12-2019

Detection of Distracted Pedestrians using Convolutional Neural Networks

Igor Grishchenko
Sheridan College

DETECTION OF DISTRACTED PEDESTRIANS USING CONVOLUTIONAL NEURAL NETWORKS

A Computer Science Thesis

Presented to

The Faculty of Applied Science and Technology, School of Applied Computing

of

Sheridan College, Institute of Technology and Advanced Learning

by

GRISHCHENKO, IGOR

In partial fulfillment of requirements

for the degree of

Bachelor of Computer Science (Mobile Computing)

December 2019

ABSTRACT

DETECTION OF DISTRACTED PEDESTRIANS USING CONVOLUTIONAL NEURAL NETWORKS

Igor Grishchenko
Sheridan College, 2019

Advisor:
Dr. El Sayed Mahmoud

The risk of pedestrian accidents has increased due to the distracted walking increase. The research in the autonomous vehicles industry aims to minimize this risk by enhancing the route planning to produce safer routes. Detecting distracted pedestrians plays a significant role in identifying safer routes and hence decreases pedestrian accident risk. Thus, this research aims to investigate how to use the convolutional neural networks for building an algorithm that significantly improves the accuracy of detecting distracted pedestrians based on gathered cues. Particularly, this research involves the analysis of pedestrian' images to identify distracted pedestrians who are not paying attention when crossing the road. This work tested three different architectures of convolutional neural networks. These architectures are Basic, Deep, and AlexNet. The performance of the three architectures was evaluated based on two datasets. The first is a new training dataset called SCIT and created by this work based on recorded videos of volunteers from Sheridan College Institute of Technology. The second is a public dataset called PETA, which was made up of images with various resolutions. The ConvNet model with the Deep architecture outperformed the Basic and AlexNet architectures in detecting distracted pedestrians.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

This work has resulted in Igor Grishchenko's undergraduate thesis of the Honours Bachelor of Computer Science (Mobile Computing). The collection of data from human participants was approved by the Sheridan Research Ethics Board. We are grateful for the insightful feedback on the work received from the thesis advisory committee members at Sheridan College.

CHAPTER ONE

1. INTRODUCTION

1.1 The Problem Context

Pedestrians are the most vulnerable objects observed by autonomous vehicles because they travel along streets, roads, sidewalks, alone and with others in both busy and idle areas. According to the Canadian Motor Vehicle Traffic Collision Statistics 2017 (Transport Canada, 2017), pedestrians accounted for 15.4% of fatalities and 14.3% of serious injuries in all motor vehicle accidents. Pedestrians can make changes in their path because many roads and streets cannot have physical constraints that ensure pedestrians use the appropriate behavior all the time. This makes planning a safe route challenging even with all the current technologies equipped to self-driving cars today.

One of the main reasons for the difficulties in detecting and predicting pedestrian behavior is attributed to the use of mobile devices while walking. Pedestrians who use handheld devices tend to walk blindly into the path of a moving vehicle. Doing so increases the likelihood of a collision. Using devices while walking limits pedestrian cognitive functions which in turn could lead to walking with high risk to cause the accident (Yogev-Seligmann, Hausdorff, & Giladi, 2012). Figure 1 shows a significant upward trend of pedestrians' injuries related to cell phone use from 2004 to 2010 based on data gathered in research conducted by Nasar et al. (Nasar & Troyer, 2013).

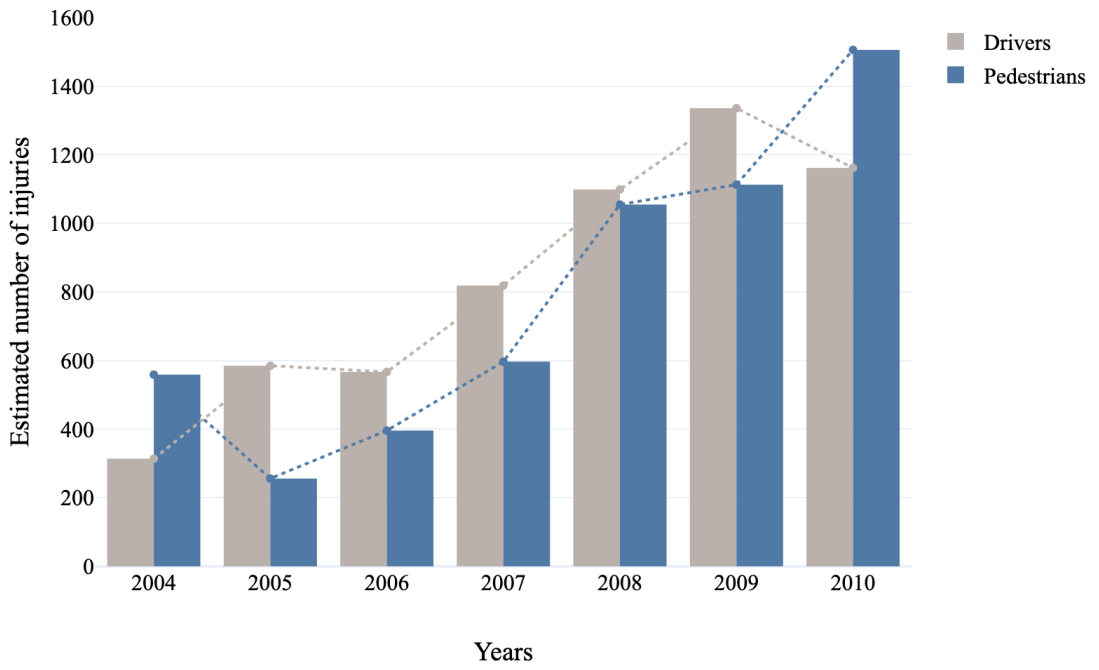


Figure 1. The growth of pedestrians and drivers' injuries associated with cell phone use

The chart in Figure 1 is a statistical analysis of injuries related to cell phone usage for both pedestrians and drivers, where the total number of traumas for pedestrians has increased yearly from 2005 to 2010. According to the study conducted by (Brenner & Smith, 2014), cell phone usage increases yearly. In support, research conducted by (Bassett et al., 2010) also found out that more and more people are beginning to walk more often and thus spend less time driving every day. Conclusively, the number of distracted pedestrians is constantly growing.

The use of handheld devices by pedestrians affects their cognitive load and the ability to pay close attention to the road, thus, increases the car accident risk. This creates a further challenge for the self-driving car to plan the safest route because the walking path of a distracted pedestrian is not related to the current road conditions. Identifying pedestrians who use cell phones during their walking will significantly decrease the number of injuries and deaths due to distracted

pedestrians. This study developed and trained a Convolutional Neural Network (CNN) to detect pedestrians who use handheld devices while crossing the road. Ultimately, this work developed the distracted pedestrian detector, based on convolutional neural networks, which is able to analyze whether the pedestrian is distracted or not in real-time.

1.2 Terms and Definitions

Table 1. Terms and Definitions

Handheld device	A mobile device, a portable computer or headset that can be used on the go.
Distracted pedestrian	A pedestrian who is using a handheld device while walking or crossing the street.
Convolutional Neural Network/CNN/ConvNet	A class of deep neural network which usually applied to image classification problems.
Detector	An algorithm implemented using convolutional neural networks to detect distracted pedestrians.
Autonomous vehicle/self-driving car	A vehicle that is capable to drive and sense its environment with little or no human supervision.
MLP	A classical type of neural network comprised of one or more layers of neurons.

Benchmark	A feature that can distinguish a distracted pedestrian from not distracted.
Fine-tuned model	A pre-trained convolutional neural network which can be continuously trained with new data to be adapted to new tasks.
Scenario	Behavior when a pedestrian is considered to be distracted based on their hands and head positioning.
Matrix	Technique for summarizing the performance of a classification algorithm.
Data Set	Collection of images of distracted and non-distracted pedestrians for model training and testing.
Python	General-purpose programming language
Keras	Open-source deep neural network library that is written in Python.

1.3 Problem Statement

The number of pedestrians, who use handheld devices while walking or crossing the road, increases every year. Distracted pedestrians are more likely to commit senseless motions that make it difficult for autonomous vehicles to predict the safe route. This work developed the distracted pedestrian detector, based on ConvNets, which identifies whether a pedestrian is using the handheld device or not. The detector can assist self-driving cars/distracted drivers in detecting pedestrians and obtaining the safest route to avoid an oncoming collision. Eventually,

this detector can further advance the general accuracy of the autonomous vehicles. The detector warns if a pedestrian is distracted or not. This is the critical piece of information that can significantly advance the safe route planning by autonomous vehicles.

1.4 Purpose

This thesis developed the detector using convolutional neural networks (CNN) for identifying distracted pedestrians. The algorithm analyzes a pedestrian video stream and identifies if the pedestrian is using a handheld device at the moment. This can assist autonomous cars in detecting pedestrians and allow vehicles to plan routes safely and efficiently. The CNN model was selected to detect distracted pedestrians. The main goal is to significantly improve autonomous vehicles' accuracy in distracted pedestrian detection.

1.5 Motivation

The motivation of this thesis was to improve the safety of pedestrians by leveraging the convolutional neural networks. The application of convolutional neural networks could improve the accuracy of detecting pedestrians and identify if they are distracted. The ConvNet investigates image structural information and builds the neural network model in a more insightful manner than non–deep neural networks (Dominguez-Sanchez, Cazorla, & Orts-Escolano, 2017). Today, research on the detection of pedestrian motions and route planning is conducted frequently with many readily available publications. However, only a few mention the fact that pedestrians can be distracted and how their behavior and movement can and may change unexpectedly due to cognitive dissonance. This study investigates the problem of distracted pedestrians by implementing the detector based on the ConvNets, which can identify

whether the observed pedestrian is holding the handheld device or not. Stakeholders who benefit from the proposed algorithm are the vehicle manufactures, smart cities project teams, and researchers. As mentioned previously, drivers are also distracted by handheld devices as well. Thus, the developed algorithm can also be applied to warn a driver if a pedestrian is distracted and the chance of accident will overall decrease. The main goal of this work is to improve the accuracy of automated vehicles to make their choices safer and minimize the possibility of injury.

1.6 Proposed Work

This thesis consists of three main stages. The first stage intended ascertaining the scenarios of distracted behavior and conducting the experiment to collect images of both distracted and non-distracted pedestrians. This stage was heavily based on identifying the cases when a pedestrian is considered to be distracted. Once we formulated the definition of distracted behavior, we asked experiment participants to mimic distracted and not distracted behavior while they were videotaped. These experimental sessions allowed us to collect high-quality data with different diversities including the age, gender, and race of participants in the variety of foreshortening. The second phase was the selection of suitable architecture for the convolutional neural network (ConvNet) model which can perform the accurate classification of distracted pedestrians. We started from the few convolutional layers and a small number of filters to understand in which direction we should expand our architecture so that the model could extract the most relevant features from images. During the third stage, we focused on training and tuning the ConvNet on collected images of both distracted and non-distracted pedestrians. The selection of hyperparameters, particularly network weight initialization, for the ConvNet model was one of

the most important parts of the research. This thesis considered three different activation functions for convolutional layers including ReLU, sigmoid, and tanh.

A ConvNet was selected for this research because it uses structural information of the image and organizes the neural network model more intelligently than the classical neural network (MLP). Another advantage of the ConvNet model is the ability to have fewer parameters which reduces training time. Therefore, CNN has enough weights to focus on small parts of the image without considering the weights of each pixel. In this case, feature extraction has been to be strongly related to the identification of distracted pedestrians.

1.7 Thesis Statement

The convolutional neural networks showed promising results in classification of distracted pedestrians in pedestrian images. Convolutional layers performed well in extracting and analyzing features from an image that empowered to focus on the important details. Extracted features from an image of a pedestrian crossing a road have been used to determine if the person is distracted. Identifying distracted pedestrians can assist in improving route planning for pedestrian safety. This study promotes pedestrian safety by improving the accuracy of detecting whether a pedestrian is distracted or not.

1.8 Contributions

This thesis has investigated the application of ConvNet models for detecting distracted pedestrians. The contributions of this work are the following:

- Optimizing the convolutional neural network architecture for detecting distracted pedestrians.

- Identifying suitable hyper-parameters which enhanced the forecasting accuracy of the model.
- Creating the dataset of images of distracted and not distracted pedestrians that can be used for further researches.

1.9 Organization of Thesis

The rest of this paper is organized as follows, the literature review chapter covers prior researches related to autonomous vehicles and the detection of pedestrians by examining different techniques such as neural nets (MLP), knowledge extractions, and model tuning. It consists of studies that focus on human cognition research and how handheld devices can lead to unwilling motions while walking. The methodology chapter focuses on describing what methodology was used and how it was applied in detail. This involves the selection of ConvNet architecture, model training and tuning as well as testing the detector on the videos of participants. Lastly, the results chapter presents the gathered experimental findings, a review of the findings with analysis and future research opportunities.

CHAPTER TWO

2. LITERATURE REVIEW

With the sharp growth of self-driving cars in the automotive industry and the increasing usage of handheld devices by pedestrians, the ability for autonomous vehicles to detect distracted pedestrians has become prevalent, hence receiving a considerable amount of attention and extensive research on determining whether the pedestrian is distracted or not. (Dominguez-Sanchez, Cazorla, & Orts-Escolano, 2017) (Tang, Ma, Liu, & Zheng, 2018).

Many research groups concentrated on the challenge of determining the limb positioning of a pedestrian for a long time and introduced a variety of models. Some studies applied classical machine learning algorithms by fitting labeled data into models, such as Gaussian process (GP) regression (Chen, Liu, Liu, Miller, & How, 2016), Support Vector Machines (SVM) (Wang, Chen, Chen, & Yang, 2012), and Mixed Markov-Chain Model (MMCM) (Asahara, Maruyama, Sato, & Seto, 2011). Other groups conducted research considering deep neural networks. Dominguez-Sanchez, Cazorla, & Orts-Escolano (2017) conducted research for the improvement of pedestrians' motions detection by leveraging convolutional neural network (CNN). Another approach proposed by Yamashita, Fukui, Yamauchi, & Fujiyoshi (2016) involves the use of Multi-Task Convolutional Neural Network for the detection of pedestrians and the position of their limbs simultaneously. The latter two approaches will be considered the closest to this study and will be the focus of this study's research.

It is essential to detect distracted pedestrians since it can help to prevent vehicle conflicts and reduce vehicle traffic due to indecisions when crossing and overall slower crossing speed (Zaki & Sayed, 2016). According to Zaki and Sayed, this type of research would benefit multiple domains which include road safety which extends the application of computer vision (CV). The

potential improvement of the current methodology for identifying distracted pedestrians would be the exploration of head and hands positional tracking (Zaki & Sayed, 2016).

With the growth of autonomous cars in the motor vehicle industry and the increasing number of distracted pedestrians, the importance of this research as well as the understanding and analysis of the distracted walking behavior of pedestrians have been more than reaffirmed. Recent studies about the exploration of pedestrians' gait benchmarks for the identification of whether they are distracted or not has been completed (Zaki & Sayed, 2016).

A survey of theory and practice in the interaction between self-driving cars and pedestrians conducted by Rasouli and Tsotsos showed that pedestrians who are distracted by handheld devices are 75% more likely to display unintentional blindness (Rasouli & Tsotsos, 2018). Another study conducted by Neider et al. investigated that distraction arising from the cell phone usage challenges pedestrians' ability to estimate the time-to-contact of traffic accurately, which increases the odds of failing to cross a road safely. Figure 2 visualizes the results gathered by Neider et al. during the research experiments and shows the percentage of attempts in which participants successfully crossed the street (Neider, McCarley, Crowell, Kaczmarski, & Kramer, 2010). Figure 2 demonstrates that pedestrians who were talking on the phone while crossing the street were less likely to successfully cross the road compared to non-distracted pedestrians (Neider, McCarley, Crowell, Kaczmarski, & Kramer, 2010).

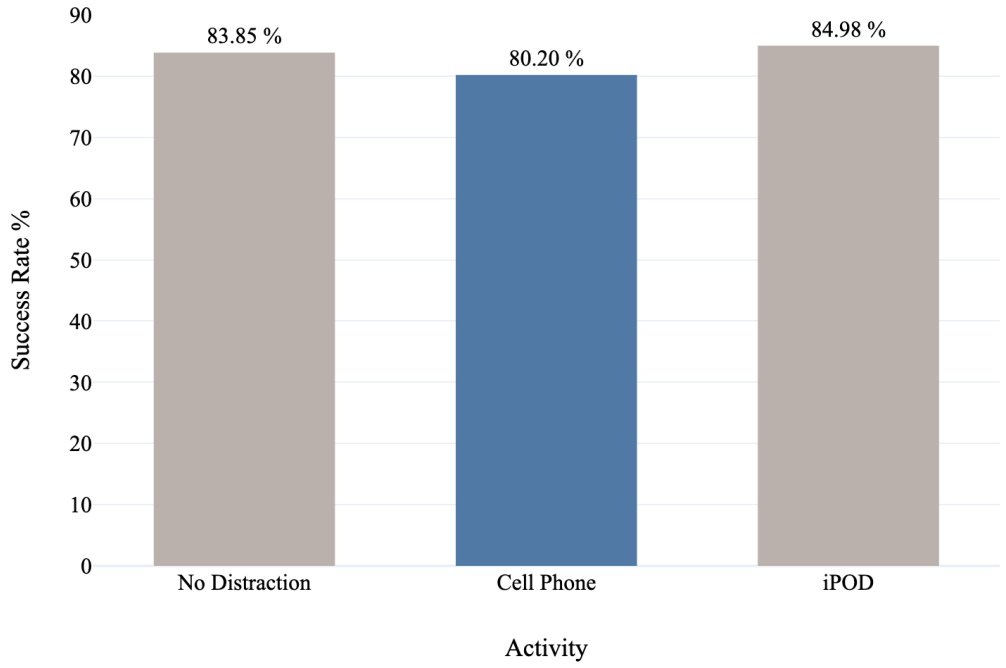


Figure 2. The percentage of pedestrians' success crossing while being distracted or not

Distracted pedestrians tend to change their walking direction more often and on average, cross the street slower than undistracted pedestrians, which can lead to unwilling accidents (Rasouli & Tsotsos, 2018) (Zaki & Sayed, 2016). The ability of autonomous cars to detect pedestrians who are not paying attention while crossing the road can improve road safety. Since the motor vehicle industry is steadily shifting towards self-driving cars, these autonomous cars must recognize if a pedestrian is not paying attention to the road, in order to prevent any hazards associated with distraction (Rangesh, Ohn-Bar, Yuen, & Trivedi, 2016). Current studies focus on analyzing pose and extracting gait parameters of pedestrians to determine whether the pedestrian(s) is distracted or not (Rangesh, Ohn-Bar, Yuen, & Trivedi, 2016) (Zaki & Sayed, 2016).

This study's intention is to improve self-driving cars' accuracy in collisions detection and path planning by identifying whether the pedestrians are distracted or not. The main goal of this work is to use a convolutional neural network model to detect distracted pedestrians by examining specific distracted behavior scenarios of pedestrians.

2.1 Convolutional Neural Networks in Computer Vision

Deep Convolutional Neural Networks (ConvNet) has demonstrated amazing performance in several computer vision tasks, including face recognition, digits recognition, and image classification, due to the ability to extract visual benchmarks from the pixel-level content (Tome, Monti, Baroffio, Bondi, Tagliasacchi, & Tubaro, 2016). However, it was a great challenge to train the deep ConvNets due to the lack of training data and computational power in the past, but many methods had been proposed to overcome this problem since 2006 (Gu et al., 2018). In 2012, Krizhevsky et al. proposed a classic ConvNet architecture, AlexNet, and demonstrated notable improvements in the image classification tasks (Russakovsky et al., 2015). AlexNet showed high levels of accuracy in image recognition applications and received considerable attention from the community, and therefore, many studies were conducted to improve or even surpass AlexNet's performance. Subsequently, more effective and deeper ConvNet architectures were proposed: ZFNet, VGGNet, GoogleNet, and ResNet (Gu et al., 2018). The typical modification of these new architectures was the increased depth in order to extract even more features from the input. Furthermore, deep ConvNets were successfully applied for pedestrians' detection problems by estimating the movement of their limbs (Hou, Song, Hao, Shen, & Qian, 2017) (Dominguez-Sanchez, Cazorla, & Orts-Escolano, 2017).

The research by Lu et al. examined the application of convolutional neural networks for player detection and team classification in group sports such as basketball, ice hockey, and soccer from broadcasting videos (Lu, Chen, Little, & Hea, 2018). They also experimented on a pedestrian dataset to evaluate the generality of their approach. Their model performed very well and was able to classify each team in different sports with 97% accuracy. Table 2 shows the confusion matrix of the percentage of players being classified by teams in the 4 different data sets (Lu, Chen, Little, & Hea, 2018). Table 2 represents the proportion of players in each team being classified into the corresponding team. Classes TA, TB, and O refer to Team A, Team B, and Others accordingly.

Table 2. Confusion matrix of team membership classification in 4 datasets

Dataset	Classes	**TA**	**TB**	**O**
Basketball	TA	**99.65**	0.18	0.17
	TB	0.91	**97.88**	1.21
	O	0.86	1.71	**97.43**
Ice Hockey	TA	**98.91**	0.72	0.37
	TB	1.33	**97.99**	0.68
	O	0.69	1.36	**97.95**
Soccer Set 1	TA	**98.63**	0.24	1.13
	TB	0.83	**98.23**	0.94
	O	2.08	1.41	**96.51**
Soccer Set 2	TA	**98.33**	0.44	1.23
	TB	0.91	**97.78**	1.31
	O	2.46	1.37	**96.17**

The study conducted by Dominguez-Sanchez et al. evaluated the ability and performance of the current convolutional neural networks and proved how CNNs can impressively perform an estimation task of determining limbs movement of a pedestrian. During the research, they trained their networks with their own novel video dataset which was processed into frames through the image preprocessing pipeline. Only one of every six frames were used for the input during experiments of pedestrians' limb position and movement detection. After the evaluation of AlexNet, GoogleNet, and ResNet architectures, they identified that ResNet was the best for pedestrians' movement recognition and demonstrated 79% accuracy in the test set. Table 3 illustrates the results obtained by the ResNet in the test set (Dominguez-Sanchez, Cazorla, & Orts-Escolano, 2017).

Table 3. Confusion matrix of results obtained by ResNet

	Front	**Left**	**Right**
Front	**0.980**	0.011	0.008
Left	0.058	**0.841**	0.100
Right	0.081	0.265	**0.652**

Abdulnabi et al. introduced a modified deep convolutional neural network architecture that enables multitasking, so different CNNs can share knowledge among each other (Abdulnabi, Wang, Lu, & Jia, 2016). Their learned Multi-Task CNN demonstrated better performance in predicting semantic binary attributes by sharing visual knowledge between tasks. The results obtained from experiments on two different datasets and multiple different CNNs shows that Multi-Task CNN used by Abdulnabi et al. outperformed single-task neural networks and

achieved 92% accuracy in attribute predictions in images (Abdulnabi, Wang, Lu, & Jia, 2016). Deep convolutional neural networks demonstrated amazing performance in pedestrians and attribute detection and were selected as the approach for this research.

2.2 Approaches

Research in computer vision for pedestrian detection has used many different methods including DCNN, Multi-Task CNN, Support Vector Machines, Kalman filter, Recurrent Neural Networks, and a combination of two or more of these approaches. Deep learning methods showed insightful results in extracting and analyzing tiny details from the image by advancing task accuracy (Tome, Monti, Baroffio, Bondi, Tagliasacchi, & Tubaro, 2016). Our algorithm intends to analyze pedestrians' images and should be able to detect key benchmarks in the frames to classify the distraction scenarios accurately what can be achieved with the application of ConvNets.

Yamashita et al. proposed a method that concurrently detects both the pedestrians and their positions by applying a regressing based Deep Convolutional Neural Network (Yamashita, Fukui, Yamauchi, & Fujiyoshi, 2016). They used Multi-Task DCNN to recognize pedestrians, their head and leg positions, and the distance between the vehicle and pedestrian applying a regression technique. They used two different datasets which contained the following: the first set consisted of 31320 positive and 254356 negative samples for training, 21790 samples for testing, and the second set consisted of 2100 positive and 50000 negative samples for training, and 1000 positive and 9000 negative samples for testing (Yamashita, Fukui, Yamauchi, & Fujiyoshi, 2016). Since both datasets had fewer positive samples than negative, they applied the augmentation technique to expand the number of positive samples. The results obtained from

their experiments showed that Multi-Task CNN performed better than Single-Task CNN in pedestrians' detection and their position estimation and achieved distance estimation error of less than 5% (Yamashita, Fukui, Yamauchi, & Fujiyoshi, 2016).

Another study by Rangesh et al. focused on the detection of distracted pedestrians due to technological factors, particularly the use of cellphones, and proposed the classification of pedestrians into 3 types of activity - a phone call, texting, and no action (Rangesh & Trivedi, 2018). As for the dataset, they gathered pedestrian frames from the camera and created their own dataset with a total of 1586 pedestrian images with labeled activities and entities. The dataset was split into training and test sets with a fraction of the occurrences of each activity remaining in both sets. Their approach included a multi-cue pipeline to detect individual parts with a pre-trained convolutional network. Lastly, they integrated each component together to create a final prediction class that will generate a probability score using an SVM classifier (Rangesh & Trivedi, 2018). Their experiments had shown promising results and they were able to achieve the overall accuracy of 94.6%. Table 4 illustrates the accuracies per each activity which includes No Action, Texting, and a Phone Call with an overall accuracy (Rangesh & Trivedi, 2018). Table 4 lists accuracies demonstrated by the model proposed by Rangesh et al. per each class and overall for types of cues Hands Only, Pose Only, Pose and Hands, as well as Pose and Hands and Gaze.

Table 4. Accuracies demonstrated by per each class and overall for types of cues

Cues	None	Texting	Phone Call	Overall
Hands Only	0.94	0.58	0.20	0.810
Pose Only	0.90	0.65	0.67	0.858
Pose + Hands	0.93	0.71	0.81	0.916
Pose + Hands + Gaze	**0.97**	**0.88**	**0.89**	**0.946**

The two approaches described above consider different clues separately in an image to make predictions, but this work analyzes the whole image of a pedestrian to detect distracted behavior. In this case, ConvNets can learn different features of increasing complexity from an image such as the position of hands in correlation to their handheld device to recognize distracted behavior.

2.3 Performance Metrics

The most established and commonly used metrics for evaluating the performance of a CNN include accuracy, precision, recall, coverage, F-Measure, failure metrics, bias metrics, and classification error. The approach in this study intended to develop a Convolutional Neural Network that can classify whether a pedestrian is or is not holding any handheld device.
One of the most common methods to estimate the performance of the classification model where output can be two or more classes is a confusion matrix (Fatourechi, Ward, Mason, Huggins, Schlögl, & Birch, 2008). The confusion matrix in Figure 3 represents a table with 4 different combinations (in the case of 2 classes) of predicted and actual values.

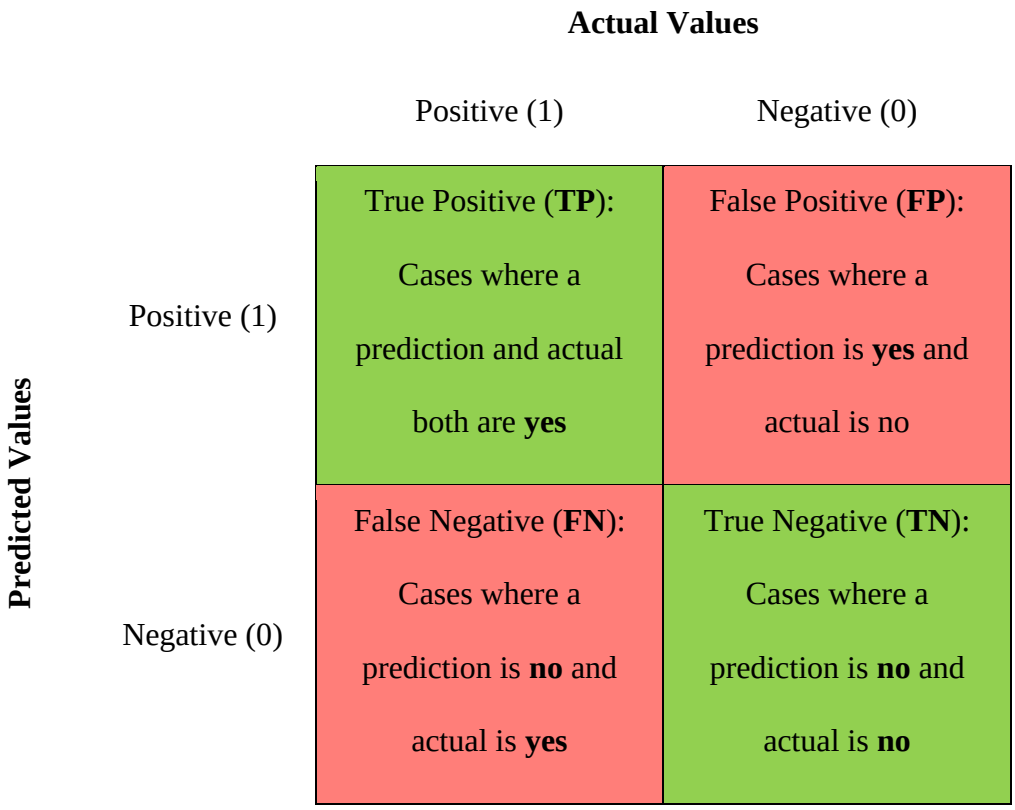


Figure 3. Example of confusion matrix and its terminology

There are a number of metrics that can be calculated from the confusion matrix for the evaluation of model performance. The first and most heuristic measure is accuracy that tells us a number of correct predictions made by a classifier over all kinds of predictions made and expressed in the following formula:

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP} + \boldsymbol{TN}}{\boldsymbol{TP} + \boldsymbol{FP} + \boldsymbol{FN} + \boldsymbol{TN}}$$

Yet, the accuracy is not an appropriate measure when the data is imbalanced (Chawla, n.d.). Therefore, we built a balanced dataset where the data gathered from the videotaped participants equally represents both distracted and non-distracted pedestrian classes. Although there are other metrics, such as precision and recall, that are useful in evaluating our model performance.

Precision is a measure that indicates the proportion of values that have been predicted as positive are actually positive. For instance, there are 100 pedestrians in our dataset and only 30 of them are distracted. If the model detected all the 30 distracted pedestrians correctly (TP) and also recognized 20 non-distracted people as distracted (FP), the precision rate would be 0.6 (60%). The result of precision can be seen in the following formula:

$$\frac{Correctly\ detected\ distracted\ pedestrians\ (30)}{Correctly\ detected\ distracted\ pedestrians\ (30) + \ Undistracted\ pedestrians\ who\ were\ classified\ as\ distracted\ (20)}$$

Precision, therefore, can be expressed as the following:

$$\boldsymbol{Precision = \frac{TP}{TP + FP}}$$

Lastly, recall measures the proportion of values that are actually positive and were classified by the model as positive. Returning to the previous example with pedestrians, if a model detected all the 25 distracted pedestrians correctly (TP) but it recognized the other 5 distracted pedestrians as non-distracted (FN), the recall rate would be 0.8333 (83.33%). This can be expressed as:

$$\frac{Correctly\ detected\ distracted\ pedestrians\ (25)}{Correctly\ detected\ distracted\ pedestrians\ (25)\ +\ Distracted\ pedestrians\ who\ were\ classified\ as\ undistracted\ (5)}$$

Therefore, recall can be represented in the following formula:

$$\boldsymbol{Recall = \frac{TP}{TP + FN}}$$

Sometimes, it is not efficient to have both Precision and Recall used estimating the performance of a classification model since they may have contradictive scoring; precision is high with low recall and vice versa. Therefore, it is better to get a single F-score as a measure that represents both Precision (P) and Recall (R) and simply calculated by taking the harmonic mean of P and R (Hripcsak & Rothschild, 2005). F-measure can be expressed in the following formula:

$$\boldsymbol{F - score = 2 * \frac{Recall * Precision}{Recall + Precision}}$$

We evaluated how ConvNets can assist us in the detection of distracted pedestrians. We also examined approaches for both distracted pedestrian and pedestrian position detections and assessed different metrics that can measure the performance of the algorithms proposed in those approaches. In the next chapter, this study will describe how to apply ConvNets in order to accurately detect distracted pedestrians in the streets.

CHAPTER THREE

3. METHODOLOGY

This chapter provides the details of the proposed research methods that include the data sources have been used, research to identify the types of scenarios and activities which will diminish a pedestrian's concentration and required cognitive effort while crossing the road, implementation process of the Distracted Pedestrians Detector (DPD), testing strategies, and overall complexity analysis of the algorithm associated with convolutional layers of the detector.

3.1 Data Sources

One of the data sources used in this research was built by recording student volunteers from the Sheridan College Institute of Technology (SCIT dataset). Recording students' videos to create a dataset was approved by the Sheridan Research Ethics Board. The total number of participants was 15 with different demographics such as gender, race, and age which allowed us to construct a good quality diverse dataset. The videos were recorded in an enclosed environment where each participant was asked to mimic a distracted/non- distracted pedestrian, based on the attributes listed in Table 5 while crossing the road. These video recordings of their walk were incorporated into the training set and further used for this study. The volunteers were recorded from three different positions for both front and rear views in order to capture every possible angle, direction, and position. Then, all the video footage was split into frames and labeled based on the participants' behavior to differentiate distracted and non-distracted scenarios. Each participant had around 350 frames per each activity, thus, we formed $350 \times 15 \approx 5{,}000$ images per activity after data preprocessing.

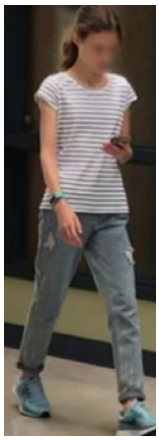
Another data source was built from Composition of PEdesTrian Attribute (PETA) dataset with 19000 images, with the image size ranging from 17 × 39 pixels to 169 × 365 pixels, which were released by Deng et al. during their research (Deng, Luo, Loy, & Tang, 2014). They also provided attribute annotations for each image in order to perform benchmarks detection. Yet, their dataset did not provide any labels whether the person on an image is distracted or not. Thus, all the images were reviewed and classified manually to fit the purpose of this research.

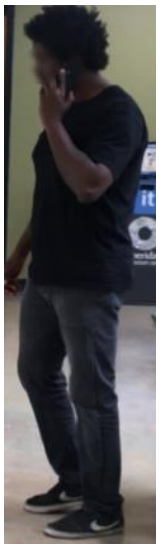
3.2 Determining Pedestrians Distracted Behavior Scenarios

After collecting data based on walking pedestrians, all the images were broken down into two classes: Distracted and Non-Distracted Pedestrians. The literature has been explored to identify what type of behavior can cause cognitive load and result in an unsafe road crossing. According to research conducted by Mwakalonge et al., 75% of pedestrians who were walking while taking on a cell phone displayed inattention blindness and failed to notice unusual activity (Mwakalonge, Siuhi, & White, 2015). Another study by Neider et al. performed the experiment in a virtual pedestrian environment and determined that participants who were distracted by music or texting were more likely to be hit by an automobile (Neider, McCarley, & Crowell, 2010). 5 different scenarios were identified where a pedestrian is considered to be distracted based on their hands and head positioning. Table 5 provides an overview of those scenarios as well as example images from the SCIT dataset. Then, PETA dataset images that fall under the identified scenarios were manually moved to a different directory to be separated from the images that were identified as non-distracted pedestrians. As for the SCIT dataset, all the videotaped volunteers were asked to mimic distracted and non-distracted behavior before the recording, thus, all the data were already structured and easily distributed in two classes. Also,

each distracted and non-distracted scenario was recorded from different views to simulate real-life situations as much as possible.

Table 5. Description of scenarios when walking pedestrian is distracted

Scenario Description	Example from the SCIT dataset
Head down and holding the phone with the left hand. A participant is chatting on the phone.	
Head down and holding the phone with the right hand. A participant is chatting on the phone.	

Head down and holding the phone with both hands. A participant is chatting on the phone.	
The left hand is near the head. A participant is speaking over the phone.	

The right hand is near the head. A participant is speaking over the phone.	

3.3 Development Steps of Distracted Pedestrian Detector (DPD)

The development phases for the proposed detector include: (i) identifying the appropriate sample size to train an accurate ConvNet image recognition classifier, (ii) datasets preprocessing to improve the quality of the data, and (iii) designing a ConvNet architecture and fine-tuning hyper-parameters to get the accurate classifier.

3.3.1 Identifying Appropriate Sample Size

The most effective dataset size to accurately train a ConvNet model is determined iteratively and can be guided by the distribution of classes and their behaviors. Therefore, it is not clearly defined which sample size would be the most suitable to train an accurate ConvNet pedestrian classifier. Li et al. used the Caltech-101 dataset which contains 9,144 images with a variety of classes to train and test their CNN image classifier and achieved 89% accuracy (Li, Peng, & Yan, 2018). The samples of 4,000 images and 30,000 images of distracted and non-distracted pedestrians were gathered from the PETA and SCIT datasets accordingly. However, the whole number of images in the SCIT dataset was not used in the experiments since this number is calculated based on the number of images for each behavior example where we have 5,000 images per scenario. Therefore, we used all the images from the non-distracted scenario set to create the first-class and randomly selected 1,000 images from each of the distracted scenarios sets to create the second class. Eventually, we constructed the dataset of 10,000 images of distracted and non-distracted classes based on the SCIT data.

3.3.2 Preprocessing of SCIT and PETA datasets

Before training the detector and conducting different experiments, people were cropped from the frames in the SCIT dataset gathered by our experiment. A pretrained Mask R-CNN object detector was used to detect people in each image and annotate their bounding boxes to perform the cropping. The resolution of the cropped pedestrian images is ranging from 62 × 224 pixels to 494 × 987 pixels in the SCIT dataset. The amount of blur in each image was also computed in order to remove images with excessive amounts of blurring that improved the dataset quality. Further, data augmentation techniques were applied to both PETA and SCIT datasets in order to increase the size of the datasets. Particularly, we augmented our data by rescaling, zoom-range, and fill-mode.

3.3.3 Determining CNN Architecture and Fine-Tuning

Convolutional Neural Networks have been selected due to their convolution layers which extract features from an input image and learn from them by exploiting small chunks of input data in order to preserve the spatial relationship between them.

Before being inputted into the CNN model, an image must be preprocessed by converting it to binary data and can be considered as a 5 x 5 matrix of pixel values. Consider an image of letter H in the English alphabet converted to pixels in which pixel values are only 0 and 1 as shown in Figure 4 and 3 x 3 feature detector matrix illustrated in Figure 5.

0	1	0	1	0
0	1	0	1	0
0	1	1	1	0
0	1	0	1	0
0	1	0	1	0

Figure 4. Letter H image converted to a 6 x 6 matrix of pixels

0	1	0
1	1	0
1	1	0

Figure 5. Feature detector matrix that will be convoluted with the letter H matrix

The convolutional operation will then be performed on the convolutional layer using these two matrices to obtain a feature map matrix.

The next layer is called the Pooling layer which subsamples or down-samples the capacity of each feature map while keeping the most important information. In the case of Max Pooling, which is used for our architecture, the pooling layer takes the largest value from the resolved feature map within the spatial neighborhood window.

Lastly, the Flatten layer goes after pooling to convert the matrix into a linear array in order to input the data into a neural network.

We proposed two architectures Basic and Deep with 3 and 5 convolutional layers accordingly to undertake the problem of distracted pedestrian detection.

The first architecture has the following structure: The first convolutional layer has 16 filters of size 3 with ReLU activation function followed by batch normalization and max-pooling layer

of size 2×2; the second convolutional layer has 32 filters of size 3 with Tanh activation function followed by batch normalization and max-pooling layer of size 2×2; the third convolutional layer has 64 filters of size 3 with ReLU activation function followed by batch normalization and max-pooling layer of size 2×2. The last max-pooling layer is followed by the dropout layer with a 25% dropout rate. After the aforementioned layers, we have flatten layer followed by two dense which also called fully connected layers. The first dense layer has 64 nodes with the ReLU activation function and the second has only 2 nodes with Sigmoid activation function since we need to find a probability of the pedestrian being distracted or not. This architecture is presented on the left side of Figure 6.

The second architecture is the modification of the above one where the second and third layers were duplicated such that two convolutional layers are stacked together before every max-pooling layer. Multiple stacked convolutional layers can be able to learn more complex features from the input before the destructive max-pooling layer (Ahire, 2018). We considered this technique to be promising in the detection of distracted pedestrian problem. The second architecture is shown on the right side of Figure 6.

We applied the same hyperparameters to both architectures; we used RMSprop optimizer with default parameters: learning rate = 0.001 and $\beta = 0.9$. The loss function we selected was the binary cross-entropy since this function better suits classification tasks with 2 classes (Lakhani, Gray, Pett, Nagy, & Shih, 2018). All the convolutional layers were preceded by the zero or "same" padding to preserve the size of post convolution. Finally, we applied the early stopping regularization technique to prevent the model from overfitting.

Figure 6. Architectures of Distracted Pedestrians Detector

3.4 Testing Strategy

The detector was tested with randomly selected images of distracted and non-distracted pedestrians which have not been seen by the model during training. Since SCIT data consists of 15 different participants, we randomly selected 4 participants and their images to generate the test set. The data of the other 11 participants were used for training. This 11/4 split is equivalent to a 75/25 data split, where 75% of data was used to train the model and the other 25% was used to test the model. This approach allowed us to always test our model on the people's data which the model had never seen before. Regarding the PETA dataset, since most of its data points represent a unique pedestrian, we randomly split data following the same 75/25 approach. Besides, the data in both datasets was always shuffled every time when we trained a new version of the model in order to reduce variance, make sure that the model remains general, and prevent overfitting. We conducted an experiment to examine how both our architectures can perform on different combinations of datasets, which drastically different in the resolution of the images. AlexNet architecture was also evaluated on the same datasets to compare it with our proposed architectures.

3.4.1 Proposed Experiment

The purpose of the experiment was to see how the quality of the images would affect the performance of the ConvNet based on different architectures. Therefore, we created three different sample sets from the SCIT and PETA datasets for this test. The first sample was made of only the SCIT dataset where all the images had high resolution (62 × 224 pixels to 494 × 987 pixels) and distraction scenarios were equally distributed. The second sample was constructed

from the PETA dataset and its images had a relatively low number of pixels (17 × 39 pixels to 169 × 365 pixels). The third data sample was created using both SCIT and PETA dataset where high and low image resolution (17 × 39 pixels to 494 × 987 pixels) were combined. The purpose of the third sample was to see whether the ConvNet accuracy would degrade or not if we feed data to it which has a huge range in quality to it.

The models with Basic and Deep architectures were trained and tested on the aforementioned datasets. We also investigated how AlexNet architecture that achieved state-of-the-art results in many computer vision tasks would tackle the distracted pedestrian detection problem (Krizhevsky, Sutskever, & Hinton, 2017). AlexNet is a much deeper network with more filters in each convolutional layer. The model with AlexNet architecture was also trained on the same data samples, so we could compare its performance with our Basic and Deep architectures. The reason why the AlexNet had been also evaluated was to examine if the deeper network with more filters would be smarter in the feature extraction related to our problem and would have better accuracy in distracted pedestrian detection. Figure 7 illustrates the design of the experiment.

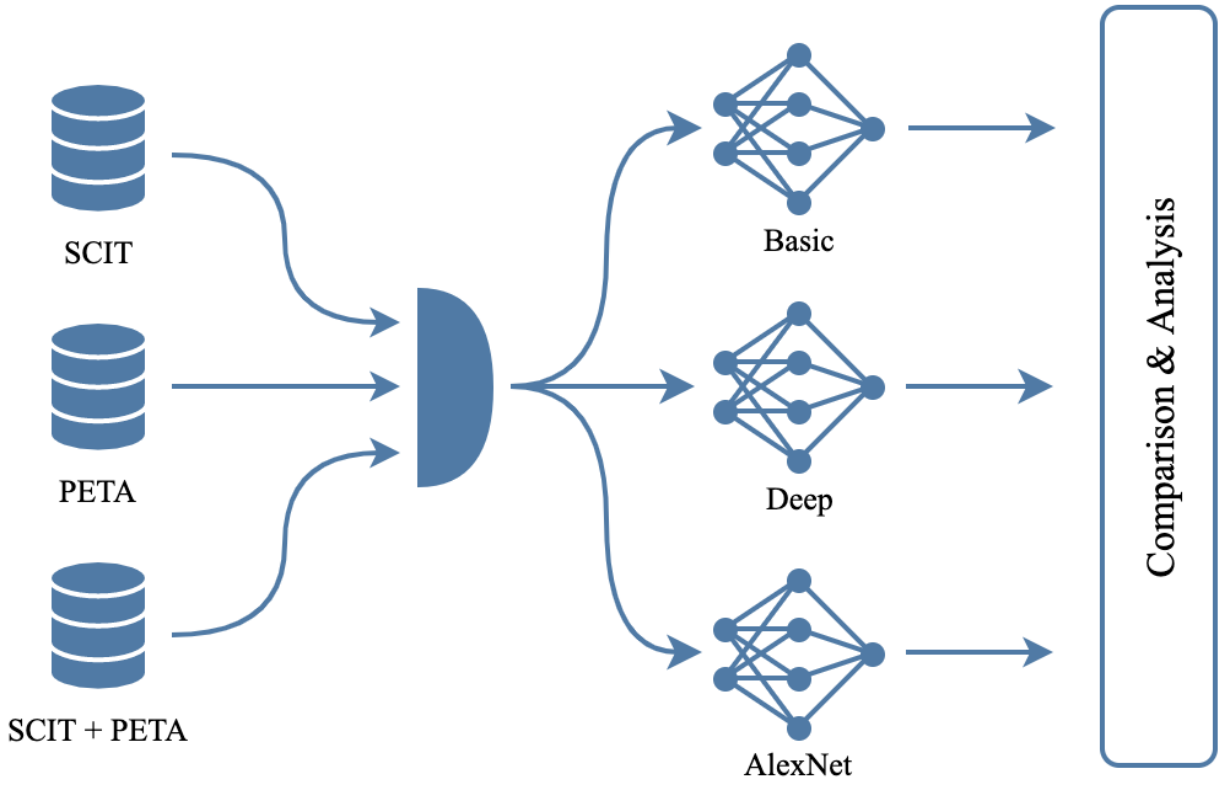


Figure 7. Experiment Design

3.4.2 Performance Metrics

Precision, recall, and accuracy metrics were used to identify the performance of the trained model for all the above experiments. By analyzing the true positive, true negative, false positive, and false negative, we determined the accuracy of the detector. These metrics helped to identify whether the model correctly classified a new image or not. These are the only performance metrics that were used in order to determine the accuracy of the classifier.

3.5 Complexity Analysis

An analytical complexity analysis was done on the CNN model to determine the feasibility of training CNN to learn how to differentiate between the distracted and non-distracted pedestrians.

The CNN time complexity has been analyzed by He et al. (He & Sun, 2015) where all the convolutional layers have the following time complexity shown in a form of Big O notation:

$$O\left(\sum_{i=1}^{d} n_{i-1} * s_i^2 * n_i * m_i^2\right) \quad (1)$$

Where i is an index of a convolutional layer, d is a number of convolutional layers, n_{i-1} is the number of input channels, s_i represents the spatial size of filter and n_i is a number of filters in a layer. Finally, m_i is the size of the outputted feature map.

This time complexity is representative of both the training and testing times but with a different scaling since the training time per image is approximately three times greater than that of the testing time per image (He & Sun, 2015). The same dimensionality for every image was used thus, the approximate time complexity of CNN is computed by applying Equation (1) and then multiplied by the number of images.

CHAPTER FOUR

4. RESULTS AND ANALYSIS

This section shows the experimental results of building the Distracted Pedestrian Detector based on different combinations of the datasets: SCIT, PETA, and a combination of both. This work tested two different ConvNet architectures. The first is called Basic, and the second is called Deep, which duplicates the second and third layers of the Basic architecture. Additionally, we examined how the AlexNet model would tackle the distracted pedestrian detection problem based on the combinations of the aforementioned datasets. The Deep ConvNet architecture was more efficient than the Basic and AlexNet architectures in detecting the distracted pedestrians based on all three datasets.

4.1 Effect of the Image Resolution on the Performance of the Detector

The highest accuracy of the Distracted Pedestrian Detector with Deep architecture for the SCIT dataset was 95.11%. Figure 8 shows the average accuracies of the Deep, Basic, and AlexNet architectures trained and tested on the SCIT data sample. Since the SCIT datasets had the highest resolution, this particular evaluation demonstrates how the architectures behave on images with a big number of pixels. The Deep architecture also showed the highest average 94.02% accuracy. The Basic architecture was the second in the accuracy and achieved 90.00% on average. Lastly, the performance of AlexNet was close to the Basic architecture but demonstrated lower average accuracy – 89.23%. Based on the high precision and recall scores, shown in Table 6, we can see that all the models trained on the SCIT data were able to correctly classify a high number of the relative data points. This is supported by the f1 score since it was also relatively high too, meaning that models were general and unbiased. This was due to the

SCIT dataset being well distributed and provided the models with balanced training and testing data. We can see that all the architectures performed relatively well on the dataset which contains images with high resolution.

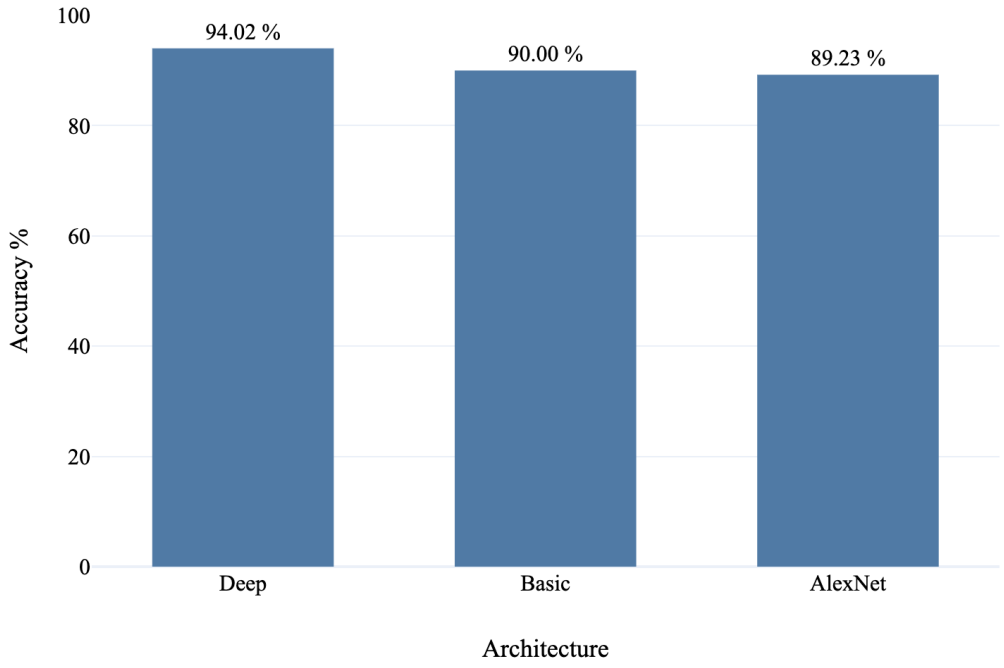


Figure 8. Average accuracy of architectures for SCIT dataset

Table 6. Average Precision, Recall, and F1 Score metrics of models for SCIT dataset

	Precision	**Recall**	**F1 Score**
Deep	0.9434	0.9374	0.9403
Basic	0.9246	0.8812	0.9024
AlexNet	0.8826	0.9000	0.8912

When trained and validated on the PETA dataset, all the architectures demonstrated lower accuracies. This can be explained by a certainly low resolution of images in the PETA dataset. Figure 9 visualizes the average accuracies achieved by the Deep, Basic, and AlexNet architectures trained and tested on the PETA data sample. The Deep architecture maintained the first place and showed an average 85.44% accuracy. The AlexNet architecture had the 83.67% accuracy on average what was close to the Deep one. Yet, the Basic architecture demonstrated the biggest reduction in accuracy and achieved 78.01% what notably different from the score of Deep and AlexNet architectures. The Basic ConvNet had the smallest number of convolutional layers and, therefore, the minimum number of filters. It performed relatively bad in distinguishing between distracted and non-distracted pedestrians. We tried to increase the number of filters in each convolutional layer by 4 times such that it had 64 filters in the first layer, 128 filters in the second layer, and 256 in the third layer. Unfortunately, this only worsened the architecture, because the high number of filters caused model overfitting since the training accuracy was 97.15% while the validation accuracy was only 76.34%. This indicates that the three convolutional layers are not enough to deal with images with a small number of pixels.

If we analyze the precision, recall, and f score metrics, demonstrated in Table 7, we can see that the recall metric significantly dropped compared to the precision metric. It means that the models evaluated on the PETA data classified more distracted pedestrians as non-distracted. We then can conclude that the data with low-quality images did not allow models to learn enough patterns, since it was relative to the distracted behavior. Also, some of the images were captured from a distance making it really difficult for the models to detect if an observed pedestrian is holding a handheld device or not.

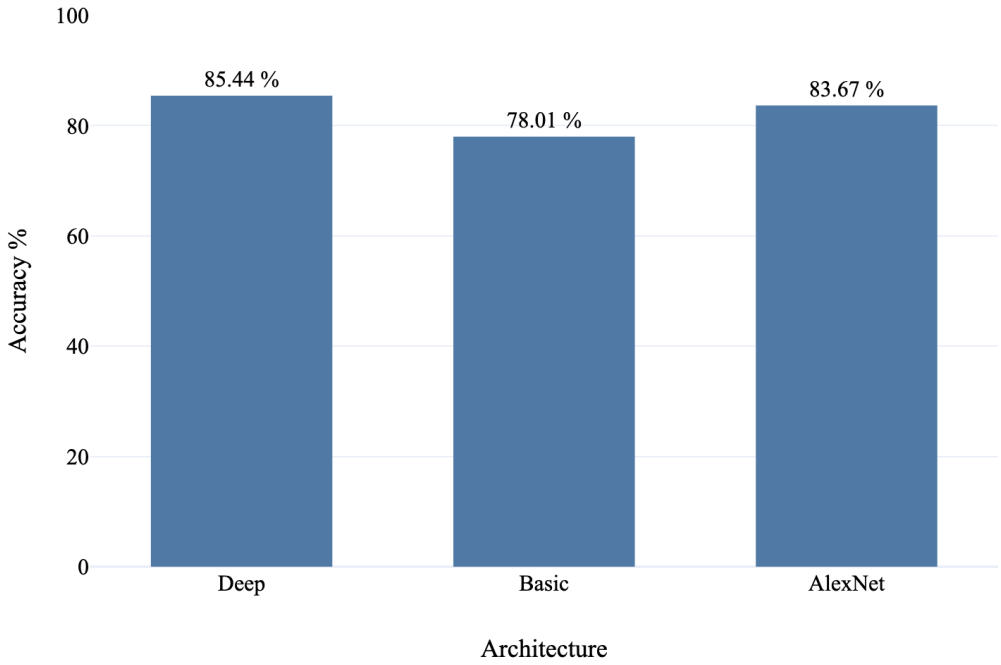


Figure 9. Average accuracy of architectures for PETA dataset

Table 7. Average Precision, Recall, and F1 Score metrics of models for PETA dataset

	Precision	**Recall**	**F1 Score**
Deep	0.8990	0.8251	0.8604
Basic	0.8780	0.7341	0.7996
AlexNet	0.8915	0.8024	0.8446

The third dataset which was used for the evaluation of the architectures was the combination of both SCIT and PETA data. The highest accuracy was demonstrated by the Deep architecture which achieved 88.78%. The average accuracy of the Deep, Basic, and AlexNet

architectures trained and evaluated on the combination of SCIT and PETA datasets is shown in Figure 10. The Deep architecture, again, showed the best average accuracy – 87.01%. The accuracies of AlexNet and Basic architectures were 84.32% and 80.56%, respectively. All the architectures did not improve much, and their average accuracies were approximately 2% better compared with the models trained and tested only on the PETA dataset. These results illustrate that even if we combine the images with low and high resolutions, the images with a low number of pixels in the set still affects the ability of ConvNet accurately detect distracted pedestrians. Besides, the big range of the resolution could also be a reason for the not significant improvement of the architectures. ConvNets could not establish a clear pattern from the extracted features to find the difference between distracted and non-distracted scenarios.

Table 8 shows the precision, recall, and f1 score metrics obtained by the ConvNet models trained and tested on the combination of both SCIT and PETA datasets. It is clear that if we add high-quality images to the dataset that contains images with a low number of pixels, the models can learn more features and distinguish distracted and non-distracted pedestrians with better accuracy. However, the following metrics are still lower compared to the obtained metrics in Table 6, which demonstrates again, that data with low-quality images has a big influence on the architectures, even if data points with a big number of pixels are dominant in this dataset.

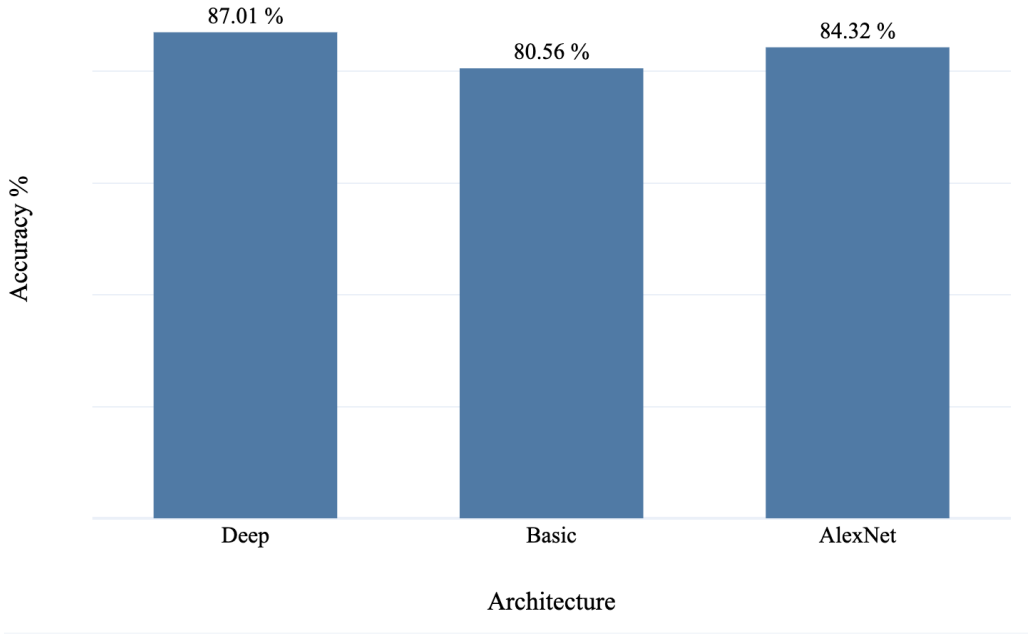


Figure 10. Average accuracy of architectures for combination of SCIT and PETA sets

Table 8. Average Precision, Recall, and F1 Score metrics for SCIT and PETA sets

	Precision	**Recall**	**F1 Score**
Deep	0.8888	0.8566	0.8724
Basic	0.8272	0.7929	0.8097
AlexNet	0.8685	0.8266	0.8470

Since the model based on the Deep architecture demonstrated higher accuracy across all three datasets, the one-way analysis of variance (ANOVA) test was used to determine if the Deep architecture's score is significantly different from the Basic and AlexNet models. The ANOVA test was conducted on three different sets of models trained on the different datasets as shown in Figure 8, Figure 9, and Figure 10. The *p-value* from the three test results was the following: 0.0003, 0.00025, 0.00027 for the sets of models trained on the SCIT, PETA, and combination of SCIT and PETA datasets, accordingly. Since the *p-value* across all the datasets was less than 0.05, this indicates that the models' accuracies were significantly different and not from the same. Thus, we can conclude that the difference in the model's scores is significant showing that the Deep actually had the highest accuracy.

4.2 Impact of Architecture Design

We also inspected the filters and feature maps during the layers' convolution of Basic and Deep ConvNet architectures. Since Deep architecture was designed to have the second and third layers combined together followed by the max-pooling layer, the third layer was able to receive a more precise feature map where we still can recognize the original image as shown in Figure 11. In contrast, all the convolutional layers in Basic architecture are split by max-pooling layer, therefore, the feature map of the third layer in the Basic architecture is less interpretable and contains high-level concepts as displayed in Figure 12. From Figure 11 and Figure 12, we can see that the feature map in the third convolutional layer of the Deep architecture still contains visual concepts like edges, which are useful for our problem since the detector needs to evaluate the position of the pedestrian limbs to differentiate distracted and non-distracted behavior. While the feature map in the third layer of the Basic architecture looks more like the abstraction of the

original image and contains high-level features that might have more information about small parts of the image such as a *mobile device in the hands*. Of course, both low and high-level features are highly important to accurately detect distracted pedestrians. Though, the design of the Deep architecture allowed filters to extract more low-level features that helped ConvNet to characterize the position of pedestrian limbs and better recognize the distractive action. This explains why ConvNet with Deep architecture outperformed the Basic ConvNet across all the three datasets since the Basic architecture could not extract enough features related to the pedestrians' actions.

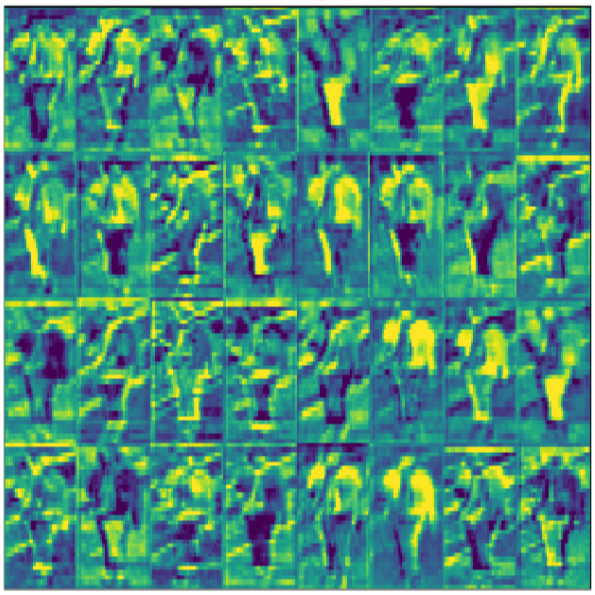


Figure 11. Visualization of the filters in the third Conv layer of the Deep architecture

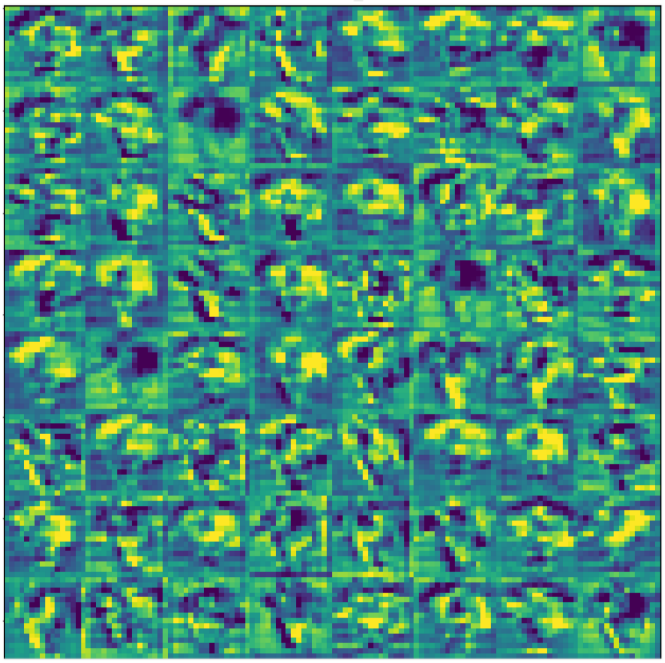


Figure 12. Visualization of the filters in the third Conv layer of the Basic architecture

Interestingly enough that Deep and AlexNet architectures had a similar design in terms of the combined convolutional layers. While the Deep architecture combined the second with third and the fourth with fifth layers, the AlexNet architecture design combined the third, fourth, and fifth convolutional layers without max-pooling layers between them. But based on the gathered results demonstrated above, the Deep architecture achieved higher average accuracies across all the three datasets. Despite the fact, that even if AlexNet has a similar structure to the Deep architecture, its combined convolutional layers focused mostly on the extraction of the high-level features since they were the last group and received feature maps that already got through multiple max-pooling layers. Therefore, AlexNet could not extract more low-level features like the Deep architecture. This derives the conclusion that the low-level features which are

responsible for the detection of edges and shapes played a very important role in the distracted pedestrian detection problem and allowed the Deep architecture to outperform the AlexNet and Basic ConvNets.

CHAPTER FIVE

5. CONCLUSION

5.1 Conclusion

This research aimed to explore the application of convolutional neural networks to address the problem of detecting distracted pedestrians automatically. This work investigated various combinations of CNN architectures and datasets to build an effective distracted pedestrian detector. A novel training dataset was created from video recordings of volunteer participants from the Sheridan College Institute of Technology when they acted as distracted and non-distracted pedestrians. This dataset is called SCIT and could be used for further research in various computer vision research problems related to human detection. Three ConvNet models were implemented with different architectures: Basic, Deep, and AlexNet. Each model was trained and tested on three different datasets: SCIT, PETA, and the combination of both. The results from the experiment had indicated that the model that utilized the Deep architecture had outperformed the other models that used the Basic and AlexNet architectures when applied to all the datasets. The developed detector could be used for autonomous vehicles and driver alert systems to identify distracted pedestrians who cross the street and minimize the probability of injury. The detector would also be useful for the variety of stakeholders including the vehicle manufactures, researchers, and smart cities project teams.

5.2 Future work

The detector currently takes an entire image and makes a prediction based on the extracted features. The next step will be to modify the algorithm so that it would extract

pedestrian limbs such as head and hands from each image and evaluate them independently instead of analyzing a complete image. This modification will increase the efficiency of the system because it will minimize the misclassification of handheld devices with other potential objects in the pedestrian's hands. An analysis of how a pedestrian's head direction changes would also create a meaningful impact on when identifying if a pedestrian is distracted.

Predicting the route of a distracted pedestrian will be another perspective addition to the system. Distracted pedestrians tend to change their route unexpectedly what increases the possibility of an accident. With the knowledge that a pedestrian is distracted, his/her long term path could be predicted more accurately. The information about pedestrians' future path and if they are distracted or not could advance the safe route planning for self-driving cars.

Sequential frame classification can be another improvement to the detector. In this case, extraction of the sequence features, which are also called temporal or time-related features, will be required in addition to the features of the images. This approach could help identify when a pedestrian had acted similar to a distracting behavior for a short period of time when the pedestrian's action was not an actual distraction. This could reduce the number of false positives that would improve the reliability of the detector.

5.3 Limitations

The significant limitation of the detector occurred when it had observed the pedestrian from the side who was talking on the phone using the opposite hand or the hand on the far side. In this case, the phone and elbow were completely hidden which make this case overlap with non-distracted behavior because the algorithm could not see any cues of distraction or obvious differences.

REFERENCES

Transport Canada (2015). Canadian Motor Vehicle Traffic Collision Statistics 2015. Government of Canada.

Yogev-Seligmann, G., Hausdorff, J. M., & Giladi, N. (2012). Do we always prioritize balance when walking? Towards an integrated model of task prioritization. *Movement Disorders,27*(6), 765-770. doi:10.1002/mds.24963x

Nasar, J. L., & Troyer, D. (2013). Pedestrian injuries due to mobile phone use in public places. *Accident Analysis & Prevention,57*, 91-95. doi:10.1016/j.aap.2013.03.021

Brenner, J., & Smith, A. (2014). 72% of Online Adults are Social Networking Site Users. Retrieved from http://www.pewinternet.org/2013/08/05/72-of-online-adults-are-social-networking-site-users/

Bassett, D., R Wyatt, H., Thompson, H., Peters, J., & Hill, J. (2010). Pedometer-Measured Physical Activity and Health Behaviors in U. S. Adults. *Medicine and science in sports and exercise*. 42. 1819-25. 10.1249/MSS.0b013e3181dc2e54.

Dominguez-Sanchez, A., Cazorla, M., & Orts-Escolano, S. (2017). Pedestrian Movement Direction Recognition Using Convolutional Neural Networks. IEEE Transactions on Intelligent Transportation Systems, 18(12), 3540-3548. doi:10.1109/tits.2017.2726140

Tang, Y., Ma, L., Liu, W., & Zheng, W. (2018). Long-Term Human Motion Prediction by Modeling Motion Context and Enhancing Motion Dynamics. Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. doi:10.24963/ijcai.2018/130

Zaki, M. H., & Sayed, T. (2016). Exploring walking gait features for the automated recognition of distracted pedestrians. IET Intelligent Transport Systems, 10(2), 106-113. doi:10.1049/iet-its.2015.0001

Chen, Y., Liu, M., Liu, S., Miller, J., & How, J. P. (2016). Predictive Modeling of Pedestrian Motion Patterns with Bayesian Nonparametrics. AIAA Guidance, Navigation, and Control Conference. doi:10.2514/6.2016-1861

Wang, J., Chen, D., Chen, H., & Yang, J. (2012). On pedestrian detection and tracking in infrared videos. Pattern Recognition Letters, 33, 775-785.

Asahara, A., Maruyama, K., Sato, A., & Seto, K. (2011). Pedestrian-movement prediction based on mixed Markov-chain model. Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS 11. doi:10.1145/2093973.2093979

Yamashita, T., Fukui, H., Yamauchi, Y., & Fujiyoshi, H. (2016). Pedestrian and part position detection using a regression-based multiple task deep convolutional neural network. 2016 23rd International Conference on Pattern Recognition (ICPR). doi:10.1109/icpr.2016.7900176

Rasouli, A., & Tsotsos, J.K. (2018). Autonomous Vehicles that Interact with Pedestrians: A Survey of Theory and Practice. CoRR, abs/1805.11773.

Neider, M.B., McCarley, J.S., Crowell, J.A., Kaczmarski, H.J., & Kramer, A.F. (2010). Pedestrians, vehicles, and cell phones. Accident; analysis and prevention, 42 2, 589-94.

Rangesh, A., Ohn-Bar, E., Yuen, K., & Trivedi, M. M. (2016). Pedestrians and their phones - detecting phone-based activities of pedestrians for autonomous vehicles. 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC). doi:10.1109/itsc.2016.7795861

Tome. D., Monti F., Baroffio L., Bondi L., Tagliasacchi M., & Tubaro S. (2016). Deep Convolutional Neural Networks for pedestrian detection. Signal Processing: Image Communication, 47, 482-489.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., … Chen, T. (2018). Recent advances in convolutional neural networks. Pattern Recognition, 77, 354–377. doi: 10.1016/j.patcog.2017.10.013

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., … Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision, 115(3), 211–252. doi: 10.1007/s11263-015-0816-y

Hou, Y., Song, Y., Hao, X., Shen, Y., & Qian, M. (2017). Multispectral pedestrian detection based on deep convolutional neural networks. 2017 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC). doi:10.1109/icspcc.2017.8242507

Lu K., Chen J., Little J. J., & Hea H. (2018). Lightweight convolutional neural networks for player detection and classification. Computer Vision and Image Understanding, 172, 77-87.

Abdulnabi, A. H., Wang, G., Lu, J., & Jia, K. (2016). Multi-Task CNN Model for Attribute Prediction. IEEE Transactions on Multimedia, 17(11), 1949-1959. doi:10.1109/tmm.2015.2477680

Rangesh, A., & Trivedi, M. M. (2018). When Vehicles See Pedestrians with Phones: A Multicue Framework for Recognizing Phone-Based Activities of Pedestrians. *IEEE Transactions on Intelligent Vehicles,3*(2), 218-227. doi:10.1109/tiv.2018.2804170

Fatourechi, M., Ward, R. K., Mason, S. G., Huggins, J., Schlögl, A., & Birch, G. E. (2008). Comparison of Evaluation Metrics in Classification Applications with Imbalanced Datasets. 2008 Seventh International Conference on Machine Learning and Applications. doi:10.1109/icmla.2008.34

Chawla, N. V. (n.d.). Data Mining for Imbalanced Datasets: An Overview. Data Mining and Knowledge Discovery Handbook, 853-867. doi:10.1007/0-387-25465-x_40

Hripcsak, G. & Rothschild, A. S. (2005). Agreement, the F-Measure, and Reliability in Information Retrieval. Journal of the American Medical Informatics Association, 296–298.

Li Q., Peng Q., & Yan C. (2018). Multiple VLAD Encoding of CNNs for Image Classification. Computing in Science & Engineering, 52–63.

Mwakalonge J., Siuhi S., & White J. (2015). Distracted walking: Examining the extent to pedestrian safety problems. Journal of Traffic and Transportation Engineering (English Edition), 327–337.

Deng Y., Luo P., Loy C. C., & Tang X. (2014). Pedestrian attribute recognition at far distance. Proceedings of ACM Multimedia (ACM MM).

He K. & Sun J. (2015). Convolutional neural networks at constrained time cost. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 5353–5360.

Ahire, J. (2018). Artificial Neural Networks: The brain behind Ai.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84–90. doi: 10.1145/3065386