# NEURAL NETWORKS FOR OBJECT DETECTION WITH UNMANNED AUTONOMOUS VEHICLES

A Project

Presented to the

Faculty of

California State Polytechnic University, Pomona

In Partial Fulfillment

Of the Requirements for the Degree

Master of Science

In

Computer Science

By

Joshua Tellez

2018

## SIGNATURE PAGE

| | |
|---|---|
| **PROJECT:** | NEURAL NETWORKS FOR OBJECT DETECTION WITH UNMANNED AUTONOMOUS VEHICLES |
| **AUTHOR:** | Joshua Tellez |
| **DATE SUBMITTED:** | Spring 2018 |
| | Computer Science Department |

Dr. Fang Tang  
Project Committee Chair  
Chair and Professor of  
Computer Science
______________________________________

Dr. Subodh Bhandari  
Project Committee  
Professor of Aerospace  
Engineering
______________________________________

# ABSTRACT

Neural networks are modern programming structures used by many to produce cutting edge technologies. These cutting-edge technologies can range from advanced medical equipment to self-driving cars.  The reason neural networks gained much momentum and became widespread in the world of technology is because they closed some of the difficult gaps that traditional computer science tools and techniques were incapable of quickly solving. These difficult tasks include computer vision, natural language processing, nonlinear prediction problems and the list continues. With the consistent widespread use of neural networks, this project aims to continue the evaluation of neural networks by applying their use on unmanned autonomous vehicles. The focus of this project will be on the computer vision tasks associated with autonomous vehicles and what kinds of neural networks can help accomplish them. The types of neural networks demonstrated in this project include the revolutionary convolutional neural network and the high performing residual neural network. A comparison of their performances on a small custom-made dataset are evaluated and analyzed for their potential on-board use in autonomous unmanned vehicles.

# TABLE OF CONTENTS

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# INTRODUCTION

Neural networks were developed with the idea of mimicking the human brain and recently skyrocketed as the leading method for producing highly intelligent programs. Ever since they started outperforming humans in recognizing objects, or helping defeat world class game players, neural networks are recognized as immensely enhancing a programmer's toolbox. Some researchers have been shouting for their potential since their infant 1958 birth as a perceptron [1]. But with the advancements in Graphics Processing Units and rise to large datasets the world can no longer deny their grand results and fascinating applications [2].

With the growing use of neural networks, research and development of autonomous vehicles have also gained much momentum over the past few years [3]. Neural networks have proven to be an incredible tool for the advancement of autonomous vehicles because they help solve many of the artificial intelligence problems associated with their development. Some artificial intelligence problems faced by autonomous vehicles and solved by neural networks include computer vision, movement control, and decision making.

The goal of this project is to develop an artificial neural network to detect objects on video obtained from three kinds of autonomous vehicles. The three vehicles include, an unmanned ground vehicle, an unmanned aerial vehicle, and an unmanned underwater vehicle. The types of objects to be detected depend on the vehicle. For instance, the aerial vehicle must detect a large red spherical object, while the ground and underwater vehicles must detect a smaller red and blue cylindrical object. Since each vehicle poses distinct computer vision difficulties, a neural network trained on data from all vehicles can unify the problems while establishing a robust method for automatic object detection.

# MISSION

This project will be complimenting a larger search and rescue mission involving the different vehicles. The overarching mission is as follows:

1. An autonomous aircraft will survey a predetermined area searching for a specified object

2. Once the object is found, another autonomous aircraft will deliver a care package to it

3. This care package is searched for and retrieved by an autonomous ground vehicle

4. If the care package is delivered over water, an autonomous underwater vehicle will search for and retrieve it

This overarching mission requires the vehicles to perform on-board processing of all the computer vision tasks. The purpose of this project, as previously mentioned, will be specifically focused on completing the automatic object detection for the vehicles.

## VEHCILES

The three types of vehicles used in the overarching mission include an unmanned aerial vehicle, an unmanned ground vehicle, and an unmanned underwater vehicle.

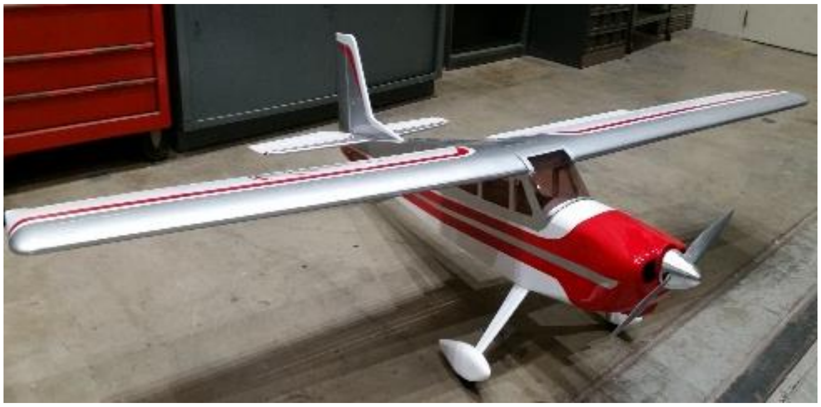

**Figure 1. Unmanned Aerial Search Vehicle**

The overarching project utilizes two types of aerial vehicles, a search vehicle and a package delivery vehicle. However, this project will only focus on the search vehicle since it will be solely tasked in completing the computer vision portion of the mission.

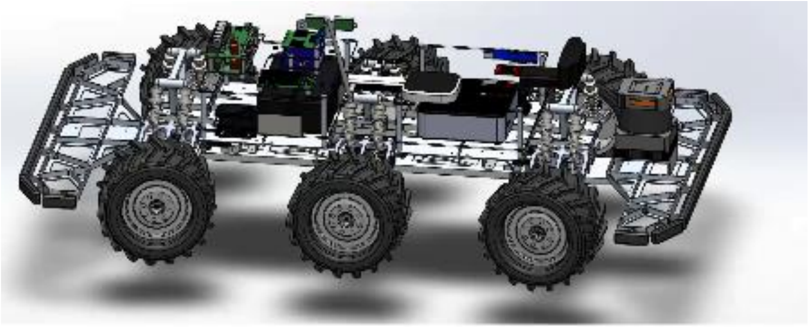

**Figure 2. Unmanned Ground Vehicle**

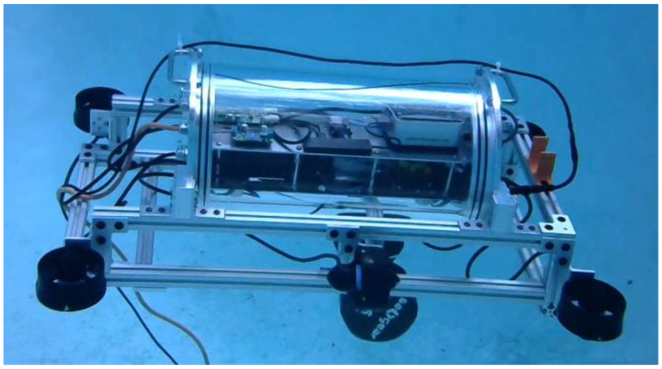

**Figure 3. Unmanned Underwater Vehicle**

The unmanned underwater and ground vehicles were built from the ground up by undergraduate engineering students from Cal Poly Pomona. The unmanned aerial vehicles were assembled, also by Cal Poly Pomona students, using fixed-wing aircraft kits. All vehicles are custom fitted with various sensors, controllers, and other electronics required by the overarching mission.

The vehicles use an intel NUC as its primary computer tasked with the execution of various programs such as communications, computer vision, movement, and other tasks necessary to the mission. Each vehicle is equipped with a different type of camera which can impose limitations on the neural network since the kind of data that will be employed will vary on each vehicle.

# DATA COLLECTION

The initial goal for this project was to obtain sufficient training data from each vehicle. Since the vehicles are tasked with objected detection in distinct environments, each vehicle was to record their individual missions. For instance, the aerial vehicle will record the surveyed area and the frames of the video will have a binary classification of whether the object is in the frame or not. The same data collection method would have been applied to the other two vehicles. The ground vehicle was to record its search for the delivered package on land and the underwater vehicle was to record its search in water. Several missions were to be recorded with each vehicle to expand the applicable data for the neural network.

Unfortunately, several complications arose during data collection with the unmanned underwater vehicle and the unmanned ground vehicle. Complications with camera gimbals, camera related software, or the camera themselves made it difficult to obtain video recordings of the vehicles during their respective missions. On the other hand, sufficient data of the aerial vehicle was collected and used for this project.

The data includes several videos from the search aircraft's point of view, surveying a predefined area. The videos were taken on separate days and for several missions to allow for greater variation among the data. Greater data variation will allow for a more robust and higher performing neural network.

Each frame of the videos collected were manually checked and sorted for whether the frame contains the target or not. Since the purpose of this project is to classify images into two categories, a binary classification was implemented as the labels for the data. If a frame contains the target object, it is given a label of ‘true’. If the frame does not contain the target, then it is given a label of ‘false’. Labeling each frame of the data is required during the training of the neural networks. A neural network requires labeled data to understand which patterns of the training data lead to which classifications. For instance, the patterns related to a red ball among a field will become associated with the ‘true’ label.

## DATA AUGMENTATION

To expand the collected data while maintaining a viable dataset, image rotations, flips, and filters are applied to the frames of the recorded missions. A labeled frame containing the object that is flipped and then rotated, will appear as a different image to the neural network but will maintain all the essential features and information of the original frame. Using this method of flips and rotations, a total of eight frames can be extracted from one.

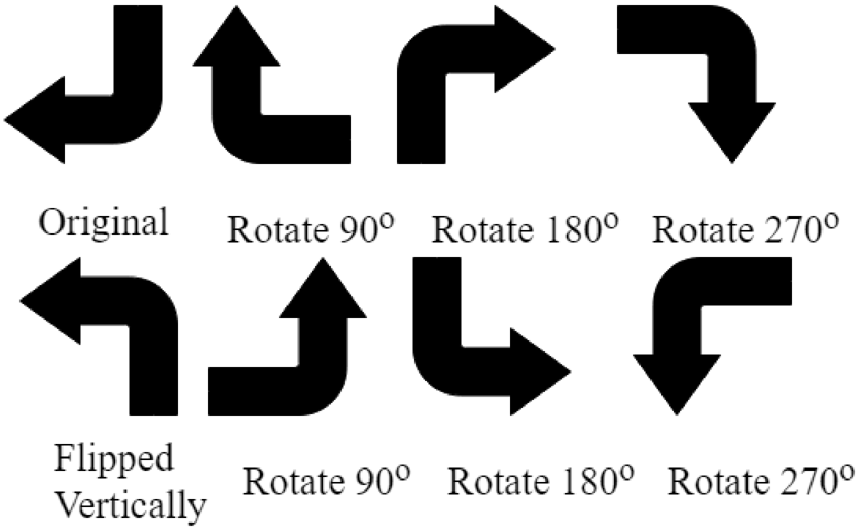

**Figure 4. Rotations and flips maintain the features of the original image**

To further increase the dataset, the frames will also be processed through different filters. The filters include increasing and decreasing exposure, brightness, saturation, and contrast. Similar to, rotations and flips, the fundamental features of a frame containing the object and a frame not containing the object will be maintained. Combining image filters with image rotations will greatly increase the dataset. However, it is worth noting that the recorded missions have a much larger number of frames without the object than frames with the object.

Therefore, these data augmentation techniques were applied to a much larger percentage of frames with the object to create a balanced training dataset.

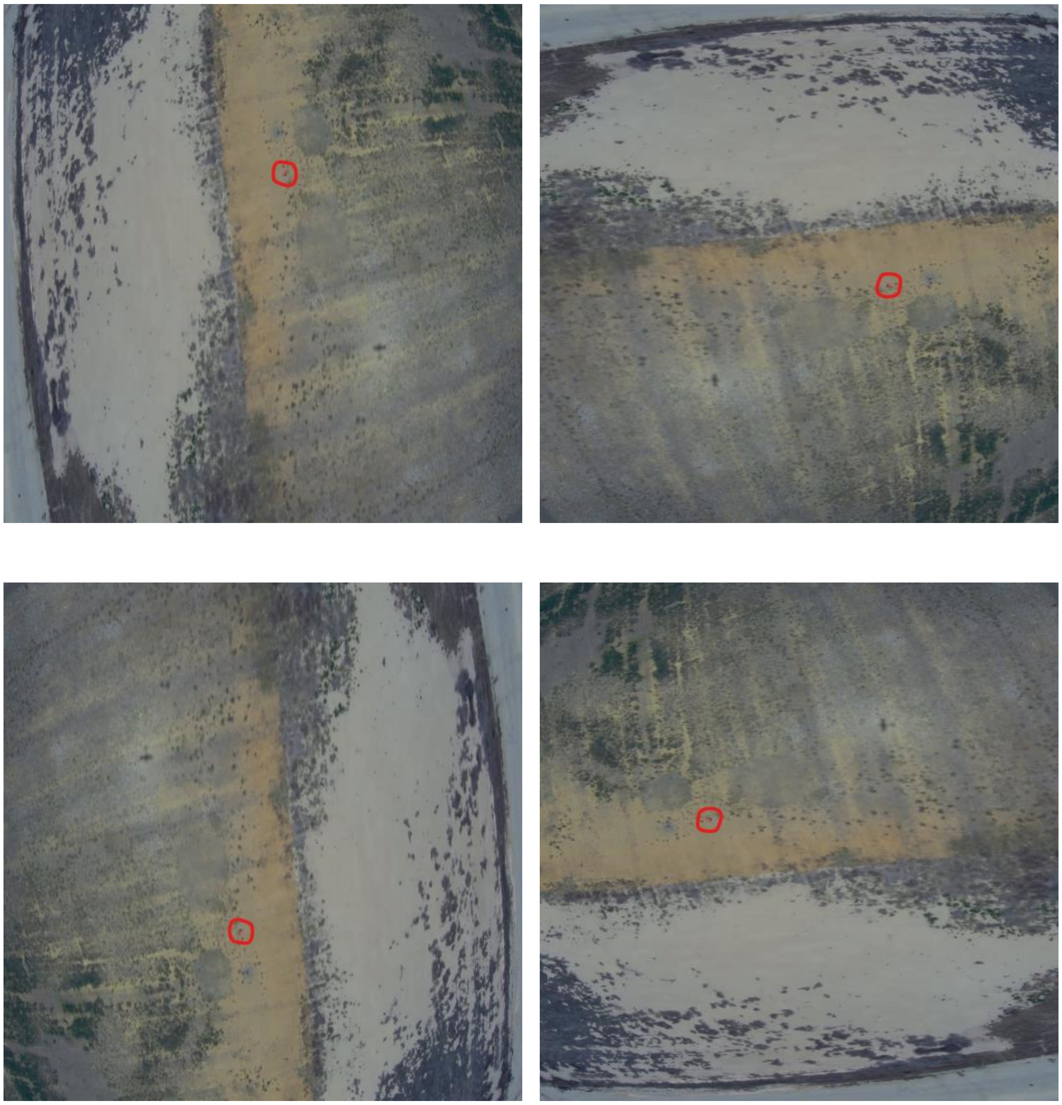

**Figure 5. Original (top left), rotate 90 degrees (top right), rotate 180 degrees (bottom left), rotate 270 degrees (bottom right)**

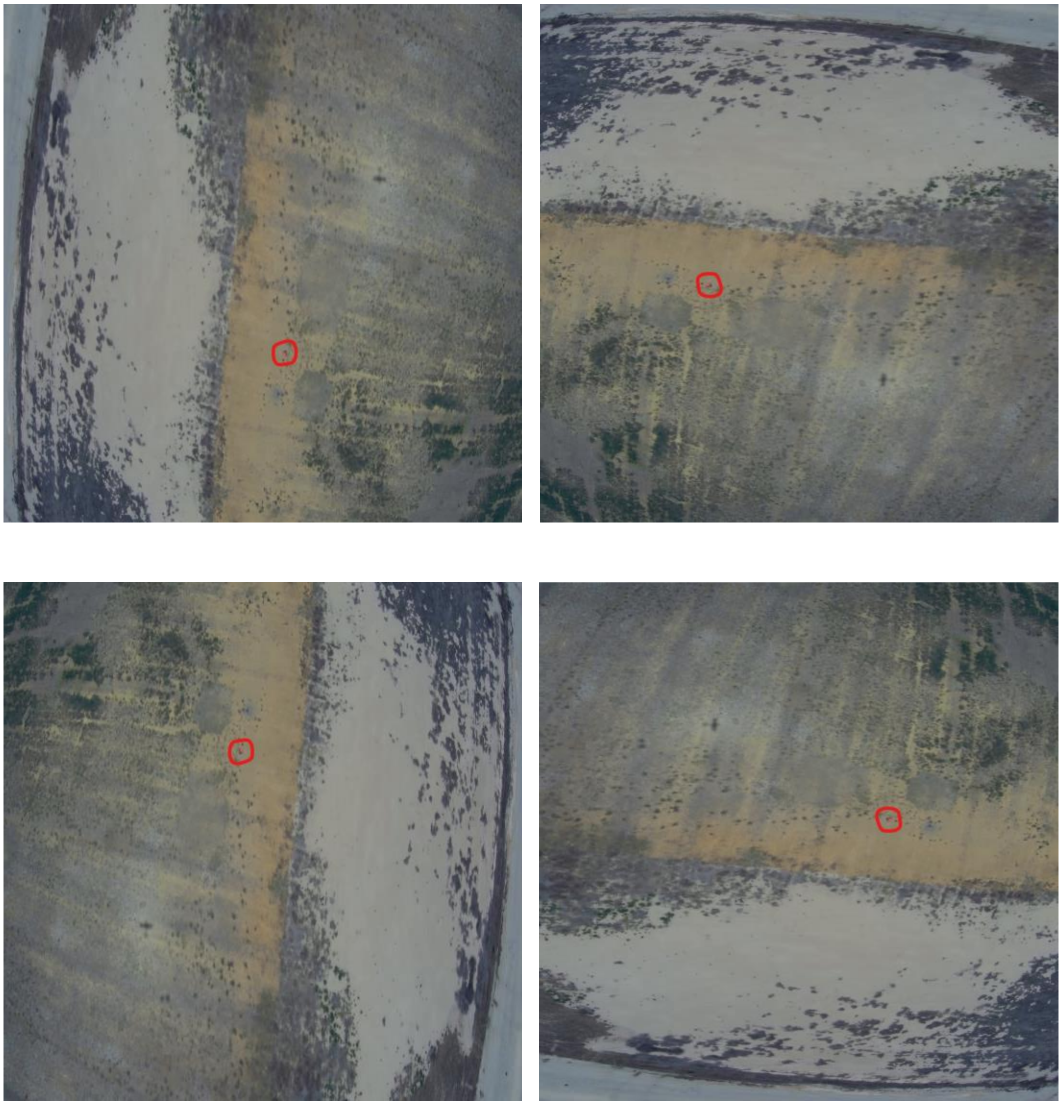

**Figure 6. Flipped Vertically (top left), rotate 90 degrees (top right), rotate 180 degrees (bottom left), rotate 270 degrees (bottom right)**

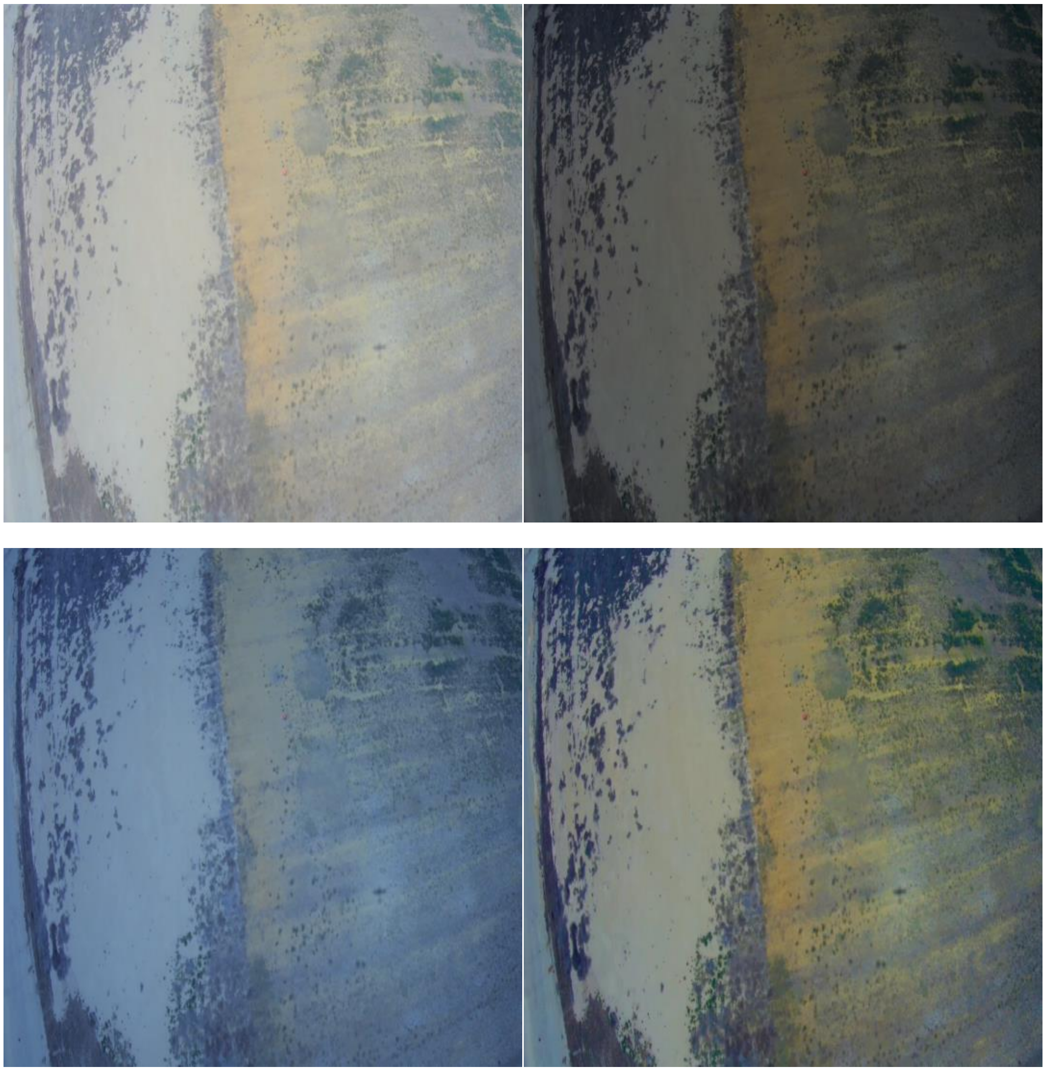

**Figure 7. Increase exposure (top left), decrease exposure (top right), color filter (bottom left), sharpness filter (bottom right)**

These data augmentation technique were applied to all the data containing the target and 20 percent of the data not containing the target. This large variation in data augmentation between data with the target and data without the target is because of the large difference with the classifications of the data collected. In total, the original data contained over 600 frames with the target and over 8,000 frames without the target. The new data with augmentation contained over 10,000 frames with the target and over 15,000 frames without the target.

# REVIEW OF NEURAL NETWORKS

Before delving further into the project and its evaluation a brief introduction of how neural networks work will be beneficial into understanding the methods utilized and the results provided.

A neural network is a type of data structure that is based on the human brain. Brains are made up of billions of brain cells known as neurons [4]. Each neuron is connected to thousands of other neurons sending electrical impulses. These neurons and electrical impulses represent our way of seeing, thinking, and behaving.

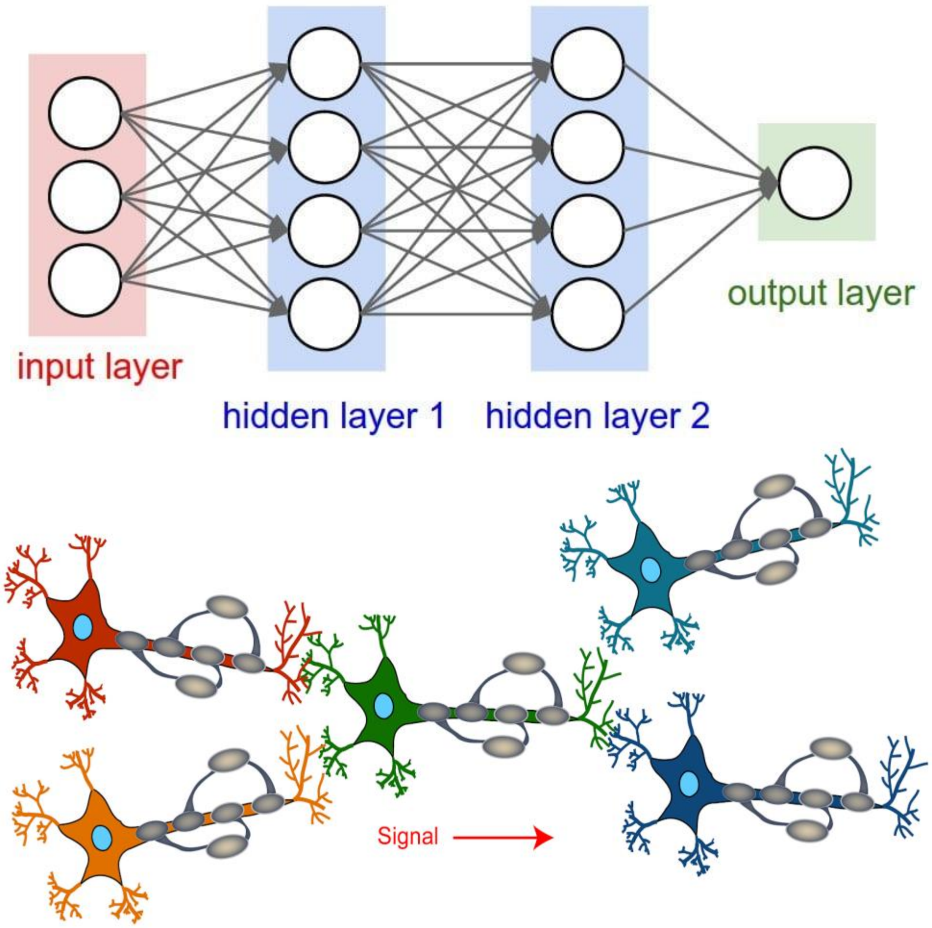

**Figure 8. Neural Network (top) brain cell structure (bottom) [4]**

A neural network consists of a set of neurons known as a layer and several layers with one-way connections. There are three main types of layers in a neural network. The first layer is known as the input layer. This layer can be thought of as the eyes and ears of the network. It is responsible for taking in data from the outside world and feeding it into the network. The second type of layer is called a hidden layer. Hidden layers are responsible for processing the data given by the input layer and recognize patterns and extract the meaningful information. The last type of layer is known as the output layer. The output layer contains the resulting value or values of the overall network [5].

A neuron in a neural network works very similar to a neuron cell. The neuron is an abstract data cell that takes in many inputs, performs varying calculations, and has a single output. This single output is sent to neurons of the subsequent layer. An entire layer of neurons sends their outputs to another layer neurons and when this pattern continues a basic neural network is formed [5].

# TRAINING NEURAL NETWORKS

There are various types of learning paradigms within in the realm of machine learning and with neural networks. The learning paradigms for neural networks may include: supervised learning, unsupervised learning and reinforcement learning [6]. This project focuses solely on the supervised learning paradigm. For a neural network to learn with supervised learning, lots of labeled data is required. For instance, a network that is developed to classify images based on whether or not there is a cat in the picture requires an abundant number of images with cats and an abundant number without cats. When these images are fed into the network, the network must be made aware of which images have cats and which do not, hence the name supervised learning. At the start of training, the network will have a very large error rate, as expected. The output of the network is compared to the expected value and the error is back propagated. That is, the weights of a layer are increased or decreased to change the output value towards the expected value. A neural network learns from a right to left standpoint. That is, the later layers are the first layers to manipulate their weights and reduce the error. This process is continued until the output of the network has a sufficiently reduced error rate [6]. In other words, training stops when a network sees a certain kind of data and knows the appropriate answer.

Evaluating the performance of neural networks requires separation of training and testing data [7]. For this project, 10 percent of data is used for testing. This allows the neural networks to be evaluated for overfitting. That is, a neural network can have very high accuracy on classifying data from the training set but very low accuracy when the network classifies data from the testing set. The testing set consists of data that has never been seen by the network and is therefore the best representation of how the network will perform with new data from other missions.

## DEEP NEURAL NETWORK

In traditional neural networks, only a single hidden layer exists but modern neural networks contain several hidden layers. In the early years of neural networks, it was believed that only one hidden layer was needed [8]. However, further research and experiments demonstrated that adding layers to a network resulted in better performance.

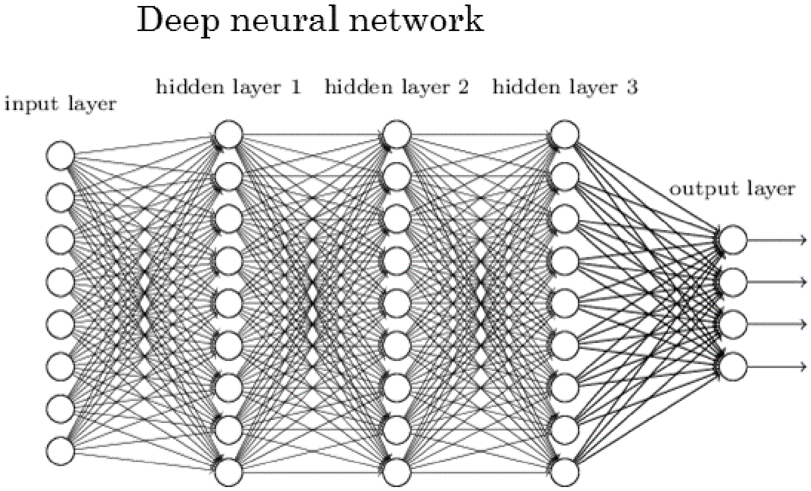

**Figure 9. A fully connected deep neural network [8]**

As mentioned before, the hidden layers of a network recognize and understand the patterns of the data that pertain to a particular output. The earlier layers of a neural network are meant to interpret the simple patterns of the data. For instance, if an image is fed into a neural network, then the early layers might observe color and shapes. The later layers, on the other hand, interpret the more complex patterns of the data, the pattern of the patterns. Going back to the image example, the later layers might recognize shadows, saturation, or depth, since their input is the output of the previous layer. A network with more layers can extract more complex information and interpret more complicated data.

# CONVOLUTIONAL NERUAL NETWORKS

A convolutional neural network is a type of deep neural network except with added kernel convolutions.

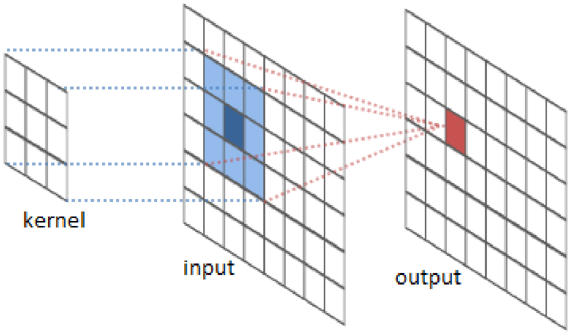

**Figure 10. Kernel convolution [9]**

A kernel convolution is simply the process of having a filter traversing an image and outputting a new image. An image is a 2-dimensional matrix of number values. Each value corresponds to either color or intensity based on the image format. A filter is a much smaller 2-dimensional matrix, usually 3x3, 5x5 or 7x7. As this filter traverses the image, some mathematical operation is performed with the filter values and the image values to output a new pixel value [9].

Convolutional neural networks combine the image manipulation and description power of kernel convolutions with the capable pattern recognition technique of a deep neural network [10]. These types of networks revolutionized the neural network industry [11].

# RESIDUAL NEURAL NETWORKS

Earlier it was stated that a neural network with more layers can learn and recognize more complex data. On the other hand, it was experimentally shown that adding too many layers can increase the error rate [12].

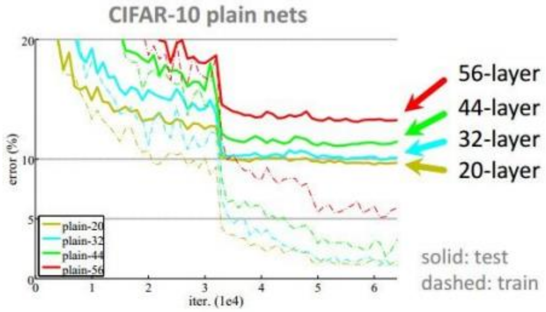

**Figure 11. Error rate of networks with varying hidden layers [12]**

There can be several reasons as to why adding more layers reduces performance. For instance, since a neural network learns through backpropagation, the later layers learn faster [12]. For instance, if the error was found to be eighty percent, the last layers in the network will manipulate their weights so that the error be reduced down to twenty percent. Then, the early layers of the network will do very little weight manipulation since the perceived error is much lower than the actual error. Any fluctuation of data from earlier layers will be handled by the later ones. This causes earlier layers to become less meaningful which is a major problem. The early layers are responsible for understanding the basic and fundamental patterns of the data [5]. If a network does not understand color and shape then it will most likely not understand contrast and depth.

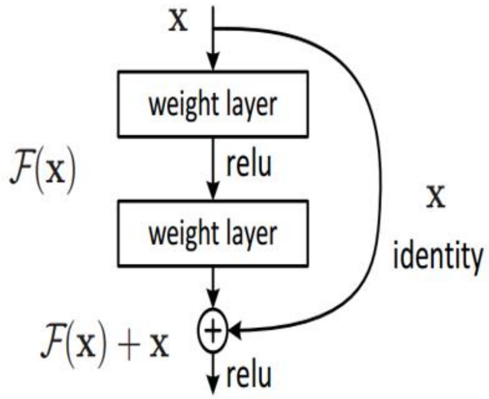

**Figure 12. Snippet of a residual network showing the identity connection (hop connection) [12]**

This was the problem that residual networks aimed to overcome. Residual networks work by adding identity connection that hop over some of the layers [12]. This allows for earlier layers to have more of an impact since their output is spread deeper into the network. Residual networks are also easier to train. Through residual connections it is easier to identify which layers had a greater impact on reducing the error and which layers increased the error. Once identified, the weights of the layers that greatly increased the error rate are changed more than those that did not. This kind of network also helps fine tune the weights since it is easier to identify error of each layer. The fine tuning of the residual is how the network got its name.

The Microsoft Residual Network (ResNet), was the first computer vision program to outperform humans at object classification for the 2015 ImageNet competition. It vastly outperformed previous winners such as AlexNet with 8 layers and a 16.4% error [11], VGG with 19 layers and a 7.3% error, GoogleNet with 22 layers and a 6.7% error [13]. ResNet had 152 layers and a 3.57% error [12].

## EVALUATION OF VGG NEURAL NEWORK

Since the goal of the networks is be object detection, robust and previously proven neural network architectures for object detection were used. Several neural network architectures were developed and trained to experimentally determine which architecture is best suited for this project with the given data. A simple and powerful neural network is the convolutional neural network known as, VGG-Net. This neural network architecture achieved an incredibly low 7.3% error rate on the widely popular ImageNet competition [14]. The architecture consists of several sets of convolutional layers separated by a max pooling layers. The last few layers of the network are fully connected layers responsible for flattening the output into a vector used for classification [11].

The first kind neural network developed and tested consisted of three convolutional layers with intermediate max pooling layers, and a single fully connected layer. The first, second and third convolutional layers have 64, 128 and 256 filters, respectively. These layers use the ReLU activation function and have a stride of 3x3. The network's architecture is based off the VGG architecture since it uses sets of convolutional layers followed by a fully connected layer. During training, this network achieved a training accuracy of 98% but a very low 51% testing accuracy. This is a clear indication of overfitting. The motivation in creating a small scaled version of VGG is because of its possible speed-up over larger VGG architectures. Unfortunately, this version is clearly shown to be a much less powerful version incapable of understanding the patterns that result in a frame containing the target.

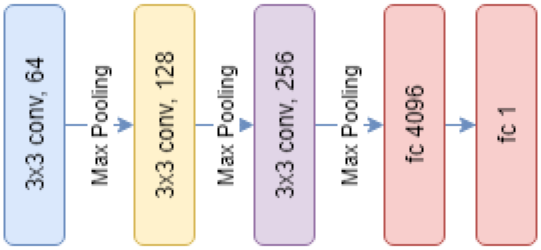

**Figure 13. VGG-5**

The second version that was developed is based exactly off the traditional VGG architecture which multiple sets of convolutional layers having multiple layers. It consists of four sets of convolutional layers, intermediate max pooling layers, and ends with three fully connected layers. Each set of convolutional layers consist of two layers. The first set of layers use 64 filters, the second, 128 filters, the third 256 filters, and the fourth have 512. Each layer uses a stride of 3x3 and the ReLU activation function. This network also had a high training accuracy of 98% but unlike the previous version this network had an incredible testing accuracy of 90%. This shows that even with a relatively small dataset the VGG architecture is powerful enough to help solve the computer vision portion of this project.

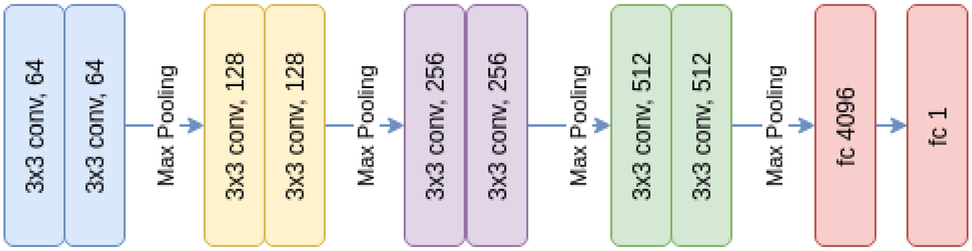

**Figure 14. VGG-10**

The last neural network developed of the VGG architecture is based on the popular neural network known as VGG-16. It consists of sixteen layers and has consistently proved to a robust and powerful architecture for understanding images. VGG-16 has five sets of convolutional layers consisting of two or three layers. The fifth and fourth set of layers consist of three convolutional layers with 512 filters, the third set has three layers with 256 filters, the second has two layers with 128 filters, and the first has two layers with 64 filters. Like the other versions of this architecture developed, each layer uses a 3x3 stride and the ReLU activation function. This version also has an additional fully connected layer.

Similar to the previous versions, this network achieved a very high training accuracy of 98%. Improving the testing accuracy of the previous architecture, this version had an astonishing 94% testing accuracy.

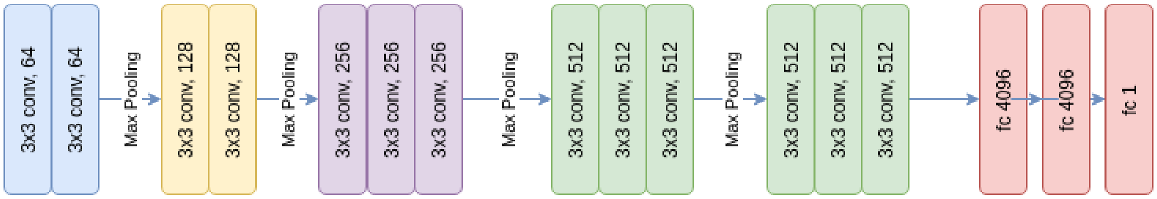

**Figure 15. VGG 16**

| **Architecture** | **Layers** | **Training Accuracy** | **Testing Accuracy** |
|---|---|---|---|
| VGG – 5 | 5 | 98% | 51% |
| VGG -10 | 10 | 98% | 90% |
| VGG -16 | 16 | 98% | 94% |

**Table 1 Comparison of VGG Architectures**

# EVALUATION OF RESIDUAL NETWORK

A residual neural network was also developed for this project. The residual network has the same structure of the VGG networks but also include identity connections between convolutional layers. This network was designed using 35 layers which was made possible because of the identity connections. The layers consisted of four sets of convolutional layers with identity mappings every two layers and a set of fully connected layers at the end of the network. The first set of layers contain six convolutional layers with 64 filters. The second contain eight convolutional layers all with 64 filters. The third set hold twelve layers with 256 filters. The last set holds six convolutional layers with 512 filters. Each convolutional layer has a stride of 3x3 and utilize the ReLU activation function. The training accuracy of this network reached 99% accuracy and executing this network on the testing set revealed an amazing 97% accuracy. Evaluation of the testing set demonstrates the immense flexibility and power of this kind of neural network.

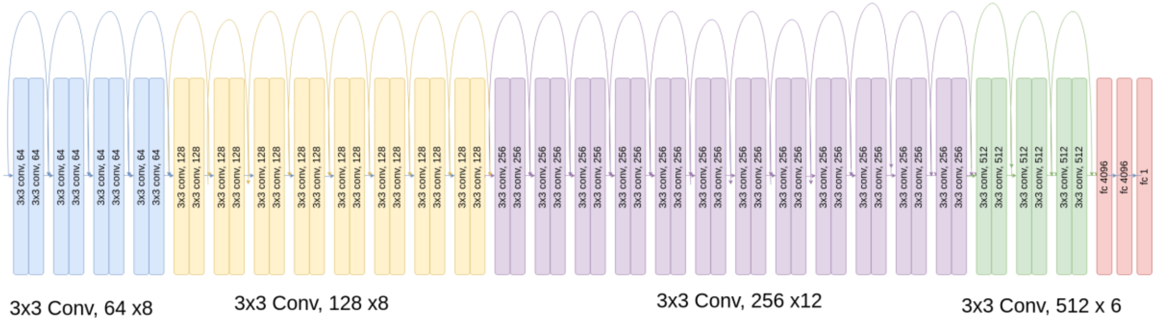

**Figure 16. Residual Neural Network – 35 layers**

| Architecture | Layers | Batch Size | Epochs | Training Time | Training Accuracy | Testing Accuracy |
|---|---|---|---|---|---|---|
| VGG – 5 | 5 | 64 | 40 | 15 mins | 98% | 51% |
| VGG -10 | 10 | 64 | 40 | 1 hour | 98% | 90% |
| VGG -16 | 16 | 64 | 40 | 2 hours | 98% | 94% |
| ResNet – 35 | 35 | 40 | 60 | 10 hours | 99% | 97% |

**Table 2 Comparison of all Neural Networks**

# CONCLUSION

In conclusion, this project shows that neural networks are great mechanisms that can help advance the future of autonomous vehicles. The simple and powerful VGG architectures help resolve some of the computer vision problems faced with autonomous vehicles. The more complex and effective residual neural network architecture improve on the performance of the VGG architectures and further advance scope of the type of problems neural networks can solve. This project demonstrates a foundation in which further research into the use of neural networks can be done. Since positive evaluation from a small custom dataset was achieved by this project with neural networks, it is worth continuing research of more complex neural networks on grander and richer datasets to improve progress on autonomous vehicle applications such as the search and rescue mission this project aimed to support.

# REFERENCES

[1] Rosenblatt, F. “The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain.” Psychological Review, vol. 65, no. 6, 1958, pp. 386–408.

[2] “History of Neural Networks.” Introduction to Neural Network Methods for Differential Equations, by Neha Yadav et al., Springer, 2015, pp. 13–15.

[3] Yang, Simon, and Max Meng. “An Efficient Neural Network Approach to Dynamic Robot Motion Planning.” Egyptian Journal of Medical Human Genetics, Elsevier, 8 Mar. 2000, www.sciencedirect.com/science/article/pii/S0893608099001033.

[4] “A Basic Introduction To Neural Networks.” A Basic Introduction To Neural Networks, 2012, pages.cs.wisc.edu/~bolo/shipyard/neural/local.html.

[5] Bishop, Christopher. Pattern Recognition and Machine Learning. Springer Verlag.

[6] Williams, Ronald J. “Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning.” Reinforcement Learning, 1992, pp. 5–32.

[7] Kesner, Ramond P, et al. “Testing Neural Network Models of Memory with Behavioral Experiments.” Egyptian Journal of Medical Human Genetics, Elsevier, 17 Apr. 2000, www.sciencedirect.com/science/article/pii/S0959438800000672.

[8] Ba, Lei J, and Rich Caruana. “Do Deep Nets Really Need to Be Deep?” Neural and Evolutionary Computing, 2014.

[9] “Generic Filters.” Convolution Matrix, 2012, docs.gimp.org/en/plug-in-convmatrix.html.

[10] Simonyan, Karen, and Andrew Zisserman. “Very Deep Convolutional Networks for Large- Scale Image Recognition.” Conference Paper at ICLR 2015, 2015

[11] Krizhevsky, Alex, et al. “ImageNet Classification with Deep Convolutional Neural Networks.”Communications of the ACM, vol. 60, no. 6, 2017, pp. 84–90.

[12] He, Kaiming, et al. “Deep Residual Learning for Image Recognition.” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[13] Liu, et al. “Going Deeper with Convolutions.” [1402.1128] Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition, 17 Sept. 2014, arxiv.org/abs/1409.4842.

[14] Deshpande, Adit. “The 9 Deep Learning Papers You Need To Know About.” 24 Aug. 2016, adeshpande3.github.io/The-9-Deep-Learning-Papers-You-Need-To-Know-About.html.