National University of Sciences & Technology

Visual Localization in Aerial Imagery using

Convolution Neural Network

Author

TAHIR SHAHZAD

Regn Number

00000118980

Supervisor

DR. HASAN SAJID

DEPARTMENT

SCHOOL OF MECHANICAL & MANUFACTURING ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY

ISLAMABAD

JULY, 2019

Visual Localization in Aerial Imagery using Convolution Neural Network

Author

TAHIR SHAHZAD

Regn Number

00000118980

A thesis submitted in partial fulfillment of the requirements for the degree of

MS Robotics and Intelligent Machine Engineering

Thesis Supervisor:

DR. HASAN SAJID

Thesis Supervisor's Signature: _____

DEPARTMENT

SCHOOL OF MECHANICAL & MANUFACTURING ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY,

ISLAMABAD

JULY, 2019

MASTER THESIS WORK

We hereby recommend that the dissertation prepared under our supervision by:
(Student Name & Regn No.) Tahir Shahzad (00000118980)
Titled: Visual Localization in Aerial Imagery using Convolution Neural Network be accepted in partial fulfillment of the requirements for the award of Masters in Robotics and Intelligent Machines Engineering degree. (Grade _____)

Examination Committee Members

1. Name: Dr. Yasar Ayaz Signature: _____

2. Name: Dr. Muhammad Naveed Signature: _____

3. Name: Ms. Sara Babar Signature: _____

Supervisor's name: Dr. Hasan Sajid Signature: _____

Date: _____

_____ _____
Head of Department Date

COUNTERSIGNED

Date: _____

_____
Dean/Principal

Declaration

I certify that this research work titled "*Visual Localization in Aerial Imagery using Convolution Neural Network"* is my own work. The work has not been presented elsewhere for assessment. The material that has been used from other sources it has been properly acknowledged / referred.

Signature of Student

TAHIR SHAHZAD

00000118980

Plagiarism Certificate (Turnitin Report)

This thesis has been checked for Plagiarism. Turnitin report endorsed by Supervisor is attached.

Signature of Student

TAHIR SHAHZAD

Registration Number

00000118980

Signature of Supervisor

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS thesis written by

Mr./Mrs. Tahir Shahzad (Registration No. 00000118980),

of Department of Robotics and Intelligent Machines Engineering (SMME)

(School/College/Institute) has been vetted by undersigned, found complete in all respects as per NUST Statutes / Regulations, is free of plagiarism, errors, and mistakes and is accepted as partial fulfillment for the award of MS/MPhil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in this dissertation.

Signature: _____

Name of Supervisor: Dr. Hasan Sajid

Date: _____

Signature (HOD): _____

Date: _____

Signature (Dean/Principal): _____

Date: _____

Copyright Statement

Acknowledgements

First and foremost, I would like to be thankful to almighty ALLAH whose blessing and guidance has been a real source of all the achievements in my life. I am Thankful to my parents, my brothers, my sisters and my family for standing beside me throughout my educational career.

It took me longer than average time to complete this research & thesis. In these four years I have been exploring other boundaries, earning money for education and working hard to make it possible. It was a tough time to do jobs and studies. Listen to bosses and clients at one time and fellows and teacher on other time. We are surrounded by people who de-motivate our struggle. I know many who couldn't complete the journey. I would like to pat my back on achieving this degree and I hope to do more.

I am really Thankful to all my well-wishers and friends for always making me smile and for helping me in studies. It wasn't very interaction time with my classmates but still there were times we made memories. I have best wishes for all, wherever they go.

This research & thesis became difficult to me at various stages in the writing of this documentation and in understanding the cutting edge algorithms and the science behind machine learning, from generating dataset and fine tuning neural networks. In this regard I owe a debt of my sincere supervisor "Dr. Hasan Sajid", always a smiling face and a kind person in the form of a teacher, who always guided me and gave me the opportunity to learn multiple tools.

A special note of Thanks to worthy teachers and staff members of the department of RIME. I would like to mention special thanks to Dr. Yasar Ayaz, Dr. Muhammad Naveed and Mr. Fahad Islam for their kind help, guidance and acknowledgment in university.

Finally, I would like to express my gratitude to all the individuals who have rendered valuable assistance to my study.

Dedicated to my family; my supporting father Ghulam Mujtaba, my caring mother and my loving brothers & sisters without whom I wouldn't have been able to complete this.

Abstract

The purpose of this research is to propose a novel approach using deep neural networks to estimate the position and orientation of aerial vehicle using vision sensor only. We will create and contribute an annotated dataset for visual odometry in aerial images. It will provide an alternative solution to locations where GPS is not working accurately. A novel DNN will be proposed and trained on data set that will cover diversity of land areas including cities, villages, forests, deserts, lakes, farms.

Key Words: *GPS Denied Environment, Travelling Distance Estimation, Convolutional Neural Networks, Deep Neural Network*

Table of Contents

List of Figures

CHAPTER 1: INTRODUCTION

This thesis is divided into 6 chapters. A brief introduction of each chapter is given here. In CHAPTER 1: we will be talking about GPS denied environment, aerial vehicle, their utilization in the real world and need of addressing challenges. **CHAPTER 2:** is about overview of similar work done before and after neural network utilization, **CHAPTER 3:** explains the methodology we used to acquire dataset, its formation and utilization, **CHAPTER 4:** we describe convolutional neural networks, end-to-end pipeline we proposed, **CHAPTER 5:** comprises of the scenarios we developed to test the proposed methodology and derive some knowledge. **CHAPTER 6:** We conclude this thesis under the light of available resources and experiments, along with a proposed pipeline to extend this work.

1.1 Navigation in UAVs

UAVs (Unmanned Aerial Vehicles) have achieved considerable attention to cater to the human needs, especially in the recent years. Also termed as drones, UAVs originated in 1849 when balloons were used by Austria to carry explosives in warfighting [1]. With crude techniques and embryonic technology, UAVs officially surfaced in 1917, during World War I. Finding extensive usage in the military applications, UAVs were regarded as the Eyes of the Army [2]. Since then, UAVs have been employed in numerous military, industrial, and domestic applications.

UAVs have propelled exponential growth in various sectors vis-à-vis disaster management, agricultural monitoring, weather monitoring, exploration of natural resources, aerial mapping, forest firefighting, freight transportation, etc.

UAVs are either autonomous or remotely controlled, i.e., via RF controllers. Equipped with a GPS (Global Positioning System) and coupled with an IMU (Inertial Measurement Unit), the autonomous UAVs have taken center stage in all domains. GPS and IMU, together, provide necessary and reliable information about time, position, and orientation of a UAV. However, GPS is not an absolute solution for UAV navigation, guidance, and control.

Certain factors contribute to blind a GPS in certain regions, referred to as GPS-denied environments. These areas typically include dense metropolitan cities and deep canyons where GPS receiver is unable to communicate with the satellites. According to the Volpe Report [3] released in 2001, GPS vulnerabilities are associated to RF interference caused by cell phone/TV/radio disruptions, ionosphere interferences, spoofing, severe weather conditions, obstacle occurrence, and jamming, etc. Being unable to establish a connection to the satellite, a UAV may follow an arbitrary path and might be lost. This makes GPS-based navigation highly unreliable and creates a need for a sensor that could locate a UAV in GPS-denied areas.

1.2 GPS Denied Environment

Generally, an IMU is integrated with GPS for velocity, orientation, and distance measurement of a UAV. An IMU is equipped with three accelerometers and three gyroscopes that provide information regarding linear and angular movements of the UAV. In the absence of a GPS, an IMU cannot act as a standalone sensor for velocity, position, and attitude estimation. IMU's readings slightly drift over time, and the aggregate error can adversely affect [4] precision maneuvers and critical military operations.

For decades, researchers have been putting efforts into finding effective and optimum solutions to UAV navigation in GPS-denied environments. Advancements in computer vision have taken the domain of UAV navigation by storm, and numerous vision-based navigation and localization techniques have surfaced to estimate the position of a UAV using an on-board camera. Vision-based navigation works on the principle of feature extraction from incoming images, whereby the position of a UAV is calculated.

SLAM (Simultaneous Location and Mapping) is a popular technique for vision-based UAV navigation, and numerous researchers [---] have employed SLAM in their proposed methods. SLAM localizes a UAV by mapping the surrounding environment and exploiting the mapped features for position estimation. Both on-board and off-board processing techniques, for position and distance estimation, have been adopted for SLAM-based UAV localization.

Alternatively, geolocalization is an effective solution for UAV localization in a GPS-denied environment by incorporating satellite images. With the evolution of deep learning, Convolutional Neural Networks (CNNs) have emerged as a powerful tool to extract features of an image. A brief review of applications of deep learning in aerial imagery has been presented by [31] that dilates upon the significance of deep learning models and their state-of-the-art performance for achieving aerial navigation tasks. Geolocalization is a very efficient approach for outdoor UAV navigation.

The aim of this paper is to propose the design of a vision-based sensor for distance estimation of UAV from a starting point. To the best of our knowledge, this is the first ever technique of UAV distance estimation that applies CNNs to two images captured by the on-board UAV camera at two consecutive instants and

extracts features from these images to measure the distance traveled during these two time instants. Initially, the model is trained using satellite images from Bing Maps [add ref], and later, real-time images are fed to the model to calculate the distance between feature points. This is a novel approach of distance estimation using a single camera as it does not require any predefined or pre-loaded maps during the flight once the model is trained. Moreover, trained in one environment, this on-board processing system can estimate the position of a UAV from a fixed point in other settings as well.

In this paper, a low-cost solution for distance estimation of UAV in GPS-forbidden regions has been presented using Convolutional Neural Networks. The rest of the paper is organized as follows: Section II presents an overview of the previous works related to UAV localization, navigation, and distance measurement. Section III briefly describes the proposed method for Visual Localization & Navigation of Aerial vehicle in GPS denied environment using Convolution Neural Network. Section IV discusses the experimental outcomes and accuracy measures. Section V concludes the entire discussion, discussing the drawbacks of this system and suggests some future implementations for vision-based UAV navigation.

CHAPTER 2: REVIEW OF RELATED RESEARCH

Navigation and localization of robots in GPS-denied environments is an extensive field of study, and researchers have proposed numerous approaches cater to this problem. We have presented a brief overview of the methods introduced by other researches to locate a UAV in GPS-denied terrains. Primarily, we have divided these approaches into two categories: SLAM-based sensor integration and geolocalization using reference maps.

2.1. SLAM based localization

A detailed study on SLAM has been presented by [8, 9]. SLAM (Simultaneous Localization and Mapping) is a process by which a mobile robot can form a map of its surrounding environment and at the same time, use it to infer its location without any prior knowledge of that place [8]. In 2009, Ahrens et al. [2] adopted an off-board vehicle tracking approach using a monocular camera for indoor MAV navigation. Visual Odometry allowed the MIT quadrotor to hover for 96s, without drifting, using EKF-based SLAM for vehicle localization. Oscillations during flight reduced persistence due to out-of-bound features. In this system, inertial sensors are employed for distance measurement from a reference point.

A downward-looking monocular camera, paired with an IMU and PD (Proportional Derivative) controller, was used by Blosch et al. [3] in 2010 that allowed MAV to navigate in unseen environments. It was the first use of visual SLAM for take-off, landing, hovering, localization, and stability gain with an estimated RMS drift error of just 2-4 cm. In 2011, the authors of [3] expanded their previous research and came forward with Swarm of Flying Objects (SFLY) project

[5] that incorporated a downward looking monocular camera, IMU, and onboard processor to perform takeoff, navigation, and landing of MAVs. During the same year, RGB-D sensors were employed by [6] for indoor autonomous aerial vehicle navigation using global SLAM. FAST feature correspondence between RGB-D frame pairs was utilized by FOVIS visual odometry system, and sparse bundle adjustment (SBA) was used for consistency. In these methods, SLAM allows the UAV to navigate utilizing building simultaneous trajectories; distance estimation is performed using inertial sensors.

In 2013, Scaramuzza et al. [18] coupled 5DOF pose of a monocular camera with inertial measurements of an IMU, using EKF for real-time localization of a lightweight MAV. This SFLY project used visual-SLAM for onboard local navigation and off-board global navigation for tracking multiple MAVs; however, distance estimation was performed through IMU. In 2014, Leishman et al. [1] demonstrated a relative navigation approach using an RGB-D sensor, IMU, and sonar altimeter, where an onboard stereo vision for visual SLAM was coupled with MEKF for localization. Image keyframes, represented as nodes, were used for landmark identification in a closed-loop flight. The data coming from the IMU was being manipulated for distance measurement.

In 2015, one of the researchers [19] fused vision with GPS sensors to perform precision maneuvers using vision-based SLAM. Initially, readings from the GPS were used for a short span of time to calibrate the metric scale of estimation. Afterward, UAV relied solely on visual information for distance estimation.

Visual Odometry (VO) has contributed a lot in the field of robotics. A detailed study on the research carried out on visual odometry has been presented in [10,11]. These articles give an extensive overview of the advancements in the field of VO

during the past few decades. The term Visual Odometry was first coined by Nister et al. [12] in 2004, and since then, VO has been employed by many researchers in different applications. (Possibly in the intro too)

A VO-based method for distance estimation of ground vehicle applications was proposed by Nister et al. [4] in 2006 that used a stereo camera for challenging outdoor environments. Researchers used a monocular approach for emphasizing feature tracking and compared VO to DGPS and INS that makes VO effectively robust. [7] estimated the egomotion of the vehicle using a series of homography relationships between features in consecutive images. Features were tracked continuously for each incoming sample. In this technique, researchers integrated vision system with IMU for distance calculation of a UAV.

A few researchers [13-15] have focused on using visual odometry for image-based distance measurement between the target and the camera. In 2012, Bourdonnaye et al. [13] computed the distance of a target from the UAV using a single camera and employing the lens equation. Texture-based background subtraction method was adopted in this method. The distance of a drone within a camera FOV was measured by [14] in 2015, where CNNs were used for drone detection and distance estimation. A vision-based approach was used for object detection; however, the mainstream of interest was object detection, and the measured distance is unreliable for any arbitrary UAV [15]. Triangular and parallel distance measurement techniques also give an intuition of distance estimation using a CCD camera [16]. In our method, a vision-based sensor allows the UAV to calculate its distance from a reference point so that it can track itself in areas where GPS does not operate.

Two years later, in 2017, [20] employed the visualization principle of a bee's compound eyes for distance estimation. The number of opaque channels, referred to as ommatidia, estimate the distance to an unknown target by calculating the target's angle at two different locations as the object moves towards it. This system proves to be highly imprecise when the target's angle goes beyond 3-10°. These researchers have contributed a great deal; however, in contrast to our method, these approaches are trying to capture a drone if it enters a particular area, just like radar.

2.2. Geo-localization using Satellite Maps

Other studies have applied intelligence modelling techniques in operational research to predict bank failures and crises. Amongst the most widely used intelligence technique is Neural Networks (NN). NN models contain mathematical and algorithmic substances that portray biological neural networks of the human nervous system. Some examples include Celik & Karatepe (2007) [8], who used artificial neural network models to forecast crises, and Alam et al. (2000) [9], whose study used fuzzy clustering and self-organizing neural network in identifying failed banks.

A research conducted by Boyacioglu et al. (2009) [10] compares numerous NN, Support Vector Machine (SVM), Multivariate Discriminant Analysis, Cluster Analysis and Logit regression analysis applied in CAMELS setting to detect bank failures in Turkey. The results indicated that Multivariate Discriminant Analysis and Logit regression analysis are better failure predicting models among all others.

A multilayer NN model, known as Back-Propagation Neural Network (BPNN) model. was used by Tam (1991) [11] to successfully predict Texas bank failures almost one to two years before the collapses. BPNN is the most commonly used

classification and prediction method as it outperforms other models. The first and last layers consist of input and output units, while the middle layer consist of hidden units. The unique feature of the BPNN model is that the errors generated by the hidden layers are calculated by back propagating the errors of the output sent by levels of the corresponding layer. Tam (1991) [11] used CAMELS variables in his research and concluded that BPNN outperformed DA, Logit and K-nearest neighbor technique, in predicting bank failures accurately.

Tam & Kiang (1992) [12] applied linear discriminant analysis (LDA), Logit, K-Nearest Neighbour, Interactive Dichotomizer 3 (ID3), feedforward NN and BPNN in predicting bank failure. Amongst all the models applied, BPNN outperformed all models for one-year prior samples while LDA outperformed the rest for two years prior samples. However, BPNN outperforms all, in both one- and two-year prior samples for holdout samples and in jackknife method. They concluded their study by indicating that NN outperforms DA method.

A study was conducted by Bell (1997) [13] for predicted bank failures using Logit and BPNN models. His findings indicate that neither Logit nor BPNN model dominates one another in terms of predictability. The methodology applied twelve input nodes, six hidden nodes and one output node in BPNN. Concluding that BPNN is better for complex decision processes.

Swicegood & Clark (2001) [14] on the other hand found BPNN outperforms other models in identifying underperforming banks. The study compared DA, PNN and human judgment in bank failure prediction.

CHAPTER 3:DATASET

Till the development of this research, no such dataset is available that contains aerial imagery along with travelled distance as the visual scene changes. A convolutional neural network can detect similarity between two images. It is required to explore if a neural network identify distance in partially overlapping images of ground. We developed a dataset that incorporate distance in captured images at time t-1 and t.

3.1. Satellite Maps and APIs

Aerial imagery along with GPS coordinates are not available. no such dataset have been developed so far. Using a drone and capturing dataset is a tedious and manual task that needs a lot of time which was not suitable under the time frame of thesis research.

With the advancements in technology and satellites, latest maps are available on the internet like Google Maps, Bings Maps, Open Street Maps and many more. These maps have precise latitude and longitude incorporated with aerial view of world.

APIs are available to extract any location of world in aerial imagery based on latitude and longitude. we selected random geo coordinates to extract available aerial images.

Each unit of data set comprises on a pair of images having some overlapping ground area with random distance and orientation between them. Let's assume we captured an aerial view at time t-1 of latitude1, longitude1. Then we captured aerial view of latitude2, longitude2 at time t. distance between two geo coordinates is

random but less than 200 meters in x-axis and y-axis. Direction is calculated by angle theta between two points.

If we consider aerial image as aerial view, our system input is a pair of aerial view that have partially common aria. Central point of second aerial view is randomly placed at a range distance of 0-200 meters in x-axis and y-axis from central point of first aerial view. Here central point represents latitude and longitude of aerial view. As displacement is always positive, it is also necessary to train network with direction of movement. To find the displacement and direction between pair we used haversine formula and bearing formula respectively for pair of latitude and longitude.

Figure 3.1: Sample Data Pair

It is important to understand the relationship of two geo coordinates in mathematical terms. Imagery latitude and longitude plotted on earth give us distribution of earth's surface. We used the following formulas to get distance and angle between two geo coordinates.

3.1.1. Haversine Formula:

To find the shortest distance between two points on a sphere we use Haversine formula. Similarly It enable us to calculate the distance between two geo coordinates.

$$a = \sin^2(\Delta\varphi/2) + \cos\varphi 1 \cdot \cos\varphi 2 \cdot \sin^2(\Delta\lambda/2)$$

$$c = 2 \cdot \operatorname{atan2}(\sqrt{a}, \sqrt{(1-a)})$$

$$d = R \cdot c$$

where $\varphi 1,\lambda 1$ are latitude, longitude of aerial view at time t, $\varphi 2,\lambda 2$ are latitude, longitude of aerial view at time t+1. R is the radius of earth. ($\Delta\lambda$ is the difference in longitude)

3.1.2. Bearing Formula:

In aircrafts we use bearing or heading angle that defines navigation of aerial vehicle. Bearing is an angle, between the north-south line of earth(meridian) and the line connecting reference and target point. It enables us to find the direction between two geo coordinates.

Bearing from latitude, longitude of time t-1 and latitude, longitude of time t can be calculated as,

$$\theta = \operatorname{atan2}(X, Y)$$

$$X = \sin \Delta\lambda \cdot \cos \varphi 2$$

$$Y = \cos \varphi 1 \cdot \sin \varphi 2 - \sin \varphi 1 \cdot \cos \varphi 2 \cdot \cos \Delta\lambda$$

where $\varphi 1, \lambda 1$ is the start point, $\varphi 2, \lambda 2$ the end point ($\Delta\lambda$ is the difference in longitude)

3.2. Challenges with the dataset

While collecting dataset with APIs, we had to ensure that geo coordinates are not in sea. In satellite maps, lakes, forests, deserts and deep sea have very low features and more similar ground patches. Picture taken at time t-1 and time t will look almost similar. Using traditional approaches we will not be able find enough features to train proposed network.

As the data gets bigger, it is a bigger challenge to clean data. We extracted millions of data pairs using APIs. It is important to filter all data pairs to get only correct dataset. While downloading aerial imagery from internet we sometimes lose packets and image is corrupt.

For the limited capacity of this degree we used simulated environment. The critical issue will be to generate actual dataset with a drone capturing aerial imagery. In that case we are dependant of GPS device mounted on aerial vehicle to generate dataset. Cross checking this dataset will be challenging or our network will be trained without addressing inaccuracies of GPS.

CHAPTER 4:PROPOSED METHODOLOGY

Experiments in recent years show promising results achieved by convolutional neural networks. CNN are suitable for spatial imagery, similar to the dataset we generated. so far convolutional neural networks are used for classification, recognition and regression problem. It extract features from given image and map with the expected output.

4.1. Artificial Neural Networks

ANNs are computing systems that are inspired by the human neurological system. Each neuron is connected with other neurons making a mesh network. Each neuron receives a signal, process it and transmits to connected neurons. In ANN this signal is a real number. Each neuron perform some non-linear function of the sum of its inputs. Neurons in artificial network have some weights that adjust based on some loss function to achieve accuracy.

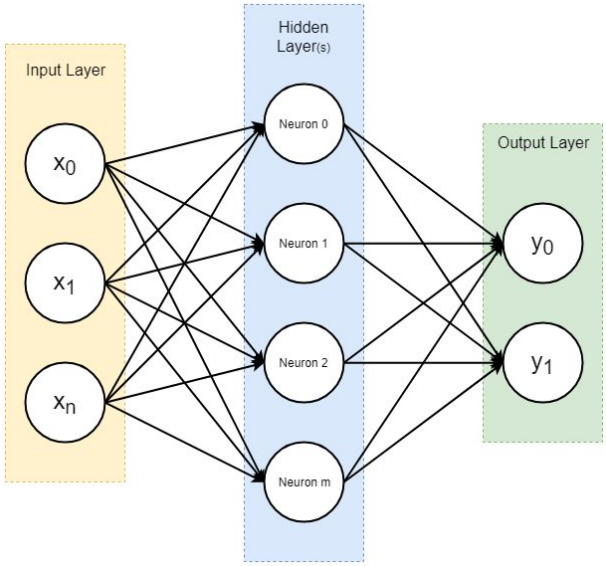


Figure 4.1: Typical ANN Network

4.2. Convolutional Neural Networks

Convolutional neural networks[33] combine computer vision with deep networks. ConvNet takes image as an input, assign importance(weights & biases) to parts of image and learns filters/features. ConvNet requires much less pre-processing in comparison with hand engineered algorithms of computer vision.

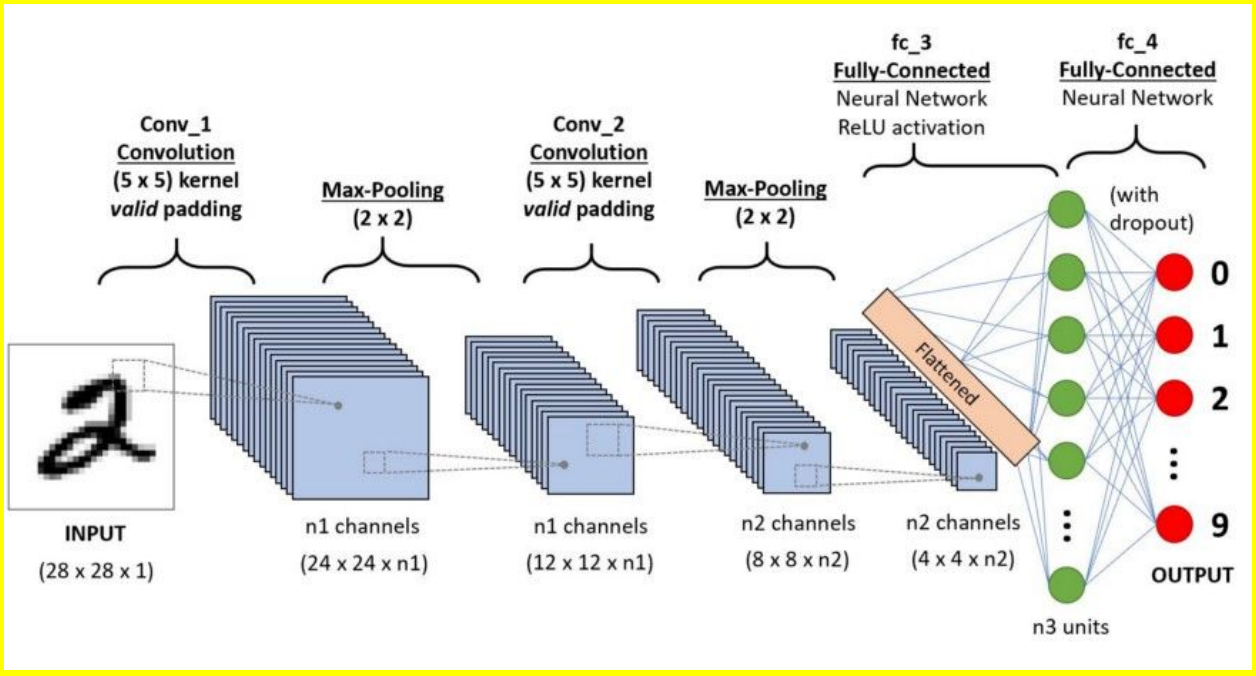


Figure 4.2: A Typical CNN Network

In its core images are matrix of pixel values. In convNet input image is a multidimensional vector of numbers.

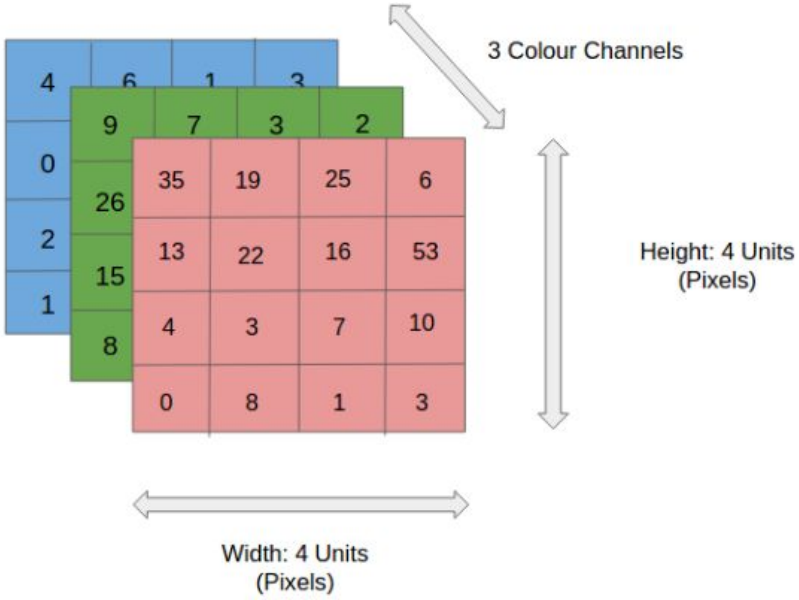


Figure 4.3: Sample 4x4x3 RGB Image

ConvNet reduces image vectors without losing image features into a form that is easier to process.

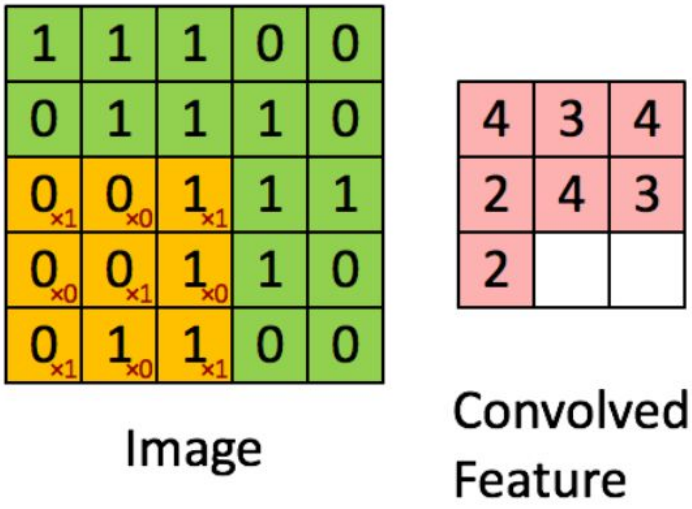


Figure 4.4: Convolution operation on a matrix

4.3. Siamese Network

Siamese network is a specialized form of ANNs where two network use same weights on input of both networks and compute comparable output.

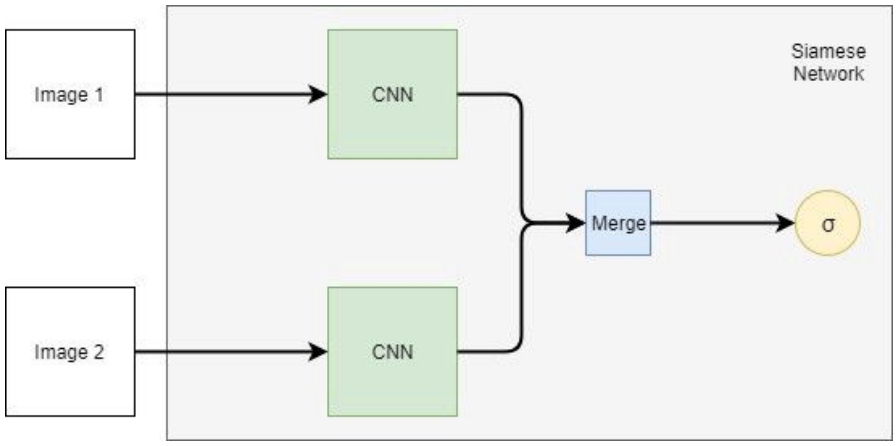


Figure 4.5: Architecture of siamese network

Siamese networks are used to identify differences in two similar instances. It can be identification of handwritten digits, face recognition and finding differences in two documents.

In case of aerial imagery we develop a similar structure but modified similarity function. Instead of finding differences we let the neural network decide to match features extracted from both images and find distance.

In case of aerial imagery at time t-1 and time t, it will have partially overlapping area. We propose the difference in both images will be helpful to identify distance covered. If there is no overlapping area in both images at time t-1 and t then distance will be infinite.

4.4. Deep Siamese Distance Estimation Network

We propose deep siamese distance estimation network to find distance covered in two aerial views that have some overlapping areas. It takes input RGB image with 224x224 height and width respectively. Neural network will have input with dimensions 224x224x3.

First part of neural network architecture consists of two CNN with 3 sets of layers. Second part of neural network consists of single CNN with two sets of layers.

In siamese part of network, image at time t-1 is input for CNN labeled as A while image at time t is input for CNN labeled as B. Each siamese neural network have same architecture.

First CNN layer takes input of 224x224x3. It is mapped with 224x224x64 at output. Second CNN layers takes the first output and result with same dimensions. The resolution of input is preserved after convolution. A 2x2 pixel window is used for pooling convolutional layer with stride 2.

Pooling layer is followed by 2 convolutional layers with feature map of 112x112x128. Another pooling layer is attached at the end of 2nd set of convolutional layers.

Third and last set of convolutional layers in siamese network consists of 56x56x256 feature map. After pooling both branches of siamese network are combined that makes a feature map of 28x28x512.

4th and fifth set of convolutional layers have feature map of 28x28x512 and 14x14x512 followed by pooling layers.

At the end of neural network we attach 3 fully connected layers that transform output of CNN into 1x1x4096, 1x1x1000 and 1x1x3. output layer is used to extract distance is x-axis, y-axis and angle.

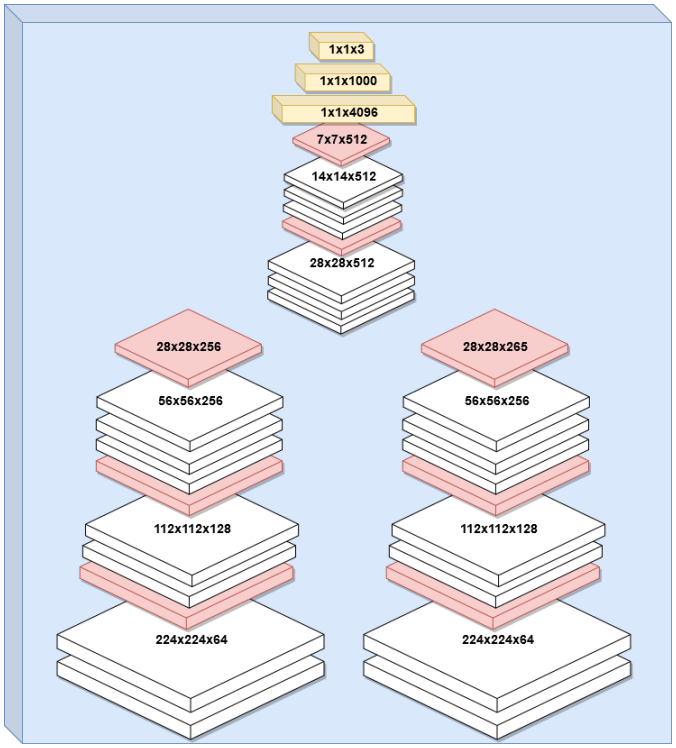


Figure 4.6: Deep Siamese Distance Estimation Network

4.5. Proposed Architecture

We initialize our system at time t-1 with aerial view, latitude and longitude. At time t aerial vehicle have another visible area of ground. These two input images are input for the distance estimation siamese network that estimates distance covered in x-axis, y-axis and angle between two aerial views.

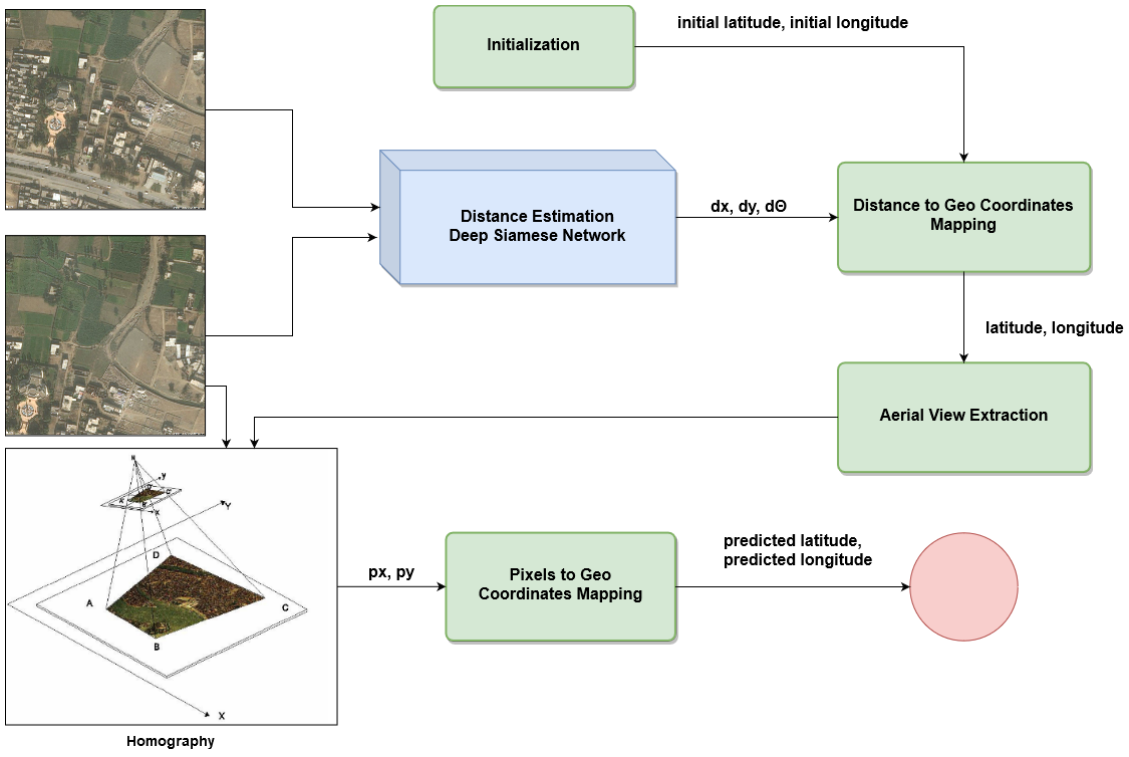


Figure 4.7: Proposed Architecture

Our system maps these distances(dx, dy and dTheta) using initial latitude and longitude of time t-1 with mathematical formulas into latitude and longitude at time t. These estimation may have error.

To accurately identify the difference between aerial view and identified location on map system use ORB detector and image homography. differences in pixels are again mapped into latitude and longitude.

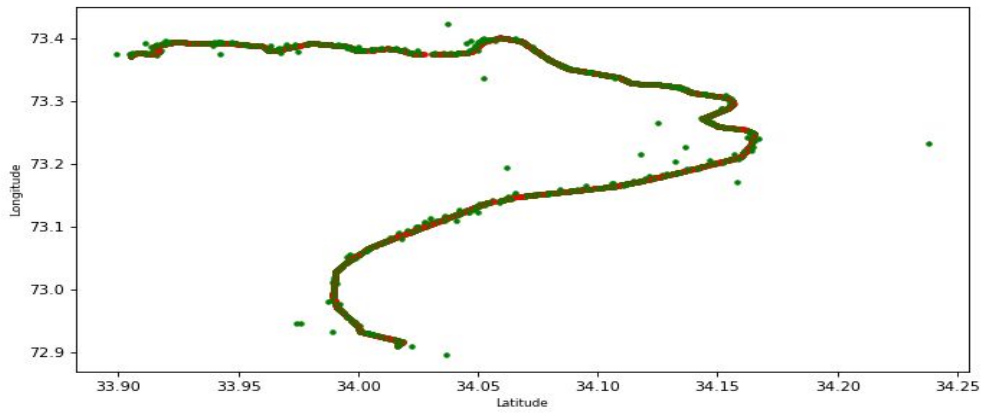
Finally our system use moving average filter to eliminate predictions that are very far from possibility.

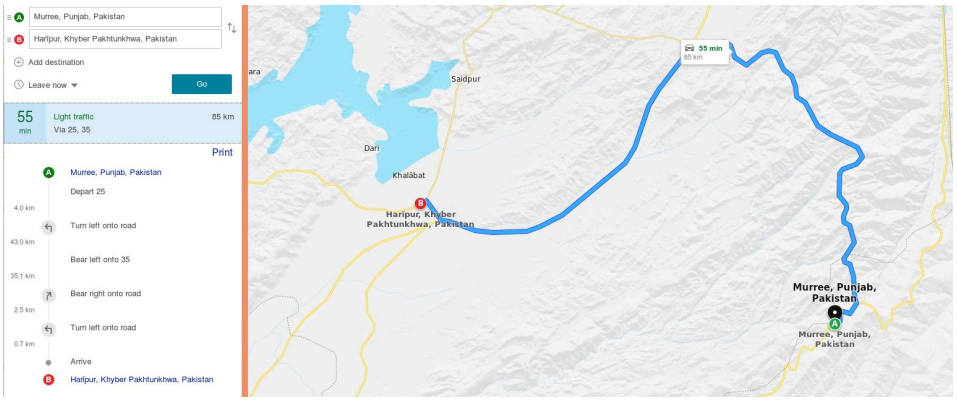
CHAPTER 5:EXPERIMENTS AND RESULTS

In this chapter we present different experiments we evaluated to compare proposed architecture and baseline implementations.

We simulated the evaluation environment with Google maps on a route from point A to point B. Distance travelled in each unit time is random.

5.1. Localization with Distance Estimation





Ground Truth .....

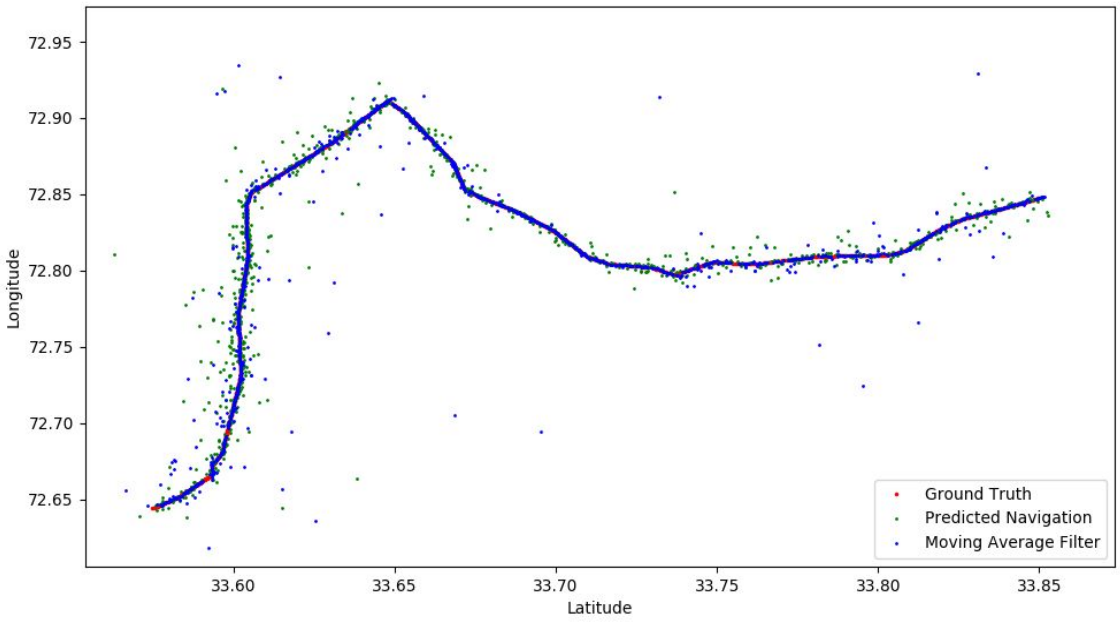
Predicted navigation .....

Figure 5.1: Simulated drone travel from Murree to Haripur

Distance estimation is purely neural network based solution that gives distance travelled in x-axis, y-axis and theta. A simulated aerial vehicle travel from Murree to Haripur covering 85KM distance with random chunks.

5.2. Localization with Distance Estimation & Moving Average Filter

In our next experiments we extent distance estimation network along with ORB filter matching and homography to achieve better results followed by moving average filter. A simulated drong traveled from Fatehjang to Khalabat covering a total distance of 81KM. It included greenery, roads, buildings, curves and urban areas. we evaluated that average error in x-axis is 71meters , in y-axis it is 93 meters and in angle it is 134 degrees.



Ground Truth .....
Distance estimation .....
Proposed architecture .....

Figure 5.2: Simulated drone travel from Fatehjang to Khalabat

CHAPTER 6:Conclusion and Future work

Results showed that proposed architecture combined with distance estimation can work in the GPS denied environment. It can use onboard camera as an additional sensor for navigation in drones.

Distance estimation can perform better based on more data diversity. We need to gather more data and train network.

In future work we need to incorporate Kalman filter. It will enable us to avoid unrealistic predictions.

6.1. Searching Neighborhood

It is possible to lose track in GPS denied environment. There are many possible reason for that. We are going to highlight two.

- When aerial imagery taken at time t-1 and time t doesn't have overlapping area.
- There are very few extracted features in aerial imagery to match like in deep forest, sea, desserts etc.

In future work, this can be addressed using conventional techniques along with neural networks to find a patch where aerial vehicle should be. Once we can find an aerial imagery in given map, it is possible to re-initialize our system.

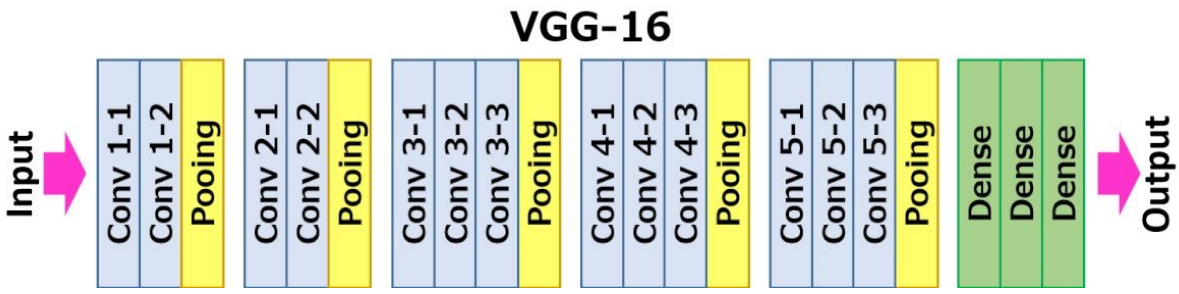
6.2. Real World Experiments

So far our training and evaluation is based on simulation environment using Google and Bing Maps. It is difficult but required to generate dataset for real drone with latitude and longitude values at each aerial view.

Appendix A: Neural Networks

VGG16

VGG16 is a convolutional neural network model proposed by K. Simonyan and A. Zisserman from the University of Oxford in the paper "Very Deep Convolutional Networks for Large-Scale Image Recognition". The model achieves 92.7% top-5 test accuracy in ImageNet, which is a dataset of over 14 million images belonging to 1000 classes. It was one of the famous models submitted to ILSVRC-2014. It makes the improvement over AlexNet by replacing large kernel-sized filters (11 and 5 in the first and second convolutional layer, respectively) with multiple 3×3 kernel-sized filters one after another. VGG16 was trained for weeks and was using NVIDIA Titan Black GPU.



VGG16 Abstract Architecture

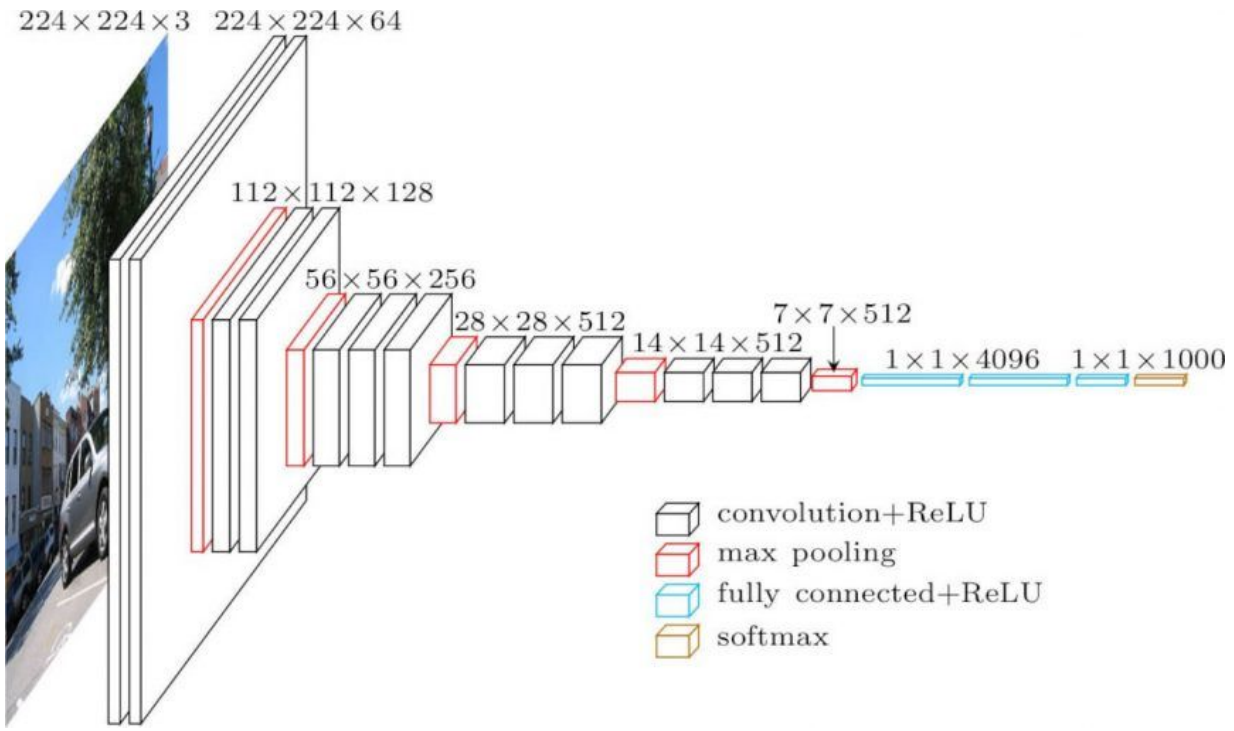
The input to cov1 layer is of fixed size 224 x 224 RGB image. The image is passed through a stack of convolutional (conv.) layers, where the filters were used with a very small receptive field: 3×3 (which is the smallest size to capture the notion of left/right, up/down, center). In one of the configurations, it also utilizes 1×1

convolution filters, which can be seen as a linear transformation of the input channels (followed by non-linearity). The convolution stride is fixed to 1 pixel; the spatial padding of conv. layer input is such that the spatial resolution is preserved after convolution, i.e. the padding is 1-pixel for 3×3 conv. layers. Spatial pooling is carried out by five max-pooling layers, which follow some of the conv. layers (not all the conv. layers are followed by max-pooling). Max-pooling is performed over a 2×2 pixel window, with stride 2.

Three Fully-Connected (FC) layers follow a stack of convolutional layers (which has a different depth in different architectures): the first two have 4096 channels each, the third performs 1000-way ILSVRC classification and thus contains 1000 channels (one for each class). The final layer is the soft-max layer. The configuration of the fully connected layers is the same in all networks.

All hidden layers are equipped with the rectification (ReLU) non-linearity. It is also noted that none of the networks (except for one) contain Local Response Normalisation (LRN), such normalization does not improve the performance on the ILSVRC dataset, but leads to increased memory consumption and computation time.

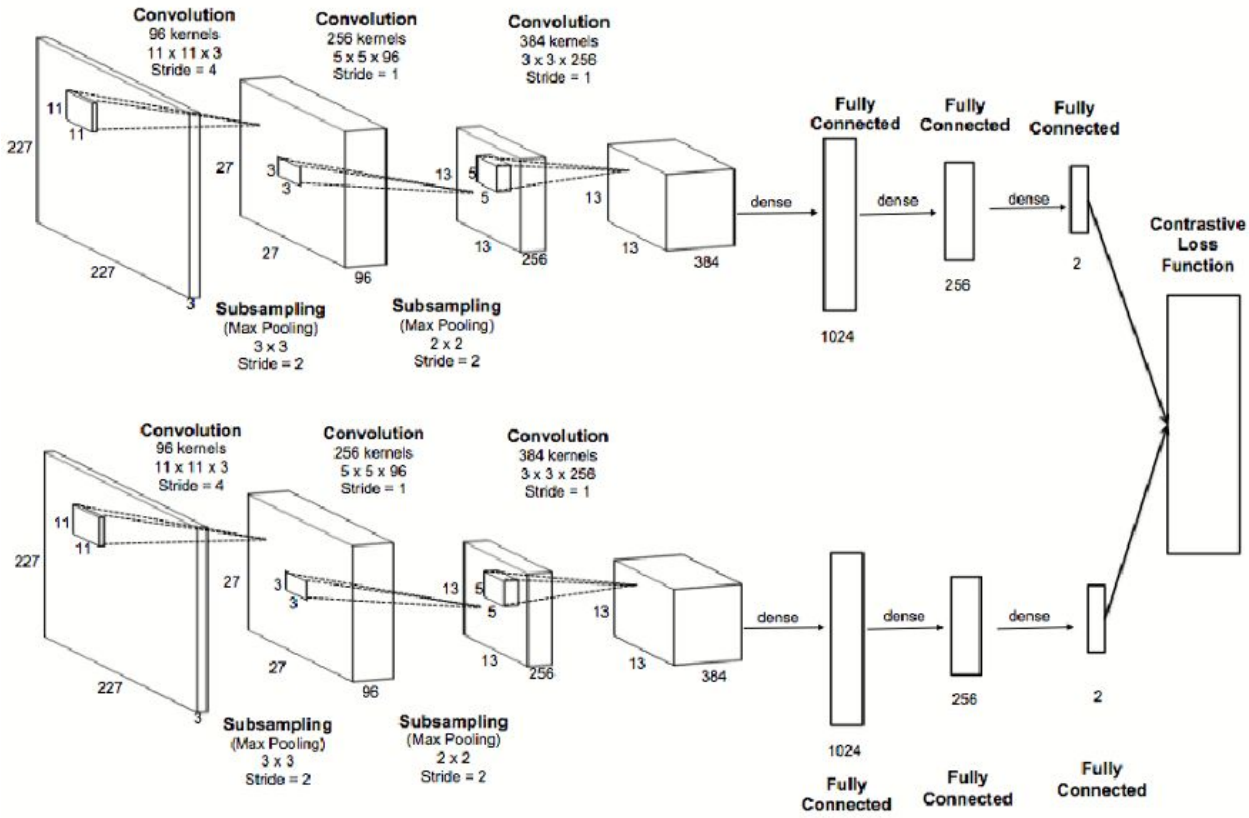
Due to its depth and number of fully-connected nodes, VGG16 is over 533MB. This makes deploying VGG a tiresome task.VGG16 is used in many deep learning image classification problems; however, smaller network architectures are often more desirable (such as SqueezeNet, GoogLeNet, etc.). But it is a great building block for learning purpose as it is easy to implement.



VGG16 Detailed Architecture

Siamese Neural Network

Siamese networks (Bromley, Jane, et al., “Signature verification using a” siamese” time delay neural network.” Advances in neural information processing systems. 1994.) are neural networks containing two or more identical subnetwork components. A siamese network may look like this:



It is important that not only the architecture of the subnetworks is identical, but the weights have to be shared among them as well for the network to be called “siamese”. The main idea behind siamese networks is that they can learn useful data descriptors that can be further used to compare between the inputs of the respective subnetworks. Hereby, inputs can be anything from numerical data (in this case the subnetworks are usually formed by fully-connected layers), image

data (with CNNs as subnetworks) or even sequential data such as sentences or time signals (with RNNs as subnetworks).

Usually, siamese networks perform binary classification at the output, classifying if the inputs are of the same class or not. Hereby, different loss functions may be used during training. One of the most popular loss functions is the binary cross-entropy loss. This loss can be calculated as

$$L = -y \log p + (1 - y) \log(1 - p)$$

, where L is the loss function, y the class label (0 or 1) and p is the prediction. In order to train the network to distinguish between similar and dissimilar objects, we may feed it one positive and one negative example at a time and add up the losses:

$$L = L_+ + L_-$$

Another possibility is to use the triplet loss (Schroff, Florian, Dmitry Kalenichenko, and James Philbin. “Facenet: A unified embedding for face recognition and clustering.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.):

$$L = \max(d(a, p) - d(a, n) + m, 0)$$

Hereby, d is a distance function (e.g. the L2 loss), a is a sample of the dataset, p is a random positive sample and n is a negative sample. m is an arbitrary margin and is used to further the separation between the positive and negative scores.

REFERENCES

[1] Leishman, R. C., McLain, T. W., & Beard, R. W. (2014). Relative navigation approach for vision-based aerial GPS-denied navigation, Journal of Intelligent & Robotic Systems, 74(1-2), 97–111

[2] S. Ahrens, D. Levine, G. Andrews and J. P. How, "Vision-based guidance and control of a hovering vehicle in unknown, GPS-denied environments," 2009 IEEE International Conference on Robotics and Automation, Kobe, 2009, pp. 2643-2648.

[3] Michael Blosch, Stephan Weiss, Davide Scaramuzza, and Roland Siegwart. Vision based ¨MAV navigation in unknown and unstructured environments. In IEEE Int. Conf. Robotics and Automation, pages 21–28, 2010.

[4] D. Nister, O. Naroditsky, and J. Bergen, “Visual odometry for ground vehicle applications,” Journal of Field Robotics, vol. 23, no. 1, pp. 3–20, 2006.

[5] S. Weiss, D. Scaramuzza, and R. Siegwart, “Monocular-SLAM-based navigation for autonomous micro helicopters in GPS-denied environments,” J. Field Robot., vol. 28, no. 6, pp. 854–874, 2011.

[6] Albert S. Huang and Abraham Bachrach and Peter Henry and Michael Krainin and Dieter Fox, Nicholas Roy}, “Visual odometry and mapping for autonomous flight using an RGB-D camera,” Proceedings of the Intl. Sym. of Robot. Research, 2011.

[7] M. K. Kaiser, N. Gans, and W. Dixon, “Vision-based estimation for guidance, navigation, and control of an aerial vehicle,” IEEE Transactions on Aerospace and Electronic Systems, vol. 46, no. 3, pp. 1064–1077, Jul. 2010.

[8] H. Durrant-Whyte and T. Bailey, “Simultaneous localisation and mapping (SLAM): Part I, the essential algorithms,” IEEE Robotics and Automation Magazine, vol. 13, no. 2, pp. 99–110, Jun. 2006.

[9] T. Bailey and H. Durrant-Whyte, "Simultaneous localization and mapping (SLAM): part II,"IEEE Robotics & Automation Magazine, vol. 13, no. 3, pp.

108-117, Sept. 2006.

[10] D. Scaramuzza and F. Fraundorfer, "Visual Odometry [Tutorial]," in IEEE Robotics & Automation Magazine, vol. 18, no. 4, pp. 80-92, Dec. 2011.

[11] F. Fraundorfer and D. Scaramuzza, "Visual Odometry : Part II: Matching, Robustness, Optimization, and Applications," in IEEE Robotics & Automation Magazine, vol. 19, no. 2, pp. 78-90, June 2012.

[12] Nister, D., Naroditsky, O., Bergen, J., (June-July 2004) "Visual odometry," Computer Vision and Pattern Recognition, 2004. Proceedings of the 2004 IEEE Computer Society Conference on , vol.1, no., pp.I-652,I-659 Vol.1, 27.

[13] A. de La Bourdonnaye, R. Doskočil, V. Křivánek, “Practical Experience with Distance Measurement Based on Single Visual Camera”, Advances in Military Technology, 7, 2, 49 – 56, 2012.

[14] Fatih Gökçe, Göktürk Üçoluk, Erol ùahin, and Sinan Kalkan. “Vision-Based Detection and Distance Estimation of Micro Unmanned Aerial Vehicles,” Sensors (Basel). 2015, 15(9): 23805–23846.

[15] Seung Yeob Nam and Gyanendra Prasad Joshi, “Unmanned aerial vehicle localization using distributed sensors”, International Journal of Distributed Sensor Networks, Vol. 13(9), 2017.

[16] M.-C. Lu, W.-Y. Wang, and C.-Y. Chu, “Image-based distance and area measuring systems,” IEEE Sensors J., vol. 6, no. 2, pp. 495–503, Apr. 2006.

[17] J. Unicomb, L. Dantanarayana, J. Arukgoda, R. Ranasinghe, G. Dissanayake, and T. Furukawa, “Distance Function based 6DOF Localization for Unmanned Aerial Vehicles in GPS Denied Environments,” in 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 09 2017.

[18] D. Scaramuzza, M. C. Achtelik, L. Doitsidis, F. Fraundorfer, E. B. Kosmatopoulos, A. Martinelli, M. W. Achtelik, M. Chli, S. A. Chatzichristofis, L. Kneip, D. Gurdan, L. Heng, G. H. Lee, S. Lynen, L. Meier, M. Pollefeys, A. Renzaglia, R. Siegwart, J. C. Stumpf, P. Tanskanen, C. Troiani, and S. Weiss, “Vision-controlled micro flying robots: from system design to autonomous navigation and mapping in GPS-denied environments,” Accepted to IEEE Robotics & Automation Magazine, 2013.

[19] Rodrigo Munguía, Sarquis Urzua," Vision-Based SLAM System for Unmanned Aerial Vehicles", MDPI, 2015.

[20] A simple algorithm for distance estimation without radar and stereo vision based on the bionic principle of bee eyes

[21] A. Yol, B. Delabarre, A. Dame, J. Dartois and E. Marchand, "Vision-based absolute localization for unmanned aerial vehicles," 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, Chicago, IL, 2014, pp. 3429-3434.

[22] Delaune, J.; Le Besnerais, G.; Voirin, T. Visual-inertial navigation for pinpoint planetary landing using scale-based landmark matching. Robotic and autonomous systems, Vol. 78, 2015, pp. 63-82.

[23] Ahmed Nassar, Karim Amer, Reda ElHakim, Mohamed ElHelw, "A Deep CNN-Based Framework For Enhanced Aerial Imagery Registration with Applications to UAV Geolocalization," IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2018, pp. 1513-1523

[24] A. L. Majdik, Y. Albers-Schoenberg and D. Scaramuzza, "MAV urban localization from Google street view data," 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, Tokyo, 2013, pp. 3979-3986.

[25] Scott Workman, Richard Souvenir, and Nathan Jacobs, 'Wide-area image geolocalization with aerial reference imagery', in Proceedings of the IEEE International Conference on Computer Vision, pp. 3961– 3969, (2015).

[26] Olivier Saurer, Georges Baatz, Kevin Köser, Marc Pollefeys, et al. Image based geo-localization in the alps. International Journal of Computer Vision, pages 1–13, 2015.

[27] T. Lin, Y. Cui, S. Belongie, and J. Hays. Learning deep representations for Ground-to-Aerial geolocalization. In IEEE International Conference on Computer Vision and Pattern Recognition, 2015.

[28] G. Conte and P. Doherty. Vision-based unmanned aerial vehicle navigation using geo-referenced information. Accepted for publication in the EURASIP Journal of Advances in Signal Processing, 2009.

[29] A. Viswanathan, B. R. Pires, and D. Huber, “Vision based robot localization by ground to satellite matching in GPS-denied situations,” in Proc. IEEE/RSJ Int’l Conf. on Intelligent Robots and Systems (IROS), 2014.

[30] Dragos Costea and Marius Leordeanu. Aerial image geolocalization from recognition and matching of roads and intersections. arXiv preprint arXiv:1605.08323, 2016

[31] A. Carrio, C. Sampedro, A. Rodriguez-Ramos, and P. Campoy, “A review of deep learning methods and applications for unmanned aerial vehicles,” Journal of Sensors, vol. 2017, 2017.

[32] Paweł Burdziakowski, Artur Janowski, Marek Przyborski, Jakub Szulwic, "A VISION-BASED UNMANNED AERIAL VEHICLE NAVIGATION METHOD"

[33] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks”, Advances in Neural Information Processing Systems 25 (Page 1097-1105), 2012