IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 22, Issue 4, Ser. III (Jul. – Aug. 2020), PP 59-72
www.iosrjournals.org

Road Extraction from Satellite Imagery Based on Fully Convolutional Neural Network

AbenezerZegeye
Aerospace Engineering Research and Development Department, Ethiopian Space Science and Technology Institute, Ethiopia

Abstract:
This research explore on studying road extraction from satellite images, based on deeplearning for semantic segmentation. Previousapproaches for extraction of roadsfromsatelliteimageryisusingmanualapproach, but itis time taking and tedious. The otherapproachisusingsemi-automatic techniques for road extraction from high resolutionsatelliteimagery. However, thisresearchfocuses on fullyautomatic road extraction based on deeplearning. Fullyconvolutional neural network isused to extractroadsfrom the satellite images. For training the CNN model,hugeamount of labeled and structureddata isneeded, but thereis no suchabigdataset for road extraction thatispubliclyavailable. Therefore data preprocessing technique isused in order to maximize the dataset size. The fullyconvolutional neural network used for theseresearchexperimentissmallfilter size architecture adoptedfrom the original Vgg architecture but the last threelayers of the network ischangedfromfullyconnectedlayers to convolutionallayers. After the feature extraction by the 16 layers of the model thendeconvolutionisused toupscale the extractedfeature. This researchwillintroduces new architecture whichisadoptedfromVgg architecture bychanging the last threefullyconnectedlayersintoconvolutionallayersand made it to have 16 trainableconvolutionallayers. Data augmentation isapplied as preprocessing in order to overcome the problem of smalldataset size limitation for the model training. Massachusetts road datasetisused for training and testing of the proposedalgorithm.

Keywords: *Road Extraction, Data Augmentation, FullyConvolutional Neural Network, Deconvolution, and ConditionalRandom Field.*

Date of Submission: 20-08-2020 Date of Acceptance: 06-09-2020

I. Introduction

In recent years, the visual world for machines has increased and machines became moreintelligent. Machine learning algorithms are becoming increasingly more important inmachine vision world. More importantly deep learning has proven to be both a majorbreakthrough and an extremely powerful tool in many fields. Deep learning proved it isa best solution on the classification and regression problems.

Deep learning (also known as deep structured learning, hierarchical learning or deep machine learning) is a branch of machine learning based on a set of algorithms thatattempt to model high-level abstractions in data. In deep Learning, unlike rudimentarymachine learning there is no need for manual feature extraction before training. Featuresare learned automatically from the input data. This makes the model much more powerfulin representation the given data. Machine learning came directly from minds of the earlyAI crowd, and the algorithmic approaches over the years included decision tree learning,inductive logic programming. Clustering, reinforcement learning, and Bayesian networksamong others.

Deep neural networks are an implementation of neural network in which a multiplehidden layers of neurons and then run massive amounts of data through the system totrain it. Today machines trained using deep neural networks in some scenarios outperformhumans with significant margin. This makes using Deep Artificial neural network idealin solving many of AI problems like Computer Vision, Natural language processing andincase of Automatic image processing both.

Machine Learning role has an effective approach for solving problems in geosciences andremote sensing. In recent years, deep learning made a great progress in solving detectingand recognizing problems in remote sensing. Accurate road extraction from high-resolutionsatellite imagery has many applications such as urban planning, trafficmanagement system, and robotics as an aid in navigation of autonomous vehicles. In orderto visualize roads from satellite images, machines must not only need to detectand recognize road regions only but also need to guarantee continuous road regions. Forthat reason, post processing must have to be done in order to minimize misclassificationerror that are occurred on road intersections and joints. Therefore, on this work I proposedto use Fully Convolutional Network (FCN) architecture for extraction of roads from high-resolutionsatellite image data.

Objective of this research is to design an effective algorithm for extraction of roads from remote sensing images.

The significance of this research is three fold:-

- To figure out the proposed concept, which is using a fullyconvolutional network to extract roads from satellite images.
- On this research a new approach is introduced for road extraction, which is asingle stream fully convolutional network that have small filter sizes in order todetect/extract local and object level information based on Vgg-Net architecture.
- Finally, it contributed a significant performance improvement with the state-ofthe-art methods.

So far, many researches has been done on extracting roads from high-resolution satellite imagery. Automatic road extraction system focuses on extracting roads based on deep learning for the feature extraction part. I compared this work with some related state of the art research that has been done.

State of the Art

In earliest time for extracting target from remote sensing images manual extraction technicalities were used. This kind of methods has high accuracy, but in terms of time and cost is not cost-effective. Then the semiautomatic approach presented to overcome a time and cost issue. By using machine learning algorithm with manual approach increases the performance and operation technicality. The new era of this research area is a fully automatic approach which is based on deep learning to extract features from digital images automatically.

Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks[1], on this research paper they tried to extract both roads and buildings simultaneously. On their CNN architecture there are five convolutional layer followed by global average pulling. They used Massachusetts road and building, and Abu Dhabi road & building datasets, and they followed patch-based approach for training their model. As a post-processing Simple Linear Iterative Clustering (SLIC) and Region Adjacent Graph (RAG) are applied to obtain the initial segmented image based on similarity in color space and proximity in the image plane, and to introduce adjacent relationships between previous segments. Finally they scoredcorrectness 91.7% for roads and 94.6% for buildings.

Multiscale road centerlines extraction from high-resolution aerial imagery[2], on this research work their aim is to extract roads center line. To address the objective of their research they followed four steps; first, convolutional neural network (CNN) is used to classify aerial imagery pixel-wise, then edge-preserving filtering is conducted on the resulting classification map, to preserve the edges and the details of the road. After that, based on shape features to extract more reliable roads some post-processing is applied.Finally, multiscale Gabor filters and multiple directional non-maximum suppression are integrated to get a complete and accurate road network. EPFL-dataset and Massachusetts Roads dataset is used for this research. Their CNN network have five layers three convolutional and two fully connected layers. They scored classification accuracy of 88.7% on a single image. And a maximum completeness on a single image 96.65%, correctness on a single image 95.70%, and quality on a single image 92.24%.

Road Structure Refined CNN for Road Extraction in Aerial Image[3], this research paper focuses on extracting of roads from aerial imagery. To obtain structured output of road extraction, both deconvolutional and fusion layers are designed in the architecture of their CNN algorithm. On this research there are 16 convolutional layers and in order to upscale the extracted features they use cropping and some deconvolutional layers. They got 92.4% classification accuracy.

Though the above researches scored a good result, in terms of classification this research scored better than the above results.

Image Resolution

Image resolution, and model density are two essential attributes of aerial maps driving the transformation from poor resolution images to updated high resolution images that are used in applications that uses direction and navigation.

There are four types of resolution when discussing satellite imagery in remote sensing: spatial, spectral, temporal, and radiometric. As described on a book called Introduction to Remote Sensing[4] these resolution types are defined as follows:

- Spatial resolution is defined as the pixel size of an image representing the size of the surface area (i.e. m^2) being measured on the ground, determined by the sensors' instantaneous field of view (IFOV).
- Spectral resolution is defined by the wavelength interval size (discrete segment of the Electromagnetic Spectrum) and number of intervals that the sensor is measuring.
- Temporal resolution is defined by the amount of time (e.g. days) that passes between imagery collection periods for a given surface location.

- Radiometric resolution is defined as the ability of an imaging system to record many levels of brightness (contrast for example) and to the effective bit-depth of the sensor (number of grayscale levels) and is typically expressed as 8-bit (0- 255), 11-bit (0-2047), 12-bit (0-4095) or 16-bit (0-65,535).
- Geometric resolution refers to the satellite sensor's ability to effectively image a portion of the Earth's surface in a single pixel and is typically expressed in terms of Ground sample distance, or GSD. GSD is a term containing the overall optical and systemic noise sources and is useful for comparing how well one sensor can "see" an object on the ground within a single pixel. For example, the GSD of Landsat is ≈30m, which means the smallest unit that maps to a single pixel within an image is ≈30m x 30m.

The resolution (is the detail an image holds) of satellite images varies depending on the instrument used and the altitude of the satellite's orbit. Depending on the resolution the aerial coverage and the image quality is defined.

A road in satellite imagery is identified as a certain width and a set of straight line segments. In addition roads are grey in color and long and thin with relative to other objects on ground. As shown image below roads can be identified as straight and mostly grey color long object.

Detecting roads from satellite imagery is to detect the corresponding straight line segments with a certain length and direction. There are different technicalities to extract roads from high resolution imagery. As described on the research paper entitled “A review of road extraction from remote sensing images”, this methods are classified as: Classification-based methods, Knowledge-based methods, Mathematical morphology methods, Active contour model, and Dynamic programming and grouping. High resolution RS images such as IKonos, QuickBird, WorldView and GeoEye create a quick and economical way to access the newly acquired geographic information, and lay a very important basis for the further applications of RS technology[5].

This research is based on extracting roads using feature based approach, specifically based on supervised learning methods. This approach, needs to be trained using labeled samples.

FCN for Road Extraction

Artificial neural networks (neural networks) are inspired by the biological neural system designed to recognize patterns. Neurons are the essential computational units in the brain. A neuron input is received from their dendrites, the produced output is sent from their axons. A synapse is a transition between the dendrite of one neuron and axon of another neuron[6].

The output production rate of an artificial neuron is modeled by the non-linear activation function f(x). The activation function of a neuronis modeled as shown in equation 1. This equation models decision-making, bydiverging the weights and activation function the output is affected.

$$a_i = f(\sum_i (w_i x_i) + b_i) \quad (1)$$

Where a_i represents the activation function of a neuron, f(x) is the non-linear activation function, w_i is the synaptic strength weight, x_i is the input neuron and b_i is the neuron bias. Neural networks were originally modeled using sigmoidal and hyperbolic tangent functions; however, Rectified Linear Unit (ReLU) functions have proven to be computationally more efficient[7].

The outputs of neurons of a certain layer become the input of neurons of the following layer; these networks are said to be feed-forward neural networks (FFNN) [7].

Overfitting is a problem that arises in neural network training[8]. When a model is overfitted to the training data, it loses its capability of generalization. In this research the model faced overfitting due to the data preprocessing technicalities used. Minimizing the processing technicalities in terms of relative similarity, by choosing the non-similar one the model overcomes the problem of overfitting.

The CNN requires adjusting and updating its kernel parameters, or weights, for the given training data. Backpropagation is an efficient method for computing gradients required to perform gradient-based optimization of the weights in neural networks[9]. The specific combination of weights which minimize the loss function (or error function) is the solution of the optimization problem.The method requires the computation of the gradient of the error function at each iteration, therefore the loss function should be both continue and differentiable at all iteration steps.

The last activation function differs from all other activation functions within the network. In an image classification problem, classes are mutually exclusive. At the final layer of the used network there is a logistic classifier, which can generate the probabilities of the output to predict the correct class. In the case of this study, since the case is dealing with pixel wise classification, the network generates the probability of each pixel belongs to the classified group that the pixel belongs to. The classification groups are “Road” and “Non-Road” on this research. The sigmoid function is used for the two-class logistic regression, whereas the softmax function is used for the multiclass logistic regression. Since this research classifies pixels into two classes, sigmoid function is used for predicting the probability of the output on this research.

Sigmoid function gives an analog activation unlike as step function and it has a smooth gradient. This activation function is always range between 0 and 1 so it makes the activation bound in range. Towards either end of the sigmoid function, the Y values tend to respond very less to changes in X which means the gradient at that region is going to be small. It gives rise to a problem of “vanishing gradients”. Gradients vanished means cannot make significant change because of the extremely small value.

$$A = \frac{1}{1+e^{-x}} (2)$$

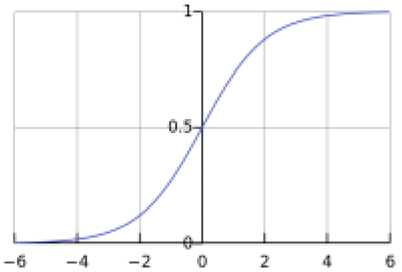


Fig 1. Sigmoid Function[10].

Extracting roads from high resolution satellite imagery by using Fully Convolutional Neural-Network needs to have a big dataset to train the model and get good performance. Nevertheless, there is no big labeled dataset for road extraction. Therefore, data augmentation technicality is used to increase the dataset size. Which means, before the CNN training is started there is preprocessing of the dataset to maximize the dataset size. On the CNN part original VggNet network is adopted and reshaped based on the proposedmethodology, which is changing the last three fully connected layers of original VggNetarchitecture into convolutional layers.

Related Works

So far, many researches are doing research on extracting roads from high-resolution aerial imagery. Manually extracting roads from digital imagery, although has high accuracy, but in terms of time and cost is not cost-effective especially when the scenes are very complex[11]. Therefore, it is crucial to design semi-automatic/automatic road extraction methods/algorithms. Semi-automatic road extraction from digital images[11], a semiautomatic method is presented to detect the roads in high-resolution satellite images. The proposed method in [11] is based on four main steps. Firstly, canny edge detector is employed to segment roads from the images. Secondly, Full Lambda Schedule merging method applied to combine adjacent segments. The third, Support Vector Machine (SVM) was used to classify entire image. Finally, the morphological operation procedure such as dilation, erosion, opening, and closing techniques are performed to remove the undesired objects.

Recently road extraction from remote sensing has scored significant improvements since the era of deep learning. Convolutional NNs (CNNs) have proven to be good at extracting mid and high-level abstract features from raw images by interleaving convolutional and pooling layers (i.e., by spatially shrinking the feature maps layer by layer)[12] .

Recent studies indicate that the feature representations learned by CNNs are highly effective in large-scale image recognition[13–15], object detection[16, 17], and semantic segmentation[18, 19]. The fully convolutional network (FCN) [19] is the most important work in deep learning for semantic segmentation, i.e., the task of assigning a semantic label to every pixel in the image. The number of papers published on this area (i.e. remote sensing based on deep learning), is increasing exponentially which shows the rapid surge of interest in deep learning for remote sensing.

There are many previously done researches, on road extraction from remote sensing images. In most previous works, semi-automatic road extraction from aerial imagery researches are the center of interest in road extraction from remote sensing images. Multiscale road centerlines extraction from high-resolution aerial imagery[2] introduces the feature learning based on deep learning to extract robust features automatically, and present a method to extract road centerlines based on multi-scale Gabor filters and multiple directional non-maximum suppression in order to address the road centerlines extraction problem, the existing algorithms have some limitations, such as spurs, time consuming. In this paper, on the architecture part they put on layer 4 and 5 a fully connected layer, which might cause over-fitting. In addition, the research they propose focuses on multi-scale road centerlines extraction using Gabor filters and multiple directional non-maximum suppression, which is not the focus of this research.

Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks[1] , on this research they propose a single patch-based Convolutional Neural Network (CNN) architecture for extraction of roads and buildings from high-resolution remote sensing data and low-level features of roads and buildings (e.g., asymmetry and compactness) of adjacent regions are integrated with Convolutional Neural Network (CNN) features during the post-processing stage to improve the performance. In post-processing stage Simple Linear Iterative Clustering (SLIC) is applied to obtain the initial segmented image based and to introduce adjacent relationships between previous segments, Region Adjacent Graph (RAG) is used to facilitate merging process between super-pixels. The main focus area of this research [1] is to extract roads and buildings simultaneously from remote sensing data using convolutional neural network. However, I argue that extracting both road and building might be difficult (or does not guarantee us on getting a good shape), since the shape of a building and shape of road is completely different. Roads are long and thin while buildings are wide and thick which might cause the irregular shape understanding by the network.

Dual local-global contextual pathways for recognition in aerial imagery[20] proposed a dual-stream deep network model to extract roads and buildings separately based on AlexNet[21] and VGG-Net[14]. Alex-Net considers information from large areas around the object of interest due to the larger filter size. VGG-Net network focuses on local and object level information due to the smaller filter size. Both networks are combined into final subnet, composed of three Fully Connected (FC) layers. CNN architectures such as Network in Network (NIN)[22] and Going deeper with convolutions (GoogLetNet) [23] proposed avoiding the use of fully connected layers to minimize the number of parameters while maintaining the high performance. In addition, even though feature dropout can be used to minimize over-fitting, fully connected layers might cause over-fitting. Object detectors emerge in deep scene CNNs[24] showed that convolution units have the ability to localize objects in convolution layers; however, this ability is lost when fully connected layers are used.

Road Network Extraction via Deep Learning and Line Integral Convolution[25], proposes a learning-based road network extraction scheme from high resolution satellite. They [25] proposed to implement the proposed research in three steps: First, the convolutional neural network (CNN), which is able to capture large context of local structures, are applied to predict the probability of a pixel belonging to road regions, and assign labels to each pixel to describe whether it is road. Then, a line integral convolution based algorithm is developed to smooth the rough map to connect small gaps. Finally, by combining with some common image processing operators, road centerlines are able to be acquired. But in fact their network is not that deep, but going deep in layer improves the performance of the network.

II. Material And Methods

The basic idea of Road Extraction from Remote Sensing Images based on Fully Convolutional Neural Network (FCN) is to design and develop a reliable, cost effective and simple algorithm to detect and extract roads from satellite images. This work is a fully automatic extraction of a road from aerial imagery, which uses Fully Convolutional Neural-Network (FCN) for feature extraction.

Procedure methodology

In this work, the research made an intensive analysis on different road extraction technicalities from aerial imagery that led me to adopt an architecture that improves the current state of the art. The model adopted from previously used architecture based on the hypothesis made, and this research proves it is suitable for the targeted purpose.

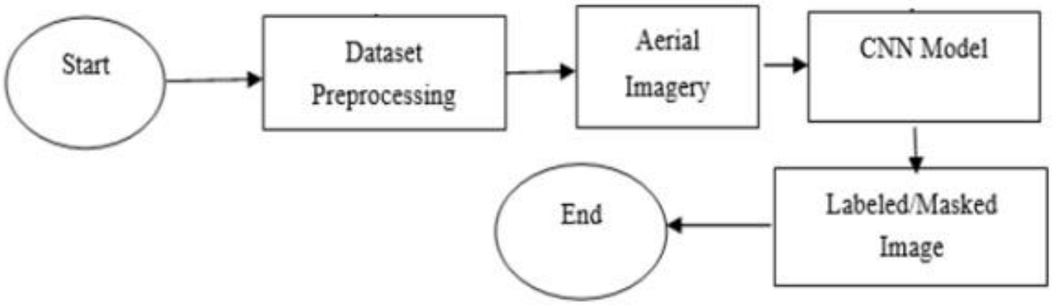
This research architecture is based on the research paper Fully Convolutional Networks for Semantic Segmentation, which take input of arbitrary size and produce correspondingly sized output with efficient inference and learning. The network produces pixel wise annotation as a matrix in the size of the image with the value of each pixel corresponding to its class. This architectural change over road extraction from remote sensing images, apart resulting unprecedented result boost, it also lit a way to different potential improvements over different areas.

As discussed in the previous review Multi-scale road centerlines extraction from highresolution aerial imagery model uses fully connected layers at the end of the architecture, which increases the size of the parameters so the mathematical complexity increases. In addition to that, using fully connected might cause over fitting and convolutional neural networks have an ability to detect the target area without the use of fully connected layers so this architecture avoids fully connected layers and gets good result.

The final output layer produces the up-scaled size of image as the original one, which is$S \times S$, size of data. Where S is the height/width of the original image, since the images used in this research are square images.

The intuition of this architecture is to show small size filters and fully convolutional network have a best performance in detecting and recognizing roads from satellite imagery. Since VggNet has a small size filter and have a deep layers, it is chose for this research believed that it is suitable for the targeted idea. Six Vgg architectures developed for large-scale image recognition. Among those architectures, the one that have a good performance is chose for this research.

On this research, two stages are applied to achieve this research goal and prove the hypothesis. The first stage is data preprocessing, on this stage the all the data that are used on training are augmented. Since there is no big road dataset of satellite imagery that is available for the deep learning training, this research used different data augmentation technicality to increase the dataset size. The second stage is training the deep learning network in order to extract roads from the satellite image. On this stage, a 16-layer CNN architecture is adopted from Vgg16 network.



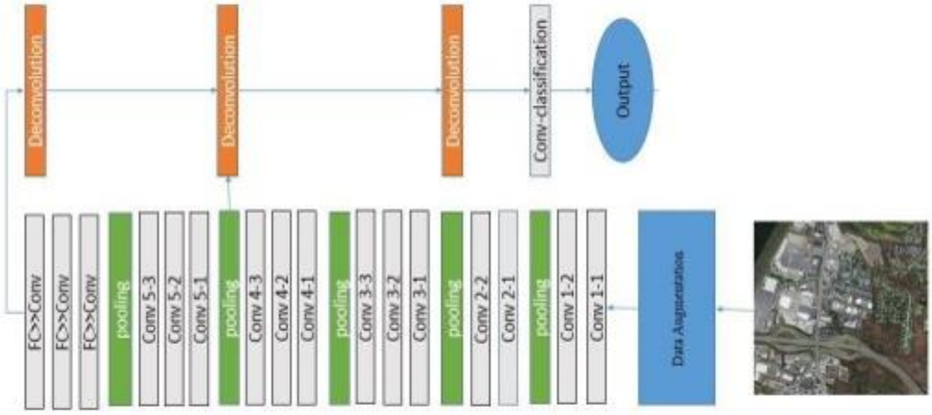


Fig 2. General system design
Fig 3. Detail of Algorithm Architecture

Data Preprocessing

The road datasets of satellite imagery that are available for deep learning training are very limited and they are small. Nevertheless, the dataset needed to train deep learning network must be big in size to increase the performance of the network to extract a reliable feature and get a good result. The dataset that is used on this research is Massachusetts road dataset of satellite imagery, which have 1171 (one thousand one hundred seventy one) images in total with the spatial resolution of 1m (one meter).

So in order to increase the size of the dataset a data preprocessing technicalities are used to get a big number of training data. Some technicalities that are used for data augmentation:

- Rotation- rotating the image with the rotation angle of 90°. Therefore, every image in the dataset is rotated by 90°. This makes the dataset to have a size of four times the original dataset.
- Width shift- shifting the pixels of the images width wise by the specified range. So shifting the image pixels width wise shifts the image within its width direction and makes a new image every time for the pixel wise classification.
- Height shift- shifting the pixels of the images height wise by the specified range. So shifting the image pixels height wise shifts the image within its height direction and makes a new image every time for the pixel wise classification.
- Shear range- this makes shear intensity (The effect of this mapping is to displace every point horizontally by an amount proportionally to its coordinate). Horizontal displacement of pixels takes place based on the shear range given.
- Zoom range- is a random zoom based on the given range. By zooming-out/zooming-in, it is possible to create a new feature for the images. It uses a different random transformation for the horizontal and vertical axis, which produces images with a random aspect ratio.
- Horizontal flip- it reverses the active layer horizontally, that is, from left to right (mirror image). It leaves the dimensions of the layer and the pixel information unchanged.

- Vertical flip- it reverses the active layer vertically, that is, from top to bottom. It leaves the dimensions of the layer and the pixel information unchanged.
- Fill mode- Points outside the boundaries of the input are filled according to the given mode.

Architecture and Design

To extract roads from satellite images, image segmentation algorithm is used. Segmentation is labeling every pixel of an image with its classification classes (“Road/Non-road”). If a single pixel belongs to road that pixel is labeled with different color value from non-road regions. Once the feature is learned, the trained feature should have to be scaled up. So, deconvolution is used to scale up the trained feature.

Based on the discussion in the preceding sections, this research focuses on extracting roads from aerial imagery based on fully convolutional neural network. So in order to extract roads automatically from the aerial images a small filter size neural network is used. In addition, there are no fully connected layers on the architecture since using fully connected layers increase parameters.

The input to convolutional neural network is a high-resolution remote sensing aerial image, which is a three-channel RGB image and the labeling is black and white image, which is a road and non-road masked image (see Fig 4.). In order to minimize the mathematical complexity and the time of training, the images are reshaped and fed into the neural network.

The images are resized from 1500*1500 to 224*224. Downsizing the image size minimizes the mathematical complexity and the time to process the images. The images that are used in this thesis are aerial imagery from Massachusetts road dataset. The resolution of the images are 1m^2.

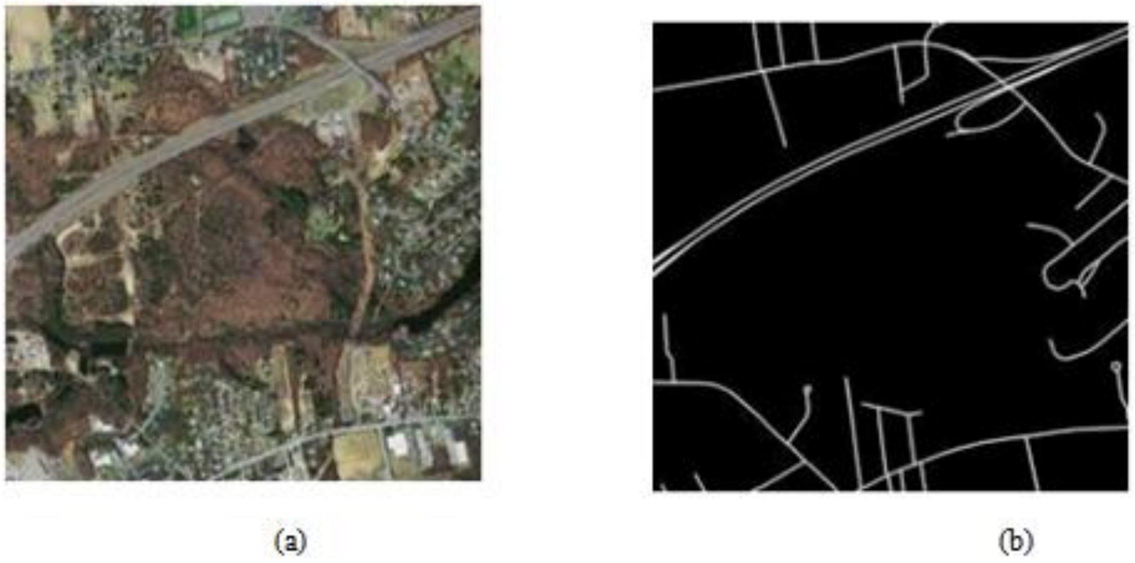
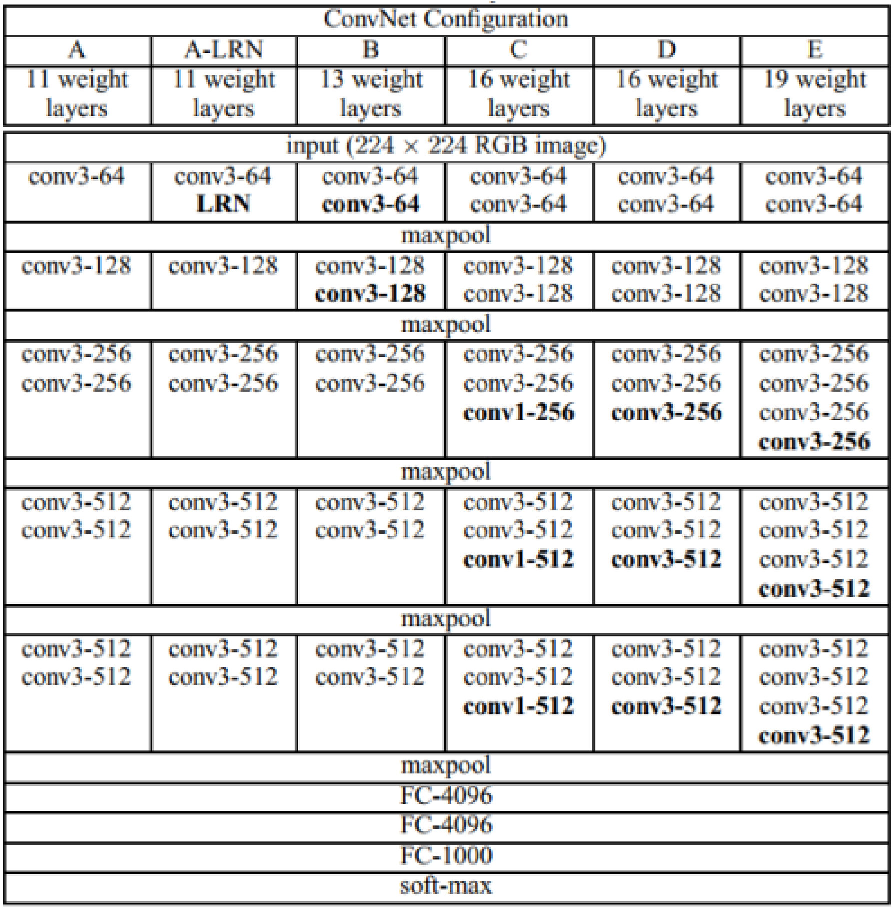


Fig 4. Image from Massachusetts road dataset. (a) Original aerial imagery. (b) Annotated labeled image that road is extracted from original image.

The architecture used for this research is fully convolutional network and the filters used on this neural network architecture are smaller. The architecture is adopted from Vgg16 original architecture. However, the fully connected layers of the original vgg16 layers are converted to convolutional layers for this research. Vgg16 is chosen for this research, because of its small filter sizes and deep layers.

The original vgg16 network have 16 layers, among them the last 3 layers are fully connected layers. There are 6(six) architectures of vgg16 networks. For this research, one is chosen depending on the performance index they showed.



ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Fig 5. The six different architectures of VGGNet[14].

On this research the D architecture is chosen since it shows the high accuracy among the six architectures. The table is from Very Deep Convolutional Networks for Large-Scale Image Recognition [14].

Image segmentation is the process of dividing an image into multiple segments (each segment is called super-pixel). And each super-pixel may represent one common entity. Segmentation is a computationally very expensive process because it need to classify each pixel of the image. Therefore, the algorithms that is used on this research, fully connected layers are removed in order to minimize the parameters that cause computational complexity. Since, neurons in a fully connected layer have full connections to all activations in the previous layer that makes computational complexity.

Convolution layer of the algorithm down samples the featured learned from the original images. However, since segmentation is about finding the class of each and every pixel of the image, down-sampled maps cannot be directly used. Thus, an upsampling convolutional layer is used, which is called deconvolutional layer or fractionally strode convolutional layer.

Fractionally strode convolution/deconvolution layer upsamples the image to get the same resolution as the input image. A simple resizing of the maps is an option as to do for resizing of an image. However, since a naive upsampling inadvertently loses details, a better option is to have a trainable upsampling convolutional layer, whose parameters will change during training.

So, for image segmentation, a deconvolutional layer is put on top of regular CNN. The down-sampled response maps from CNN are upsampled through this deconvolution layer, producing the feature that can be used to predict class labels at all the pixel locations. These predictions are compared with the ground truth segmentation labels available, and a loss function is defined which guides the network towards correct prediction by updating the parameters involved in backward propagation as usual.

The general intuition is that deconvolution is a transformation that goes in the opposite direction of normal convolution, hence the name. Therefore, in deconvolution, output of convolution becomes the input of deconvolution and input of convolution becomes output of deconvolution. The table below (table 3.1) shows the summary of the FCN architecture used on this thesis. The architecture has 16 convolutional layers and 3 deconvolution layers to upscale the extracted feature as the original image size. Since the architecture has any layers the parameters are also many. 2,164,305 total parameters and among them 2,161,361 are trainable parameters, and 2,944 non-trainable parameters.

III. Result

Dataset Description

In order to accomplish the proposed system using the deep learning approach, a suitabledataset was needed to train the model and evaluate its performance. The lack of getting a big dataset that is annotated for road extraction led me to do data preprocessing to get a larger dataset for training. Even though, the dataset that is used in this research is the biggest from allavailable datasets which are annotated for road extraction from satellite images, but it is not enough to get a good result. Therefore, I tried to increase the dataset size by using different technicalities that helps us to increase the performance of the CNN network. The training, validation, and test data of the dataset are merged before data preprocessing and then the data is dividedbetween training data and validation data. And for the final testing of the trained model, randomlyselected images are used from test set. Afterward, the preprocessing of the data takes place on both the training data and the validation sets. Eight (8) technicalitiesare used for preprocessing and all those technicalities applied after several change of technicalities, since some of them cause overfitting.

For this research Massachusetts Road dataset is used. The Massachusetts Roads Dataset consists of 1171 aerial images of the state of Massachusetts. As with the building data, each image is 1500×1500 pixels in size, covering an area of 2*:*25 square kilometers. The dataset is randomly split into a training set of 1108 images, a validation set of 14 images and a test set of 49 images. The dataset covers a wide variety of urban, suburban, and rural regions and covers an area of over 2600 square kilometers. With the test set alone covering over 110 square kilometers, this is by far the largest and most challenging aerial image labeling dataset.

And after augmentation each image gives us eleven different type of new images. Whichis, there are 8 types of data preprocessing technicalities that were used and among themthe image rotation technicality made four (4) times of the original image since rotation angle of 90° is used. So, the training became 11*1108 = 12188, whereas the validationdata became 14 * 11 = 154, the test images that are left to apply testing are not augmented.

Experiment

For the implementation of the proposed method, Keras open source deep learninglibrary that isrunning on top of TensorFlow as backend is used. The framework experimented on amachine with specifications: Intel ® Core i5-7500 CPU at 3.40GHz 4-core with NVIDIAGeForce 1080 GTX GPU and 16GB RAM. The research experiments are performed on aMassachusetts road dataset which is augmented, since the annotated images that areavailable are not enough for deep learning training.The batch size to5 is set and because of the classes that each image is classified, binary cross-entropy lossfunction is used as the cost function. The network is trained for 40 epochs; hence, themodel with the best validation accuracy was selected and used for evaluation.

In the experiment, the hyperparameters used for training the model has been presented insection architecture and design. The deepness of the proposed CNN model used for feature extractioncomprises of sixteen trainable convolutional layers for road extraction.As described on architecture and design, the architecture have 16 trainable CNN layers and there are3 deconvolutional layers. Since, research is extraction of roads from aerial imagery fortraining pixel wise classification is applied. Under pixel wise classification the extractedfeatures must be rescaled to the size of the original image to classify the original imageas Road and Non-road region. Therefore deconvolutional layers are used for scale up, thatcan smoothly scaled it up so as not to lose the extracted features.

This section presents the results and evaluation metrics of the performance and accuracy of the proposed system. As discussed above, the proposed model is trained and testedon preprocessed dataset and achieved promising results of **95.4%** classification accuracy of road extraction. Fig 6 and Fig 7 effectively demonstrate that the accuracy increases and the loss decrease with different input vector sizes. The value of training and validation loss declines gradually during training. This is a very nice indication of how good the learning is and how good the predictions of the model is.

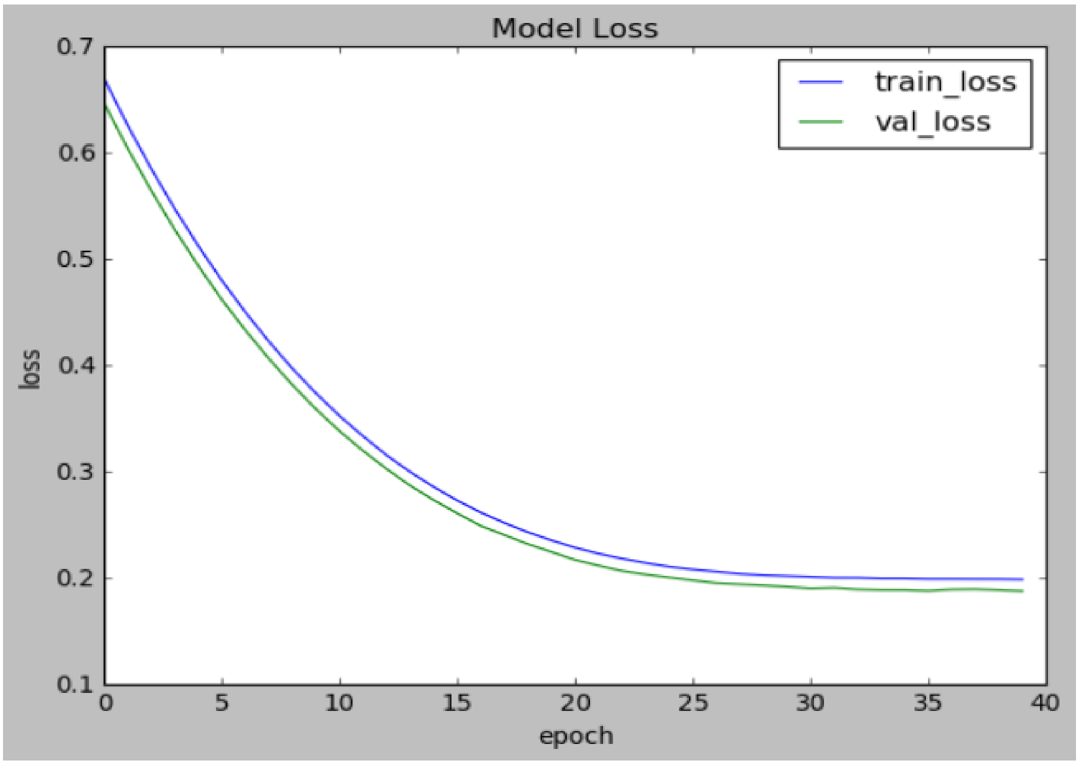


Fig 6. Training and validation loss of the model

The accuracy ofthe model is shown on theFig 7. The training and validation accuracy graph below shows the flow of the model accuracy.From the accuracy graph the validation accuracy is greater than the training accuracy which represents good model performance. As it can be seen from the loss graph the validation loss is greater than training loss in addition from the accuracy graph the validation accuracy is greater than the training accuracy which shows the model have a good performance and there is no everfitting.

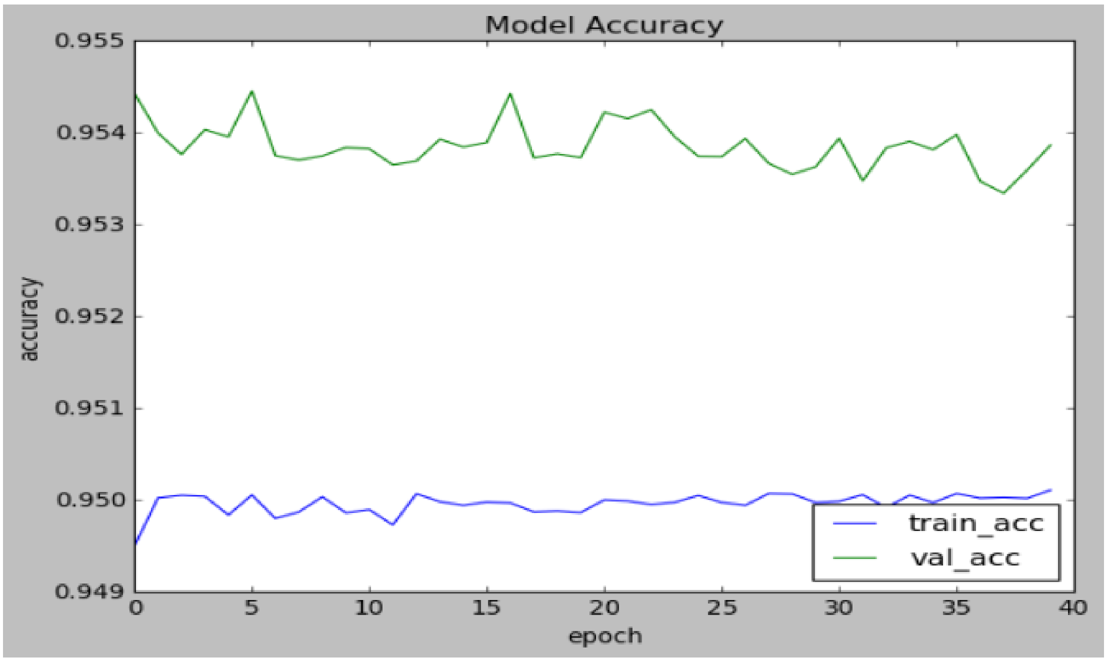


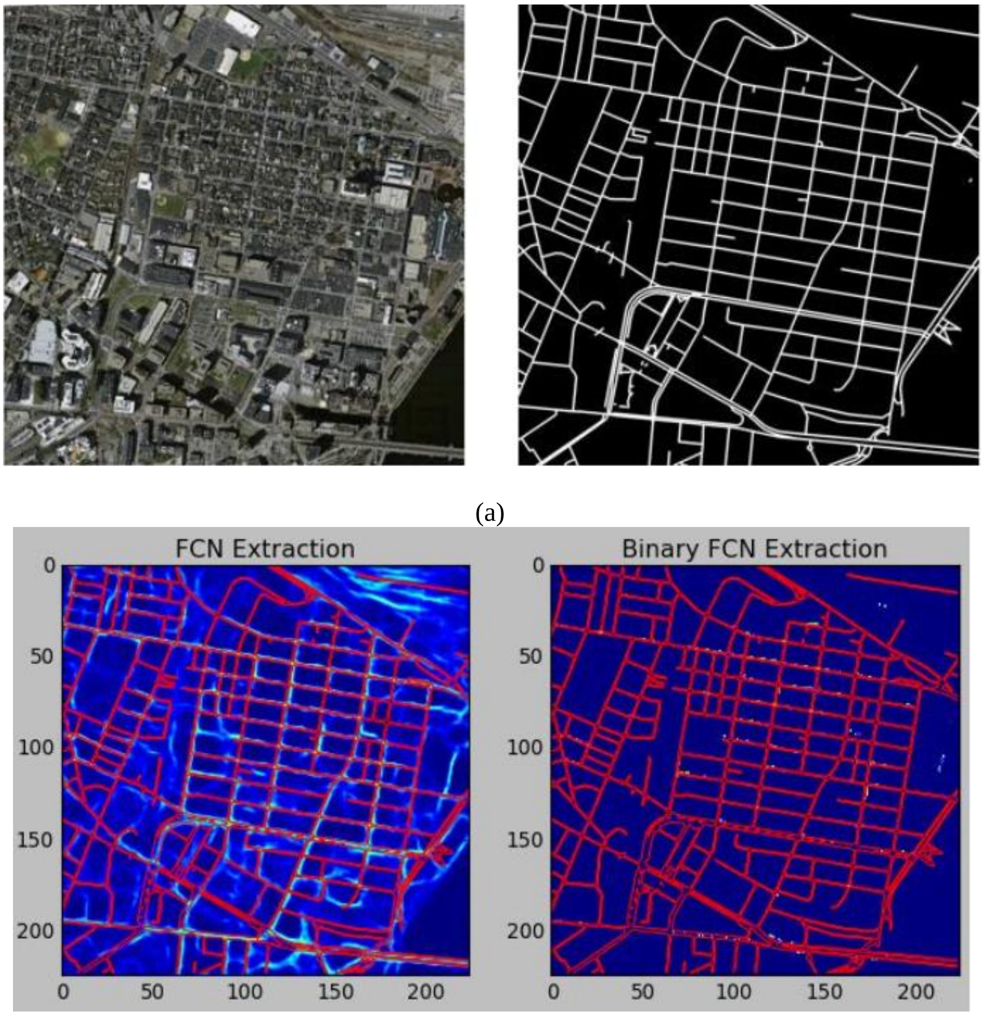
Fig 7. Training and validation accuracy of the model.

Table 1: Shows different models with different data sizes and there relative accuracy.

Model	Dataset Size	Accuracy
Unet Model	1108 training, 14 validation, 49 test images	64%
Adopted Vgg without Augment.		66%
Unet with 90° Image Rotation	4432 training, 56 validation,196 test images	73.5%
Adopted Vgg with 90° image Rotation		74%
Unet with data augmentation	**12188 training, 154 validation, 49 test images**	93.7%
Adopted Vgg with data augmentation (proposed)		**95.4%**

The above table (Table 1) shows, how the proposed architecture and algorithm is successful with relative to different model and with different data size. The table shows that when the data size increases the classification accuracy increases and the proposed model scores highest classification accuracy than Unet. On the first experiment I used the original data size of the dataset both on the Unet and Vggnet and the result shows that Vgg have higher accuracy. In the second experiment, I made a rotation on every images in the dataset and the dataset size becomes 4 fold of the original size, due to the that the accuracy increases for both Unet and Vgg. On the final experiment data augmentation is used and the size of the data increased (N.B. data augmentation is applied only on the training and validation dataset.) as a result the accuracy of both Adopted Vgg and Unet increased dramatically. It is visible that on all experiments the adopted Vgg model beats the Unet. From the above table it can conclude that, using data augmentation to increase the dataset size improves performance and the proposed model achieves higher accuracy.

This research shows that the proposed algorithm works well for extraction of roads from satellite imagery. Bellow, there are images that are extracted using the proposed algorithm.

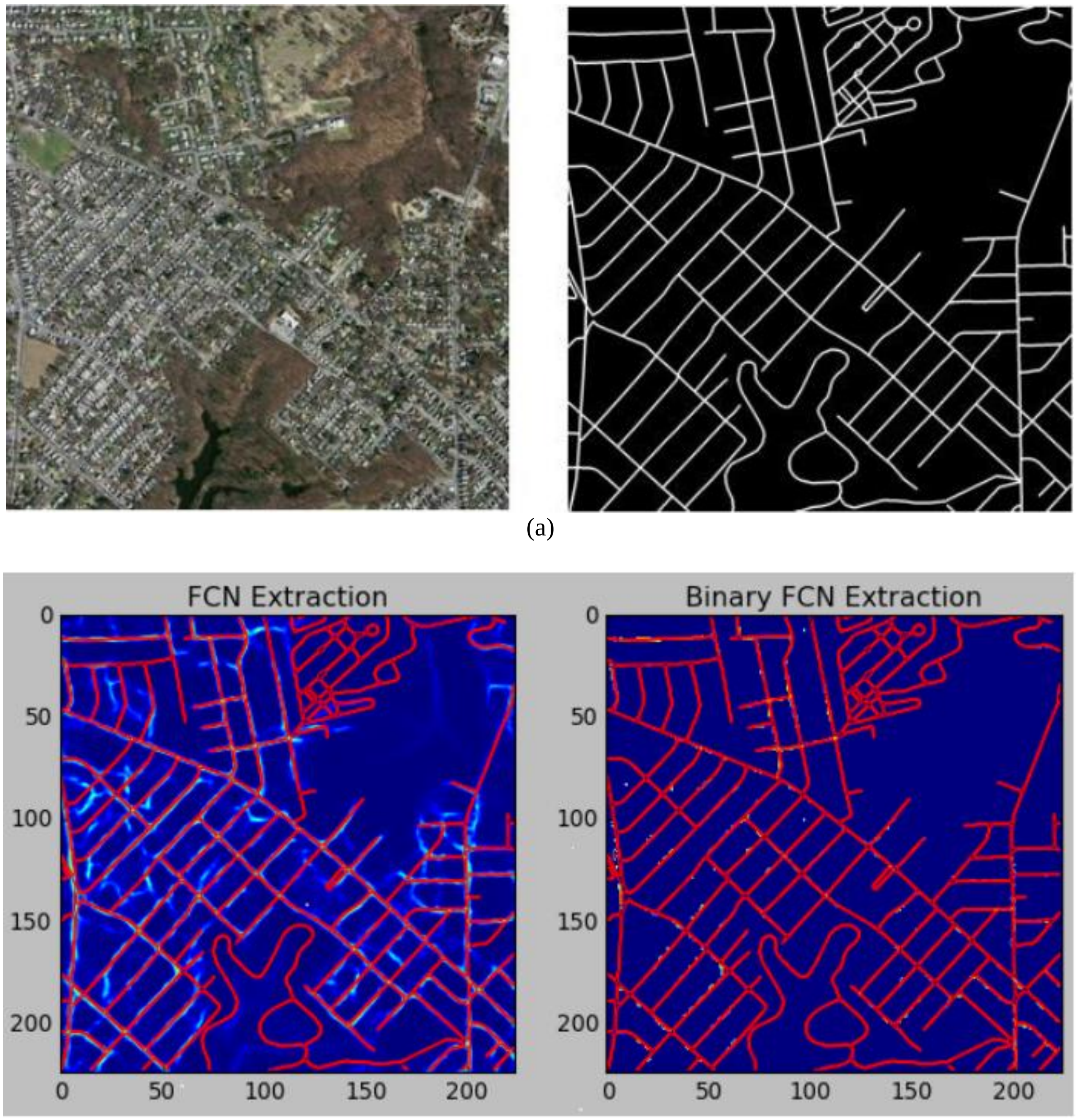


(a)

(b)

Fig 8. Extracted road from training images. (a) Training original image and its label.
(b) Extracted road from grayscale image and binary extraction image.

On the above image (Fig 8.) shows the extraction roads from the training images when the training is done. The image is selected randomly and the image in (a) are the original image and the second image is the label of the original image. The images in (b) are gray scale extraction of the road from the original image whereas the forth image is a binary extraction of the image.

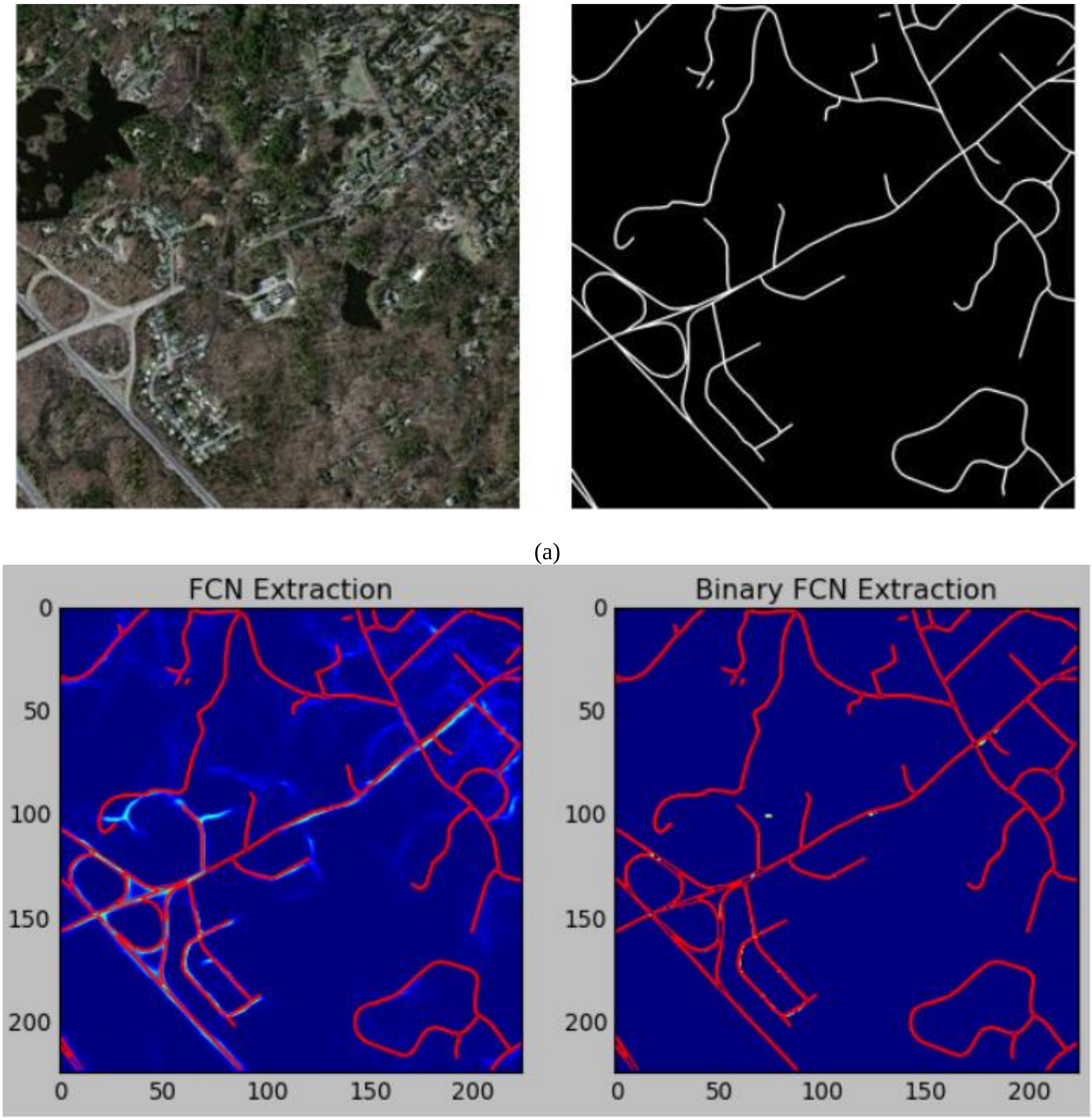


(a)

(b)

Fig 9. Extracted road from the validation images. (a) Validation original image and its label. (b) Extracted road from grayscale image and binary extraction image.

Fig 9. shows the image the road extracted from a randomly selected image from the validation dataset. The first image (a) is the original image and a label of it whereas the second image (b) is the gray extraction (prediction) and binary extraction of the road from the original image respectively.



(a)

(b)

Fig 10. Extracted road from the test images. (a) Testing original image and its label. (b) Extracted road from grayscale image and binary extraction image.

Fig 10. shows the extracted image on randomly selected image from the test dataset. As it is seen the extracted road is clear and reliable. The first image (a) is the original image and its label. The second image is the road extracted from the test image and binary extraction of the road from the original image respectively.

IV. Conclusion

In this study, FCN based extraction of roads from high resolution satellite imagery is designed and developed. The systems accept satellite images as input and extracts roads from the aerial imagery then it returns segmented image of road and non-road classification. To achieve the promising result, different techniques such as image preprocessing, feature extraction using fully convolutional neural network and classification is used, in order to get a reliable result.

In this work, image preprocessing techniques are applied to maximize the dataset size since the available dataset for road extraction from the satellite image is not big enough to train the CNN network from scratch. Data augmentation technicality such as image flipping, image rotation, bit shifting, image zooming, and shear range methods for image preprocessing is used. Depending on the similarity of data augmentation technicalities, only some of augmentation technicalities are chosen prevent overfitting. During the first few tests overfitting was a problem due to the above mentioned reason. Fully convolutional network is used for the extraction of the road features from the input satellite images. On this architecture there are 16 trainable convolutional layer and there are deconvolutional layers to upscale the feature as the input size. There are 2,161,361 trainable parameters and 2,944 non-trainable parameters in total there are 2,164,305 parameters through the whole network architecture.

On FCN part the last three fully connected layers of the Vgg architecture are changed to convolutional layer. Then sigmoid activation function is used at final classification layer and binary cross entropy is used for the loss. Since the classification is binary, sigmoid activation function is chosen for the final classification layer and the binary cross entropy chosen for the loss. On classification layer pixel wise classification is takes place. The input image is classified as road and non-road region.

And finally the result I got is a promising result that is above 95.4% accuracy. Which implies that the proposed algorithm have a promising result for further study. So three things can be conclude from this research. First, since the deep learning needs much data to train the network and in this research condition there is no big dataset that is publicly available, using data preprocessing techniques to maximize the dataset size increases the performance of the algorithm. The second is, using small size filters helps to get detail feature extraction and fills the small breaks (gaps) which is helpful for the continuity. Finally, using only convolutional layer can extract features and there is no need to use fully connected layers in order to minimize the parameters which in return minimizes the computational complexity.

References

[1]. RashaAlshehhi, Prashanth Reddy Marpu, Wei Lee Woon, Mauro Dalla Mura, “Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks” ISPRS Journal of Photogrammetry and Remote Sensing 130 (2017) 139–149.

[2]. Ruyi Liu, Qiguang Miao, Jianfeng Song, YiningQuan, Yunan Li, PengfeiXu, Jing Dai “Multiscale road centerlines extraction from high-resolution aerial imagery”, Neurocomputing 329 (2019) 384–396.

[3]. Yanan Wei, Zulin Wang, and Mai Xu, “Road Structure Refined CNN for Road Extraction in Aerial Image”, IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, VOL. 14, NO. 5, MAY 2017, 709-713.

[4]. Campbell, J. B. 2002. Introduction to Remote Sensing. New York London: The Guilford Press.

[5]. Weixing Wang, Nan Yang, Yi Zhang, Fengping Wang, Ting Cao, PatrikEklund “A review of road extraction from remote sensing images”, ScienceDirect (journal of traffic and transportation engineering), 2016.

[6]. Fei-Fie, Li. Stanford University computer science class cs231n: Convolutional neural networks for visual recognition. 2017. http://cs231n.github.io/convolutional-networks/ (accessed 02 09, 2017).

[7]. Chen, Lin, Zhong Zhao, Wu Wei, and Yan Junjie. "Synaptic Strength for Convolutional Neural Network." Machine Learning (cs.LG); Artificial Intelligence (cs.AI); Computer Vision and Pattern Recognition (cs.CV); Machine Learning (stat.ML). Semantic Scholar, 2018. 1-10.

[8]. N., Srivastava, Hinton G., Krizhevsky A., Sutske I., and Salakhutdinov R. "Dropout: A simple way to prevent neural networks from overfitting." Journal of Machine Learning Research 15, 2014.

[9]. D., Rumelhart, Hinton G., and Williams R. Learning representations by back-propagating errors. Nature, 1986.

[10]. https://medium.com/the-theory-of-everything/understanding-activation-functions-in-neuralnetworks-9491262884e0

[11]. Hamid Reza RiahiBakhtiari, AbolfazlAbdollahi, Hani Rezaeian, “Semi-automatic road extraction from digital images”, The Egyptian Journal of Remote Sensing and Space Sciences, 2017.

[12]. XIAO XIANG ZHU, DEVIS TUIA, LICHAO MOU, GUI-SONG XIA, LIANGPEI ZHANG, FENG XU, FRIEDRICH FRAUNDORFER “Deep Learning in Remote Sensing”, ieee Geoscience and remote sensing magazine, December 2017.

[13]. A. Krizhevsky, I. Sutskever, and G. Hinton, “Imagenet classification with deep convolutional neural networks,” in Proc. Advances in Neural Information Processing Systems (NIPS), 2012, pp. 1097–1115

[14]. K. Simonyan and A. Zisserman. (2015). Very deep convolutional networks for large-scale image recognition. arXiv. [Online]. Available: https://arxiv.org/pdf/1409.1556.pdf

[15]. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[16]. R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Region-based convolutional networks for accurate object detection and semantic segmentation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 1, pp. 142–158, 2016.

[17]. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779–788.

[18]. H. Noh, S. Hong, and B. Han, “Learning de-convolutional network for semantic segmentation” in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2015, pp. 1520–1528

[19]. J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3431–3440.

[20]. Marcu, A., Leordeanu, M., Dual local-global contextual pathways for recognition in aerial imagery. Computing Research Repository (CRR) abs/1605.05462.

[21]. Alex Krizhevsky, IlyaSutskever, Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In: Neural Information Processing Systems (NIPS) Conference.

[22]. Lin, M., Chen, Q., Yan, S., 2014. Network in network. In: International Conference on Learning Representations, pp. 1–10.

[23]. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9

[24]. Zhou, B., Khosla, A., Lapedriza, À.,Oliva, A., Torralba, A., 2015. Object detectors emerge in deep scene cnns. In: International Conference on Learning Representations (ICLR)

[25]. Peikang Li, Yu Zang,Cheng Wang, Jonathan Li, Ming Cheng, LunLuo, Yao Yu, July 2016. “ROAD NETWORK EXTRACTION VIA DEEP LEARNING AND LINE INTEGRAL CONVOLUTION”, DOI: 10.1109/IGARSS.2016.7729408