Department of Computer Science, Namal Institute Mianwali

FYP Mid-Report

Autonomous Checkout System

Students :
Muhammad Hazma
2016-UET-NML-CS-28
Mubashir Nisar
2016-UET-NML-CS-05

Supervisor :
Dr. Junaid Akhtar
Assistant Professor at Namal
Dr. Muhammad Hussain
Professor at King Saud University

Version 2
February-2020

Contents

1 Abstract

A system is proposed for the retailers with a computer-vision based system that keeps track of buyers and items they pick up in order that the customers may be charged rightly and save the retailers from shoplifting. The system will detect and classify the products customer picked and keep track of the customers and finally generate a bill for each customer once they reach the checkout point. The dataset used for the training of the Deep Convolutional Neural Network is curated using there different kind of images from real environment videos. CCTV video with the field of view of 45 degrees (Parallel to the products rack), Street view imagery (Perpendicular to the products rack) and CCTV captured videos with the field of view of 45 degrees (Diagonal to the products rack). One solution, this solution on the local system showing 99.9 percent accuracy for human detection and tracking, 89 percent accuracy for picked or unpicked products. Publicly available packages for deep learning like TensorFlow, Keras, OpenCV and Streamlit are used to construct an Autonomous Checkout System from a CCTV video. Pertinent to mention is the lack of expensive sensing devices (CCTV camera, real products, Products rack etc.) in our implementation, a stark shift from many of the previous studies along with demonstrable detection results with excellent potential for scalability of the proposed solution to work with hyper-spectral sensory data.

2 Acknowledgement

First of all, we are grateful to Almighty Allah, who helped us persevere and persists in the face of hardships and turmoils. Followed by our dedicated group of companions and friends who pushed us literally to achieve what our so easily would have given up on. Without you, we would not be here, thank you. This prose shall remain incomplete without expressing my gratitude to Dr Junaid Akhtar and Dr Muhammad Hussain for their unconventional yet charismatic and insightful mentorship throughout the course of this a semester... project. We are humbled and shall remain forever indebted. A faithful thanks to our parents who deserve acclaim for putting up with our tantrums, and whose sincere prayers and well wishes have propelled us to achieve great things. They shall always remain our shimmering light. Last but not the least, to all the marvellous individuals that remain anonymous on StackOverflow and other internet communities working hard and out of the passion to help advance the scientific cause and the people who write tutorials, and make exhaustive YouTube videos to make knowledge sharing easy. It is truly impossible to progress without the intellect and genius of such contributors. Thank you very much, whoever you are and wherever you are.

3 Problem Statement

According to reports, there are a total of 5,327 utility stores in the country out of which 4,470 are incurring losses. In Islamabad 494, Abbottabad 428, Peshawar 791, Quetta 265, Sargodha 607, Sukkur 280, Karachi 364, Lahore 508 and Multan has 679 of these stores which are incurring losses. These are the losses that facing every utility store in Pakistan. Also, everyone knows the mismanaged queues in Pakistan on every store. This project envisions the possibility to walk in, grab whatever you like and leave with a self-checkout system. But for starters, we want to confine the project to shops and bakeries that have a few products like Lays chips on racks and glass cased cold drink freezers located outside the shops. Given Pakistani society's ethical standing and economic situation of a common shopkeeper, shoplifting is a legitimate and serious issue. Some retailers do install cameras to hold a check on their products however that's not enough because there needs to be someone who's manually looking at the display screen all the time. Keeping in view these serious issues there is a need to develop an intelligent camera-based system that keeps track of customers and the items they picked from outside the store, to help both the retailers as well as shoppers.

4 Objectives

The objective of this project is to provide retailers with a computer-vision based system that keeps track of buyers and items they pick up in order that the customers may be charged rightly and save the retailers from shoplifting. First of all, the system will detect the human from CCTV video also track the customer in front of the camera. If the customer interacted with the products rack then the system will detect and classify the products customer picked and keep track of the customers and products picked the customer and finally generate a bill for each customer once they reach checkout point.

5 Introduction

This system includes our models for person detection, entity tracking, item detection, item classification, and ownership resolution, all working together to visualize which person has what item in real-time.

6 Methodologies

6.1 Computer Vision

Computer vision, in its simplest form, is the ability of a computer to "see" and process visual data. More formally, it is an interdisciplinary field of study aimed at understanding why the human visual system works and ultimately

replicating it. It is the process of extracting, analyzing and understanding visual information present in images or videos. A digital image can be defined as a continuous sample of light radiations reflected from different objects onto a rectangular array of pixels. Digital signal processing can be applied to extract information and generate meaning out of these numbers.

With the deep learning revolution, the domain of computer vision has progressed exponentially and can now boast about object detection and scene comprehension from videos and images, alike. In order to understand an image, three primary levels of abstraction are utilized to make sense of the sensory data, at the bottom includes the basic components of an image i.e. edges, texture elements or regions, a higher level of abstraction is inclusive of boundaries, surfaces and volumes, while the highest level of comprehension includes objects, scenes, or events recognition.

6.2 Neural Networks

A neural network mimics the human brain. The neural network learns via an algorithm by incorporating new data. A neuron also is known as a node or unit and is the basic unit of computation in a neural network. These neurons are then connected into a large mesh network. The nodes receive inputs from an external source or from some other nodes and compute an output. All these nodes have some unique weights (w), which is assigned based on its relative significance to other inputs. The nodes apply a function to the weighted sum.

So x1, x2 and x3 are the three inputs, where w1, w2 and w3 are the associated weights to the three inputs respectively. Additionally, another input is 1 with weight b, known as the bias. The bias provides a trainable constant value to every normal input that the node receives. This bias function value allows moving the activation function to the left or right. The above f function is a non-linear activation function. An activation function introduces non-linearity into the output of a node because most of the real-world problems are non-linear.

In practice, there are several activation functions e.g. sigmoid, tanh and ReLU.

6.3 Methodologies Data Augmentation

Any neural network can be thought of as a network of "neurons" which is organized in layers and those layers are defined as follows:

6.3.1 Input Layer

It takes predictors/inputs and attaches coefficients to them which are called "Weights".

6.3.2 Hidden Layer

It makes the neural network non-linear and improves performance. Neural-Network can be linear with no hidden layer.

6.3.3 Output Layer

It outputs the results generated from weighted predictors passed through hidden layers, if any.

6.3.4 Back-propagation

A procedure to repeatedly adjust the weights so as to minimize the difference between actual output and desired output.

6.3.5 Error

It is calculated by taking the difference between desired output and output predicted by neural network. It creates a gradient descent which can help in altering weights.

6.3.6 Weights

These are altered by method back-propagation. In this way, an error can be removed. Then gives multiple iterations to this process and finally an accurate model is trained.

There are several types of neural networks, each having specific use and flexibility. The most common neural networks are; feed-forward neural network, recurrent neural network, and convolutional neural network and Boltzmann machine network.

6.4 Difference between Traditional ML and ANN

The major difference between traditional ML techniques and ANN is 'feature extraction'. In traditional ML techniques, features are extracted manually by using different methods but in ANN features are extracted automatically by adjusting the weights. Each layer in ANN learns particular features and uses it to identify objects. Figure-no below shows the major difference between traditional ML and ANN.

A Rectified linear unit (ReLU) takes a real-valued input and threshold it with zero, replacing every negative value with zero. This operation has been used after every convolution operation. As convolution is a linear operation but most of the real-world problems are non-linear, so for non-linearity, ReLU is introduced. For a processing task in deep learning, neural networks cannot be programmed directly like other algorithms. Neural networks use some strategies to learn information e.g. supervised learning, unsupervised learning and reinforced learning.

The sigmoid function produces similar results to step function in that the output is between 0 and 1. The curve crosses 0.5 at z=0, which we can set up rules for the activation function, such as: If the sigmoid neuron's output is larger than or equal to 0.5, it outputs 1; if the output is smaller than 0.5, it outputs 0.

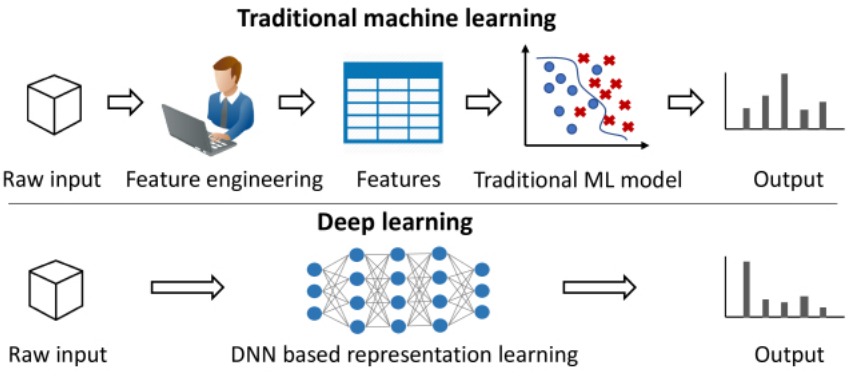

Figure 1: Traditional ML vs ANN

Neural Networks' behaviour is determined by its neurons' particular transfer functions, by learning rule and architecture. Transfer function, also called Activation functions, defines the output of given input or set of inputs to a node of neurons is shown in Figure below:

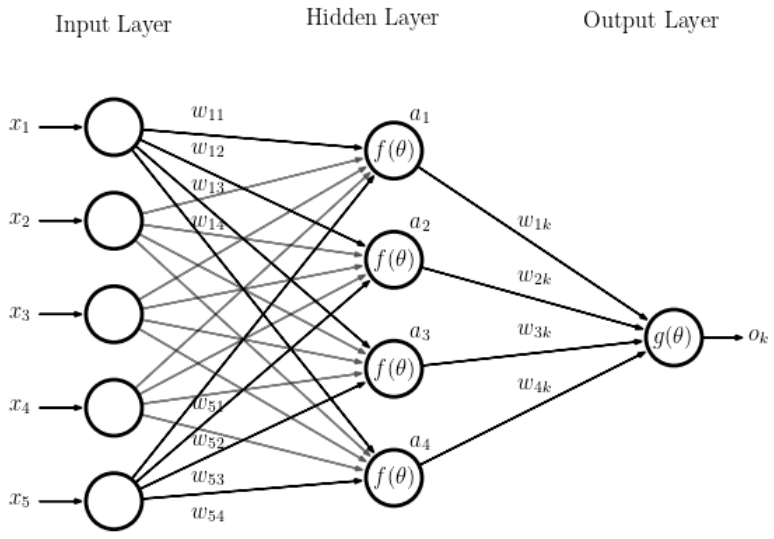

Figure 2: Basic Functioning of Neural Network

Applications of the neural networks can be summarized into classification or pattern recognition, prediction and modelling.

6.5 Convolutional Neural Networks (CNN/ConvNets)

Convolutional neural networks are the same as other neural networks, made up of neurons and synapses (weights). Each neuron receives an Input, multiplies it with weights and applies an activation function on it. Error is calculated and weights are altered according to the desired output. The change between CNN and other neural networks is that ConvNet architecture. ConvNet is assumed that it takes inputs in the form of images. CNN/ConvNet architecture has three main layers; convolutional layer, pooling layer and fully-connected layer.

6.6 Convolutional layer

It is a mathematical operator which multiplies the image matrix with kernels/filters to extract features from an input image. Figure-no below shows the multiplication of the image pixels matrix and filter matrix.

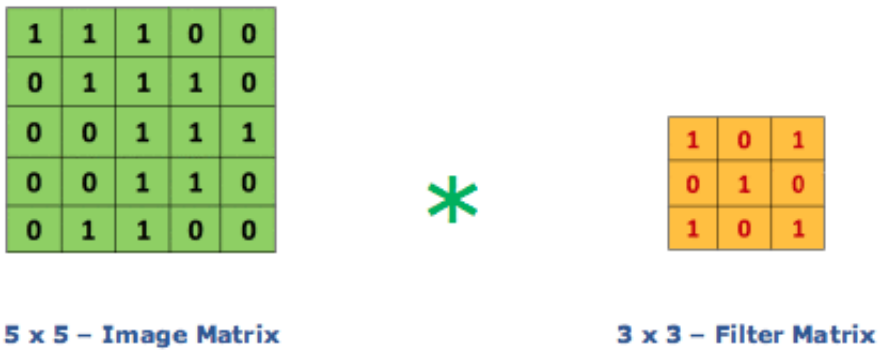

Figure 3: Convolutional Layer Operation

Filter matrix move across the whole matrix of image and gives convolved feature matrix as shown in Figure-no

At this layer, Rectified Linear Unit (ReLU) such as f(x)=max(0, x) is applied element wise to introduce non-linearity.

6.7 Pooling Layer

In a case of large or complex images, pooling layer reduces the number of features without losing important features information. It has three types; Max pooling, Average pooling and Sum pooling.

6.8 Fully Connected Layer

Fully connected layer combines all the features coming from convolutional layer and pooling layer and develops a trained model. Then the activation functions like sigmoid or softmax are applied to detect and classify objects.

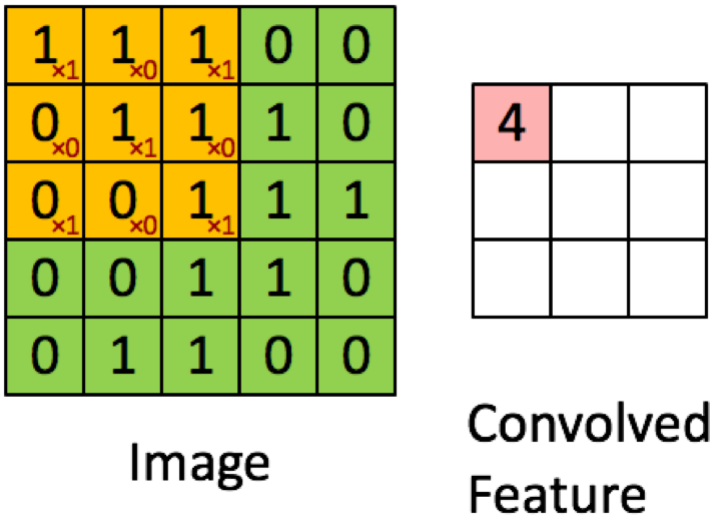

Figure 4: Convolved Features

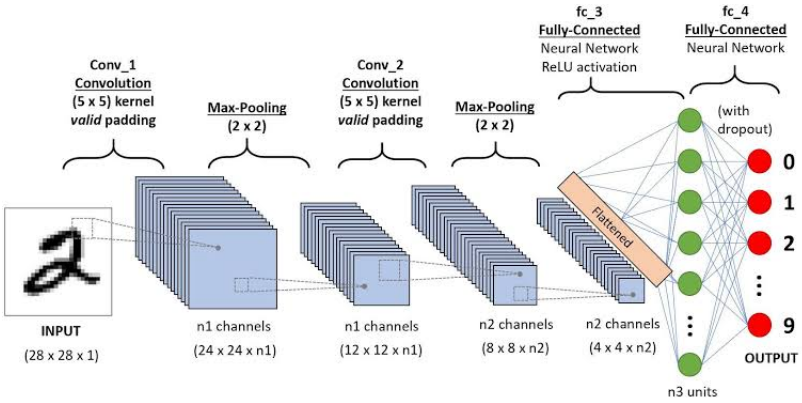

Figure 5: The whole process of Convolutional Neural Networks

7 Literature Review

7.1 Machine Learning

Since the beginning of the industrial revolution, machines have been the helping hand of humans. We use various machines from start to the end of the day. Almost two decades before, machines were dumb and operating manually

but now they are more flexible and intelligent. Understanding texts, driving cars, identifying specific person etc. all are the tough tasks which machines are doing now. This is all happening because of 'Machine Learning'. According to PaperName machine learning is a branch of artificial intelligence which makes machines to learn and act like humans and improve their learning in an autonomous way by feeding them data in the form of observations, statistics and real-world interaction. It has been categorized into the following groups.

7.2 Supervised Learning

In supervised learning, machine concentrates on learning patterns by connecting the relationship between variables and known outputs. It works by feeding machine with input features(X) and the correct value of output(Y). Then the machine learning algorithm makes patterns and develops a model which will give an intelligent output on new data. A very simple example is a training system on images and labels. Then in future, if you'll give a new image, it will intelligently recognize the new image.

7.3 Unsupervised Learning

In unsupervised learning, we do not feed machines with all the variables and known outputs. Instead, we let the machines to uncover hidden patterns and generate labels through unsupervised learning algorithms. One of the examples is the K-means clustering algorithm which groups the data points which have similar features. Figure-2 below shows the difference between supervised and unsupervised learning.

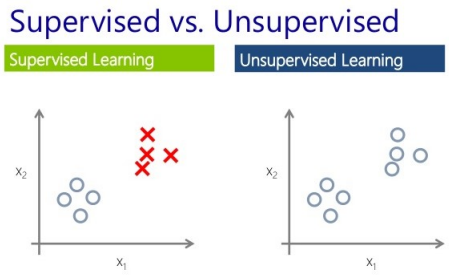

Figure 6: Supervised and unsupervised learning

7.4 Reinforcement Learning

This type of learning is quite different and advanced as compared to supervised and unsupervised learning. In reinforcement learning, machine continuously improves its model by leveraging feedback from the previous iteration. It can be explained with an example of a self-driving car. If in any case, car avoids

crash then it will learn the value of this action. Next time, if the car will face the same situation then it will take the same action to avoid that accident.

7.5 Machine Learning Contributions

Machine learning plays a significant role in our daily lives. From transportation to health care, education to public safety, it has improved our lives. In transportation, autonomous cars are in the market. Cars are becoming better drivers than men and saving a lot of time for other activities. In Healthcare, we have monitoring devices and mobile applications which keep the track of our daily health routine. Moreover, robots are also assisting doctors in various surgeries. In Education, we have interactive tutoring machines capable of teaching science, math and other disciplines. Further, classrooms have an intelligent system which continuously informs the teacher about the interest rate of students on the basis of facial expressions. In Public Safety, smart cameras and drones are being used for surveillance. Moreover, algorithms for detecting and avoiding financial fraud and privacy interference are also available. Besides these, machine learning is contributing in other areas also. One of them is in advertising.

8 Proposed Framework

The proposed system includes person detection, picked and unpicked classification, product detection, product tracking, product classification, and owner resolution, these all models will work together to visualize which person has what product in real-time.

9 Work being done

The work that we have done so far is real-time person detection and product picked and unpicked classification. For the purpose of person detection, we have used the MobileNet-SSD pre-trained model, and for our second part picked and unpicked classification we have used CNN Binary Classifier trained on our own dataset that we have described ahead.

10 Dataset Collection

We have collected datasets of two different environments, one using DSLR and the other using mobile-cam. In the first environment, we had lays chips rack that was filled with chips packets, for this environment we have shot two videos. In the first video, only one person goes and picks up a packet of chips, while in the second video, more than one person used to pick up chips packets.

In the second environment, we had a shelf in which two different kinds of products were placed. For this environment too, we have shot two videos, In

Figure 7: Dataset 1

the first video, only one person goes and picks up a product, while in the second video, more than one person used to pick up products.

Figure 8: Dataset 2 Frames

11 Dataset Labeling

Image labelling or annotation is the process of defining regions in an image and giving them a textual label for creating ground truth. One of the many magnificent features of the humans' visualization is to create a rich description of a scene within a glance at an image. They can tell what are the objects and their associated attributes in the image. For example, an image can be defined as "containing a person, picking a product from the rack". Moreover, humans can effortlessly describe each object in the image. During the past few decades, one of the fundamental objectives of computer vision research has been to replicate this ability, resulting in in-depth studies of computer vision problems including detection of objects in a scene by describing them using their attributes and creating the ground truth.

OpenCV is the most common library for Computer Vision problems. We have extracted 2400 frames from both videos using OpenCV. After extracting frames from the videos we labelled these frames manually in bases of product

picked or unpicked. Our dataset split for training is 80 percent and 20 percent for testing.

12 The Streamlit

Streamlit is an open source app framework specifically designed for ML engineers working with Python. It allows you to create a stunning looking application with only a few lines of code (www.streamlit.io). It supports hot-reloading, so your app updates live as you edit and save your file. No need to mess with HTTP requests, HTML, JavaScript, etc. All you need is your favorite editor and a browser.

We used Streamlit for live demo. Using Streamlit we show the sidebar which represents the all testing images using a simple slider. In center vertical order we show the live demo of videos first single person in frame, two person in one frame, and Single person in different environment dataset video. After videos we show the complete procedure on single frame. Before processing a single frame, Person detection from a single frame, person cropping, and then show result picked or unpicked.

13 Flow Chart

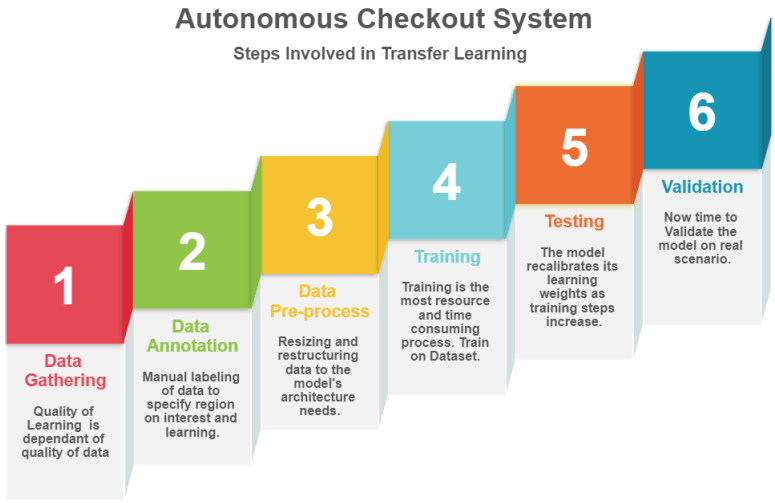

Figure 9: Steps we are following

14 Picked and unpicked classification

Our essential undertaking is to make an algorithm to classify whether a picture contains a person with a picked product or unpicked. The input for this task is pictures of picked or unpicked from the training dataset, while the output is the classification accuracy on the test dataset. The given dataset for this task is gathered by ourselves, clarified previously. Our training set contains 2,400 pictures, including 1,200 pictures of picked and 1,200 pictures of unpicked, while the test dataset contains 400 pictures.

Our principle task is to learn a classification model to determine the decision boundary for the training dataset. The input for the learning task is pictures from the training dataset, while the yield is the learned classification model. Our end task is to apply the learned model to classify pictures from the test dataset weather picked or unpicked, and then evaluate the classification accuracy.

15 Deep Neural Networks

The CNN is a sort of deep architecture that has accomplished extraordinary performance in tasks like document recognition and image recognition. Not the same as customary BP Neural Networks, which contains input layer, hidden layers, and output layer, CNN likewise contains Convolutional layers and Max Pooling layers.

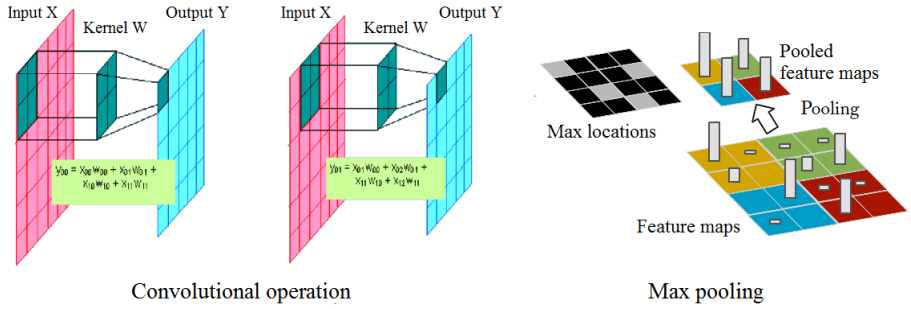

Figure 10: Convolutional Operation Max Pooling

Convolutional layers contain many feature maps, which are two dimensional hidden nodes. Each feature map claims a weight matrix called kernel, and different feature maps owns different kernels. Kernels do convolutional activity with each feature map in previous layer (layer j), at that point we summarize them and put it into sigmoid function. The output is the pixel value in layer j+1. With various kernels, we can learn various representations of information and the measure of parameters isn't expanded exponentially with the number of hidden nodes and layers.

When a feature has been detected, its exact location becomes less significant. Just its approximate position comparative with different features is relevant. Not exclusively is the exact position of every one of those features unimportant for recognizing the pattern, yet it is additionally conceivably unsafe in light of the fact that the positions are probably going to change for various instances of the pattern. Max pooling layers do sub-sampling procedure on feature maps. For each four pixels, we just hold the maximum value, with the goal that the size of feature maps will be half of the first size. Sub-sampling lessens the resolution of feature maps and decreases the affectability of the output to movements and distortions with the goal that the model will be increasingly robust. Max pooling activity can be joined into convolutional layers, and we needn't bother with extra layers to do sub-sampling.

16 Model

The original model contains 4 layers and the last two layers are one fully connected layer and output layer. We added one more fully connected layer and Dropout layer to avoid over fitting. This deep convolutional neural network trained by ourselves on our own dataset.

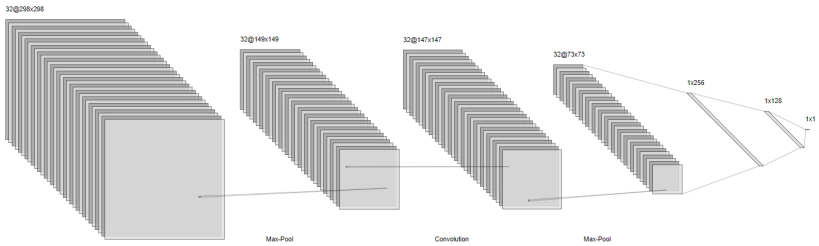

Figure 11: Model View

The architecture of this model is outlined in the above figure. In the first place, input pictures are normalized into 298 x 298 then the first convolutional layer filters the 298 x 298 x 3 input picture with 32 kernels of size 3 x 3 with a stride of 4 pixels. The second convolutional layer takes as input the pooled output of the first convolutional layer and filters it with 32 kernels of size 3 x 3. The Dense layers have one layer with 256 neurons and second layer with 128 neurons and finally the output layer. In the output layer we have 1 neuron and sigmoid activation, which means this will give us out as 0 or 1 (binary ouput).

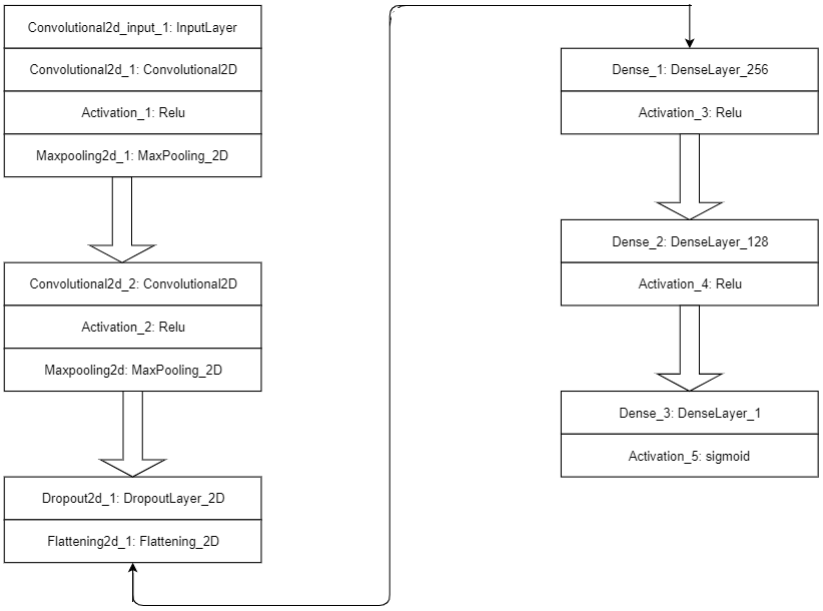

Figure 12: Picked-Unpicked Model Architecture

17 Model Architecture

17.1 Convolutional Layers

17.1.1 2 Layers

1. 32 Filters of 3x3 size on First Layer 2. 32 Filters of 3x3 size on Second Layer All convolutional layers have pooling of 2. Dropout Layer with keep probability of 0.5.

17.2 Fully Connected Layers

17.2.1 3 Layers

1. 256 Neurons with Relu Activation Function at first layer 2. 128 Neurons with Relu Activation Function at second layer 3. 1 Neuron with Sigmoid Activation Function at output layer

18 Results and Analysis

By using this model we have got very impressive results with the training accuracy 98 percent and testing accuracy 88 percent. Following are graphs of training and validation accuracy and training loss and validation los respectively.

Figure 13: Number of Epochs

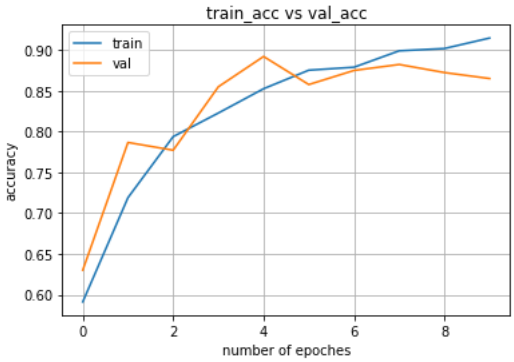

Figure 14: Number of Epochs

As mentioned above that we have dataset of two different environments for product picked or unpicked classification, but only one environment dataset was used for training purpose while data of the other environment used for testing purpose on which the model have given very good results.

Figure 15: Results

References

[1] Beaulieu, K. and Dalisay, D. (2020). Machine Learning Mastery. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/ [Accessed 25 Feb. 2020].

[2] Khan, M., Awais, M., Shamail, S. and Awan, I. (2011). An empirical study of modeling self-management capabilities in autonomic systems using case-based reasoning. Simulation Modelling Practice and Theory, 19(10), pp.2256-2275.

[3] Lecun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), pp.2278-2324.

[4] LV, J. (2019). Visual Attentional Network and Learning Method for Object Search and Recognition. Journal of Mechanical Engineering, 55(11), p.123.

[5] Marino, D., Ireifej, S. and Loughin, C. (2011). Micro Total Hip Replacement in Dogs and Cats. Veterinary Surgery, 41(1), pp.121-129.

[6] Radovic, M., Adarkwa, O. and Wang, Q. (2017). Object Recognition in Aerial Images Using Convolutional Neural Networks. Journal of Imaging, 3(2), p.21.

[7] Ren, S., He, K., Girshick, R. and Sun, J. (2017). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6), pp.1137-1149.

[8] Datasciencemasters.org. (2020). The Open Source Data Science Masters by datasciencemasters. [online] Available at: http://datasciencemasters.org/ [Accessed 25 Feb. 2020].

[9] thika, R. and lakshmi, S. (2017). Object Detection and Semantic Segmentation using Neural Networks. International Journal of Computer Trends and Technology, 47(2), pp.95-100.

[10] Triana, E. and Pasnak, R. (1981). Object permanence in cats and dogs. Animal Learning Behavior, 9(1), pp.135-139.

[11] Youtube.com. (2020). YouTube. [online] Available at: https://www.youtube.com/ [Accessed 25 Feb. 2020].