processing Systems*. 2014.

[21] Yue-Hei Ng, Joe, et al. "Beyond short snippets: Deep networks for video classification." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

[22] Oquab, Maxime, et al. "Learning and transferring mid-level image representations using convolutional neural networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.

[23] Kang, Kai, et al. "Object detection from video tubelets with convolutional neural networks." *arXiv preprint arXiv:1604.04053* (2016).

[24] Jiang, H., & Wang, S. Object Detection and Counting with Low Quality Videos. Retrieved November 17, 2016, from http://cs231n.stanford.edu/reports2016/287_Report.pdf

[25] Forczmański, Paweł, Adam Nowosielski, and Paweł Marczeski. "Video Stream Analysis for Fish Detection and Classification." *Soft Computing in Computer and Information Science*. Springer International Publishing, 2015. 157-169.

[26] Fabic, J. N., et al. "Fish population estimation and species classification from underwater video sequences using blob counting and shape analysis." *Underwater Technology Symposium (UT), 2013 IEEE International*. IEEE, 2013.

[27] LeCun, Yann A., et al. "Efficient backprop." *Neural networks: Tricks of the trade*. Springer Berlin Heidelberg, 2012. 9-48.

[28] Kingma, Diederik, and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* (2014).