Development and Evaluation of a Visual Attention Model with Python and Tensorflow

Oleg Yarin

A thesis presented for the degree of
Bachelor of Science

First supervisor: Prof. Dr. Christian Herta
Second supervisor: Dr. Vera Hollink

Applied computer science
HTW Berlin
Germany
31-05-2017

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Neural network approaches have received much attention in the last several years. It's becoming a popular choice for performing various tasks like speech and image recognition, object detections etc. as these methods have dramatically increased accuracy compared to traditional machine learning approaches. However, achieving high accuracy on recognition tasks is still computationally expensive and needs improvements in performance. This study will be a close resemblance of the recurrent neural network of visual attention which is able to extract necessary information from an image by looking at it in low resolution, and then adaptively select parts that are most relevant for a task [1].

The idea of visual attention was inspired by how human perception works. Humans do not perceive a visual scene as a whole but focus on parts of the scene that gives the most useful information to them. Humans are also capable of combining information from different parts of a picture. They then connect it to build a subjective knowledge of the picture (or sequence of pictures) [2]. Taking into account these properties, researchers from google Deepmind build a model which can be described as follows:

> Instead of processing an entire image or even bounding box at once, at each step, the model selects the next location to attend to; based on past information and the demands of the task.... The model is a recurrent neural network (RNN) which processes inputs sequentially, attending to different locations within the images (or video frames) one at a time, and incrementally combines information from these fixations to build up a dynamic internal representation of the scene or environment.[1]

One of the main advantage of this model, is that the computation required is controlled and is independent of the input image size. Deepmind's researchers evaluated their model on several image classification and dynamic visual control problems which showed a better performance when compared with convolution neural network[2].

The evidence from this study suggests application of this model on large scale object recognition as well as classification of sequence of images, which will be a great fit since the model's performance is not dependent on the size of an input object.

The main aim of this study is to extend the current knowledge of the work mentioned above and build a model which will be able to classify a set of images and develop appropriate prototype system since it can be useful in a variety of areas. However, the current work is limited by low-resolution images and mostly will concentrate on classifying a group of objects as this restriction will reduce complexity of the task and therefore reach better results on a task of classifying a group of images.

1.1 Motivation

This approach to classify a group of images has a potential to help with automated detection and classification of breast cancer metastases, which is the main concern of camelyon challenge [3]. Camelyon challenge is an inspiration for this work since pathologist's efforts along with the assistance of automated detection system will reduce significantly not only the workload of pathologists but the human error rate in diagnosis as well.

This work will be the first step in building software that will be capable of classifying whole-slide images of histological lymph node at the patient level. That is, bringing together estimations from multiple lymph node slides into a single outcome.

Digital pathology is a very attractive field for machine learning researchers since whole-slide images have a very high resolution and are typically about 200000 x 100000 pixels. To give you some sense of data, camelyon challenge provides data for 200 patients, where each patient has 5 different slides. It means that in total they release about 1000 slides and that is 55.88gb of uncompressed data [3].

It is quite clear that using CNN for this task is computationally very expensive. Applying model of visual attention promises to solve the issue of high-resolution pictures at a computational level. Therefore making an extensible piece of software, that will allow further improvements is also one of the main concerns of this work.

Chapter 2

Theory

2.1 Artificial Neural Networks

Why Neural Networks? The problem that we trying to solve in this paper is known as pattern recognition problem. The problem can be described as searching for patterns in data like image, text, sound and etc. An example of the problem can be to find a tree in a picture. Artifical Neural networks have recently shown good performance and accuracy at recognizing objects [4].

What is Neural Network? Artificial Neural Network(ANN), often referred just as Neural Network(NN), is in simply words a computational model, which was inspired by how human/animal brain works. Artificial NN is modeled after the neuronal structure if the brain's cortex[5]. Though the inspiration was from the brain, it's indeed much much simpler than brain in terms of number of neurons that is used in ANN.

What is meant here by computation.

Basic concept of neural network There are a good variety of

Bibliography

[1] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent Models of Visual Attention," *Nips-2014*, pp. 1–9, jun 2014.

[2] P. Goldsborough, "A Tour of TensorFlow Proseminar Data Mining,"

[3] "CAMELYON17 - Home."

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances In Neural Information Processing Systems*, pp. 1–9, 2012.

[5] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.