Identification and Classification in Food Images

Tom Barrett

December 13, 2017

Abstract

HI

Contents

Chapter 1

Introduction

1.1 Overview

This project explores the use of indentification and classification of food images for use in a calorie measurement 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. A mobile application that could help keep track of a user's calorie intake by taking pictures of their meals would be a great help. The area of Machine Vision is a very difficult topic to address as it is a very hard task for computers to undertake. We, as humans, take vision for granted as we can soon see, from the study of Machine Vision, that there are many difficult steps that have to be made for full identification and classification of an image.

When looking into calorie measurement using an image, there are three questions that have to be answered:

- Where are the Regions of Interest (ROI) in this food image?
- What food types are in these ROI's?
- What is the portain size of each food type?

In this project, my main focus will be on the first question, 'Where are the Regions of Interest (ROI) in this food image?'.

Many researchers in various machine vision labs have attempted to solve this problem using different methodologies. There has been promising results from some papers but these are mostly under highly constrained circumstances. When mixed foods are introduced to the problem, many of the methods fail. Convolution Neural Networks (CNN) have had very promising results in the field of image classification in the recent years but to get to the classification step, image segmentation is first needed, otherwise known as image identification.

I have researched many different methods of image segmentation but it seems that CNN's have had the best results for multiple objects in one image and I hope to apply these results for many foods in an image.

The application that I am proposing to solve the problem statement above is an easy to use Android mobile phone application. The idea is, that when a user is about to eat their meal, they can simply take a picture of their meal for computation. From here, the application would take the image, find the objects (ROI) in the image and take note of them. Concurrently, the application would attempt to classify each object detected. Once this is done, 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. I will focus on finding the objects for this application as I would not be able to implement the full system due to time constraints.

1.2 Objectives

I have a few objectives for this project and I will explain each one in detail.

Understanding of Convolutional Neural Networks

In the project, I will be using Convolutional Neural Networks (CNN's) for object identification in Food Images. I will be using an API for this due to time

constraints but it is a key objective for me to develop a deep understanding of CNN's as they are quite pivotal in the current Machine Vision Industry and I find bio-inspired systems very interesting.

Learn about different image identification and classification techniques

Although, I will be using CNN's for my implementation but I will not be turning a blind eye to other methods of indentification and classification. I have done extensive research on many different methods prior to my decision to use CNN's. I think it is very important to learn about other methods as different methods are better for some situations and it would be best to know about these methods due to the inevitability of their use.

Develop real world skills in Machine Vision

Machine Vision is a growing field in computer science and I think it is a very interesting field to study. My main objective is develop the skills necessary in order to partake in Machine Vision projects in industry or to do further research in academia.

1.3 Methodology

The following methodology were adapter for this project:

1.3.1 Define the research question

The first step to this project was to define the research question. I knew that I wanted to look into the general area of 'Food Identification and Classification' but the scope of this is too broad for an FYP. Therfore I decided that the research question would be to look at the food identification aspect ie. region of interest detection.

1.3.2 Literature Review

Once I had defined the question, 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, I had to follow quite a few references to different papers that focused solely on image segmentation.

1.3.3 Explore different image identification methods

I had collected various image identification from the literature review that I carried out so there were many options to evaluate. I wanted to try something with Convolutional Neural Networks so more traditional methods of identification using colour and texture was not so strenuously explored.

1.3.4 Select an image identification method

1.3.5 Research technologies and develop skills in these technologies

As I used a Convolutional Neural Network (CNN) in this project, I leveraged many resources to enrich my understanding of the process. I decide to use Tensorflow as my main resource in creating a CNN so I followed online tutorials for this technology. I also enrolled in a Deep Learning Course on Udacity to enhance my understanding and skills.

1.3.6 Build a prototype of the application

1.3.7 Compare and analyse results to other implementations

1.4 Overview of Report

This report is broken down into various main headings:

Introduction

This section is to give on overview of what this project is about, how I will approach this project and why I am doing it.

Background

I will be giving some information on the background of the subject that I am focusing on.

Design and Implementation

In this part of the report I will discuss the design of the protoype that I developed and I will also explain the implementation.

Empirical Studies

This section will analyse the results I have obtained and compare them against various metrics and other implementations.

Discussion and Conclusion

In this section, I will discuss my results from the empirical studies and conclude my 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 Convolutional 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 percent 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.

Chapter 2

Background

2.1 Introduction to Machine Learning

In Mitchell [4], 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.

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, medical diagnosis and many more. There may be fear that Machine Learning will start to take away many jobs from humans but this may not be the case. Imagine a doctor, having to diagnose a patient, a machine can offer suggestions based on very large datasets of what the diagnosis is. This is not to say that a Machine would be perscribing patients, but merely act as an assistant to the doctor.

Machine Ethics is a large problem that comes hand in hand with Machine Learning. There is a very important question of who takes the blame when things go wrong, that is why I think it is important that we only use Machine Learning to advise and not to determine but this can be very difficult in a world where, for example, an autonomous car has to decide between crashing into a vehicle beside them with an unknown number of people inside or the

two children playing in the street.

One of the most exciting avenues in Machine Learning, in my opinion, is Computer Vision. Computer Vision 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 skin diseases in an image. The applications are nearly limitless and that is without taking into account other uses.

2.2 Machine Learning Paradigms

There are many Machine Learning paradigms, all of which I will discuss briefly below, but the main area of my focus for this project is in Artificial Neural Networks (ANN). This is because I have researched extensively into ConvolutionalNeural Networks which are based on ANN's.

2.2.1 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. The ANN uses this model to build a very complex web of connected units called artificial neurons. These nuerons are connected by certain weights which determines the processing capacity of the network and these weights are created by learning a dataset.(Malachy) An ANN has a set of inputs that take in a value, sometimes from network outputs and produce a single result or classification. While an ANN is bio-inspired from the human brain, there are many elements of the brain that are not present in ANN and many new elements in ANN that are not modelled from the human brain.

Before I can talk about Convolution Neural Networks which are vital the image processing, I will have to talk about the perceptron learning algorithm, the multi layer percepton, and backpropogation.

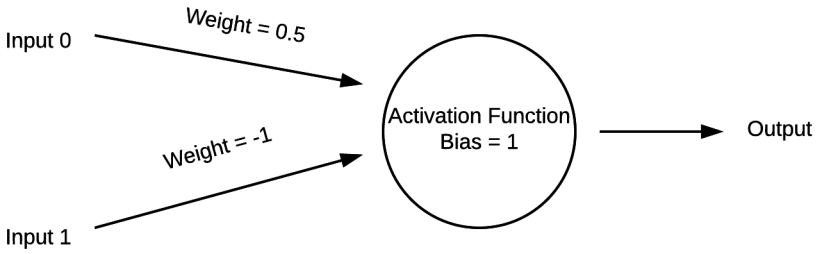


Figure 2.1: Perceptron

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 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.

This Perceptron Learning Rule assumes that there are two sets of instances, a positive and negative set, and each of these has an input and output domain.

Multi Layered Perceptron

Multi Layer Perceptrons (MLP) are made up of multiple layers of perceprons connected together. 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 layer.

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.

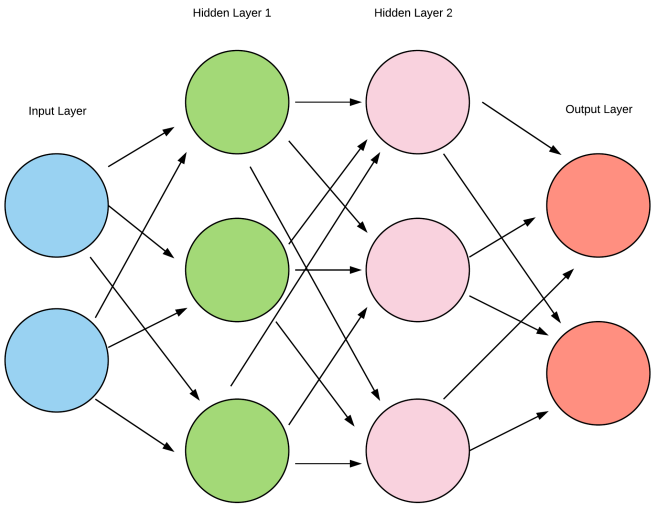
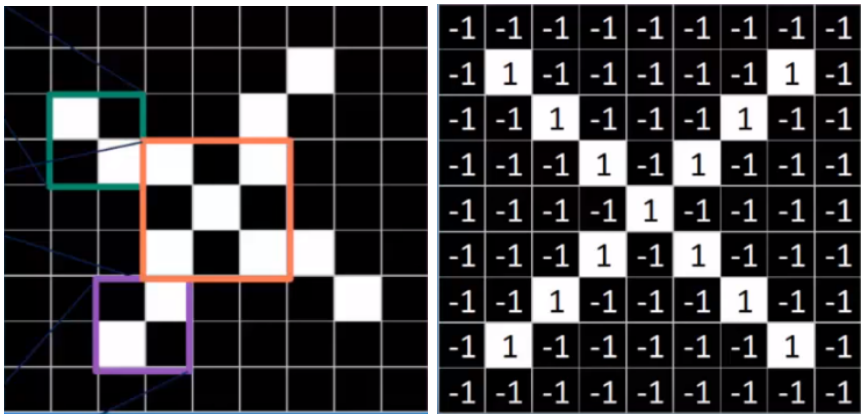


Figure 2.2: Multi Layer Perceptron



(a) Image to Classify (b) Image to Compare

Gradient Descent and Backpropogation

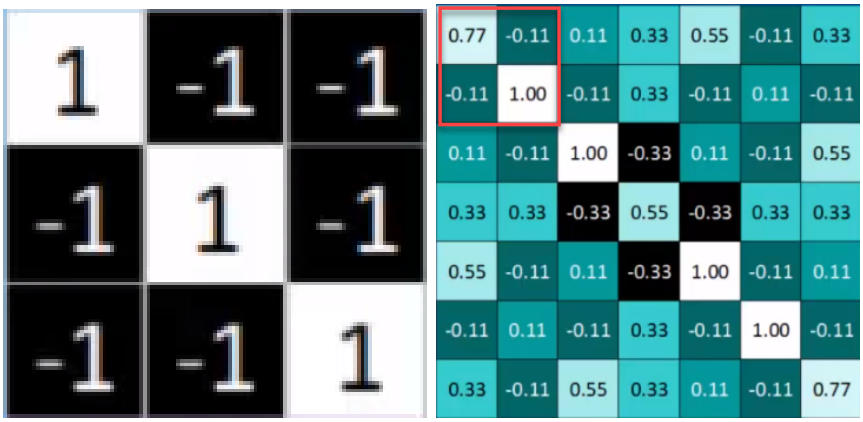
Convolutional Neural Networks

Convolutional Neural Networks (CNN's) are essentially a Multi Layered Percetron with a special structure. CNN's have one major difference from a MLP, they have extra layer of convolution and pooling.

Figure **??** show an image that we want to compare against Figure **??**. 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 **??** and compare it to a subsection of Figure **??**. The CNN multiplies the feature and a section of Figure **??**, adds up the results and divides by 9. 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 2.4b. This is called filtering. The Convolutional layer is composed of carrying out this filtering for every single possible location in Figure **??**.

Next is the Pooling Layer, what this layer does, is it takes the convoluted layer output, you can use Figure 2.4b as reference, and from a user defined size ie. 2x2, gets either the highest decimal value (Max pooling) or the average value (Mean 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 2.5 we know have a much



(a) Image Feature to Search (b) Convoluted Image

Figure 2.5: Pooled Image

smaller image stack in which to classify, thus making the computation easier.

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 2.4b, it changes all minus values to zero.

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.

2.2.2 Other Paradigms

There are many other Machine Learning paradigms apart from Neural Networks, each of which I will give a brief introduction.

Decision Trees

"Decision tree learning is method for approcimating disrete-valued target functions, in which the learned function is represented by a descision tree" Mitchell [5]. There are many different classification algorithms that can create decision trees from supervised learning.

Meta/Ensemble Classifiers

Logistic Regression

Support Vector Machine

Regression Analysis

Unsupervised Learning

Reinforcement Learning

2.3 Overview of Machine Vision Approaches to Identification and Classification

There have been many attempts of identification and classification by many different researchers over the last number of years. There some approaches that decouple the two tasks of identification and classification from one another but mostly, researchers have attempted the two together. Sometimes, by semantic segmentation and in others simply building a classifier for the image without taking into account, the need for object identification.

2.3.1 Region Based Convolutional Neural Networks

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. I will analyse this papers, their results (2.1) and the changes made through each iteration.

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 auxilary task, followed by domain-specific fine-tuning, yields a significant performance boost" Girshick et al. [2].

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 2.1.

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 [1].

This paper analyses why RCNN Girshick et al. [2] 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 47s per image Girshick [1].

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 [1].

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. [6].

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 seperately, 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. [6]. This paper follows this approach.

Table 2.1: 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%

- Approximate joint training.
- Non- approximate joint training.

Mask RCNN

The most recent paper on this topic was also written by Ross Girshik while working with Facebook AI Research He et al. [3]. Mask RCNN "extends Faster RCNN by adding a branch for predicting an object mask in parallel with the existing branch for bounding box regression" He et al. [3].

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.

2.3.2 Fully Convolutional Neural Networks for Semantic Segmentation

2.3.3 Image Segmentation

Graph Based Segmentation

Sift/Surf

2.3.4 Convolutional Neural Networks for Classification

2.4 Technologies

2.4.1 Tensorflow

2.4.2 Jupyter

2.5 Evaluating the Output

Chapter 3

Experiments

Chapter 4

Empirical Studies

Chapter 5

Discussion and Conclusion

Bibliography

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

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

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

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

[5] Tom M. Mitchell. *Machine Learning*. International Edition 1997. NewYork: The McGraw-Hill Companies, Inc., 1997, pp. 52–78.

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