Identification and Classification in Food Images

Tom Barrett

April 2, 2018

Abstract

Health conditions in the modern age are progressively getting worse with a high percentage of the population being obese. In order to combat this problem, a dietary assessment smartphone application would be invaluable. The task of recording ones food intake could be quite tedious and therefore, utilizing computer vision to automatically classify and calculate nutritional value of a users food image could help to make the process more practical for use. One approach that could be attempted, is by using Convolutional Neural Networks (CNNs) for the classification of food images. This is due to the recent high success in using CNNs for image classification. Retraining of the Inception-V3 model architecture, using tensorflow, the Food-101 dataset and the ImageNet dataset was carried out, resulting in a classification model. A Top 1 accurcay of 66.6% and a Top 5 accuracy of 85.96% was achieved using a dataset of 108 classes. It was found that CNNs are very successful in classifying images with over a 100 classes and that network fine tuning could result in even better results.

Acknowledgements

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The chapter outlines an introduction to the subject area of this Final year Project (FYP), the objectives of this project, the methodology used to complete this project, an overview of the report and an explanation of the motivation behind this FYP.

1.1 Overview

This project explores identification and classification of food images for use in a nutritional assessment android application. Food calorie consumption is a huge problem in the modern world. Over 25% of the population in Ireland is obese and this figure is likely to rise over the coming years. There has been extensive research carried out into nutritional assessment. Many of these studies utilise smart phone applications to record calorie intake while physical food diaries have also been used in the past. A study was carried out by (Goodman et al., 2015) whereby they used a mobile application to keep track of Vitamin D intake for young Canadians ranging from teenagers to early adulthood. (Arens-Volland, Spassova, and Bohn, 2015) also carried out a review of current computer supported nutritional assessment applications. They applications ranged from web based applications to smart phones. The problem with many of these 'food diary' applications is that they can be very tedious and time

intensive for users. Leveraging artificial intelligence to decrease the effort of the user or dietician would be very beneficial and many even encourage use of said nutritional assessment applications. Artificial intelligence could be utilised to classify food images and calculate the calorie count of an image so that a user would not have to input the information themselves.

The classification of images is the process whereby an algorithm predicts what objects are in an image. As humans, we take vision for granted. Computers (which find what are complex task for us, seemingly simple i.e mathematical computations) find it very difficult to interpret images. A black and white image is simply an array of numbers with values of 1 or 0 as seen in Figure 1.1. For color images, there are multiple layers of arrays for each of three colour channels red, green and blue. It can be quite difficult to extract feature information from these arrays.

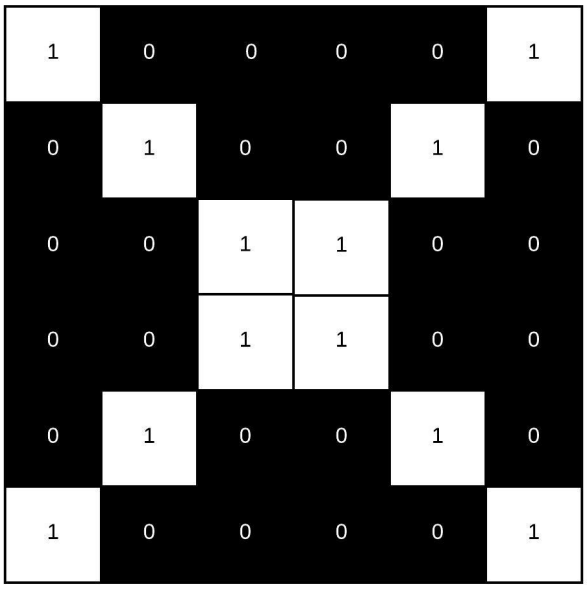

Figure 1.1: Representation of an Image as an Array

Even when classification can be achieved by analysing these arrays of numbers, it is most successful when the object in question takes up most of the image. For example, the banana is clearly a very prominent object in Figure 1.2.

Figure 1.2: Banana

Alternatively when the object of the image is occluded or not the focus of the image (Figure 1.3 and Figure 1.4), classification becomes very difficult. In contrast it can be clearly seen to humans that a banana is still present in both images. If we take a filled roll for example, it is difficult for the human eye to see the contents nevermind a computer so this is another issue that arises. Illumnination is also a factor that influences classification greatly as colour can play an important aspect in classifciation of food images as discovered by (Pouladzadeh, Villalobos, et al., 2012).

Traditionally, Support Vector Machines (SVMs) have been used to classify food images, many examples of which are covered in Chapter 2. The aim behind SVMs is to cast the information to a higher dimension, for example from a 2D to 3D space, so that the classes can be separated. SVMs will be explored further in Chapter 2. The features used by SVMs for food image classification are normally colour, texture, shape and size as in (Pouladzadeh, Shirmohammadi, and Al-Maghrabi, 2014) and (Pouladzadeh, Villalobos, et al., 2012). SVMs have their problems however and these problems of long training times and low accuracy with many classes have resulted in SVMs

Figure 1.3: Occluded Banana Image

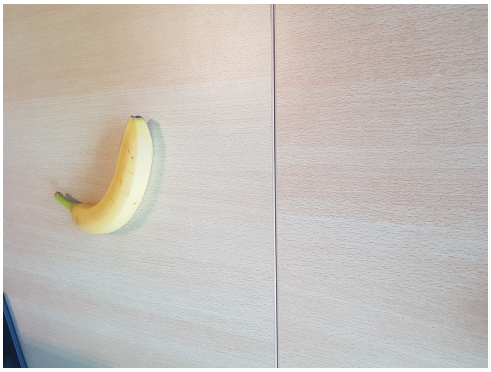

Figure 1.4: Banana at Small Scale

losing popularity in recent years.

Another reason for the shift from the use of SVMs is due to the recent success of utilizing neural networks for image classification. Neural networks are bio-inspired algorithms based off the structure of the human brain. A special type of neural networks, Convolutional Neural Networks (CNNs), have proven to be very successful in image classification (Krizhevsky, Sutskever, and G. E. Hinton, 2012). CNNs differ from other neural networks due to their extra layers of convolution and pooling. Convolutional layers filter the images for features while pooling refines the information in the arrays to reduce computational time and to make the features in the image clearer. CNN's have been around since the 1980's (Aurélien, 2017) but were not used by many until 2012 when CNNs became popular again due to the work carried out in that time, for example by (Krizhevsky, Sutskever, and G. E. Hinton, 2012). CNNs have been proven to be very successful in facial recognition and have been applied to food images with promising results as outlined in various studies in Chapter 2 such as (Yanai and Kawano, 2015) and (Mao, Yu, and Wang, n.d.).

For a nutritional assessment application, aided by computer vision, there are a few steps that must completed. Firstly, the food images are classified with prior segmentation in some cases. Image segmentation is the process by which an image is segmented into multiple parts with usually a segment for each different object or as in the case of Figure 1.5 each different food item. Secondly, size

estimation of each food type is carried out. Once data on the food type and size has been collected, calorie count estimation can be undertaken. Size and calorie estimation is outside the scope of this FYP.

Figure 1.5: Segmented Food Image (*Meanshift Algorithm for the Rest of Us (Python)* 2016)

The full system proposed to solve the problem statement would be able to integrate with an Android mobile phone application. The concept is, that when a user is about to eat their meal, they can simply take a picture of their meal for computation. The application would take an image, and send it to a server to be analysed. On the server side the image would be classified using a Tensorflow model. Tensorflow is a machine learning library that can be used to create CNNs. Once the classification step is completed, the size of each food type would be measured and through this, an overall calorie count would be displayed for the user. This could be logged for user metrics. The full system may not be possible to implement due to time constraints so therefore, classification will be the furthest step looked in to.

1.2 Objectives

Primary Objectives

Use Convolutional Neural Networks for Food Image Classification

There has been a large paradigm shift in food image recognition in recent years to using convolutional neural networks. This paradigm will be used to answer the problem statement.

Tune the CNN, replicating previous work

There are many different approaches to food identification and classification, many of which we will see in Chapter 2. An approach will be selected that has shown promising results in the past and replication of these results will be attempted. In addition to this, there are many different network architectures and parameters that can be adjusted when building CNNs and these will be tuned in order to get the best outcome possible.

Develop a mobile application to leverage the CNN

Another objective for this project is to develop an application that can be used for dietary assessment. This application would be able to take an image and then send the image to a server to identify and classify the foods within the image. Size estimation will not be explored for this project. Alternatively, due to time constraints, a previously developed application could be used to demonstrate the CNN developed.

Secondary Objectives

Understanding of Convolutional Neural Networks

In the project, Convolutional Neural Networks (CNNs) will be used for object identification in Food Images. A machine learning library will be used for this

due to time constraints but it is a key objective to develop a deep understanding of CNN's as they are quite pivotal in the current computer vision industry and bio-inspired systems are very interesting.

Learn about different image identification and classification techniques

Although, CNN's will be used for implementation, other methods of identification and classification will not be ignored. It is very important to learn about other methods as different methods are better suited for some situations and it would be best to know about these methods due to the inevitability of their use.

1.3 Methodology

The following methodology were adapter for this project:

Define the research question

The first step to this project was to define the research question. The general area of a computer vision supported nutritional assessment application was known at conception but the scope of this is too broad for an FYP. Therefore it was decided that the research question would be to look at the food identification and classification aspect.

Literature Review

Once the research question was defined, finding related work was the next milestone. There are many attempts at Food Image Classification and these were not difficult to find, using Google Scholar, but many of these papers glossed over the segmentation aspect and relied on third parties for this step.

Because of this, quite a few references to different papers had to be followed that focused solely on image segmentation and classification.

Explore different image identification methods

Various image identification methods were collected from the literature review that was carried out, so there were many options to evaluate. Convolutional Neural Networks were the clear choice due to recent popularity and successful results, so more traditional methods of identification using colour and texture was not so strenuously explored.

Select an image identification and classification method

After the various methods to identify and classify food images had been collated in the literature review, one of these had to be selected.

Research technologies and develop skills in these technologies

As Convolutional Neural Networks (CNN) are used in this project, many resources have been leveraged to enrich understanding of the process. Tensorflow is the main resource utilised in creating a CNN so on-line tutorials for this technology were greatly beneficial. A Deep Learning Course on Udacity was also used to enhance understanding and skills.

Build a prototype of the application

A prototype application had to built for demonstration for this project.

Compare and analyse results to other implementations

Once a model had been successfully trained, comparison to previous work was carried out.

1.4 Overview of Report

This report is broken down into various chapters. The introduction chapter is to give on overview of what this project is about, how the project will be approached and why it is being carried out. Following this, some information on the background of the subject of this Final Year Project (FYP) will be outlined. Once the chapters of introduction and background have been completed, a chapter outlining the learning of Tensorflow will be explored. Tensorflow is a machine learning library that is used to develop CNNs. In this chapter, the goal is to demonstrate the activities undertaken to learn about the technologies used and show tutorials that have been completed to aid with this learning. In chapter titled 'Training Using the Inception-V3 Model Architecture' , the purpose is to outline experiments that have been carried out in relation to training Tensorflow models. A Tensorflow model is the term used to describe the neural network produced by Tensorflow. The Inception-V3 architecture is proven to be a very successful architecture for classifying images and will be explained in detail in said chapter. Once the model has been trained, the following chapter 'Analysing the Trained Model' will consist of experiments that analyse and test the models trained. A prototype smart phone application was developed to demonstrate the efficacy of the system and an overview of the process will be explained in the chapter titled 'Prototype Application.' The final chapter's purpose is to provide a discussion on the results obtained and offer a conclusion to the findings.

1.5 Motivation

I find the topic of Computer Vision a very interesting one. It excites me, to be able to 'teach' a machine how to see as we do. For this reason, I really wanted to learn about Neural Networks and this was a large motivator for this project.

Once I had a topic that I wanted to research, I needed a focus or problem statement for this research. I find that it is much more rewarding to work on

something that positively impacts both myself and other people so I decided that I wanted to research something that fit this requirement.

Food calorie consumption is a very big problem in the modern world. Over 25 % of the population in Ireland are obese. A mobile application that could help keep track of a user's calorie intake by taking a picture of their meals would be a big help to combat this problem. This problem statement works very well for me because of it's application use and because of its complexity. Identifying and recognising food is much more difficult than say recognising faces as it has no uniform shape. Therefore, this problem would also be very beneficial to developing skills in the computer vision area.

I would like to develop these real world skills so that I can partake in Computer Vision projects in industry or to do further research in academia. I have also been offered a graduate role in Jaguar Land Rover and these skills would be very relevant in the autonomous driving industry. This is because machine learning has really taken off in the last few years and is used by many in industry. While machine learning as a whole has become very popular, computer vision is probably the most prominent that has come out of it. Face detection, for one, is being researched extensively for use in personalised advertising and also secure access to devices and systems.

Chapter 2

Background

This chapter outlines background information on the subject of nutritional assessment using computer vision. An overview of neural networks is explored, followed by a literature review of the subject area. Analysis into evaluating the efficacy of CNNs is covered and finally a conclusion to the information gained.

2.1 Introduction to Machine Learning

In (Mitchell, 1997a), machine learning is defined as "the question of how to construct computer programs that automatically improve with experience". Machine learning has blossomed in recent years with applications across multiple domains using vastly different paradigms and technologies. Some of the different approaches used in machine learning are: Artificial Neural Networks, Genetic Algorithms, Decision Tree Learning and Bayesian Learning (Mitchell, 1997a).

There are many ways in which machine learning can be used in the modern world, many of which are being utilised to great affect. Some of these applications are image recognition, natural language processing and many more. These applications can be applied to many different domains such as security (face detection in airports) or object detection (autonomous driving). (Erick-

son et al., 2017) carried out research into the area of using machine learning to aid diagnosis of medical conditions. Machine learning algorithms could suggest possible diagnosis for medical professionals to interpret and therefore reduce the time spent diagnosing a patients condition. There may be fear that machine learning will start to take away many jobs from humans but this may not be the case as in the example above as computers will only be aiding professionals.

One of the most exciting avenues in machine learning is computer vision. This is due to the application domains mentioned above. Computer vision is the process of extracting high-dimensional data from an image to produce useful information, which in terms of classification usually results in labeling. It can be used in many areas to improve our lives. As mentioned earlier, autonomous cars are only possible when a machine can determine what objects are around it. Computer Vision can allow a machine to recognise medical conditions in an image such as breast cancer using mammography images (Erickson et al., 2017). The applications are nearly limitless.

2.2 Neural Computing

The main area of my focus for this project is in Artificial Neural Networks (ANN). This is because extensive research has been carried out into Convolutional Neural Networks (CNNs) which are based on ANNs.

Artificial Neural Networks

An Artificial Neural Network is a bio-inspired system that is used to model the human brain in how it learns from experience. ANNs were first introduced in 1943 and were studied until the 1960's when the concept was shelved and revived gain in the 1980's (Aurélien, 2017). The ANN uses this model to build a very complex web of connected units called artificial neurons. These neurons are connected by certain weights which determines the processing capacity of the network and these weights are created by learning a dataset (Eaton, 2017).

An neuron has a set of inputs that take in a value, sometimes from network outputs and produce a single result or classification.

As stated above, an ANN is bio-inspired from biological neurons (Aurélien, 2017). Biological neurons are cells located in animals brains. These neurons receive signals (electrical impulses) from other neurons and fires their own signal thereafter. While an individual neuron is very simple, millions of these neurons working together can result in very complex computations. From this, an artificial neuron was proposed through a simple model which had a binary input (1 or more) and a single binary output.

Before Convolution Neural Networks can be explored, which are vital to image processing, the perceptron learning algorithm, the multi layer perceptron, backpropogation must be analysed.

Perceptron Learning - Artificial Neuron

In our Artificial Neural Network a Perecptron is an Artificial Neuron. It is called an Artifical Neuron because it is a bio-inspired neuron which models a neuron in the human brain in terms of inputs and output.

In Perceptron learning, we can take two inputs which are put towards an activation function with a bias attached as seen in Figure 2.1. These inputs are multiplied by the weights that connect the input to the activation function and depending on the result, the activation function may fire an output. These inputs are either 1 or -1.

The Perceptron Training Rule is the means by which weights are selected to produce the correct output during training. As in (Mitchell, 1997a), a common way to train a perceptron is to start with random weights and change them during training as per the training rule. This rule follows the formula below:

$$w_i \leftarrow w_i + \Delta w_i$$

where xi is the input and the equation below is valid:

$$\Delta w_i = n(t - o)x_i$$

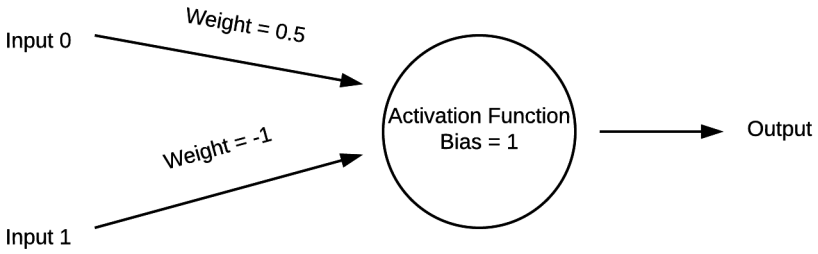

Figure 2.1: Perceptron

In the above formula, "t is the target output for the current training example, o is the output generated by the perceptron, and n is the positive constant called the learning rate" (Mitchell, 1997a). The ouput of a neuron is claculated using the activation function. This Perceptron Training Rule assumes that there are two sets of instances, a positive and negative set (class x and - in Figure 2.2), and that they are linearly separable, as in demonstrated in Figure 2.2.

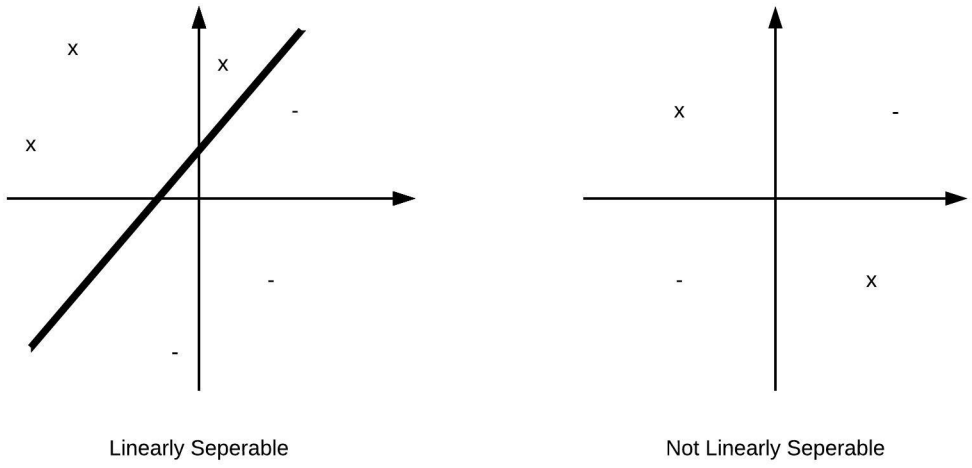

Figure 2.2: Linearly Separable, adapted from (Mitchell, 1997a)

A perception is trained using supervised learning. When the perceptron classifies a results, it is told if it is correct or not. If the result is incorrect, weights

are changed in value so that this error can be reduced (Luger, 2005).

The activation function used in a neural network calculates the output for a neuron.

There is one major problem with perceptron learning and that is, it can't solve a problem if there is not a clear linear separation between the classes. There is a way in which we can attempt to solve this, through the delta rule. The delta rule utilizes gradient descent to find the best weight for the training samples (Mitchell, 1997a). We will discuss gradient descent in the next section.

Multi Layered Perceptron

Multi Layer Perceptrons (MLP) are made up of multiple layers of perceptrons connected together and are used to combat non-linearly separable classes. While the delta rule can solve problems on non-linearity when there are two classes, MLPs can solve non-linearity when there are more than two classes. Firstly, we have an input layer, followed by one of more hidden layers and then finally an output layer. Any Neural Network with more than three hidden layers is categorised as a deep neural network.

The input layer of your network consists of the data you feed into the network in order to classify it. The input layer passes this data to a hidden layer whose purpose is to transform this data into something that the output layer can understand. This transformation results in a linearly separable space that can be classified. The output layer normally consists of a class prediction.

Multi Layer Perceptrons are a class of feed forward Artificial Neural Networks. These means that the output of each perceptron feeds into an input in the next layer of the network, example seen in Figure 2.3.

During training, using backpropogation, for each step, the output of every neuron is calculated in each layer and then passed to the next layer. Then the error of the output is calculated and the network calculates how each neuron in the last hidden layer effected the error. This is continued back through all the layers until the input layer (Aurélien, 2017). The weights are then altered to try and reduce this error.

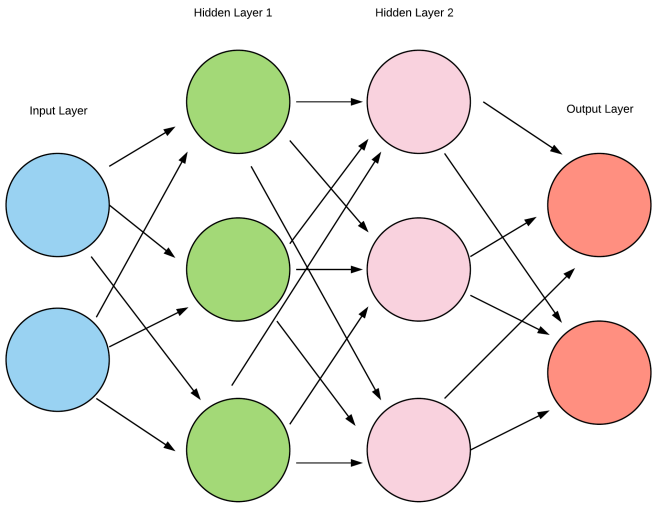

Figure 2.3: Multi Layer Perceptron

There is one large problem with MLP's and this is why Convolutional Neural Networks (CNN) were created. If you attempting to classify images with a MLP then each pixel in that image would have to be a separate input. This creates a massive amount of neurons through all the layers and this isn't feasible. CNN's solve this problem which we will discuss later.

Gradient Descent and Backpropogation

Gradient Descent is an algorithm used to find the optimal weights to produce the smallest prediction error. It is used to overcome problems of non linearly separable classes. Gradient descent search selects a random weight value and then modifies it gradually to minimize the error. "At each step, the weight vector is altered in the direction that produces the steepest descent along the

surface" (Mitchell, 1997a). This step is iterated until the lowest value is met.

There is an error function used for the perecptron which finds the lowest error for that neuron, but it can't be used here because, since we have many neurons, there could be an error in multiple neurons. Gradient Descent is mathematically based on the derivitive of a function. The gradient of a function can be calculate by differentiating it. As the weights are what is being controlled, "they are what we differentiate in respect to" (Marsland, 2015). The negative gradient of this function is followed to find the lowest possible point, hence the name gradient descent (Marsland, 2015).

One problem with Gradient Descent is that if we look at Figure 2.4, we may never get to the optimal point, point B. This is because we will find point A without too many problems but when the weights change we will get too high a slope of error and therefore will never reach point B.

Another variation of Gradient Descent is Stochastic Gradient Descent (SGD). SGD is different because it updates "weights incrementally, following the calculation of the error of each individual example" (Mitchell, 1997a).

"The Backpropagation algorithm learns the weights of a multilayer network, given a network with a fixed set of units and interconnections" (Mitchell, 1997a). Backpropgation attempts to minimise the mean squared error between the target output and the output of a network.

Backpropagation works by starting at the output layer of the network and going back through previous hidden layers, updating weights as it goes ie. it propogates back through the network, updating the weights to try and reduce the error.

(Mitchell, 1997a) defined a walkthrough of the backpropgation algorithm. For every value of

$$\vec{x}, \vec{t}$$

, in the training set where x is a vector of inputs and t is a vector of output values to act a target:

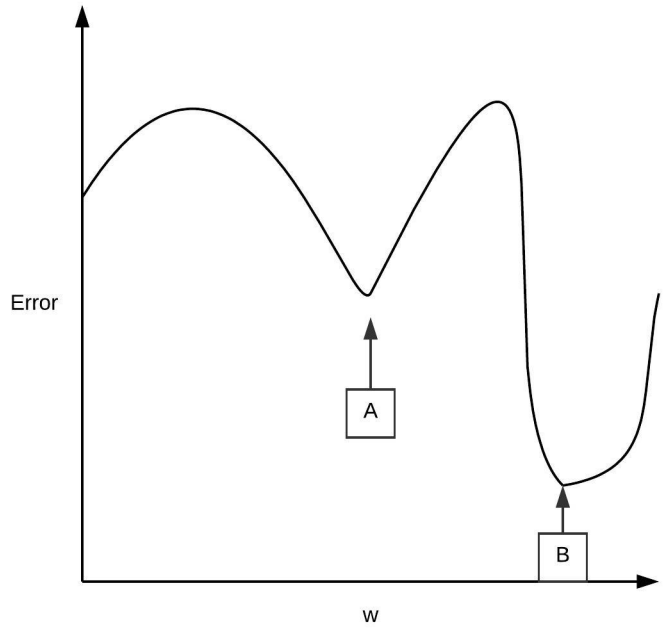

Figure 2.4: Gradient Descent

- Run x through the network and output

  o_u

- For each output k, calculate the error by:

  $\delta_k \leftarrow o_k(1 - o_k)(t_k - o_k)$

- For every hidden unit, calculate the error by:

  $\delta_h \leftarrow o_h(1 - o_h) \sum_{k \in outputs} w_{kh}\delta_k$

- Updates weights by:

  $w_{ji} \leftarrow w_{ji} + \Delta w_{ji}$

  where

  $\Delta w_{ji} = \delta_j x_{ji}$

2.3 Convolutional Neural Networks Overview

Convolutional Neural Networks (CNN's) are essentially a Multi Layered Perceptron with a special structure. CNN's have one major difference from a MLP, they have extra layers of convolution and pooling. The architecture of a convolution network can be seen in Figure 2.5.

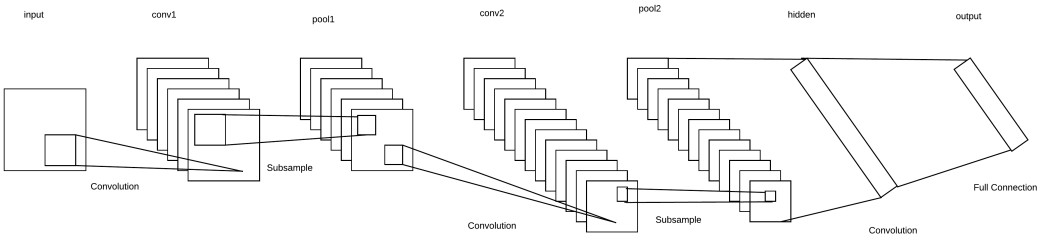

Figure 2.5: CNN Architecture

Each convolutional layer in the network extracts features from the images. The closer the image is to the input layer, more primitive features are extracted while the closer the image is to the output layer, more complex features are extracted.

Figure 2.6 (a) shows an image that we want to compare against Figure 2.6 (b). For humans, it is quite easy to determine that these images are very similar but for a computer this task is surprisingly difficult.

So what a CNN does, to combat this problem, is to take a small feature from Figure 2.6 (a) and compare it to a subsection of Figure 2.6 (b). The CNN multiplies the feature and a section of Figure 2.6 (c), adds up the results and divides by 9 (The number of pixels in the feature). This then gives a decimal value of how likely it is that the feature is in the part of the image, as seen in Figure 2.6 (d). This is called filtering. The convolutional layer is composed of carrying out this filtering for every single possible location in Figure 2.6 (a).

Next is the Pooling Layer that takes the convoluted layer output (you can use Figure 2.6 (d) as reference) and from a user defined size ie. 2x2, gets either the highest decimal value (Max pooling) , the average value (Mean pooling)

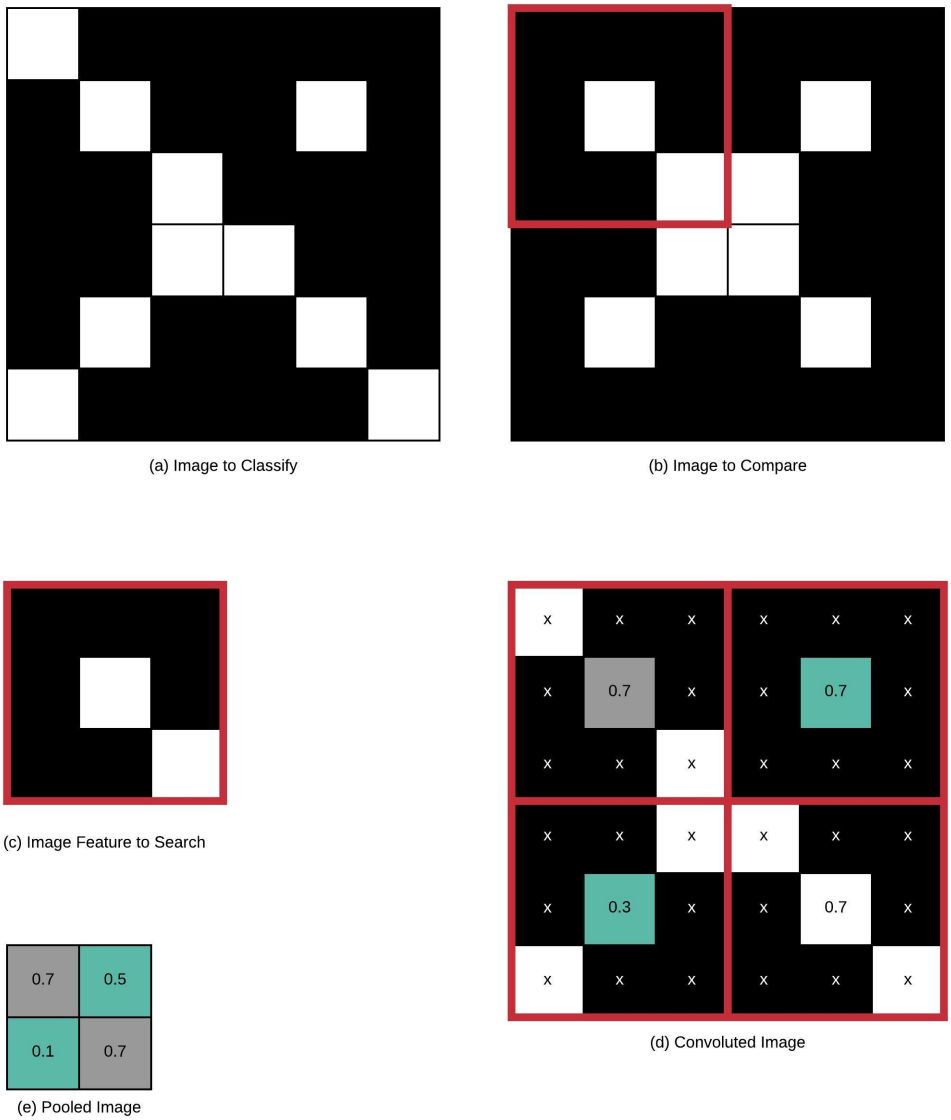

Figure 2.6: CNN Overview

or the lowest decimal value (Min pooling) and records that as the new value for the section. This is then applied to the entire image. As we can see in Figure 2.6 (e) we now have a much smaller image stack in which to classify, thus making the computation easier. Pooling also makes features of the image clearer which helps to result in more accurate classification.

In between the convolution and pooling layer, there is sometimes a normalization layer. This normalization layer creates Rectified Linear Units (RLU's). In other words, if we take Figure 2.6 (d), it changes all minus values to zero.

There are some problems with CNN's however. One of the main problems is that you need a very large dataset in order to produce am accurate model and the training can be very time consuming even with a GPU. This is because each training image (or batch) has to be ran through the network and go through backpropogation.

Receptive Field and Kernel

In the previous section where an overview of CNN's was addressed, there was a discussion focused on filtering where features were searched for across the images. This filtering resulted in windows sliding across the image, looking for common features. This window is called the receptive field and this is the section of the image that is being looked at.

As the image is convoluted and processed through the network, the changed image can be called the kernel. The kernel of the CNN represents what features are extracted from the image throughout the network.

Activation Function

The activation function used in a neural network calculates the output for a neuron. The most commonly used activation function in CNN's is the ReLu activation function or the rectified linear unit (Aurélien, 2017). The softmax activation function is sometimes used for the output layer (Aurélien, 2017).

Dropout

Dropout is a type of regularisation technique used to prevent overfitting in a network (Aurélien, 2017). Overfitting in a network is when the network cannot generalise to new data after over learning the training data. In dropout, each neuron is given a probability (normally set to 0.5) and this is the likelihood that the neuron will be used for that training step. It has proven quite successful for increasing a networks accuracy (Aurélien, 2017).

Different Sizes of Convolutions

Zero padding can be added to an output so that it is same size as the previous layer by adding zeros around the feature (Aurélien, 2017).. Through training, it is "possible to connect a large input layer to a much smaller layer by spacing out the receptive fields" (Aurélien, 2017). The stride between two different receptive fields is the spacing between them (Aurélien, 2017).

Fully Convolutional Networks

A Fully Convolutional Network is one that does not have a fully connected layer and in a fully connected layers place is another convolution layer. This can be achieved by making the size of the receptive field equal to the size of the input (*Image Segmentation Using DIGITS 5* n.d.). They are proven to be very successful in object detection.

Inception-V3 Model Architecture

The Inception V3 model network architecture was used for this experiment. The Inception V3 architecture was created by building on the existing Inception model aimed at efficient image classification (Szegedy et al., 2016). This research was carried out due to the popularity of convolutional neural networks. The main aim of this study was to produce a model that would "scale up networks in ways that aim at utilizing the added computation as efficiently

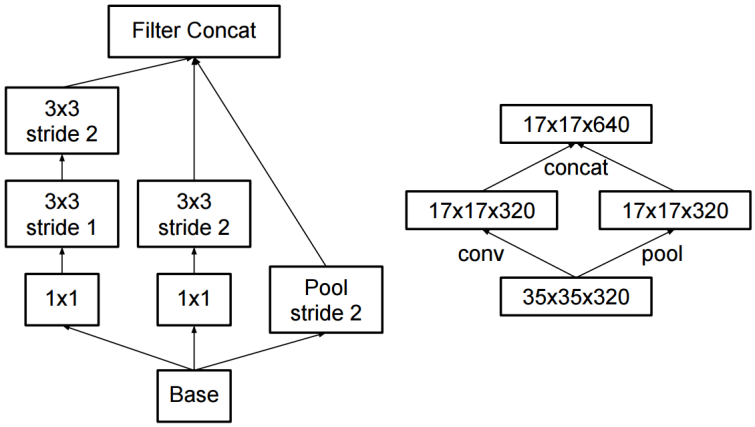

Figure 2.7: Inception module (Szegedy et al., 2016)

as possible by suitable factorized convolutions and aggressive regularization" (Szegedy et al., 2016). The team did not want to simply rely on larger models for better results.

The are 4 main design principles that the research team followed in the paper (Szegedy et al., 2016):

- Avoid representational bottlenecks
- It is easier to process higher dimensional representations locally
- Without much loss in power, spatial aggregation can be carried out over lower dimensional embeddings
- The width and depth of the network should be balanced and when one is increased it can be beneficial to increase the other.

The main idea behind the network is that different strands of the network are followed and the best result is used as seen in Figure 2.7.

In order to combat the problem of low resolution input, three separate experiments were carried out based on the fact that "One simple way to ensure constant effort effort is to reduce the strides of the first two layer in the case of lower resolution input, or by simply removing the first pooling layer of the network" (Szegedy et al., 2016).

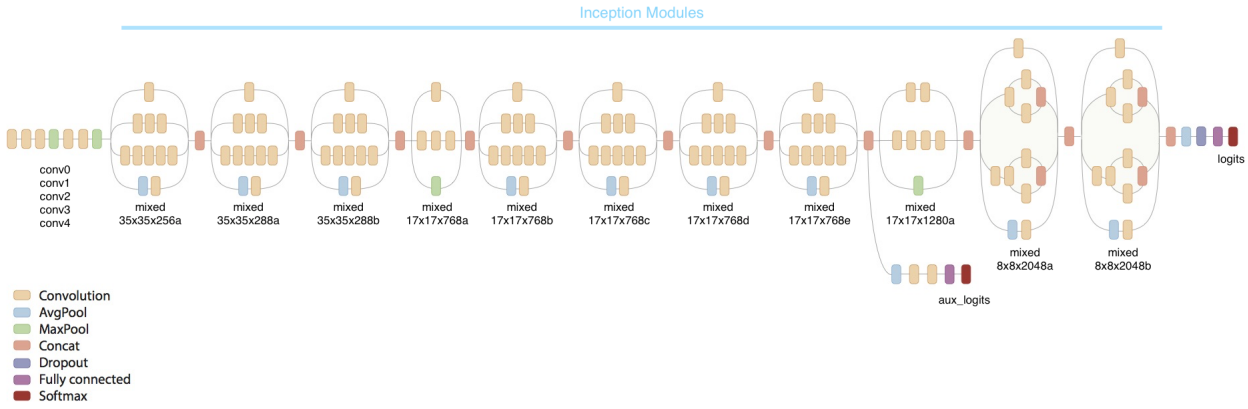

Figure 2.8: Inception V3 Architecture (Szegedy et al., 2016)

These experiments are as follows, with the first resulting in the most accurate:

- "299 x 299 receptive field with stride 2 and maximum pooling after the first layer." (Szegedy et al., 2016)
- "151 x 151 receptive field with stride 1 and maximum pooling after the first layer." (Szegedy et al., 2016)
- "79 x 79 receptive field with stride 1 and without pooling after the first layer." (Szegedy et al., 2016)

The best results of the Inception-V3 model achieved 78.8% top-1 accuracy and 94.4% top accuracy of the ILSVR 2012 dataset (Szegedy et al., 2016). This model also was the least computationally expensive of other published and successful models. The full model can be viewed in Figure 2.8.

Mobilenet Model Architecture

The network architecture used for this experiment is MobileNet (Howard et al., 2017). A parameter can be set to retrain this model in the code, as mentioned in previous experiments, instead of inception V-3 (*How to Retrain Inception's Final Layer for New Categories* n.d.). This architecture is designed to be

smaller and much more efficient so that it can be used on smartphones which have less powerful resources available.

Mobilenet is designed for various different use cases on mobile phones such as image classification, landmark recognition, object detection and face attributes. It is based on "depthwise separable convolutions" whose purpose is the split a convolutional layer into a depthwise convolution and a pointwise convolution (1x1 convolution) (Howard et al., 2017). The purpose of the former is to apply "a single filter to each input channel" (Howard et al., 2017) while the latter combines these outputs. This in turn reduces the computational time. The first layer of the network is a standard convolutional layer and does not get split up.

Unfortunatley there is a significant decrease in accuracy in using Mobilenet as it is a much thinner model.

History

Convolutional Neural Networks are a bio-inspired approach to solving intensive tasks. They have been applied to image recognition, along with natural language processing and voice recognition, with image recognition being explored since the 1980's (Aurélien, 2017). In 2012, CNNs increased in popularity due to increased computational power in the hardware, availability in datasets and breakthrough research carried out in the field (Krizhevsky, Sutskever, and G. E. Hinton, 2012). CNNs have been recently applied to many complex tasks with exceptional results such as facial recognition, and many more.

2.4 Support Vector Machine

A Support Vector Machine (SVM) is a machine learning algorithm that has been very popular before the use of a CNN was mainstream. Many of the texts that will be analysed later use SVMs for their classification.

A SVM works by creating an n-dimensional space, with n as the number of inputs you have (*Support Vector Machines for Machine Learning* n.d.). The SVM algorithm finds the hyperplane that splits this space. This hyperplane can then be used for classification. An SVM casts the problem to a higher dimensional space and this can solve a problem when the classes are not linearly separable. This can be seen in Figure 2.9, where on the left there is no clear way to separate the classes but once the problem is cast to another dimensional, the separation is clear.

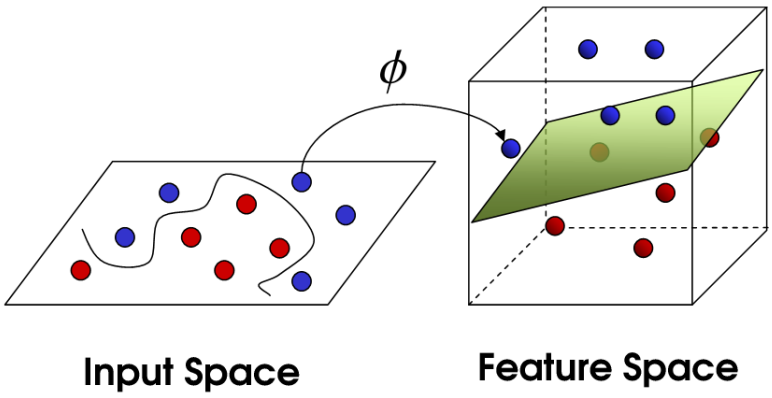

Figure 2.9: Support Vector Machine Applicability sourced from https://stackoverflow.com/questions/9480605/what-is-the-relation-between-the-number-of-support-vectors-and-training-data-and

A kernel is used to implement a SVM. There are a few types of kernels that can be used in the SVM algorithm such as:

- Linear Kernel
- Polynomial Kernel
- Radial Kernel

The kernel is a vector of how similar two vectors are. In terms of a linear kernel, the dot product of two vectors is the kernel.

There are benefits and liabilities to using a SVM. They can be very accurate and can work efficiently with small datasets but unfortunately it can take a large amount of time to train.

2.5 Evaluating the Output

There are various different metrics that can be used for evaluating the output of a classifier or segmentation algorithm, many of which we have seen in previous sections. These metrics will be examined below.

Research into Diagnosing Errors in Object Detectors

There has been some research into the question of how to evaluate object detectors, one of which be discussed in detail (Hoiem, Chodpathumwan, and Dai, 2012). This paper in question "analyzes the influences of object characteristics on detection performance and the frequency and impact of different types of false positives" (Hoiem, Chodpathumwan, and Dai, 2012). They found that there were many effects that had influence on detectors as follows:

- occlusion
- size
- aspect ratio
- visibility of parts
- viewpoint
- localization error
- confusion with semantically similar objects
- confusion with other labeled objects
- confusion with background

The research team goes on to analyse false positives in object detectors. Localization errors were a large factor. This is were bounding boxes overlap to other objects in the image. Confusion with similar objects had a large influence on false positives also by which, for example, a dog detector had a high score for a cat (Hoiem, Chodpathumwan, and Dai, 2012). Confusion with dissimilar objects and confusion with background are the categories of the rest of the false positives they measured.

In conclusion the team would that "Most false positives are due to misaligned detection windows or confusion with similar objects" (Hoiem, Chodpathumwan, and Dai, 2012). They had some recommendations towards improves detectors as follows:

- Smaller objects are less likely to be detected
- Localization could be improved
- Reduce confusion with similar categories
- Robustness to object variation
- More detailed analysis

Detection Average Precision

The average detection accuracy of a system. See **??**.

Mean Average Precision

The mean accuracy of a system across all results. See **??**.

Distribution of top-ranked false positives

This metric is used to tell where most errors occur when false positives are evident. These errors are broken down into four categories of localization, similar objects, background confusion and others. See **??**.

Segmentation Mean Accuracy

This is the mean accuracy of segmentation. See **??**.

Per-category segmentation accuracy

This metric measures the accuracy of segmentation at a category level. See **??**.

Per-class segmentation accuracy

The accuracy of segmentation at a class level is analysed. See **??**.

Top-1 and Top-5 Accuracy

When a classifier is given an image it normally responds with a list of predictions along with a decimal representation of the likelihood that it is of that class. The Top-1 accuracy is the top valued prediction and Top-5 accuracy is the top 5 predictions.

2.6 Dietary Assessment using Computer Vision

Convolution Neural Networks

Many researchers have used convolutional neural networks for image classification with various network architectures and many have used a food image dataset. Some of these papers will be looked at below. A summary of their results can be seen in Table 2.2.

(Mao, Yu, and Wang, n.d.) focused on a deep learning approach to food image recognition based their neural network architecture on Inception-ResNet

and Inception V3. Deep learning is a term usually given to algorithms based on neural networks. They also used the Food-101 dataset. For this system, Google's Tensorflow was used for image preprocessing. Preprocessing was needed as the environmental background is different in many food images. Because of these "Grey World method and Histogram equalization" (Mao, Yu, and Wang, n.d.) were used. Amazon Web Services (AWS) Graphics Processing Units (GPU) instances were used for training. AWS instances are cloud servers. The results on completion were quite impressive with a Top-1 Accuracy of 72.55% and a Top-5 Accuracy of 91.31%.

Another research team in Japan, (Yanai and Kawano, 2015) researched this topic. This was built off previous research they had carried out in the field (Kawano and Yanai, 2014). They were aware of how difficult the problem was and therefore employed many techniques to solve the problem such as "pre-training with the large-scale ImageNet data, fine-tuning and activation features extracted from the pre-trained DCNN". In conclusion, they found that the "fine-tuned DCNN which was pre-trained with 2000 categories" from ImageNet was the best method. A DCNN is a Deep Convolution Neural Network. A network can become a DCNN when the number of hidden layers is larger than three. While many of the CNNs discussed have been DCNN, most are just labeled as CNNs. The achieved results of 78.77% for Top-1 Accuracy in the UECFOOD100 dataset.

(Kagaya, Aizawa, and Ogawa, 2014) also employed the use of convolutional neural networks for image detection. They used a CNN for the "tasks of food detection and recognition through parameter optimization". They found that a CNN is much better suited to the task than a Support Vector Machine (SVM). They achieved an overall classification accuracy of 93.8% against their baseline accuracy of 89.7%. This accuracy was calculated using a dataset that they created specifically for this task. When they had completed the task they analysed the trained convolutional kernels and came to an interesting conclusion. They found that "color features are essential to food image recognition".

The last paper that will be analysed, (Liu et al., 2016), oriented around using a Convolutional Neural Network for food image recognition, focuses on

developing a dietary assessment application for use on a smart phone. They used the UEC-256 and Food-101 dataset for their experiments and achieved impressive results. They used a Convolutional Neural Network but "with a few major optimizations, such as optimized model and an optimized convolution technique". They used the Inception module for their CNN. After the inception module was complete, they made the GoogleNet architecture by combining modules. In total, the network had 22 layers. They achieved the results shown in Table 2.1.

Table 2.1: DeepFood Results

Dataset	**Top-1**	**Top-5**
UEC-256	54.7%	81.5%
UEC-100	76.3%	94.6%
Food-101	77.4%	93.7%
UEC-256 With Bounding Box	63.8%	87.2%
UEC-100 With Bounding Box	77.2%	94.8%

(Mezgec and Koroušić Seljak, 2017) developed a new neural architecture specifically for detecting food and drink images using deep convolutional neural networks called NutriNet. The trained network was to be used to aid patients with Parkinson's disease in monitoring their diet. NutriNet was created based off of the AlexNet architecture. The dataset used for this study consisted of approximately 500 images for each of over 500 classes. Through this dataset a Top-1 accuracy of 86.72& and a Top-5 accuracy of 94.47% was recorded. A smart phone application was used for real world testing which brought a Top-5 accuracy of 55%. The application also saved these real world images from the smart phone to increase their dataset size. In conclusion, the team found that there were modifications that could be made to the NutriNet architecture as real world images didn't perform incredibly well due to occlusion and background noise in the images. The detection and recognition steps were separated in this architecture. The team also acknowledges that joining these steps into a single DCNN may be successful and should be explored.

Table 2.2: Summary of results in CNN based methods

Title	**Dataset**	**Top-1 Accuracy**
(Mao, Yu, and Wang, n.d.)	Food 101	72.6%
(Yanai and Kawano, 2015)	UECFood101	78.8%
(Kagaya, Aizawa, and Ogawa, 2014)	Own dataset	93.8%
(Liu et al., 2016)	Food 101	77.4%
(Mezgec and Koroušić Seljak, 2017)	Own dataset	86.7%

Support Vector Machines

While Convolutional Neural Networks have proven very successful in recent years, there are many other methods of food image identification and classification that have been employed by food image recognition researchers, such as SVMs. A summary of the results of these methods can be seen in Table 2.3.

A Food Image Recognition System with Multiple Kernel Learning

In this paper, (Joutou and Yanai, 2009), a practical use for food image recognition in the form of a mobile phone application was proposed. In order to classify the images, multiple kernel learning(MKL) was used. MKL is similar to am SVM expect that instead of a single kernel during training, MKL "treats with a combined kernel which is a weighted linear combination of several single kernels" (Joutou and Yanai, 2009). The idea behind this is that different food types are distinguishable by different factors and using this method, the best of these factors can be used for classification of that food type. In the experiment carried out by (Joutou and Yanai, 2009), three different factors were used for learning:

- Color Histograms:
- Gabor Texture Features
- Bag-of-Features using Scale Invariant Feature Transformation (SIFT)

50 different classifiers were created in a SVM using MKL with "one category as a positive set and other 49 categories as a negative set" (Joutou and Yanai, 2009). For each of these categories, a web scrape was carried out and then the best 100 images for each scrape was manually selected.

Five-fold cross validation was utilised in the paper. When cross-validation is utilised, the test set is used for gradient descent while a separate validation set is used to measure error (Mitchell, 1997a). Five-fold cross-validation is where five separate validation sets are created. These five different validation sets are used to test the model and the results are averaged.

MKL proceeded to yield results of 61.34% on the 50 food types and a Top-3 accuracy of 80.05% (Joutou and Yanai, 2009). The prototype mobile phone application resulted in a 37.55% user accuracy.

A Novel SVM Based Food Recognition Method for Calorie Measurement Applications

Another quite successful study was carried out using a SVM. (Pouladzadeh, Villalobos, et al., 2012) had established that both colour and texture are very important but they also decided that shape and size are vital features to analyse. The proposed system has two main parts, segmentation followed by classification. In order to create a 'robust' system, a 'Robust Handling of Different Lighting Conditions' module is added to the system (Pouladzadeh, Villalobos, et al., 2012). This is so that various lighting conditions don't cause color data to be distorted.

Since this paper calls for calorie estimation, the first step of the system calculates the size of the food portion. In order to do this, a coin or the users thumb is included in the image taken so that the pixel count of the thumb and the food can be compared to estimate the size. Following this the image is segmented into various portions. The following step classifies each segment of the image by extracting color, texture and shape features and inputing these into a SVM (Pouladzadeh, Villalobos, et al., 2012).

12 different food types were trained for this SVM with an average classification accuracy of 92.6%.

Measuring Calorie and Nutrition From Food Image

Another study that employed both a SVM and an emphasis on colour, texture and shape features, was carried out by (Pouladzadeh, Shirmohammadi, and Al-Maghrabi, 2014). Size was also a factor in the calorie measurement module of the system. It was found that using all four of these features increases the overall accuracy.

In order to segment the image successfully, Gabor filters were applied to separate texture features while color was also utilised. For each segment established, size, shape, color and texture features were extracted and using a SVM, a classification was made. The SVM used the radial basis function kernel (Pouladzadeh, Shirmohammadi, and Al-Maghrabi, 2014).

Calorie estimation was also a large part of this paper, and the users thumb was taken with the food in order to calculate food size.

In the prototype application, once the classification had been made, the user can confirm or change the prediction. Another feature of the application was in regards to "Partially Eaten Food" (Pouladzadeh, Shirmohammadi, and Al-Maghrabi, 2014). This was accomplished by taking a picture before and after consumption and as a result only calculated the size of the food eaten and therefore more accurate calorie counts can be produced.

15 food types were trained using the SVM with 3000 images. The accuracy for the classifier averaged at 90.41% using 10 fold cross-validation (Pouladzadeh, Shirmohammadi, and Al-Maghrabi, 2014). There was also a calorie count accuracy of 86%. The best classification results were on single foods followed by non mixed and finally mixed foods produced the worst results.

Segmentation Assisted Food Classification for Dietary Assessment

(Zhu et al., 2011) had a strong focus on the segmentation aspect of a dietary assessment system. The segmentation of the food images was achieved "using Normalized Cuts based on intensity and colour" (Zhu et al., 2011). Normalized Cuts is a graph based segmentation method. To aid the segmentation aspect of this study, a common background colour was introduced to the images. Segmentation refinement was also an important module in the experiment. This is the process by which neighbouring segments with the same classification label are merged together. This also helps calculate a more accurate size estimation.

The classification of the segmented image was processed by using a SVM calculating colour and texture features. Gabor filters were used for the texture feature extraction (Zhu et al., 2011).

In the experimental results for this study, it was found that segmentation was not always successful "when the region of interest is camouflaged by making its boundary faint" (Zhu et al., 2011). In their case, it was a can of coke that wasn't segmented correctly. The classification accuracy was of 56.2% and 95.5% with ground truth segmentation data (Zhu et al., 2011).

Large Scale Learning for Food Image Classification

(Abbirami.R.S et al., 2015) proposed a food image recognition system using a Bag of Features model. This study used over 5000 images separated into 11 classes.

A clustering algorithm was employed on this study before classification. For the classification step, experiments were carried out using different methods:

- SVM
- ANN
- Random Forests

The final accuracy of the system was 78% (Abbirami.R.S et al., 2015).

A Personal Assistive System for Nutrient Intake Monitoring

Similar to other approaches seen thus far, (Villalobos, Almaghrabi, Hariri, et al., 2011) employs the use of the users thumb in the image for size estimation.

Once a photo has been taken by the user with their thumb present, the system segments the food on the plate using shape, colour and texture detectors. The system then classifies the food type based on these features.

In this paper, it was decided to allow the users to change the prediction by the system. The thumb of each user is calibrated upon first use of the application so that size estimation can be as accurate a possible (Villalobos, Almaghrabi, Hariri, et al., 2011).

Toward Dietary Assessment via Mobile Phone Video Camera

Another study into using computer vision for dietary assessment was carried out by (N. Chen et al., 2010). They had a unique medium for the topic by using a video of the dishes in question and extracting frames from these videos to get the food from different angles.

(N. Chen et al., 2010) then formed a region of interest in the image, where there were the most food items and extracted colour and image features. These image features were extracted using Maximally Stable Extremal Regions (MSER), Speeded Up Robust Features (SURF) and Star detector.

This research team also uses k-means clustering to build a bag-of-words model (N. Chen et al., 2010).

The system had results as seen below across 20 categories using five images out of each video taken of the food:

- MSER - 95%
- SURF - 90%
- STAR - 90%

Automatic Chinese Food Identification and Quantity Estimation

There was a study carried out on food identification through a smart phone application (M.-Y. Chen et al., 2012). This study resulted in an application that allows a user to send an image of their food to a server which can give them an automatic response in 12 seconds (M.-Y. Chen et al., 2012). This back end service can have 34 threads working concurrently as stated at the time the paper was published (M.-Y. Chen et al., 2012).

A SVM is used to classify the image across 50 categories trained on around 100 images each. The SVM uses SIFT and Local Binary pattern feature extractors (M.-Y. Chen et al., 2012). A separate SVM was trained for each of these extractors and was merged together using a "Multi-class AdaBoost algorithm" (M.-Y. Chen et al., 2012).

The study produced a top-1 accuracy of 68.3%. Accuracy of 80.6%, 84.8% and 90.9% were recorded using top-2, top-3 and top-5 accuracy respectively (M.-Y. Chen et al., 2012).

An Image Processing Approach for Calorie Intake Measurement

(Villalobos, Almaghrabi, Pouladzadeh, et al., 2012) researched the question of using a computer vision approach to this topic. They focus mostly on the segmentation and region of interest calculation of the system in their study.

The system in question requires two images of the food, one from above and one from the side. This helps with size estimation. The users thumb is required to be in the image for accurate size estimation. The application also requires an image after consumption as to not calculate calories for uneaten food.

The system segments the image and then extracts colour, size and shape information from each segment. This data is then used by a SVM for classification along with a nutritional database for calorie information.

Multiple segmentation methods were tested such as:

- Semi automatic contour definition

- Watershed transformation
- Colour rasterization
- Edge accentication

The first two were dismissed due to poor results but the second two were used in conjunction for the segmentation aspect of the system.

Food Recognition and Nutrition Estimation on a Smartphone

An application called "Snap-n-Eat" was proposed by (Zhang et al., 2015).

When an image is taken using this application,the system finds saliency regions to remove the background of the image. If the image has multiple food types present, hierarchical segmentation takes place before proceeding to a SVM. Similar segments are merged together. These are found by using colour, texture and size.

SIFT and Histogram of Oriented Gradients (HOG) feature extractors are used on the image and these features are used by the SVM for classification. The SVM is trained using Scholastic Gradient Descent. (Zhang et al., 2015) also uses a Bog of Visual Words model along with k-means clustering.

An accuracy of 85% on 15 classes was recorded using this method (Zhang et al., 2015).

Merging dietary assessment with the adolescent lifestyle

(Schap et al., 2014) proposed a system used by smart phones which sends an image of a users food to a back end system for computation.

Once this has been completed, the image is segmented, features are extracted from each segment and these segments are classified. Colour and texture features are used for classification. The user has the ability to confirm or amend predictions of the food type.

Table 2.3: Summary of accuracy in dietary assessment methods

Title	**Classes**	**Accuracy**
Food Image Recognition with Multiple Kernel Learning	50	61.3%
A Novel SVM Based Food Recognition Method	12	92.6%
Measuring Calorie and Nutrition From Food Image	15	90.4%
Segmentation Assisted Food Classification	N/A	77.4%
Large Scale Leaning for Food Image Classification	11	78.0%
Toward Dietary Assessment via Mobile Phone Video Camera	20	92.0%
Automatic Chinese Food Identification and Quantity Estimation	50	68.3%
Food Recognition and Nutrition Estimation on a Smartphone	15	85.0%

Size estimation is also an important aspect of this system. In contrast to previous studies, (Zhang et al., 2015) uses food type shape and then those shape's geometric properties to estimate size.

This study produced results of 94% out of 32 test cases.

Other Methods

Methods outside of CNNs and SVMs are outlined below. A summary of the results can be seen in Table 2.3 along with SVM method results.

Large Scale Learning for Food Image Classification

(Abbirami.R.S et al., 2015) proposed a food image recognition system using a Bag of Features model. This study used over 5000 images separated into 11 classes.

A clustering algorithm was employed on this study before classification. For the classification step, experiments were carried out using different methods:

- SVM
- ANN

- Random Forests

The final accuracy of the system was 78% (Abbirami.R.S et al., 2015).

A Personal Assistive System for Nutrient Intake Monitoring

Similar to other approaches seen thus far, (Villalobos, Almaghrabi, Hariri, et al., 2011) employs the use of the users thumb in the image for size estimation.

Once a photo has been taken by the user with their thumb present, the system segments the food on the plate using shape, colour and texture detectors. The system then classifies the food type based on these features.

In this paper, it was decided to allow the users to change the prediction by the system. The thumb of each user is calibrated upon first use of the application so that size estimation can be as accurate a possible (Villalobos, Almaghrabi, Hariri, et al., 2011).

Toward Dietary Assessment via Mobile Phone Video Camera

Another study into using computer vision for dietary assessment was carried out by (N. Chen et al., 2010). They had a unique medium for the topic by using a video of the dishes in question and extracting frames from these videos to get the food from different angles.

(N. Chen et al., 2010) then formed a region of interest in the image, where there were the most food items and extracted colour and image features. These image features were extracted using Maximally Stable Extremal Regions (MSER), Speeded Up Robust Features (SURF) and Star detector.

This research team also uses k-means clustering to build a bag-of-words model (N. Chen et al., 2010).

The system had results as seen below across 20 categories using five images out of each video taken of the food:

- MSER - 95%

- SURF - 90%
- STAR - 90%

Merging dietary assessment with the adolescent lifestyle

(Schap et al., 2014) proposed a system used by smart phones which sends an image of a users food to a back end system for computation.

Once this has been completed, the image is segmented, features are extracted from each segment and these segments are classified. Colour and texture features are used for classification. The user has the ability to confirm or amend predictions of the food type.

Size estimation is also an important aspect of this system. In contrast to previous studies, (Zhang et al., 2015) uses food type shape and then those shape's geometric properties to estimate size.

This study produced results of 94% out of 32 test cases.

Possible Applications in Region Based CNNs

Ross Girshik and other contributers had some very positive results in the area of object detection using region based convolutional neural networks. There were four iterations of papers based on this work by Ross and groups in UC Berkley, Mircosoft and Facebook. A PHD student at the time of Ross's first paper also completed his dissertation on the subject. These papers, their results (Table 2.4) and the changes made through each iteration will be analysed throughly in the proceeding sections.

RCNN

In the first paper written by Ross Girshik, while researching at UC Berkeley, focused on two main insights. These were that "one can apply high-capacity convolutional neural networks (CNNs) to bottom-up region proposals in order

to localize and segment objects" and that "when training data is scarce, supervised pre-training for n auxiliary task, followed by domain-specific fine-tuning, yields a significant performance boost" (Girshick et al., 2014).

The system that they developed followed these steps:

- Take image as input
- Extract approximately 2000 region proposals from the image
- Compute fixed length vectors of features for the regions using a convolutional neural network
- Use a Support Vector Machine (SVM) to classify these regions
- Bounding box regression for final region proposals

This system utilised selective search to gather these region proposals but they mention that a sliding-window detector is also an option. Ross Girshik and his team used the open source Caffe CNN library for this system. The system is quite efficient and scalable. It is scalable because of the fixed length vector of features which will remain constant regardless of inputs and additional outputs.

The team evaluated their results on a few metrics and test sets as seen in Table 2.4.

Fast RCNN

Ross Girshik's next iteration of work on region based convolution neural networks took place in Microsoft Research. This paper was titled "Fast R-CNN" as it's aim was to decrease training and testing time "while also increasing detection accuracy" (Girshick, 2015).

This paper analyses why RCNN (Girshick et al., 2014) was slow and therefore how it could be improved. RCNN was classified to be slow because of three main factors:

- There are multiple stages to training as both a CNN and a SVM need to be trained.
- In training of the SVM, each region proposal must be written to disk and is therefore expensive.
- Object detection takes 47 seconds per image (Girshick, 2015).

Due to these problems with RCNN, a new algorithm, titled Fast RCNN was proposed. The architecture is as follows. An image is taken as input along with a proposals for regions. The image is pushed through convolutional and pooling layers (using max pooling). A fixed-length vector of features is then extracted from each region proposal. These vectors are inputted to fully connected layers for bounding box location prediction (Girshick, 2015).

At detection time, a pass through of the net is all that is needed so this runtime is significantly less than RCNN.

Faster RCNN

Due to the success of RCNN and Fast RCNN, Faster RCNN was introduced to combat the problem of region proposal computation (Ren et al., 2015).

The architecture for this system comprises of two modules. These consist of a convolutional neural network for region proposals (RPN) which the feeds into a Fast RCNN detector. These combine to produce a single neural network for object detection.

Instead of training these networks separately, the team had to look at how to share layers between the two networks. There were three option available:

- Alternating training whereby RPN is trained, and then used to train Fast RCNN. The Fast RCNN network is then used to initialise RPN and the process is iterated (Ren et al., 2015). This paper follows this approach.
- Approximate joint training.
- Non- approximate joint training.

Table 2.4: Results from Region Based CNN Research

	VOC07	**VOC10**	**VOC11**	**VOC12**	**COCO15**	**COCO16**
RCNN	58.5%	53.7%	47.9%	N/A	N/A	N/A
Fast RCNN	70.0%	68.8%	N/A	68.4%	N/A	N/A
Faster RCNN	78.8%	N/A	N/A	75.9%	42.7%	N/A
Mask RCNN	N/A	N/A	N/A	N/A	N/A	63.1%

Mask RCNN

The most recent paper on this topic was also written by Ross Girshik while working with Facebook AI Research (K. He et al., 2017). Mask RCNN "extends Faster RCNN by adding a branch for predicting an object mask in parallel with the existing branch for bounding box regression" (K. He et al., 2017).

Mask RCNN has two modules, similar to Faster RCNN, where the first module is the Region Proposal Network. In the second module, in parallel to classification, a binary mask is outputted for each region. Bounding box regression and classification are done in parallel.

2.7 APIs and Libraries

Tensorflow

Tensorflow is a deep learning software library for various machine learning paradigms. Tensorflow will be used to create neural networks. Tensorflow has two utilisations, through a Graphics Processing Unit (GPU) and also through a Central Processing Unit (CPU). GPU computation is recommended for CNN training.

Central Processing Unit Computation

It is quite easy to get tensorflow up and running if you are only using a CPU to train. Tensorflow CPU has been successfully installed on both Windows and Ubuntu, for the purpose of this project. For Windows you can download and install using the tensorflow website and on ubuntu you can using apt-get. Once installed, tensorflow can be imported into any python shell or script for use. Tensorflow can also be used in C++. There will be various python implementations of neural networks in Chapter 3.

Graphics Processing Unit Computation

For use with a GPU, the set up for tensorflow is a bit more complicated. Firtsly you use check that the GPU in your machine is compatible for CUDA 8.0 using the NVIDIA website. If your GPU is compatible, you must install CUDA after signing up as an NVIDIA developer. CUDA 8.0 is compatible with tensorflow. You also need to install cudnn6. The NVIDIA website contains tutorials to install these. Once these are installed, download and install tensorflow-gpu. This can imported into python similar to CPU computation.

Caffe

OpenCV

OpenCV is an industry wide, open source library for computer vision and machine learning (*About - OpenCV library* n.d.). It has over 2500 algorithms that are available for use (*About - OpenCV library* n.d.). OpenCV is supported across multiple languages and platforms such as Python, C++, C, Java, Matlab, running on Windows, Android, Mac OS and Linux (*About - OpenCV library* n.d.).

There is not much of the library utilised in this project due to the nature of Tensorflow but some algorithms for image reading, writing and resizing were used due to the ease of use.

```
image = cv2.imread('image.jpg')

resized = cv2.resize(image, (299, 299))

cv2.imwrite('imageResized.jpg', resized)
```

Numpy

Jupyter

2.8 Conclusion

Chapter 3

Introduction to Using Tensorflow

The purpose of this chapter is to give an overview of what has been completed in order to gain understanding of how to create CNNs and also how to work with datasets. Three experiments are explored in this chapter, the first two of which are tutorials completed as an introduction to Tensorflow and CNNs. The final experiment explores how to feed datasets from file directories into CNNs.

3.1 Template for Experiments

The template followed for all experiments in chapters three, four and five is outlined in Figure 3.1.

3.2 Udemy Tutorial

Overview

There are many on line resources that are geared towards helping deep learning novices. One of these resources is a course on Udemy titled 'Complete Guide to

Table 3.1: Experiment Template

Section	**Rationale**
Overview	An explanation of the purpose of the experiment along with how it is carried out.
Network Architecture	An explanation of the network architecture used in the experiment. If the architecture has been explained in another experiment, the architecture will only be referenced by name.
Dataset	The dataset used for the experiment, with its source and an overview of its details. A brief mention will be made in following experiments.
API's	Reference to the technologies used as outlined in Chapter 2.
Script	Snippets of the script used for the experiment.
Results	The results acquired from the experiment, usually in the form of a percentage accuracy.
Empirical Analysis	States any information gained from the experiment and speculation for the reasoning of the results.

Tensorflow for Deep Learning with Python' (*Complete Guide to Tensorflow for Deep Learning with Python* n.d.). This course has a section on Convolutional Neural Networks which has been followed and completed.

Network Architecture

The architecture for this network is very simple as it is an introductory CNN. It consists of two convolutional layers, two max pooling layers and a fully connected layer.

Dataset

This CNN is trained on the MNIST dataset. This dataset consists of 70000 handwritten digits (LeCun and Cortes, 2010).

API's

This experiment was carried out in a jupyter notebook using tensorflow.

Script

```
#Helper Functions

#INIT WEIGHTS
def init_weights(shape):
    init_random_dist = tf.truncated_normal(shape, stddev = 0.1)
    return tf.Variable(init_random_dist)

#INIT BIAS
def init_bias(shape):
    init_bias_vals = tf.constant(0.1, shape = shape)
    return tf.Variable( init_bias_vals )
```

```
12
13 #CONV2D
14 def conv2d(x, W):
15     #x -> [batch, H, W, Channels]
16     #W -> [filterH, filterW, ChannelsIn, ChannelsOut]
17     return tf.nn.conv2d(x, W, strides = [1,1,1,1], padding = 'SAME')
18
19 #POOLING
20 def max_pool_2by2(x):
21     #x -> [batch, H, W, Channels]
22     return tf.nn.max_pool(x, ksize = [1,2,2,1], strides = [1,2,2,1],
           padding = 'SAME')
23
24     #NORMAL (FULLY CONNECTED)
25 def normal_full_layer(input_layer, size):
26     input_size = int(input_layer.get_shape()[1])
27     W = init_weights([input_size, size])
28     b = init_bias([size])
29 return tf.matmul(input_layer, W) + b
```

```
1 #32 features for every 5 x 5 patch with 1(grayscale)
2 convo_1 = convolutional_layer(x_image, shape = [5,5,1,32])
3 convo_1_pooling = max_pool_2by2(convo_1)
4
5 convo_2 = convolutional_layer(convo_1_pooling, shape = [5,5,32,64])
6 convo_2_pooling = max_pool_2by2(convo_2)
7
8 convo_2_flat = tf.reshape(convo_2_pooling, [-1,7*7*64])
9 full_layer_one = tf.nn.relu(normal_full_layer(convo_2_flat, 1024))
```

```
1 with tf.Session() as sess:
2     sess.run(init)
3
```

```
    for i in range(steps):
        batch_x, batch_y = mnist.train.next_batch(32)
        batch_test = mnist.test.next_batch(32)
        sess.run(train, feed_dict = {x:batch_x, y_true:batch_y,
            hold_prob:0.5})
        if i%500 == 0:
            print("ON STEP: {}".format(i))
            print("ACCURACY: ")
            match = tf.equal(tf.argmax(y_pred, 1), tf.argmax(y_true, 1))
            acc = tf.reduce_mean(tf.cast(match, tf.float32))
            print(sess.run(acc, feed_dict = {x:batch_test,
                y_true:mnist.test.labels, hold_prob:1.0}))
            print('\n')

    sess.run(acc, feed_dict = {x:mnist.test.images} )

#Final Accuracy 0.973
```

Results

The Final Accuracy for this experiment was of 97.3%.

Empirical Analysis

3.3 Udemy Tutorial 2

Overview

Similar to the previous experiment 3.3, this experiment is a Udemy course exercise (*Complete Guide to Tensorflow for Deep Learning with Python* n.d.). In contrast, this does not use a dataset built into Tensorflow so therefore, there is extra configuration to be done on the dataset.

Network Architecture

Same architecture as in 3.2.

Dataset

The CIFAR-10 dataset is used here. This has 60000 images split into 10 classes and a test set of 1000 images (Krizhevsky, Nair, and G. Hinton, n.d.).

Script

```
def unpickle( file ):
    import pickle
    with open( file , 'rb') as fo:
        cifar_dict = pickle.load(fo, encoding='bytes')
    return cifar_dict

dirs = ['batches.meta','data_batch_1','data_batch_2','data_batch_3','data_batch_4','data_batc

all_data = [0,1,2,3,4,5,6]

for i,direc in zip(all_data,dirs):
    all_data[i] = unpickle(CIFAR_DIR+direc)

batch_meta = all_data[0]
data_batch1 = all_data[1]
data_batch2 = all_data[2]
data_batch3 = all_data[3]
data_batch4 = all_data[4]
data_batch5 = all_data[5]
test_batch = all_data[6]
```

```
def set_up_images(self):

        print("Setting Up Training Images and Labels")

        # Vertically stacks the training images
        self.training_images = np.vstack([d[b"data"] for d in
            self.all_train_batches])
        train_len = len(self.training_images)

        # Reshapes and normalizes training images
        self.training_images =
            self.training_images.reshape(train_len,3,32,32).transpose(0,2,3,1)/255
        # One hot Encodes the training labels (e.g. [0,0,0,1,0,0,0,0,0,0])
        self.training_labels = one_hot_encode(np.hstack([d[b"labels"] for
            d in self.all_train_batches]), 10)

        print("Setting Up Test Images and Labels")

        # Vertically stacks the test images
        self.test_images = np.vstack([d[b"data"] for d in self.test_batch])
        test_len = len(self.test_images)

        # Reshapes and normalizes test images
        self.test_images =
            self.test_images.reshape(test_len,3,32,32).transpose(0,2,3,1)/255
        # One hot Encodes the test labels (e.g. [0,0,0,1,0,0,0,0,0,0])
        self.test_labels = one_hot_encode(np.hstack([d[b"labels"] for d in
            self.test_batch]), 10)


    def next_batch(self, batch_size):
        # Note that the 100 dimension in the reshape call is set by an
```

```
            assumed batch size of 100
        x =
            self.training_images[self.i:self.i+batch_size].reshape(100,32,32,3)
        y = self.training_labels[self.i:self.i+batch_size]
        self.i = (self.i + batch_size) % len(self.training_images)
    return x, y

cross_entropy =
    tf.reduce_mean(tf.nn.softmax_cross_entropy_with_logits(labels =
    y_true, logits = y_pred))
optimizer = tf.train.AdamOptimizer(learning_rate = 0.001)
train = optimizer.minimize(cross_entropy)
init = tf.global_variables_initializer()
with tf.Session() as sess:
    sess.run(tf.global_variables_initializer())

    for i in range(10000):
        batch = ch.next_batch(100)
        sess.run(train, feed_dict={x: batch[0], y_true: batch[1],
            hold_prob: 0.5})

        # PRINT OUT A MESSAGE EVERY 100 STEPS
        if i%1000 == 0:

            # Test the Train Model
            matches = tf.equal(tf.argmax(y_pred,1),tf.argmax(y_true,1))

            acc = tf.reduce_mean(tf.cast(matches,tf.float32))
            testSet = ch.next_test_batch(100)
            print('Accuracy: ')
            print(sess.run(acc, feed_dict={x:testSet[0]
                ,y_true:testSet[1], hold_prob:1.0}))
```

```
print('\n')
```

Results

A Final Accuracy of 71% was reached.

Empirical Analysis

3.4 Using the Food 101 dataset

Overview

For the next experiment, it was decided to use the Food-101 dataset. An on line tutorial was used to create a dataset in tensorflow using image directories on disk (*Build an Image Dataset in Tensorflow* n.d.). Tutorials used previously were also utilized for this experiment (*Complete Guide to Tensorflow for Deep Learning with Python* n.d.) and (Krizhevsky, Nair, and G. Hinton, n.d.).

Network Architecture

A similar network architecture to the previous two experiments (3.2 and 3.3) was used here.

Dataset

The Food-101 datset was used for this experiment (Bossard, Guillaumin, and Van Gool, 2014).

Script

```
# Reading the dataset
# 2 modes: 'file' or 'folder'
```

```
3  def read_images(dataset_path, mode, batch_size):
4      imagepaths, labels = list(), list()
5      if mode == 'file':
6          # Read dataset file
7          data = open(dataset_path, 'r').read().splitlines()
8          for d in data:
9              imagepaths.append(d.split(' ')[0])
10             labels.append(int(d.split(' ')[1]))
11     elif mode == 'folder':
12         # An ID will be affected to each sub-folders by alphabetical order
13         label = 0
14         # List the directory
15         try:  # Python 2
16             classes = sorted(os.walk(dataset_path).next()[1])
17         except Exception:  # Python 3
18             classes = sorted(os.walk(dataset_path).__next__()[1])
19         # List each sub-directory (the classes)
20         for c in classes:
21             c_dir = os.path.join(dataset_path, c)
22             try:  # Python 2
23                 walk = os.walk(c_dir).next()
24             except Exception:  # Python 3
25                 walk = os.walk(c_dir).__next__()
26             # Add each image to the training set
27             for sample in walk[2]:
28                 # Only keeps jpeg images
29                 if sample.endswith('.jpg') or sample.endswith('.jpeg'):
30                     imagepaths.append(os.path.join(c_dir, sample))
31                     labels.append(label)
32             label += 1
33     else:
34 raise Exception("Unknown mode.")
```

```
# Convert to Tensor
    imagepaths = tf.convert_to_tensor(imagepaths, dtype=tf.string)
    labels = tf.convert_to_tensor(labels, dtype=tf.int32)
    # Build a TF Queue, shuffle data
    image, label = tf.train.slice_input_producer([imagepaths, labels],
                                                 shuffle=True)

    # Read images from disk
    image = tf.read_file(image)
    image = tf.image.decode_jpeg(image, channels=CHANNELS)

    # Resize images to a common size
    image = tf.image.resize_images(image, [IMG_HEIGHT, IMG_WIDTH])

    # Normalize
    image = image * 1.0/127.5 - 1.0

    # Create batches
    X, Y = tf.train.batch([image, label], batch_size=batch_size,
                          capacity=batch_size * 8,
                          num_threads=4)

    return X, Y
```

```
# Start training
with tf.Session() as sess:

    # Run the initializer
    sess.run(init)

    # Start the data queue
    tf.train.start_queue_runners()
```

```
9
10      # Training cycle
11      for step in range(1, num_steps+1):
12
13          if step % display_step == 0:
14              # Run optimization and calculate batch loss and accuracy
15              _, loss, acc = sess.run([train_op, loss_op, accuracy])
16              print("Step " + str(step) + ", Minibatch Loss= " + \
17                    "{:.4f}".format(loss) + ", Training Accuracy= " + \
18                    "{:.3f}".format(acc))
19          else:
20              # Only run the optimization op (backprop)
21              sess.run(train_op)
22
23  print("Optimization Finished!")
```

Results

The final accuracy for this model was 18.8% after 5000 steps.

Empirical Analysis

The poor accuracy of this model is to be expected due to the simple architecture of the network and the fact that 101 classes is a lot to process.

Chapter 4

Training Using the Inception-V3 Model Architecture

This chapter consists of experiments which train various Tensorflow models. There are parameters that can be changed in the code that creates a Tensorflow model based on the Inception-V3 model architecture. These parameters were changed in order to see how the model accuracy would be affected. In addition to this, change in dataset size was explored in relation to model accuracy.

4.1 Retrain ImageNet Inception V3 Model

Overview

For the fourth experiment undertaken, it was decided to take inspiration from (Yanai and Kawano, 2015), where pre-training was used training a model for food classification. In order to achieve this, the final layer of the Inception V3 model which was trained on the ImageNet dataset had to be retrained. This is called Transfer Learning. A tutorial, created by Google, on the tensorflow website was followed for direction on this process (*How to Retrain Inception's Final Layer for New Categories* n.d.).

Firstly, in order to retrain the final layer of a model, a dataset must be prepared

in the correct way. The Food-101 dataset (Bossard, Guillaumin, and Van Gool, 2014) was used for this experiment, which will be analysed below. This dataset is in contrast to (Yanai and Kawano, 2015) as they used the UECFOOD100 dataset. The food-101 dataset was chosen due to larger number of images per class, a difference between 100 and 1000. The dataset must be structured so that there is a separated directory for each class with the directory name as the class name. These directories should contain all the images for this class.

Once this dataset has been set up correctly, a directory can be found on github which contains the necessary files for this tutorial. When the directory has been downloaded, the following command can be ran:

```
python tensorflow/examples/image_retraining/retrain.py \ --image_di
~/dataset_directory
```

The first thing that the script will do is create bottleneck files for the images. A bottleneck is a term used to define the final layer before the output layer. This is so that for each image, we do not have to push it through the entire network during training (*How to Retrain Inception's Final Layer for New Categories* n.d.).

After, the bottlenecks are created, the training can be completed. The images are split into three sub directories of training, testing and validation. By default, these images are split into percentages of 80%, 10% and 10% respectively. The model is trained at a default of 4000 steps.

At the final stage of the script, the model is run on a batch of test images not yet seen and a final test accuracy is displayed. This can be seen in the Script section below.

The command used for using this model once it is trained is:

```
python tensorflow/examples/label_image.py --graph=/tmp/output_graph
--labels=/tmp/output_labels.txt --input_layer=Mul --output_layer=fi
--input_mean=128 --input_std=128 --image=~/image_directory
```

Network Architecture

Dataset

The dataset used for this experiment is the Food-101 dataset (**food 101**). This dataset has 101 classes with 1000 images for each class.

Libraries

Tensorflow and Numpy were used to run this scipt.

Script

The following snippets of code are from the retrain.py script.

```
# Add the new layer that we'll be training.
(train_step, cross_entropy, bottleneck_input, ground_truth_input,
final_tensor) = add_final_training_ops(
        len(image_lists.keys()), FLAGS.final_tensor_name,
        bottleneck_tensor,
        model_info['bottleneck_tensor_size'],
        model_info['quantize_layer'])

# Create the operations we need to evaluate the accuracy of our new layer.
evaluation_step, prediction = add_evaluation_step(
final_tensor, ground_truth_input)

# Merge all the summaries and write them out to the summaries_dir
merged = tf.summary.merge_all()
train_writer = tf.summary.FileWriter(FLAGS.summaries_dir + '/train',
                                     sess.graph)

validation_writer = tf.summary.FileWriter(
```

```
      FLAGS.summaries_dir + '/validation')

# Set up all our weights to their initial default values.
init = tf.global_variables_initializer()
sess.run(init)

# We've completed all our training, so run a final test evaluation on
# some new images we haven't used before.
test_bottlenecks, test_ground_truth, test_filenames = (
    get_random_cached_bottlenecks(
        sess, image_lists, FLAGS.test_batch_size, 'testing',
        FLAGS.bottleneck_dir, FLAGS.image_dir, jpeg_data_tensor,
        decoded_image_tensor, resized_image_tensor, bottleneck_tensor,
        FLAGS.architecture))
test_accuracy, predictions = sess.run(
    [evaluation_step, prediction],
    feed_dict={bottleneck_input: test_bottlenecks,
        ground_truth_input: test_ground_truth})
tf.logging.info('Final test accuracy = %.1f%% (N=%d)' %
                (test_accuracy * 100, len(test_bottlenecks)))
```

Results

The final test accuracy for this retrained model was 54.8%.

For example, an image of pizza Figure4.1, was fed into the model with the followng results:

- pizza 0.925
- pancakes 0.008
- nachos 0.007
- beef carpaccio 0.006

Figure 4.1: Pizza - sourced from https://www.cicis.com/menu/pizza/bbq-pork

- tiramisu 0.004

In contrast, Figure 4.2 was not classified as a pizza.

Empirical Analysis

These poor results are not that surprising. This is because we have many classes to train for, 101, and no parameter tuning has been carried out on the running of this code.

Figure 4.2 was not classified correctly, this is most likely due to the fact that the pizza does not take up much of the image.

4.2 Retrain with Extended Dataset

Overview

As the Food-101 dataset mostly consisted of meals (Bossard, Guillaumin, and Van Gool, 2014), it was decided to extend the dataset slightly by including some single foods such as:

- cheese

Figure 4.2: Pizza not classified correctly by the model

- grapes
- banana
- apple
- orange
- spaghetti
- roll

In order to collect these images, the ImageNet repository was utilised to search for these foods individually and then download the subset of images to be included with the Food 101 dataset (Deng et al., 2009). The retrain.py script was run on the extended dataset as in 4.1.

Network Architecture

The Inception V3 Model was used for this experiment.

Dataset

In this experiment the extended Food-101 dataset was used.

Libraries

Tensorflow and numpy are used in the retrain.py script.

Script

As seen in 4.2.

Results

For this model, an accuracy of 55.3% was achieved.

For example, an image of a banana (Figure 4.3) was fed into the model with the followng results:

- banana 0.9962
- orange 0.0009
- cheese 0.0003
- frozen yoghurt carpaccio 0.0002
- churros 0.0001

Empirical Analysis

The slight increase in accuracy in 4.1, from 54.8% to 55.3%, makes sense. Since the model was pre-trained using the ImageNet dataset and all the new images used were from ImageNet, we would expect a higher classification accuracy on the new additions to the dataset. This would overall increase the average classification accuracy.

Figure 4.3: Banana - sourced from http://www.ciaoimports.com/

4.3 Retrain with Parameter Tuning

Overview

Within the retrain.py script (*How to Retrain Inception's Final Layer for New Categories* n.d.), as mentioned in previous experiments, there are various parameters that can be set and changed. Various combinations of these parameters were changed to see if it would increase the test accuracy of the model.

Network Architecture

The Inception V3 architecture was used for this model.

Dataset

The Food-101 dataset (Bossard, Guillaumin, and Van Gool, 2014) with additional classes, as per 4.1, was used for this model.

Libraries

The libraries in use for this experiment are tensorflow and numpy.

Script

Script as seen in 4.1 but with some additions to calculate Top 5 accuracy as seen below:

```
python retrain_top5.py \ --image_dir
~/dataset_directory \ --how_many_training_steps 4000 \ --learning_ra
--testing_percentage 10 \ --validation_percentage 10
```

Top 5 accuracy of the model was als calcultaed using the code below.

```
print('Top 5 Evaluation')
#Variables used to store a list of all classes, the amount of images tested,
# the amount of images with a top 5 accuracy and the sum of the highest probabilities
classes = list(image_lists.keys())
image_count = 0
top_5_count=0
total_of_top1_probs = 0
i = 0
class_counter = 0

#Loops through all test images, selecting 10 images per class
#and running them through the method in 'label_image.py'.
while(i < len(test_bottlenecks)):
    class_counter += 1

    if class_counter > 10:
        i = int(round((i + len(test_bottlenecks)/len(classes)) - 10))
        class_counter = 0
    else:
        image_count += 1
        results_from_classifier = label_image.runModel(test_filenames[i])
        results = results_from_classifier[0]
```

```
23        probabilities = results_from_classifier[1]
24
25        if classes[test_ground_truth[i]] in results:
26          top_5_count += 1
27        else:
28          print("Expected: " + classes[test_ground_truth[i]])
29          for result in results:
30            print("Classes: " + result)
31          print("")
32
33        total_of_top1_probs += max(probabilities)
34        i = i + 1
35  print(str(top_5_count))
36  average_probabilities = total_of_top1_probs/(len(classes*10))
37
38  #Prints out the amount of test images used, the top 5 accuracy
39  # and the average probability of predictions.
40  print("Amount of test images: " + str(image_count))
41  print("Top 5 Accuracy: " + str((top_5_count/image_count)*100))
42  print("Average probability: " + str(average_probabilities))
```

Method from label_image.py.

```
1  def runModel(file_name):
2    model_file = \
3      "/tmp/output_graph.pb"
4     label_file = "/tmp/output_labels.txt"
5    input_height = 299
6    input_width = 299
7    input_mean = 128
8    input_std = 128
9    input_layer = "Mul"
10   output_layer = "final_result"
```

```
11
12    graph = load_graph(model_file)
13    t = read_tensor_from_image_file(file_name,
14                                    input_height=input_height,
15                                    input_width=input_width,
16                                    input_mean=input_mean,
17                                    input_std=input_std)
18
19    input_name = "import/" + input_layer
20    output_name = "import/" + output_layer
21    input_operation = graph.get_operation_by_name(input_name)
22    output_operation = graph.get_operation_by_name(output_name)
23
24    with tf.Session(graph=graph) as sess:
25       results = sess.run(output_operation.outputs[0],
26                          {input_operation.outputs[0]: t})
27    results = np.squeeze(results)
28
29    top_k = results.argsort() [-5:][::-1]
30    labels = load_labels( label_file )
31
32    setIndex = False
33
34    top5_results = [None] * 5
35    index = 0
36    for i in top_k:
37       top5_results[index] = labels[i]
38       index += 1
39
40     final_results = [top5_results, results]
41    return final_results
```

Some further parameters could be set such as:

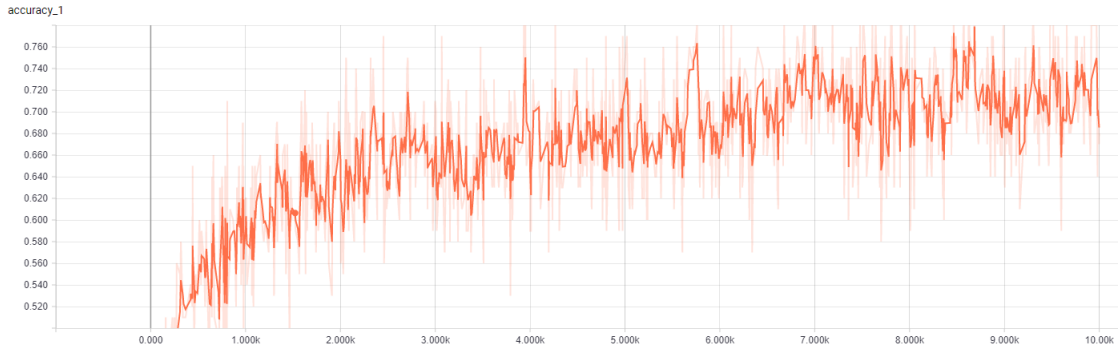

Figure 4.4: Graph of accuracy of the test dataset during training

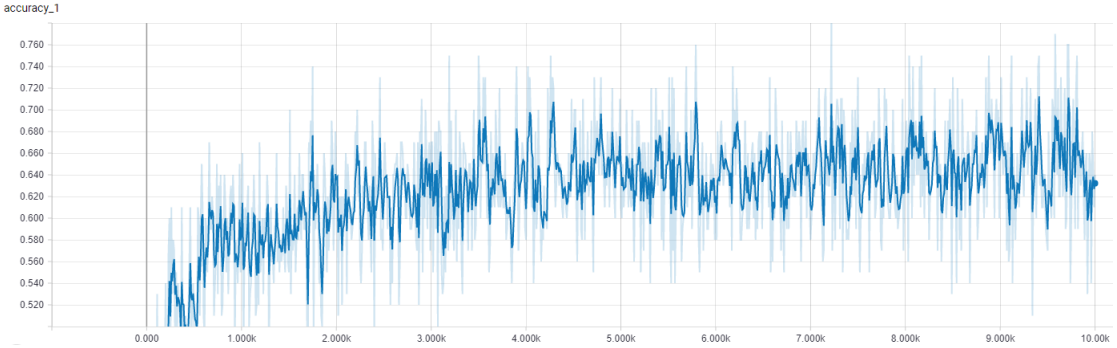

Figure 4.5: Graph of accuracy of the validation dataset during training

- –flip_left_right
- –random_crop
- –random_scale
- –random_brightness

Results

The results of each set of parameters can be seen in Table 4.1. Using the parameters of 10,000 steps, and a learning rate of 0.1 and therefore a final test accuracy of 66.3%, the model achieved a Top 5 accuracy of 85.96%. This figure was calculated from 1090 images in the test dataset and the average probability of the predictions were at 0.62.

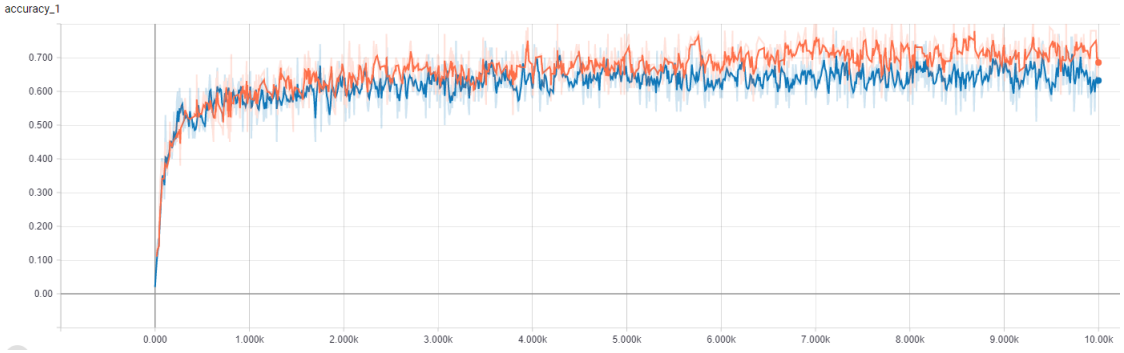

Figure 4.6: Comparison of accuracy

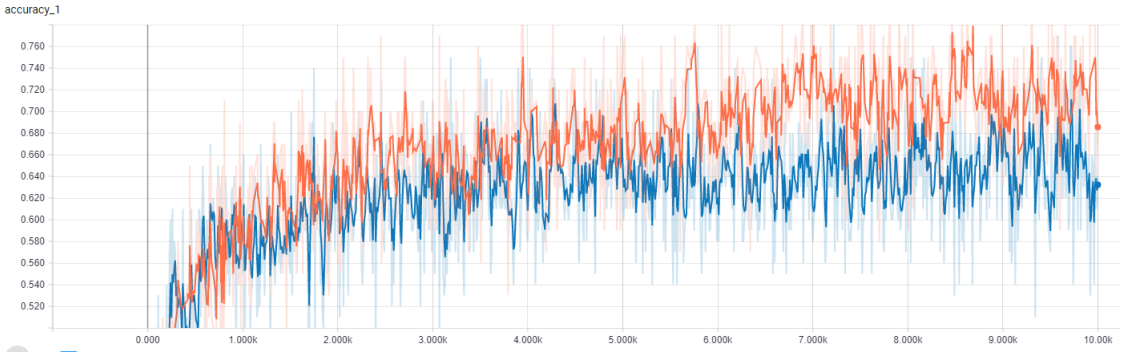

Figure 4.7: Comparison of accuracy

Table 4.1: Comparison of parameters

Parameter Tuning	**Steps**	**Learning Rate**	**Test %**	**Validation %**	**Results**
Configuration 1	8,000	0.01	10	10	59.1%
Configuration 2	8,000	0.10	10	10	65.8%
Configuration 3	10,000	0.10	10	10	66.3%
Configuration 4	12,000	0.10	10	10	66.6%
Configuration 5	10,000	0.20	10	10	66.0%
Configuration 6	10,000	0.10	15	15	66.3%

Empirical Analysis

The set of parameters that seem to be the most effective are 10000 steps with a 0.1 learning rate. Graphs of this model can be see in Figures 4.5 and 4.4. These are based on the validation set and then the test set respectively. A side by side comparison can also be seen in Figures 4.6 and 4.7 where orange is for during training and blue for the validation set.

There were two separate factors that each increased classification accuracy of about 5% each. These were training steps and learning rate.

Training steps are related to the number of images so before, when the training steps were at 4000, not all of our training images were being used. As the steps were increased twofold we saw a 3.8% increase in accuracy.

Another parameter that increased accuracy significantly was learning rate. The default learning rate is 0.01 which was increased to 0.1. This resulted in an increase of 6.7%. This is most likely due to the fact that since we are only looking at the last layer, we can afford to change the weights more significantly.

The Top 5 accuracy of the model was quite good but it was not run on the same number of images as the final test accuracy. This is due to the fact that tensorflow does not have an API for calculating Top 5 accuracy. As a result, it had to be calculated manually and this is very time intensive. The average probability of 0.62 is interestingly close to the overall Top 1 accuracy.

4.4 MobileNet

Overview

Due to the fact that the end goal for this project is to have a smartphone application that a user can use to keep track of their calorie measurement, there are a couple of options in how to achieve this. Firstly, an image can be taken on the phone and sent to a server to run a classification algorithm. Secondly, a model can be stored on the phone for computation. Transfer

learning was once again used but on a different, smaller architecture called MobileNet (Howard et al., 2017).

Network Architecture

Dataset

The Food 101 dataset (Bossard, Guillaumin, and Van Gool, 2014) with added classes was used for this experiment.

Libraries

Tensorflow and numpy.

Script

The retrain.py script (*How to Retrain Inception's Final Layer for New Categories* n.d.) was used, with a different command paramater.

```
python tensorflow/examples/image_retraining/retrain.py \ --image_di
~/dataset_directory \ --architecture mobilenet_1.0_224 \
--how_many_training_steps 10000 \ --learning_rate 0.1
```

Results

The final test accuracy of this model came to 50.2%.

Empirical Analysis

There was a decrease of 16.1% in this model to the highest accuracy from 4.4. This is due to the smaller architecture which is aimed to be faster and smaller with an expected decrease in accuracy.

4.5 Food 101 subset

Overview

In many of the papers that have been researched where food image classification was carried out, they attempted to classify a lot less than 108 food types as has been the case for experiments previously shown. (Pouladzadeh, Villalobos, et al., 2012) used 12 classes, (Pouladzadeh, Shirmohammadi, and Al-Maghrabi, 2014) had 15, (Abbirami.R.S et al., 2015) attempted to classify 11 classes of food, 20 classes were used in (N. Chen et al., 2010) and (Zhang et al., 2015) predicted 15 classes. Due to the lower number of classes in these papers, it was decided to retrain inception on a subset of the food-101 extended dataset to benchmark results. 13 classes were selected from food-101 for training.

Network Architecture

Retrained Inception model.

Dataset

A subset of the Food 101 dataset was used for this experiment (Bossard, Guillaumin, and Van Gool, 2014).

Results

A Final test Accuracy of 92.6% was recorded for this experiment which performs quite high in comparison to the data in Table 4.2. A Top 5 accuracy of 100% was calculated with an average prediction probability of 0.89 using 130 images. This was calculated using the script defined in 4.3.

Table 4.2: Accuracy of other studies

Reference	**Classes**	**Accuracy**
Novel SVM	12	92.6%
Measuring Calorie and Nutrition	15	90.4%
Large Scale Learning	11	78.0%
Toward Dietary Assessment	20	91.7%
Snap-n-eat	15	85.0%

Empirical Analysis

It would make sense the accuracy of our model would increase when the number of classes are reduced as the margin of error is decreased. The Top 5 accuracy of the model was very successful. There were only 10 images per class tested though, totaling at 130 images so if the number of images increased, this accuracy would probably decrease slightly. If the code is run again, the same results cannot even be replicated for sure. On another run of this code a Top 5 accuracy of 96% was recorded. While this is still a very good accuracy, it does not compare to the first run of the script.

Chapter 5

Analysing the Trained Model

The purpose of this chapter is to analyse the models trained previously to gain understanding and applicability to the problem statement.

5.1 Sliding Window

Overview

In the preious experiments, the one-shot approach to food image classification has been looked at. That is, the model will give a prediction of the most likely food item in that image. This is a problem when there is a composite image i.e are multiple foods in an image, see Figure 5.1. There are a few options to combat this problem. Firstly, one could detect objects in the image, segment the image according to these objects and then run each segment through the model. A simple approach to this would be to segment the image into a number of sections and then run each section through the model. In order to follow the latter approach, a sliding window approach was used. This sliding window would move across the image and classify the segment of the image in the window. There are three options for window sliding shape as defined by a command line argument.

```
python sliding_window.py --image=~/image_dir --window_shape grid
```

Figure 5.1: Bowl of fruit - sourced from https://www.geo.de/natur/

There are three options for window shape:

- Grid based window as per Figure 5.2
- Row based window as per Figure 5.3
- Column based window as per Figure 5.4

Libraries

For this experiment Tensorflow provided the classification of each segment while also helping with resizing along with Numpy. OpenCv was used to implement the sliding window as per (*Sliding Windows for Object Detection with Python and OpenCV* n.d.)

Script

There were four main elements to the script. Firstly, extracting a window to be classified. Secondly, resizing the image to be compatible with the Tensorflow

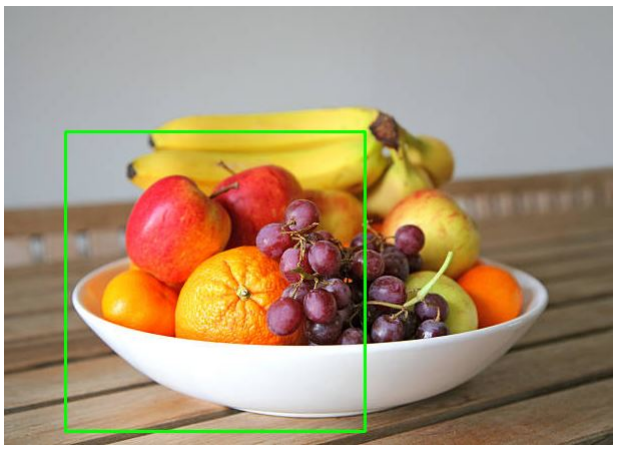

Figure 5.2: Grid based window

Figure 5.3: Row based window

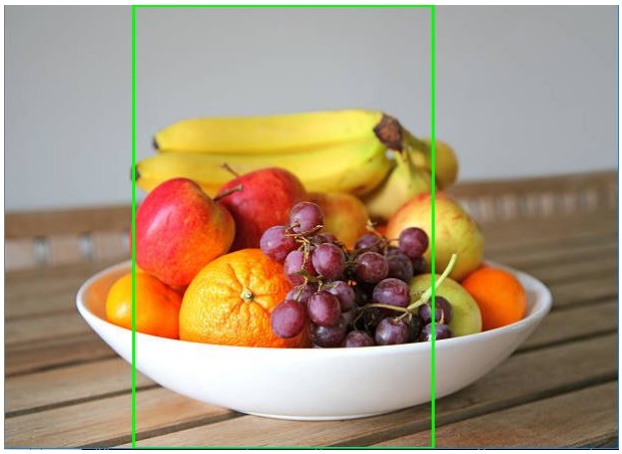

Figure 5.4: Column Based Window

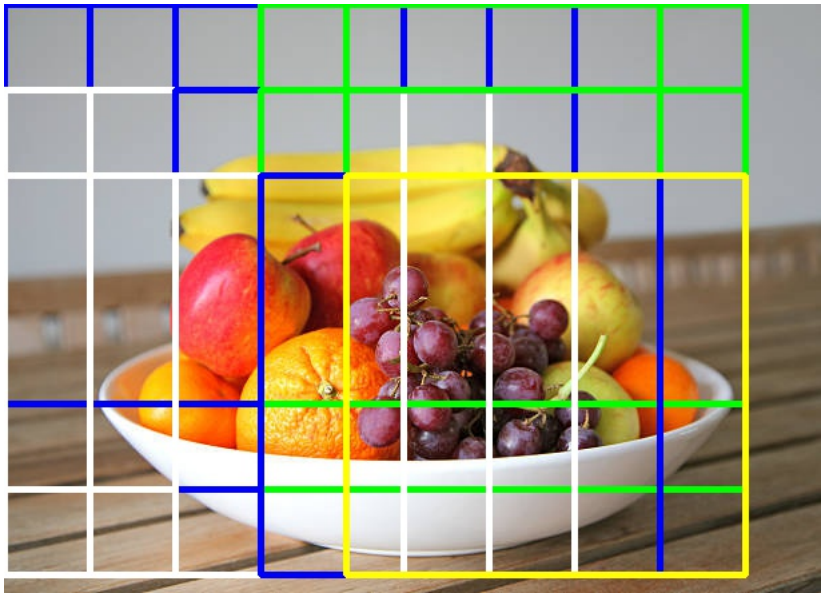

Figure 5.5: Fruit with Color Overlay

model. Thirdly, running the window through the Tensorflow model and finally saving a new image with a coloured overlay of classifications.

Extracting the window from the image

```
for (x, y, window) in sliding_window(resized, stepSize=64,
    windowSize=(winW, winH)):
    # if the window does not meet our desired window size, ignore it
    if window.shape[0] != winH or window.shape[1] != winW:
        continue
```

```
def sliding_window(image, stepSize, windowSize):
    # slide a window across the image
    for y in xrange(0, image.shape[0], stepSize):
        for x in xrange(0, image.shape[1], stepSize):
            # yield the current window
            yield (x, y, image[y:y + windowSize[1], x:x + windowSize[0]])
```

Resizing the window

```
window = cv2.resize(window, (299, 299))
```

```
resized_image = tf.reshape(image, [1, input_height, input_width, 3])
resized = tf.image.resize_area(resized_image, [input_height, input_width])
normalized = tf.divide(tf.subtract(resized, [input_mean]), [input_std])
```

Running the Tensorflow model

```
with tf.Session() as sess:
    numpy_image = sess.run(normalized)

with tf.Session(graph=graph) as sess:
     results = sess.run(output_operation.outputs[0],
                        {input_operation.outputs[0]: numpy_image})
    probabilities = np.squeeze(results)
```

Saving the image with colour overlay

As seen in Figure 5.5, each square represents a window and each colour is for a different classification. Blue is for an apple, yellow for banana, green for grape, white for orange and black if an unexpected prediction is made.

```
cv2.rectangle(display_image, (x, y), (x + winW, y + winH),
    colour_dict.get(top1,
(0,0,0)), 4)
```

Results

Grid based window

The grid based window resulted in fifteen separate classification. As seen in Figure 5.1, there are multiple fruits in the image. Of these fruits, our model

Table 5.1: Grid Based Sliding Window Results

Food type	**No. of Top-1 Classifications**
Apple	5
Banana	1
Grape	4
Orange	5

Table 5.2: Row Based Sliding Window Results

Food type	**No. of Top-1 Classifications**
Apple	1
Banana	0
Grape	0
Orange	0
Other	3

is trained on four, apple, banana, orange and grapes. This method classified all four to Top-1 accuracy at least once each. This method took 42.8 seconds to run.

Row based window

The row based method resulted in four predictions as follows in Figure 5.2. Out of these four predictions, only one classified a known fruit at Top-1 accuracy,

Table 5.3: Column Based Sliding Window Results

Food type	**No. of Top-1 Classifications**
Apple	3
Banana	1
Grape	0
Orange	0
Other	1

an apple. An apple was also predicted to Top-5 accuracy in another instance. The runtime of this method was 16.1 seconds.

Column based window

The column based window approach had five total predictions and ran for a total of 13 seconds. As seen in 5.3, two out of four known fruits were classified to a Top-1 accuracy with all other fruits predicted to Top-5 accuracy. Only one Top-1 prediction did not contain a correct fruit.

Empirical Analysis

These results are very interesting because while a banana was only predicted to Top-1 accuracy once in grid based, once in column based and zero times in row based, if the whole image is ran through the model, a banana is at the Top-1 accuracy.

5.2 Recursive Refinement

Overview

After the sliding window code was run on Figure 5.1 in 5.1, it was observed that a sliding window was predicting grapes correctly in regions that contained a bunch of grapes. Since it would make sense that the model would be able detect an individual grape, it was decided that recursive refinement would be ran on a window that contained a grape. Due to the model requiring a 299 x 299 image size, the window could only be refined once as very small segments could not be resized up to 299 x 299. A window of 70 x 70 size was used.

Script

As you would think with recursive refinement, a recursive function would used, but this was unnecessary due to image size restrictions. Instead, a conditional

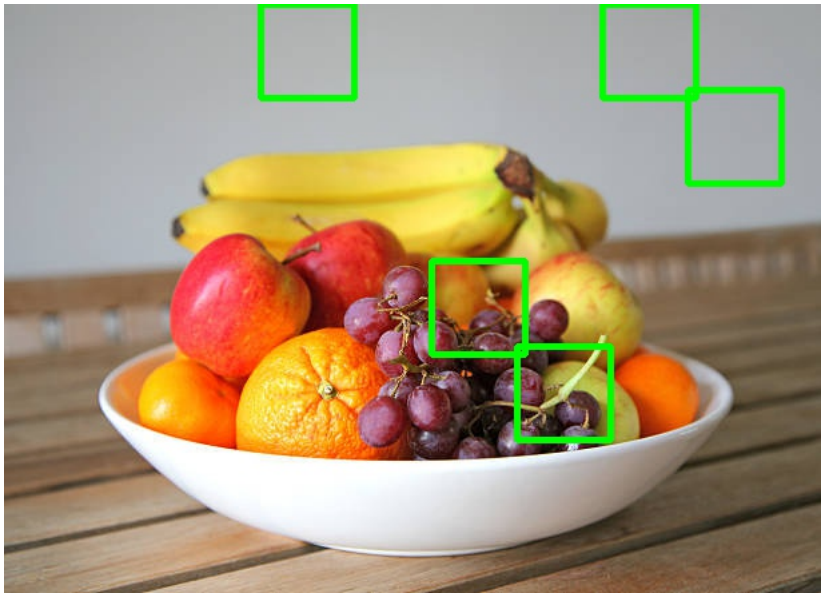

Figure 5.6: Recursive refinement 1

for loop was added to the existing code.

```
if top1 == "grape" and window_shape == "grid" and rr_grape:
        for (x_grape, y_grape, grape_window) in
            sliding_window(window_resized, stepSize=64, windowSize=(70,
            70)):
            #reshape to square
            grape_window_resized = cv2.resize(grape_window, (299, 299))
            top1_grape = subSample.classify(grape_window_resized,
                window_shape)
            if top1_grape == "grape":
                cv2.rectangle(display_image, (x_grape + x, y_grape + y),
                    (x_grape + x + 70, y_grape + y + 70),
                    colour_dict.get(top1, (0,0,0)), 4)
                #cv2.imshow("Window", grape_window_resized)
                cv2.waitKey(1)
time.sleep(0.025)
```

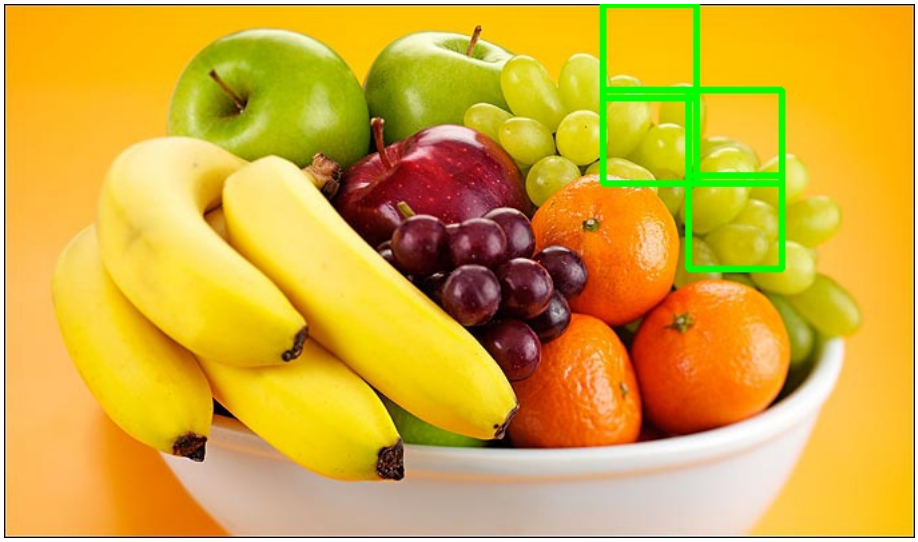

Figure 5.7: Recursive refinement 2 - sourced from http://www.travispta.org/news/2017/9/4/fruit-bowl-challenge

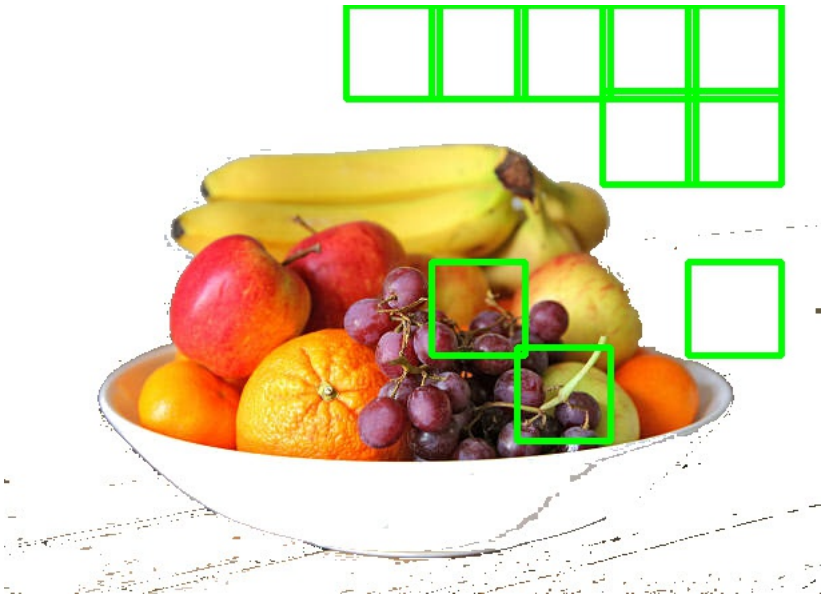

Figure 5.8: Recursive refinement 3

Figure 5.9: Bowl of fruit with background removed

Results

Some very interesting results were recorded on three separate images. Two new images are seen here which we will explore in future experiments. In all three images, while we are getting some expected predictions, grapes are being classified in locations that have nothing resembling a grape. These can be viewed in Figures 5.6, 5.7 and 5.8.

Empirical Analysis

The instances where false positives were recorded are mind boggling. An analysis to why this may be occurring is outlined in the section 5.5.

Table 5.4: Comparison of fruit image sliding window results with and without bakground

Food type	Grid	Row	Column	White Grid	White Row	White Column
Apple	5	1	3	10	0	4
Banana	1	0	1	0	0	0
Grape	4	0	0	1	0	1
Orange	5	0	0	3	0	0
Other	0	3	1	1	4	0

5.3 Impact of Background

Overview

As we can see in section 5.1, using sliding windows to classify many sections of an image, there were some cases where some unexpected predictions were made. Due to this, the decision was made to analyse the effect the background of the image on its classification. The sliding window code was then ran on a new image. This new image was the same fruit bowl as used previously but the background was filled in as white as per Figure 5.9.

Results

Grid

For grid based sliding window approach, the results turned out to be less successful than with the background. In this experiment, fourteen out of fifteen of top-1 classification were of an expected food type rather than fifteen out of fifteen with the background present. We expected the food types of apple, orange, grape and banana to appear in this image but while a banana was detected to a top-5 accuracy on a few occasions it was never predicted to a top-1 accuracy. The contrast between the image results can be seen in Table 5.4.

Table 5.5: Comparison of fruit bowl images

Food type	Grid	Row	Column	New Grid	New Row	New Column
Apple	5	1	3	4	1	0
Banana	1	0	1	5	0	5
Grape	4	0	0	2	1	0
Orange	5	0	0	0	0	0
Other	0	3	1	1	2	1

Row

The row based sliding window again had worse result than its counterpart, with zero out of four correct classifications as opposed to one. In this case, an orange appeared at top-5 accuracy once. The most common prediction was ice-cream which appeared at top-1 accuracy in three out of four instances.

Column

In contrast to our previous two methods of sliding window, this method outperformed its counterpart with correct predictions of all five windows while before we only had four out of five. In this experiment, a apple was predicted four times and a grape once, with all correct fruits appearing to top-5 accuracy.

Empirical Analysis

It is quite interesting that removing the background to the image reduced our accuracy overall. Many white foods were classified instead which makes sense due the impact of colour expected.

Figure 5.10: Alternative Bowl of fruit

5.4 Alternative Test Image

Overview

In our previous sliding window oriented experiments, we had only used a single image. In order to see whether this image had biases unknown to us, another fruit bowl image had to be tested. This image was selected as fruit took up a larger portion of the image as seen in Figure 5.10.

Results

Grid

The performance of this experiment was slightly worse than with the previously used image. When the grid based sliding window was executed on Figure 5.10, fourteen out of fifteen predictions had an expected value. Out of the fourteen predictions orange was not predicted to top-1 accuracy at all. This can be seen, in comparison to previously used image, in Table 5.5.

Row

In the column based window for the new fruit image, the results were not very successful as has been the trend for most row based classification. Two out of four predictions had an expected value at top-1 accuracy.

Column

The column based approach had a similar result to its counterpart in that only one of its predictions was unexpected. Although, due to the size of the new image, another column was created and thus has a better overall accuracy.

Empirical Analysis

A possible reason that an orange was not classified in any of these images is because in Figure 5.10, a more mandarin food is displayed.

5.5 Analysing Results of Recursive Refinement Further

Overview

In Section 5.2, some interesting results were obtained when attempting to run recursive refinement on an image with grapes in it. Originally it was meant to see if individual grapes could be recognised. On evaluation of the images used for training, it was found that bunches of grapes were used for training, not individual grapes. Therefore the results expected where never feasibly going to be obtained. Even though this was the case, the script resulted in finding grapes in portions of the background, as seen in Figure 5.6. The purpose of this experiment is to look into why this is occurring. A new image of a fruit bowl (Figure 5.11), taken on a mobile phone was selected for this experiment. The image was used to run the script as defined in 5.2 but on two separate

models. The model with tuned parameters, trained on 108 classes, with a final test accuracy of 66.3% and the model trained on 13 classes, with a final accuracy of 92.6%.

Figure 5.11: Fruit image taken on a mobile phone

Results

The outputs of the script can be seen in Figure 5.12 and Figure 5.13. The output using the model trained on 13 classes had less false positive predictions. As the script ran (on both models), the segments predicted as grapes were saved to disk and then were each classified using 'label_image.py'. The probabilities of the predictions were recorded with results as seen in Table 5.6.

Table 5.6: Results of recursive refinement segment classifications using two models

Fruitbowl Segment	**13 class model**	**108 class model**
False Positive 1	grape 0.31476045 apple 0.25866747 orange 0.11228518 apple pie 0.10234257 chocolate cake 0.09062083	grape 0.31820437 panna cotta 0.091400646 macarons 0.08560496 roll 0.08188102 apple 0.03981942
False Positive 2	grape 0.29391274 apple 0.2371384 orange 0.16196558 chocolate cake 0.10039616 apple pie 0.096121624	panna cotta 0.21783036 grape 0.105718106 apple 0.087030284 macarons 0.07753955 orange 0.037592527
False Positive 3	grape 0.22402291 chocolate cake 0.16412543 apple 0.14523567 orange 0.12525927 apple pie 0.099948086	grape 0.2835378 panna cotta 0.088551655 macarons 0.06166248 orange 0.041896477 roll 0.040348493
False Positive 4	apple 0.37422368 grape 0.28592423 orange 0.10230055 chocolate cake 0.085367255 apple pie 0.053731006	grape 0.20476209 macarons 0.17529227 panna cotta 0.16139281 roll 0.04359601 orange 0.036991216
False Positive 5	N/A	grape 0.22574392 macarons 0.14840269 panna cotta 0.08670294 roll 0.08650105 orange 0.046900786
False Positive 6	N/A	grape 0.31369162 panna cotta 0.11833602 macarons 0.09010604 apple 0.057512555 roll 0.047116004

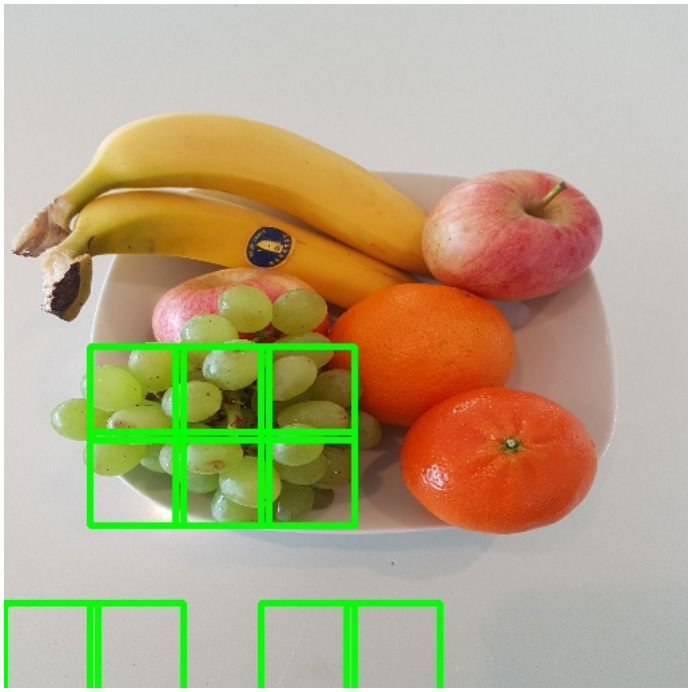

Figure 5.12: Image after sliding window - 13 classes

Empirical Analysis

Comparing the two models

As we can see evidence of in Table 5.6, along with Figure 5.12 and Figure 5.13, the model trained on 13 classes predicted only four false positives while the larger model predicted 6. This is most likely due to the larger model not being able to separate grapes as well as the smaller model. It is worth mentioning that the grape class has the lowest number of training images of all the other classes.

Analysing Probabilities

The probabilities of predictions for grapes in these false positive classifications are all low, in both models. In contrast, some of the correctly classified segments were run through the model and the results all had probabilities in the high nineties. When these background segments are run through the model,

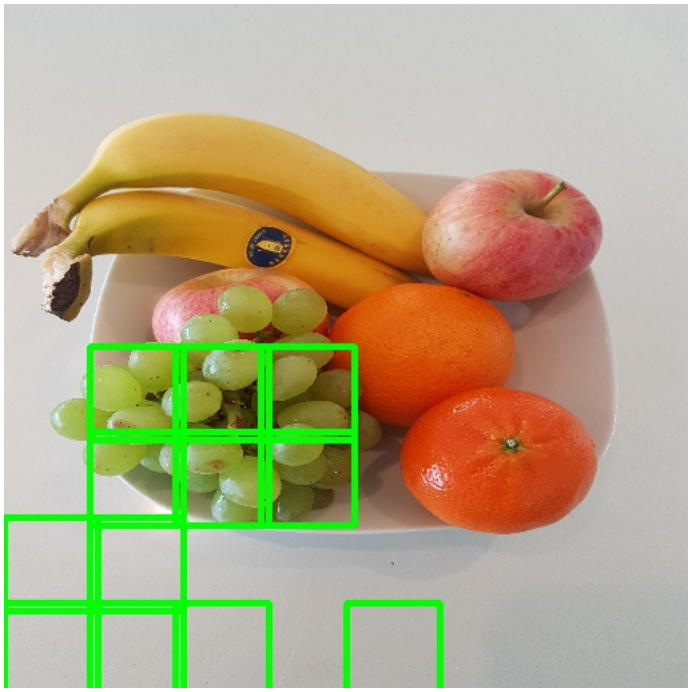

Figure 5.13: Image after sliding window - 108 classes

some prediction has to be made and none of these false predictions have a probability of over 0.4. The average probability of these segments being grapes is in fact 0.257 (rounded to three decimal places). A counter measure to this problem could be to disregard any predictions with a probability under a certain threshold, perhaps 0.4.

5.6 Scaling Down Images

Overview

In this experiment, it was decided to look into how scaling down the size of the image would influence its classification. Three images were scaled down to 10% of their size for this purpose, Figure 5.14, Figure 5.15 and Figure 5.16. Each of these images were run through the model before and after image resizing and the predictions were recorded.

Figure 5.14: Banana Image Pre-Resolution

- sourced from https://www.bettycrocker.com/recipes/scrumptious-apple-pie.png - sourced f

Figure 5.15: Apple Pie Image Pre-Resolution

Figure 5.16: Pizza Image Pre-Resolution

Table 5.7: Results of Down Scaled Images

Food Image	**Pre-Scaled Image**	**Scaled Image**
Banana	Top 1 : 0.9971	Top 1 : 0.9995
Apple Pie	Top 1 : 0.7986	Top 5 : 0.0836
Pizza	Top 1 : 0.9753	Top 1 : 0.9705

Network Architecture

Inception V-3 architecture. Parameter tuned model was used for this.

Dataset

Food-101 extended dataset.

Results

The results of the images before and after were somewhat contrasting as seen in Table 5.7. The 'Top 1' notation in the table suggests that the food was predicted as the number one prediction. The decimal value is the probability of that food type being the correct classification.

Empirical Analysis

As we can see in Table 5.7, the banana (Figure 5.14) has a very similar prediction and probability. This would indicate that the classifier does not need clear quality images for bananas but maybe shape, texture and colour are important factors.

The apple pie (Figure 5.15) has drastically different results in Table 5.7. Originally the apple pie was correctly predicted with a probability of 0.7986 but after downscaling, the apple pie was the fifth food predicted with a likelihood of 0.0836.

Figure 5.17: Banana Image Greyscale

The image of the pizza (Figure 5.16), had very similar results before and after resizing. This might indicate that pizza is determined more by shape, texture and colour that fine quality.

5.7 Effect of Colour

Overview

As seen in the last experiment, it is possible that colour plays a significant part in the overall classification of some food types, along with shape and texture. Due to this, what would happen if colour was removed from the image? The three images, Figure 5.14, Figure 5.15 and Figure 5.16, were all converted to greyscale to test this and the resulting images can be seen in Figure 5.17, Figure 5.18 and Figure 5.19. Each of these images were ran through the model before and after converting to greyscale and the results recorded.

Figure 5.18: Apple Pie Image Greyscale

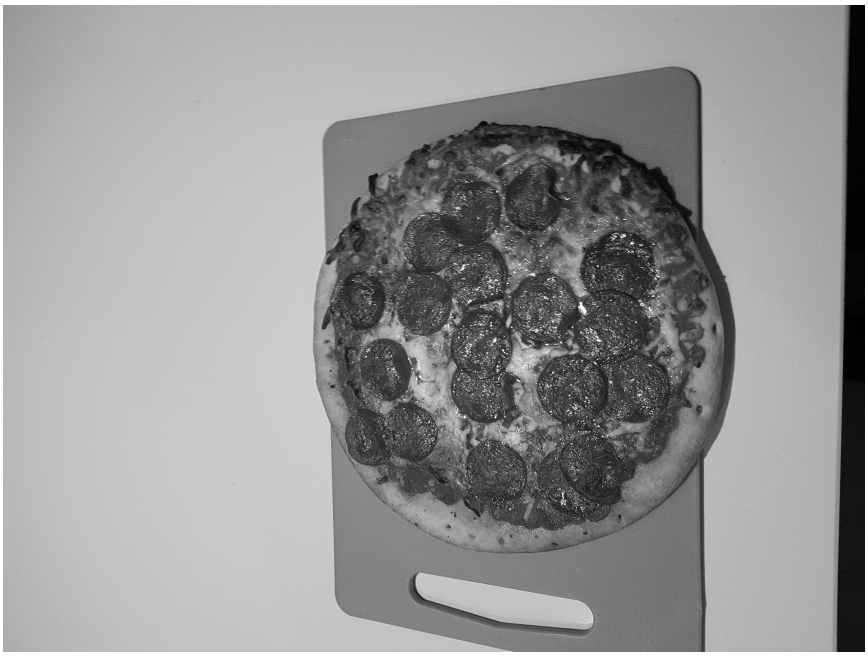

Figure 5.19: Pizza Image Greyscale

Table 5.8: Effect of Colour

Food Image	**Pre-Scaled Image**	**Greyscale**
Banana	Top 1 : 0.9971	Top 1 : 0.9986
Apple Pie	Top 1 : 0.7986	Top 1 : 0.3462
Pizza	Top 1 : 0.9753	Top 2 : 0.1299

Network Architecture

Inception V-3 architecture. Parameter tuned model was used for this experiment.

Dataset

Food-101 extended dataset.

Results

The results for this experiment can be seen in Table 5.8. The 'Top 1' notation in the table suggests that the food was predicted as the number one prediction. The decimal value is the probability of that food type being the correct classification.

Empirical Analysis

Colour

In regards to the images of the banana (Figure 5.14 and Figure 5.17), there is very little difference in classification. They were both classified to a Top 1 accuracy and their probabilities differing by only 0.0015 with the grey scale image being of a higher probability interestingly enough. This would indicate that colour is not an important factor for classifying bananas and the coloured image may even have noise. This can be seen in Table 5.8.

The apple pie images (Figure 5.15 and Figure 5.18) were also both classified correctly but with a large gap in the likelihood of that classification being correct. As per Table 5.8, the coloured image had a probability of 0.7986 as opposed to one of 0.3462. This would indicate that while colour is important, there is enough unique data from the rest of the image to result in correct classification.

Finally, the pizza images (Figure 5.16 and Figure 5.19)were the most contrasting in their results. While the original image was classified to Top 1 accuracy with a likelihood of 0.9753, the coloured image was classified to Top 2 accuracy while a probability of 0.1299. From this we can deduce that colour is vital fro the classification of pizza.

Since this experiment only compared three images, we cannot say for sure if this analysis is biased or not.

Colour vs Scale

In the last experiment, it was seen that the pizza and banana images (Figure 5.16 and Figure 5.14) were not really affected by image quality. While this seems to be true, the image of apple pie, Figure 5.15, was greatly influenced by only being classified to a Top 5 accuracy when down scaled.

For the greyscale images, bananas were not effected greatly and neither was an apple pie but a pizza was greatly influenced.

From this, it is clear that the prominent unique features for each of these food types are different. The banana (Figure 5.14) may have focus on shape and texture, the pie (Figure 5.15) on image quality and some influence of colour while the pizza (Figure 5.16) may have a focus on shape, texture and colour.

Chapter 6

Prototype Application

In order to demonstrate practical use of the models trained in Chapter 4, a prototype application was developed. This application was developed for a smartphone running on the Android operating system in the Java programming language. A backend service was created to receive an image from the smartphone application, process the image using the CNN previously developed and return a response to the application in the form of a prediction.

6.1 Requirements

6.2 Design

The design process for this prototype application consisted of user interface design, android application system design and the activity of choosing resources for a backend host.

User Interface

The user interface mock ups were created using (*fluid* 2018). The aim of this design was to provide a simple, easy to use interface so that a user could use the application with minimal effort. The main benefit of using computer

vision for this type of application is to reduce the effort of the user or dietician keeping track of food intake. Therefore, this application had to be very quick and easy to use.

Three main activities were needed for this application. The first activity, as seen in Figure 6.1, would consist of options to either take an image of food or to view the food logs of the user. If the user decides to take an image of food, they will be brought to the second activity which can be seen in Figure6.2. The user is required to take a picture before reaching the activity in Figure 6.2. Once the user presses the send button the image is classified and the content of the activity changes to display the classification and calorie count as in Figure 6.3. Alternatively, the user can cancel the process and return to the first activity. The user would then submit the food classification for logging and automatically return to the first activity. The final activity is displayed when a user views their food logs. This activity has the ability to view a list of the food logs taken by the user by day, week or month. The calorie count of the selected time frame is to be displayed on this activity also.

System Architecture

Architectural Patterns

The Model-View-Presenter (MVP) architectural pattern was adapted for this application (*Model-View-Presenter: Android guidelines* 2017). This pattern is very similar to the Model-View-Controller (MVC) architecture which is quite popular in the software development industry. The main difference between MVC and MVP is that whereas in MVC the controller is responsible for which view is used, in MPV the presenter is called through the view. This is due to the architecture of Android applications in general, where the view takes primary control. In the view classes, there should be no logic whatsoever in an MVP architecture and all logic should be called in the presenters. There is also a one to one dependency between views and presenters.

As illustrated in Figure 6.5, the views in the application have a dependency

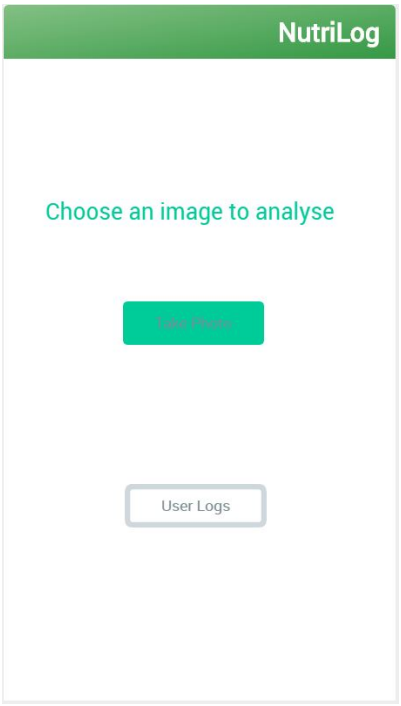

Figure 6.1: Landing Activity

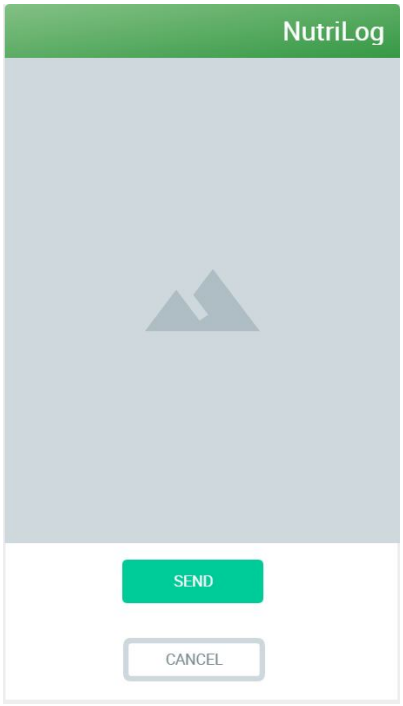

Figure 6.2: Image Submission Activity

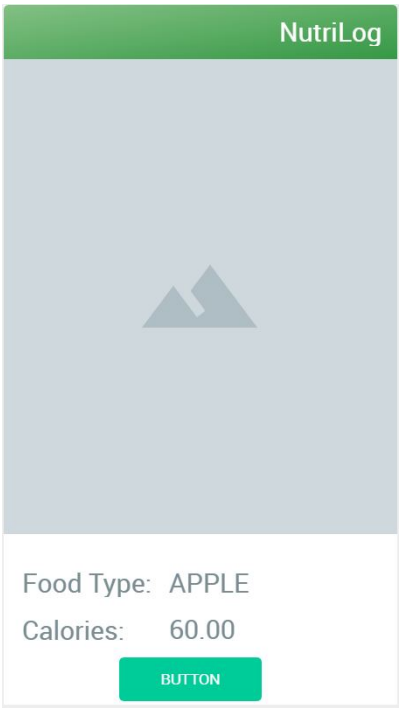

Figure 6.3: Classification Activity

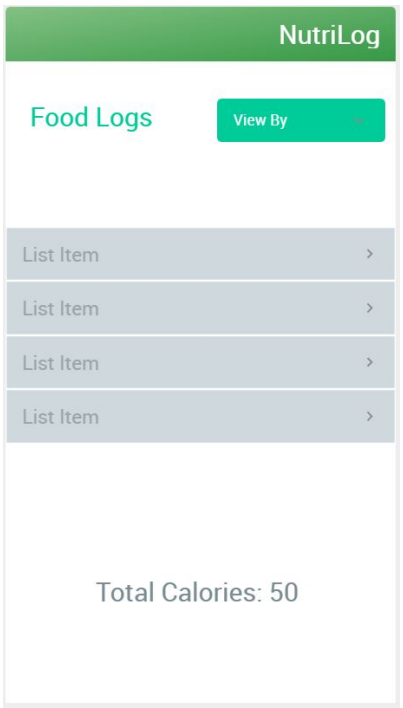

Figure 6.4: Food Logs Activity

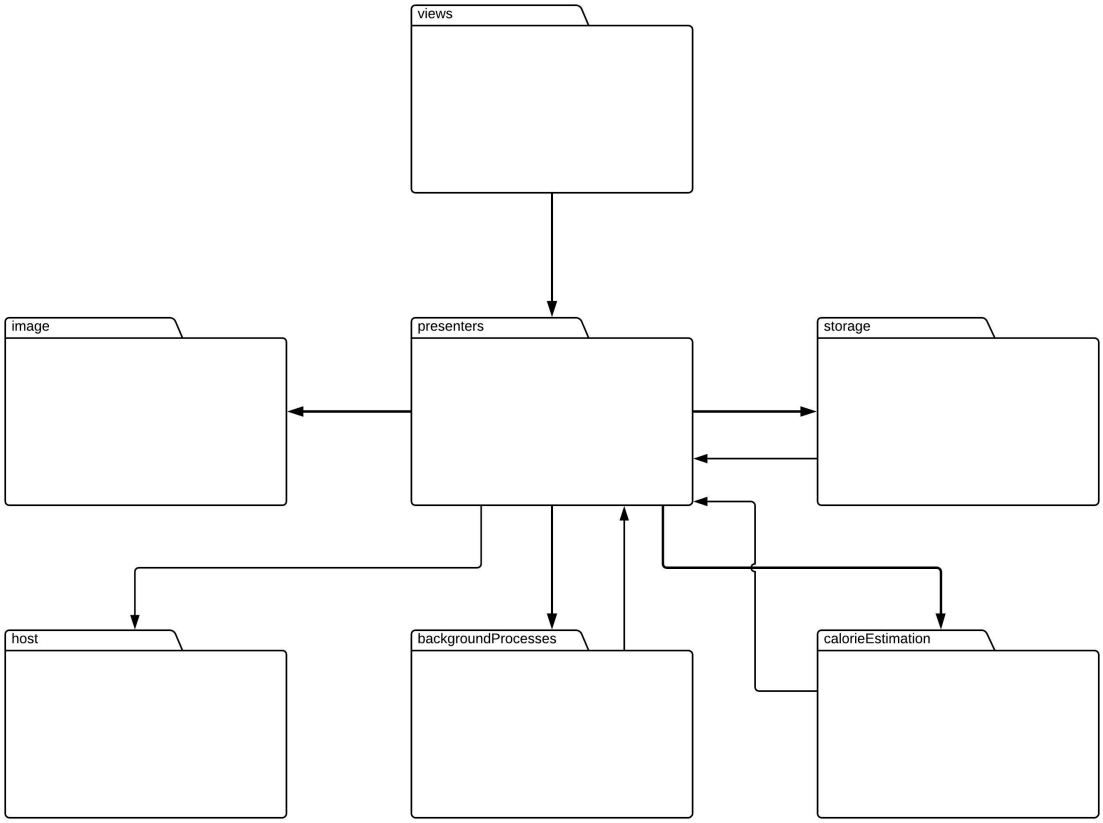

Figure 6.5: Package Diagram

only on their corresponding presenter. The presenters contain the business logic of the application and have a dependency on all other packages. The packages of backgroundProcesses, calorieEstimation and storage have a dependancy on the presenters package for callback purposes and also database creation in Android requires context of the activity information.

AWS Architecture

The application interfaces with an Amazon Web Services (AWS) server called an EC2 instance. In order to keep the application secure AWS was used to provide a secure architecture. A Virtual Private Cloud (VPC) was created to host the instances. In a production environment a load balancer could be used to evenly distribute traffic across a fleet of backend instances but for the purposes of this prototype application, a single instance as in Figure 6.6 could be used.

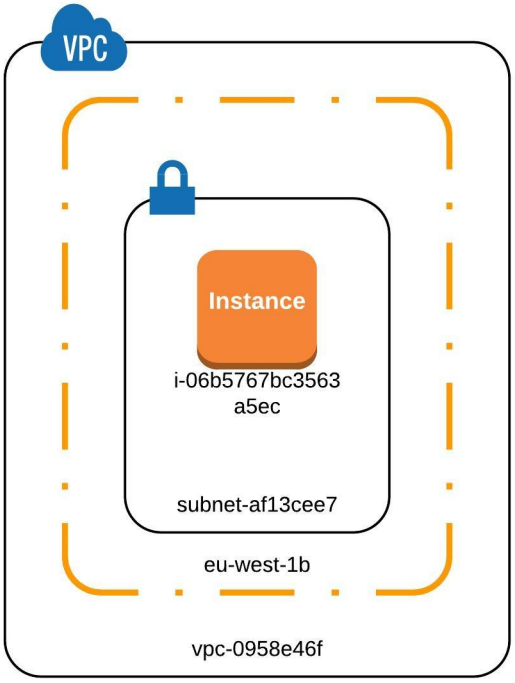

Figure 6.6: AWS Network Diagram

Security groups were utilised to restrict the type of traffic that would be allowed to reach the instance, so that the instance would only receive http requests at port 5000 and receive ssh connections at port 22 as per Figure 6.7.

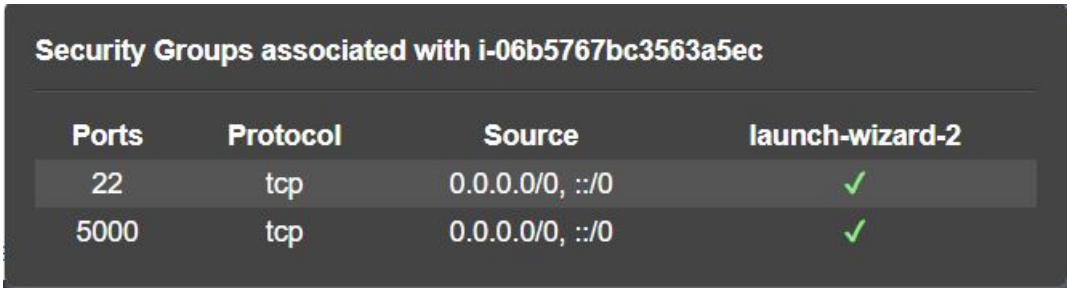

Figure 6.7: AWS Security Group

6.3 Implementation

Outlined below are some coding fragments categorised under: Design Patterns, Asynchronous Tasks, Presenters, Views and Interesting Coding Fragments.

Design Patterns

Various design patterns were used in the implementation of this application such as the Factoy, the Builder and the Singleton which are outlined below.

DAOFactory class

The Factory design pattern was used to retrieve a DAO (Data Access Object). This was used so that if at a future time the storage type of the application is to be changed, the devloper would only have to change one instance of the codebase and return a new implementation of the DAO interface in the method getDAO().

```
1 public class DAOFactory {
2
```

```
3       public DAO getDAO(Context context){
4           return SqlLiteDAO.getInstance(context);
5       }
6   }
```

HostFactory class

The Factory design pattern was used to retrieve a Host object that acts as an endpoint for the backend server. As this application is a lightweight proof of concept prototype only one backend server was used. In a future iteration, it may be decided to use a more complex method of Host selection and this factory class allows for simple modification to the codebase in this instance.

```
1   public class HostFactory {
2
3       public Host createHost() {
4           Host host = new
               Host.HostBuilder("52.214.205.157")
5           .withDns(
6               "ec2-52-214-205-157.eu-west-1.compute.amazonaws.com"
7           .withPort(5000)
8           .withRoute("/classifyImage/")
9           .build();
10          return host;
11      }
12  }
```

HostBuilder

A Host builder was used to create a Host object. This builder is a static inner class in the Host class file.

```
1   public static class HostBuilder {
```

```
2       private final String ipv4;
3       private String dns;
4       private int port;
5       private String route;
6
7       public HostBuilder(String ipv4) {
8           this.ipv4 = ipv4;
9       }
10
11      public HostBuilder withDns(String dns) {
12          this.dns = dns;
13          return this;
14      }
15
16      public HostBuilder withPort(int port) {
17          this.port = port;
18          return this;
19      }
20
21      public HostBuilder withRoute(String route) {
22          this.route = route;
23          return this;
24      }
25
26      public Host build() {
27          return new Host(this);
28      }
29
30  }
```

FoodLogBuilder

A FoodLog builder was used to create a FoodLog object. This builder is a static inner class in the FoodLogImpl class file.

```
1  public static class FoodLogBuilder {
2          private int id = 0;
3          private String food;
4          private double calories;
5          private Date timestamp;
6
7          public FoodLogBuilder(String food) {
8              this.food = food;
9          }
10
11         public FoodLogBuilder withId(int id) {
12             this.id = id;
13             return this;
14         }
15
16         public FoodLogBuilder withCalories(double
               calories) {
17             this.calories = calories;
18             return this;
19         }
20
21         public FoodLogBuilder withTimestamp(Date
               timestamp) {
22             this.timestamp = timestamp;
23             return this;
24         }
25
26         public FoodLog build() {
27             return new FoodLogImpl(this);
```

```
28          }
29
30  }
```

Singleton DAO Object

A Singleton instance of the DAO implementation was used in this application. This was mainly due to the best practice of keeping database access objects as singleton instances.

```
1  public static DAO getInstance(Context context) {
2      if(instance == null) {
3          instance = new SqlLiteDAO(context);
4      }
5      return instance;
6  }
```

Asynchronous Tasks

Asynchronous Tasks in Android are used to run tasks in the background or are sometimes used to simply distribute processess off the main thread. The AsyncTask class must be extended on creation of the background processes.

UploadImage

An AsyncTask was used to send the food image to the backend host.

```
1  @Override
2  protected String doInBackground(Void... params) {
3      String result = "";
4      OkHttpClient client = new OkHttpClient();
5      String imageToSend = image;
```

```
6       RequestBody requestBody = new
           MultipartBody.Builder()
7                .setType(MultipartBody.FORM)
8                .addFormDataPart("image", imageToSend)
9                .build();
10
11      Request request = new
           Request.Builder().url(host.getUrl())
12               .post(requestBody).build();
13
14      Response response = null;
15      try {
16          response = client.newCall(request).execute();
17          result = response.body().string();
18          response.body().close();
19      } catch (IOException e) {
20          e.printStackTrace();
21      }
22
23      return result;
24  }
```

CalorieEstimation

An AsyncTask was used to call the Nutritionix API. This API can be used to retrieve nutritional information on query strings. In this case, the calorie count of the food image was returned. The calorie count was extracted from a JSON object. JSON is explored in detail in a following section.

```
1   @Override
2   protected String doInBackground(Void... params) {
3       String result = "";
4       OkHttpClient client = new OkHttpClient();
```

```
5
6       Request request = new Request.Builder().url(url)
7               .build();
8
9       Response response = null;
10      try {
11          response = client.newCall(request).execute();
12          result = response.body().string();
13          response.body().close();
14      } catch (IOException e) {
15          e.printStackTrace();
16      }
17      try {
18          JSONObject jsonObject = new
               JSONObject(result);
19          JSONArray jsonArray =
               jsonObject.getJSONArray("hits");
20          JSONObject hits = jsonArray.getJSONObject(0);
21          JSONObject fields =
               hits.getJSONObject("fields");
22          String calories =
               fields.getString("nf_calories");
23          result = calories;
24      } catch (JSONException e) {
25          e.printStackTrace();
26      }
27      return result;
28  }
```

Presenters

In the MPV architecture as described earlier, the presenter for each activity contains all the logic for that view.

Presenter for Main Activity

This presenter was task with creating intents to new activities (views).

```
1  public class MainPresenter {
2
3      private Intent intent;
4      private Context context;
5
6      public MainPresenter(Context context) {
7          this.context=context;
8      }
9
10     public void takePhoto() {
11         intent = new Intent(context,
               CaptureImageActivity.class);
12         context.startActivity(intent);
13     }
14
15     public void userLogs() {
16         intent = new Intent(context,
               FoodLogsActivity.class);
17         context.startActivity(intent);
18     }
19
20 }
```

Views

In the MPV archtecture that this application uses, the views are responsible for interacting directly with the user interface.

Main Activity

This view responds to button clicks and calls to its presenter to carry out the logic of the required task.

```
1  public class MainActivity extends AppCompatActivity {
2
3      private MainPresenter mainPresenter;
4
5      @Override
6      protected void onCreate(Bundle
           savedInstanceState) {
7          super.onCreate(savedInstanceState);
8          setContentView(R.layout.activity_main);
9
10         mainPresenter = new MainPresenter(this);
11     }
12
13     public void onClickTakePhoto(View view) {
14         mainPresenter.takePhoto();
15     }
16
17     public void onClickUserLogs(View view) {
18         mainPresenter.userLogs();
19     }
20 }
```

Storage

SqlLite was used for storing data in this application and the class that handled creation, input and output to this database implemented the DAO interface.

DAO interface

An interface for storage devices so that each implementation would have the ability to add food logs, retrieve food logs, delete food logs and remove the data store completely.

```
1 public interface DAO {
2     void addFoodLog(FoodLog foodLog);
3     List<FoodLog> getLogsByDay(Date date);
4     List<FoodLog> getLogsByWeek(Date date);
5     List<FoodLog> getLogsByMonth(Date date);
6     void deleteFoodLogs(List<FoodLog> foodLogs);
7     void deleteDb();
8 }
```

Get Food Logs per Date

This method was used to execute a query and return a List of FoodLog objects.

```
1 @Override
2 private List<FoodLog> selectQuery(String query) {
3         foodLogs = new ArrayList<>();
4         Cursor resultSet = database.rawQuery(query,
             null);
5         while(resultSet.moveToNext()) {
6             try {
7                 foodLogs.add(new FoodLogImpl
8                 .FoodLogBuilder(resultSet.getString(1))
9                 .withId(resultSet.getInt(0))
```

```
                    .withCalories(resultSet.getDouble(2))
                    .withTimestamp(dateFormat
                        .parse(resultSet.getString(3)))
                    .build());
            } catch (ParseException e) {
                e.printStackTrace();
            }
        }
    return foodLogs;
}
```

Delete Food Logs

This method was used to delete food logs from the database that were in the List passed to the method.

```
@Override
public void deleteFoodLogs(List<FoodLog> foodLogs) {
    SQLiteStatement stmt =
       database.compileStatement("DELETE FROM
       FoodLog WHERE rowid = ?");
    for(FoodLog foodLog: foodLogs) {
        stmt.bindString(1,
           String.valueOf(foodLog.getId()));
        stmt.execute();
    }
}
```

Interesting Coding Fragments

Some interesting coding fragments are outlined below which merit inclusion in this report.

Encode Bitmap to base64 String

This method was used to encode the image taken on the phone to a string. This was used to send the image to the backend with minimal latency. The bitmap size was also reduced as in line 8 to decrease response time also.

```
1  public String toBase64() {
2      BitmapFactory.Options options = new
           BitmapFactory.Options();
3      options.inSampleSize = 8;
4
5      Bitmap imageBitmap =
           BitmapFactory.decodeFile(image.getAbsolutePath(),
           options);
6
7      //resize image for faster upload to server
8      Bitmap.createScaledBitmap(imageBitmap, 300, 400,
           false);
9
10     ByteArrayOutputStream byteArrayOutputStream =
           new ByteArrayOutputStream();
11     imageBitmap.compress(Bitmap.CompressFormat.PNG,
           100, byteArrayOutputStream);
12     byte[] byteArray =
           byteArrayOutputStream.toByteArray();
13
14     return Base64.encodeToString(byteArray,
           Base64.DEFAULT);
15 }
```

Map of Runnables

A map of strings to runnables was used to query the database for food logs. This was used as the select query for the database had to dynamically change depending on how the food logs were being viewed, by day, week, or month.

```
1 listViewOptions = new HashMap<>();
2 listViewOptions.put("Day", () -> getLogsByDay());
3 listViewOptions.put("Week",() -> getLogsByWeek());
4 listViewOptions.put("Month", () -> getLogsByMonth());
```

User Interface

User interface implementation as in Figures 6.8, 6.9, 6.10, 6.11, 6.12, 6.13, 6.14 and 6.15.

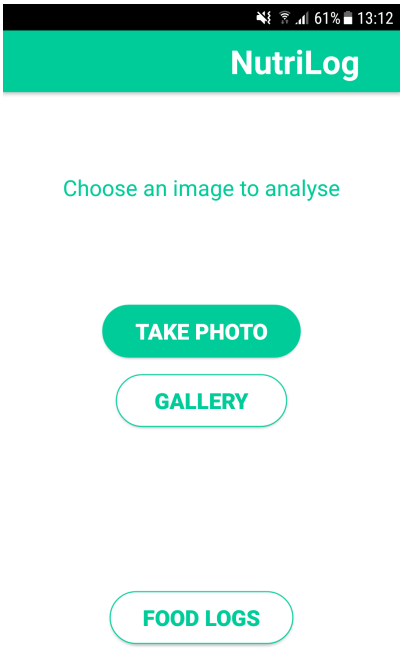

Figure 6.8: Landing Activity

Figure 6.9: Image Capture Activity

Figure 6.10: Image Send Activity

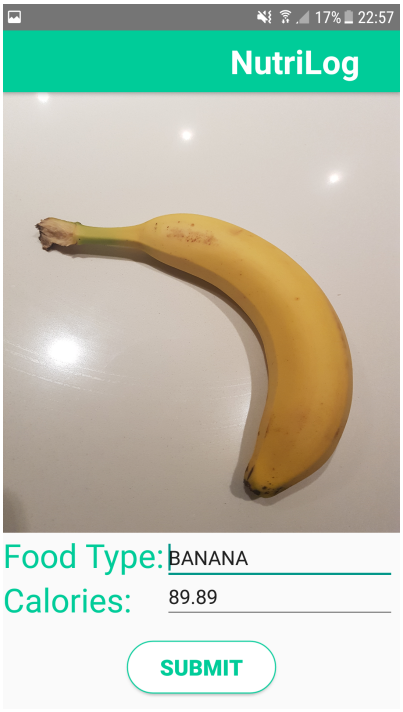

Figure 6.11: Image Submit Activity

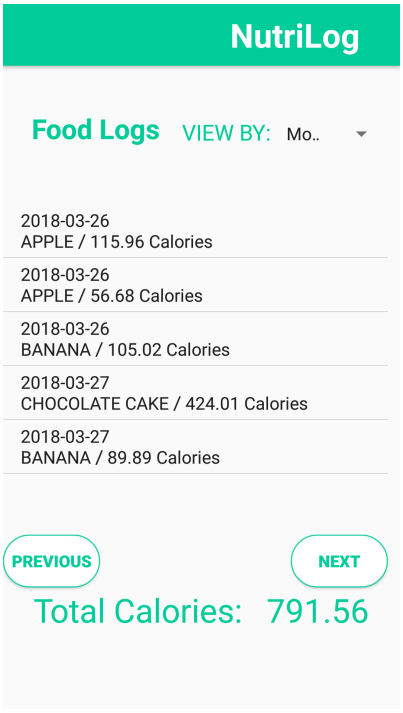

Figure 6.12: FoodLog Month Activity

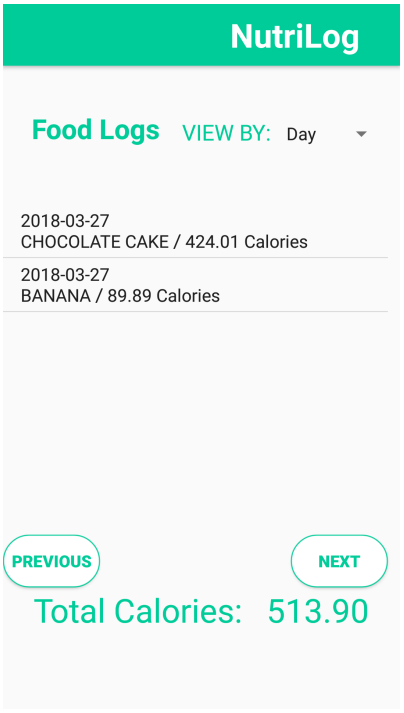

Figure 6.13: FoodLog Day Activity

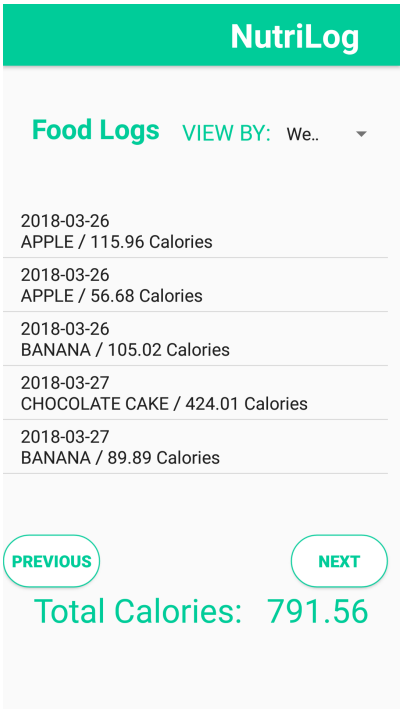

Figure 6.14: FoodLog Week Activity

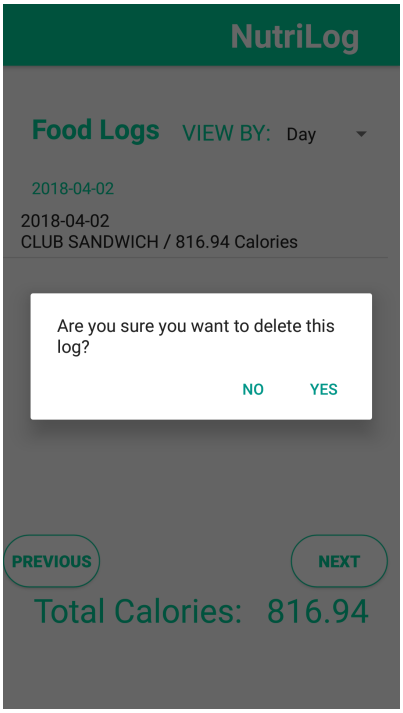

Figure 6.15: Food Log Deletion

Version Control

Github was used as the version control backup system for this project. Commit hisory can be seen in Figure 6.16

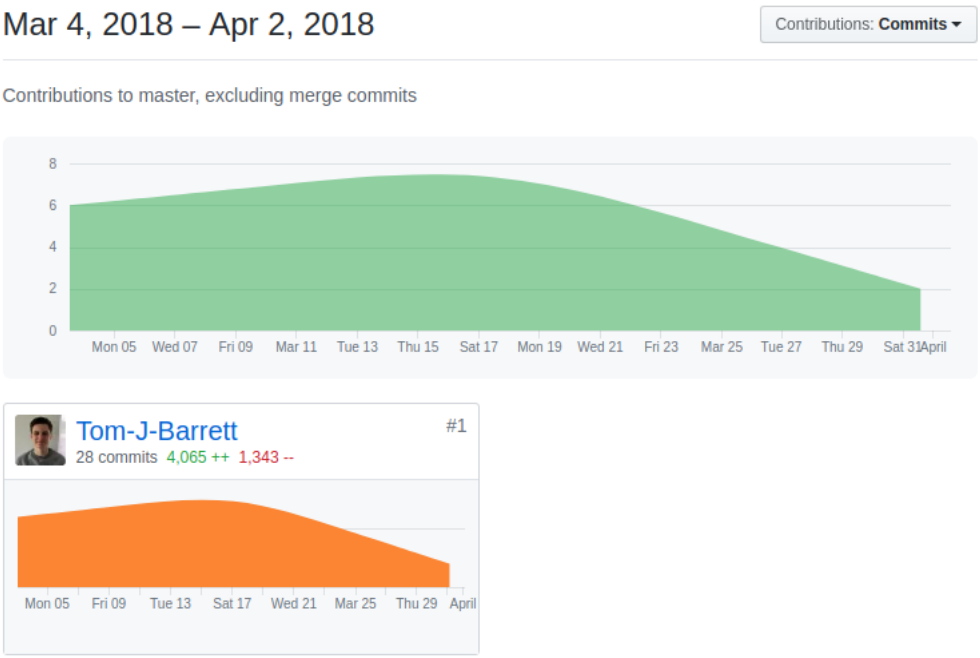

Figure 6.16: Github Contributions to Master

JSON

6.4 Testing

Junit

Mockito

Unit Tests

Test to ensure FoodLog object stores information correctly

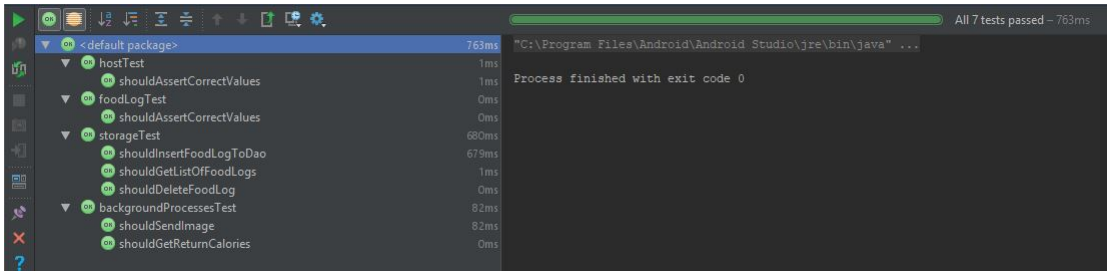

Figure 6.17: Automated Unit Tests

```
@Before
public void setUp() {
    foodLog = new FoodLogImpl(0, "banana", 59.03,
       new Date());
}

@Test
public void shouldAssertCorrectValues() {
    assert food.equals(foodLog.getFood());
    assert calories == foodLog.getCalories();
    assert timestamp.equals(foodLog.getTimestamp());
    assert id == foodLog.getId();
}
```

Test to ensure Host object stores information correctly

```
@Before
    public void setUp() {
        host = new Host.HostBuilder("52.214.205.157")
                .withDns("ec2-52-214-205-157.eu-west-1.compute.a
                .withPort(5000)
                .withRoute("/classifyImage/")
                .build();
    }
```

```
9
10 @Test
11 public void shouldAssertCorrectValues() {
12     assert this.ip.equals(host.getIpv4());
13     assert this.dns.equals(host.getDns());
14     assert this.port == host.getPort();
15     assert this.route.equals(host.getRoute());
16     assert this.urlString.equals(host.getUrl());
17 }
```

Test to ensure DAO object stores and removes data correctly

```
1  @Before
2  public void setup() {
3      dao = mock(SqlLiteDAO.class);
4      date = new Date();
5      foodLog = new FoodLogImpl(0, "test", 0.0, date);
6  }
7
8  @Test
9  public void shouldGetListOfFoodLogs() {
10     foodLogs = dao.getLogsByDay(date);
11     foodLogs = dao.getLogsByWeek(date);
12     foodLogs = dao.getLogsByMonth(date);
13 }
14
15 @Test
16 public void shouldInsertFoodLogToDao() {
17     dao.addFoodLog(foodLog);
18     setTestLog();
19     assert foodLogs.contains(testLog);
20 }
```

```
21
22 @Test
23 public void shouldDeleteFoodLog() {
24     List<FoodLog> foodLogs = new ArrayList<>();
25     foodLogs.add(testLog);
26     dao.deleteFoodLogs(foodLogs);
27     setTestLog();
28     assert testLog.equals(null);
29 }
30
31 private void setTestLog() {
32     foodLogs = dao.getLogsByDay(date);
33     foodLogs.forEach(f -> {
34         if(f.getFood().equals("test")) {
35             testLog = f;
36         }
37     });
38 }
```

6.5 Backend

Python Flask

Implementation

Backend service code

```
1 import base64
2 import label_image
3 import time
4
5 app = Flask(_name_)
```

```
6
7  @app.route('/classifyImage/', methods=['POST'])
8  def classify():
9      imgdata = base64.b64decode(request.form.get('image'))
10     filename = 'images/' + str(time.strftime("%Y%m%d-%H%M%S")) +
           '.jpg'
11     with open(filename, 'wb') as f:
12         f.write(imgdata)
13
14     results = label_image.runModel(filename)
15     predictions = results[0]
16
17     result_String = ""
18     for i in predictions :
19         result_String += str(i) + ","
20     return result_String
21
22 if __name__ == '__main__':
23     app.run(host='0.0.0.0', threaded=True)
```

Method called from 'label_image.py'

```
1  def runModel(file_name):
2    model_file = \
3      "models/output_graph13.pb"
4     label_file = "labels/output_labels13.txt"
5    input_height = 299
6    input_width = 299
7    input_mean = 128
8    input_std = 128
9    input_layer = "Mul"
10   output_layer = " final_result "
11
```

```
12    graph = load_graph(model_file)
13    t = read_tensor_from_image_file(file_name,
14                                    input_height=input_height,
15                                    input_width=input_width,
16                                    input_mean=input_mean,
17                                    input_std=input_std)
18
19    input_name = "import/" + input_layer
20    output_name = "import/" + output_layer
21    input_operation = graph.get_operation_by_name(input_name)
22    output_operation = graph.get_operation_by_name(output_name)
23
24    with tf.Session(graph=graph) as sess:
25       results = sess.run(output_operation.outputs[0],
26                          {input_operation.outputs[0]: t})
27    results = np.squeeze(results)
28
29    top_k = results.argsort() [-5:][::-1]
30    labels = load_labels( label_file )
31
32    setIndex = False
33
34    top5_results = [None] * 5
35    index = 0
36    for i in top_k:
37       top5_results[index] = labels[i]
38       index += 1
39
40     final_results = [top5_results, results]
41    return final_results
```

Deployment

In order to deploy the Flask application to an AWS EC2 instance the following steps had to be followed:

- Upload the output_graph.pb file (Tensorflow model), the output_labels.txt file and the source code to the server using SFTP
- SSH into the server
- Create a folder called 'images' in the home directory to store uploads images
- Place model and label files into folders called 'models' and 'labels' respectively
- Ensure Tensorflow, NOHUP (Ignores the hangup signal in Linux) Python and Flask are installed on the server
- Ensure no processes are running on port 5000 using the command 'lsof -i :5000'
- Run the server.py code using the command 'nohup python server.py &'

Chapter 7

Discussion and Conclusion

Bibliography

Abbirami.R.S et al. (2015). "Large Scale Learning for Food Image Classification". In: *International Journal on Recent and Innovation Trends in Computing and Communication* 3.3, pp. 973–978. URL: http://dx.doi.org/10.17762/ijritcc2321-8169.150317.

About - OpenCV library. https://opencv.org/about.html. Accessed: 2017-02-18.

Arens-Volland, Andreas G, Lübomira Spassova, and Torsten Bohn (2015). "Promising approaches of computer-supported dietary assessment and managementfffdfffdfffdCurrent research status and available applications". In: *International journal of medical informatics* 84.12, pp. 997–1008.

Aurélien, Géron (2017). *Hands-On Machine Learning with Scikit-Learn & TensorFlow*. Offfdfffdfffdreilly.

Bossard, Lukas, Matthieu Guillaumin, and Luc Van Gool (2014). "Food-101 – Mining Discriminative Components with Random Forests". In: *European Conference on Computer Vision*.

Build an Image Dataset in Tensorflow. https://github.com/aymericdamien/TensorFlow-Examples/blob/master/examples/5_DataManagement/build_an_image_dataset.py. Accessed: 2017-02-15.

Chen, Mei-Yun et al. (2012). "Automatic chinese food identification and quantity estimation". In: *SIGGRAPH Asia 2012 Technical Briefs*. ACM, p. 29.

Chen, Nicholas et al. (2010). "Toward dietary assessment via mobile phone video cameras". In: *AMIA Annual Symposium Proceedings*. Vol. 2010. American Medical Informatics Association, p. 106.

Complete Guide to Tensorflow for Deep Learning with Python. https://www.udemy.com/complete-guide-to-tensorflow-for-deep-learning-with-python/learn/v4/overview. Accessed: 2017-02-15.

Daugherty, Bethany L et al. (2012). "Novel technologies for assessing dietary intake: evaluating the usability of a mobile telephone food record among adults and adolescents". In: *Journal of medical Internet research* 14.2.

Deng, J. et al. (2009). "ImageNet: A Large-Scale Hierarchical Image Database". In: *CVPR09*.

Donahue, Jeffrey (2017). "Transferrable Representations for Visual Recognition". PhD thesis. EECS Department, University of California, Berkeley.

Easterbrook, Steve et al. (2008). "Selecting empirical methods for software engineering research". In: *Guide to advanced empirical software engineering*. Springer, pp. 285–311.

Eaton, Dr. Malachy (2017). *Lecture notes in Intelligent Systems*.

Erickson, Bradley J et al. (2017). "Machine learning for medical imaging". In: *Radiographics* 37.2, pp. 505–515.

fluid (2018). https://www.fluidui.com/internify-android-app-prototype. Accessed: 2018-03-27.

Girshick, Ross (2015). "Fast R-CNN". In: *Proceedings of the International Conference on Computer Vision (ICCV)*.

Girshick, Ross et al. (2014). "Rich feature hierarchies for accurate object detection and semantic segmentation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Goodman, Samantha et al. (2015). "Vitamin D intake among young Canadian adults: validation of a mobile vitamin D calculator app". In: *Journal of nutrition education and behavior* 47.3, pp. 242–247.

He, Kaiming et al. (2017). "Mask R-CNN". In: *arXiv preprint arXiv:1703.06870*.

He, Ye et al. (2013). "Food image analysis: Segmentation, identification and weight estimation". In: *Multimedia and Expo (ICME), 2013 IEEE International Conference on*. IEEE, pp. 1–6.

Hoiem, Derek, Yodsawalai Chodpathumwan, and Qieyun Dai (2012). "Diagnosing Error in Object Detectors". In: *Computer Vision – ECCV 2012:*

12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part III. Ed. by Andrew Fitzgibbon et al. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 340–353.

How to Retrain Inception's Final Layer for New Categories. `https://www.tensorflow.org/tutorials/image_retraining`. Accessed: 2018-01-24.

Howard, Andrew G et al. (2017). "Mobilenets: Efficient convolutional neural networks for mobile vision applications". In: *arXiv preprint arXiv:1704.04861*.

Image Segmentation Using DIGITS 5. `https://devblogs.nvidia.com/image-segmentation-using-digits-5/`. Accessed: 2017-03-06.

Inception in Tensorflow. https://github.com/tensorflow/models/tree/master/research/incept
Accessed: 2018-02-05.

Joutou, Taichi and Keiji Yanai (2009). "A food image recognition system with multiple kernel learning". In: *Image Processing (ICIP), 2009 16th IEEE International Conference on*. IEEE, pp. 285–288.

Kagaya, Hokuto, Kiyoharu Aizawa, and Makoto Ogawa (2014). "Food detection and recognition using convolutional neural network". In: *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, pp. 1085–1088.

Kawano, Yoshiyuki and Keiji Yanai (2014). "Food image recognition with deep convolutional features". In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. ACM, pp. 589–593.

Khanna, Nitin et al. (2010). "An overview of the technology assisted dietary assessment project at Purdue University". In: *Multimedia (ISM), 2010 IEEE International Symposium on*. IEEE, pp. 290–295.

Krizhevsky, Alex, Vinod Nair, and Geoffrey Hinton. "CIFAR-10 (Canadian Institute for Advanced Research)". In: URL: `http://www.cs.toronto.edu/~kriz/cifar.html`.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*, pp. 1097–1105.

LeCun, Yann and Corinna Cortes (2010). "MNIST handwritten digit database". In: URL: http://yann.lecun.com/exdb/mnist/.

Liu, Chang et al. (2016). "Deepfood: Deep learning-based food image recognition for computer-aided dietary assessment". In: *International Conference on Smart Homes and Health Telematics*. Springer, pp. 37–48.

Luger, George F. (2005). *Artificial Intelligence. Structures and Strategies for Complex Problem Solving/Luger GF*. Pearson Education Limited.

Mao, Dongyuan, Qian Yu, and Jingfan Wang. "Deep Learning Based Food Recognition". In:

Marsland, Stephen (2015). *Machine learning: an algorithmic perspective*. CRC press, pp. 71–110.

Meanshift Algorithm for the Rest of Us (Python) (2016). http://www.chioka.in/meanshift-algorithm-for-the-rest-of-us-python/. Accessed: 2018-03-30.

Mezgec, Simon and Barbara Koroušić Seljak (2017). "Nutrinet: A deep learning food and drink image recognition system for dietary assessment". In: *Nutrients* 9.7, p. 657.

Mitchell, Tom M. (1997a). *Machine Learning*. International Edition 1997. New York: The McGraw-Hill Companies, Inc., pp. 81–126.

— (1997b). *Machine Learning*. International Edition 1997. NewYork: The McGraw-Hill Companies, Inc., pp. 52–78.

Model-View-Presenter: Android guidelines (2017). https://medium.com/@cervonefrancesco/model-view-presenter-android-guidelines-94970b430ddf. Accessed: 2018-03-27.

Pouladzadeh, Parisa, Shervin Shirmohammadi, and Rana Al-Maghrabi (2014). "Measuring calorie and nutrition from food image". In: *IEEE Transactions on Instrumentation and Measurement* 63.8, pp. 1947–1956.

Pouladzadeh, Parisa, Shervin Shirmohammadi, and Abdulsalam Yassine (2014). "Using graph cut segmentation for food calorie measurement". In: *Medical Measurements and Applications (MeMeA), 2014 IEEE International Symposium on*. IEEE, pp. 1–6.

Pouladzadeh, Parisa, Gregorio Villalobos, et al. (2012). "A novel SVM based food recognition method for calorie measurement applications". In: *Multimedia and Expo Workshops (ICMEW), 2012 IEEE International Conference on*. IEEE, pp. 495–498.

Ren, Shaoqing et al. (2015). "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". In: *Neural Information Processing Systems (NIPS)*.

Schap, TE et al. (2014). "Merging dietary assessment with the adolescent lifestyle". In: *Journal of human nutrition and dietetics* 27.s1, pp. 82–88.

Shelhamer, Evan, Jonathan Long, and Trevor Darrell (2017). "Fully Convolutional Networks for Semantic Segmentation". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 39.4, pp. 640–651.

Sliding Windows for Object Detection with Python and OpenCV. `https://www.pyimagesearch.com/2015/03/23/sliding-windows-for-object-detection-with-python-and-opencv/`. Accessed: 2018-01-30.

Support Vector Machines for Machine Learning. `https://machinelearningmastery.com/support-vector-machines-for-machine-learning/`. Accessed: 2017-12-20.

Szegedy, Christian et al. (2016). "Rethinking the inception architecture for computer vision". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826.

Villalobos, Gregorio, Rana Almaghrabi, Behnoosh Hariri, et al. (2011). "A personal assistive system for nutrient intake monitoring". In: *Proceedings of the 2011 international ACM workshop on Ubiquitous meta user interfaces*. ACM, pp. 17–22.

Villalobos, Gregorio, Rana Almaghrabi, Parisa Pouladzadeh, et al. (2012). "An image procesing approach for calorie intake measurement". In: *Medical Measurements and Applications Proceedings (MeMeA), 2012 IEEE International Symposium on*. IEEE, pp. 1–5.

Yanai, Keiji and Yoshiyuki Kawano (2015). "Food image recognition using deep convolutional network with pre-training and fine-tuning". In: *Multi-*

media & Expo Workshops (ICMEW), 2015 IEEE International Conference on. IEEE, pp. 1–6.

Zeiler, Matthew D and Rob Fergus (2014). "Visualizing and understanding convolutional networks". In: *European conference on computer vision*. Springer, pp. 818–833.

Zhang, Weiyu et al. (2015). "Snap n Eat - Food Recognition and Nutrition Estimation on a Smartphone". In: *Journal of diabetes science and technology* 9.3, pp. 525–533.

Zhu, Fengqing et al. (2011). "Segmentation assisted food classification for dietary assessment". In: *Computational Imaging IX*. Vol. 7873. International Society for Optics and Photonics, 78730B.

Appendix A

Appendix

A.1 Image Segmentation

Fully Convolutional Neural Networks for Semantic Segmentation

There has been a very interesting paper from UC Berkeley focused on using Full Convolutional Networks for semantic segmentation (Shelhamer, Long, and Darrell, 2017). Fully Convolutional Networks (FCN) do not have any fully connected layers. They are replaced with more filtering layers. Nvidia Digits have a semantic segmentation implementation based off the work of this paper.

They took this approach because "feedforward computation and backpropogation are much more efficient when computer layer-by-layer over an entire image instead of independently patch-by-patch" (Shelhamer, Long, and Darrell, 2017). This was also because they were focused on object detection. Normal classifiers do not work very well when they are to classify more than one subject in an image and image segmentation was a way to solve this.

There are a set of steps you can follow to turn a CNN into a FCN for semantic segmentation as follows ie. change to a convolutional layer from a fully connected one:

- The size of the filters must be set to the size of the input layers.
- For every neuron in the fully connected layer, have a filter.

Segmentation

Graph Based Segmentation

Graph cut segmentation has been used extensively in image segmentation. OpenCV has an implementation of a graph cut algorithm called grabcut which has been used to segment food on occasion (Pouladzadeh, Shirmohammadi, and Yassine, 2014).

Table A.1: FCN Resulys (Shelhamer, Long, and Darrell, 2017)

	FCN
VOC11 mean IU	62.7
VOC12 mean IU	62.2
PASCAL VOC10 pixel acc.	67.0
PASCAL VOC10 mean acc.	50.7
PASCAL VOC10 mean IU	37.8
PASCAL VOC10 f.w. IU	52.5

Table A.2: Results

	Single Food Portion	Non-mixed Food	Mixed Food
Color and Texture	92.21	N/A	N/A
Graph Based	95 (3% increase)	5% increase	15% increase

According to (Pouladzadeh, Shirmohammadi, and Yassine, 2014), "Graph cut based method is well-known to be efficient, robust, and capable of finding the best contour of objects in n image, suggesting it to be a good method for separating food portions in a food image for calorie measurement". Along with the graph cut segmentation algorithm, this research team also used color and texture segmentation. Gabor filters were used to measure texture features (Pouladzadeh, Shirmohammadi, and Yassine, 2014). When color and texture segmentation was applied, the method came into difficulty with mixed foods but by applying graph cut segmentation, clearer object boundaries were shown.

In conclusion, the accuracy of the classification increased when using graph based segmentation rather than color and texture as seen in A.2.

Local Variation Framework

Another paper was published in which the research team attempted to create a food calorie extimation system (Y. He et al., 2013). This system would compromise of three steps, image segmentation, image classification and weight estimation. For the segmentation module, a local variation approach to seg-

mentation was performed. Local variation is by which intensity differences between neighbouring pixels is measured. This is a type of graph based segmentation.

The team also carried out some segmentation refinement when the segmentation algorithm had been performed. This consisted of removed small segments (defined as less than 50 pixels) and trying to prevent over and under segmentation. After classification was performed on each segment, segments with low confidence values were removed (Y. He et al., 2013).

Conclusion

Both of the above papers of (Pouladzadeh, Shirmohammadi, and Yassine, 2014) and (Y. He et al., 2013) used a graph based segmentation. The first paper used a more generic implementation while the second used a local variation framework. Both methods provided successful results in the image segmentation process.