Applying deep learning to develop endoscopic imaging technology

C H Weng* and C Y Tan

Medical Device Innovation Center, National Cheng Kung University, 138 Shengli Rd, North District, Tainan City 704, Taiwan
*E-mail: z9809260@mail.ncku.edu.tw

Abstract. As an important technology means in the arena of image categorisation, deep learning has broad application prospects, to carry out research on image classification technology has important practical significance and theoretical value for promoting the development of artificial intelligence and computer vision, this article gives a comment for the application of deep learning in image designation. Analyse the research findings obtained in recent years in the domain of image determination such as medical image identification and remote sensing image recognition, discuss the development trend and existing research of deep learning in the arena of image categorisation, points out that the effective use of deep learning technology to classify large sample data, with using the supervised and unsupervised learning to designate images, how to effectively determinate medical images and strengthen the theoretical nature of the models are further directions of research in this field. This paper presents a study based on deep learning, with the aim of practical research being by using Convolutional Neural Networks to identify endoscopic groupings in datasets. The proposed method can automatically detect the features of endoscopy images under the condition recognised by the algorithms and achieve the correct and fast categorisation results.

1. Introduction

Applications of image classification has always been the best known and most stunning applications of deep learning including handwritten digit recognition, object identification and so on. In the image-related applications, the most critical technology is the Convolutional Neural Network (CNN), therefore the convolutional neural networks are become synonymous with deep learning in the minds of general people. Convolution neural network is a hierarchical classification type structure, each of its module is composed of convolution layer and pooling layer (Lecun, 1989), by the addition of multiple layers of deep neural network continuous superposition of modules to form a deep model. The entire convolutional neural network through three important ideas to help improve the system of mechanical learning: equivariant representation, parameter sharing and sparse interaction (Goodfellow, 2016). We say that a function having an equivariant property, meaning that when the input changes, the output changes in the same way. Parameter sharing refers to the same kernel, each element acts is the same weight in different locations. This means that in the process of convolution, the model only needs to learn a fixed set of parameters, rather than the weights used by each element in different locations is independent. Therefore, this will also greatly improve the training effectiveness of the model. While the mechanism of parameter sharing coupled with the appropriate pooling strategy, also contributed to the convolution neural network with some invariant features for local translation which can be applied to image processing or speech recognition, especially when concerned with a certain feature whether is there appear, rather than about a specific location where it appears. Sparse interaction is also called as sparse connectivity, mainly using several kernels based on the self-
defining kernel size to locally link the two-tier network, such that the parameters in the entire model to be stored are less, can effectively improve the computational efficiency and reduce the amount of computation. Currently convolution neural network has been successfully applied to image analysis (Ciresan, Gambardella & Schmidhuber, 2012), natural language processing (Wallace & Zhang, 2015) and speech recognition (Abdel-Hamid, Deng & Yu, 2013). [1]

The CNN is constituted of convolution layer, pooling layer and full connection layer, the convolution layer is like the attribute mask in image classification, to know what the features of a picture, the filter can be used to do convolution operation on the picture, convolution operation is to perform inner convolution operation on two pictures of the same size. In the CNN the machine will generate a random number of feature masks, these features mask called kernel in the neural network, by convoluting each block of every picture in the training set, can know which blocks have this feature, after the convolution operation the activation function will be used, to allow the neural network can solve complex problems better, the common activation functions are rectified linear unit, sigmoid; the pooling layer is to reduce the picture size, so that the next layer of the convolution layer can capture finer features and reduce noise; the fully connected layer is flattening the outcomes of the previous layer to fully link to the deep neural network, which can classify the captured feature vectors, in the full connection layer the dropout layer will be added for the prevention of overfitting, it will randomly make several neurons stop functioning, but still maintain the weight values, the number of neurons in the last layer will be the same as the number of classifications, that will also normalise the sigmoid function through the activation function, which normalise all the results of the last layer, with the values between 0 and 1. [2]

2. Related works

2.1. The operating principles of convolutional neural network

When there is a breakthrough in deep learning, these advances are in all probabilities related to Convolutional Neural Network. CNN also known as CNNs or ConvNets, is currently the main development force in the field of deep neural network, it can even be more accurate than humans in image identification. If there is any way to live up to the expectations of deep learning, CNN is definitely the first choice.

Whenever CNN distinguishes a new picture, CNN will compare anywhere in the picture, without knowing where the various parts of the features are. To calculate how many consistent features there are in the whole picture, we create a screening mechanism here. The mathematical principle behind this mechanism is called convolution, which is the origin of the CNN name. [3]

Another powerful tool used by CNNs is pooling. Pooling is a method of compressing picture and retaining important information, its operation principle to be understood only need by the mathematical level of elementary school. Pooling will choice different frame on the image and select a maximum value in this window range.

After the original picture is pooled, the number of pixels it contains will be reduced to one-fourth of the original picture, but because the pooled picture contains the maximum value of each range in the original picture, it still retains the degree of consistency among each range and every feature. In other words, the pooled information is more focused on whether there is a presence of matching features in the picture, rather than where these features exist in the picture.

Ultimately CNNs also have a secret weapon - the fully connected layers. The fully connected layers aggregate the filtered pictures from the higher order layers and convert this feature information into vote.

3. Research methodology

Chieflly, TensorFlow is the open-source software function library for mainly to establish of machine learning model, which is primarily used for data flow graphs to perform numerical computation, the edges of the graph represent the communication between multidimensional data arrays, while the nodes in the graph represent the mathematical operations. The flexible architecture allows using a single API (Application Programming Interface) to deploy operations to one or more CPUs or GPUs
on a desktop, mobile or server device, so there are three main priorities, these are respectively: (1) Data flow graphs (2) Library for numerical computation (3) Open source.

Then the main goal of TensorFlow is not to offer an existing machine learning solution, but to provide a set of suites that allows users through mathematical methods to define models from scratch, so that to facilitate users with a certain technical background can quickly create customised and high flexible models. Further the operational model of TensorFlow is primary directed graph, each edge in the graph represents a matrix, number or tensor, each node in the graph represents a function or operation. [4]

3.1. Concept of model (VGG network structure)
Artificial intelligence especially deep learning has been extremely successful in many applications such as object classification, face recognition, target identification etc. The architecture of the common deep learning network is shown in the figure below. The input data will get high level feature after multiple layers of computation, and then the result will be obtained through the fully connected layer. The number of convolution kernels in each convolutional cluster is the same, the more backward of the convolutional groups the more the number of convolution kernels: 64 - 128 - 256 - 512 – 512.

3.2. Hyperparameter tuning
Each set of training samples contains two items, each network with parameters and structures are identical. This turns the image classification problem transformed into a binary classification problem. [5]

In this paper image categorisation is used to determine whether the target is oesophagus or not, it is a binary classification problem, which is opting the sigmoid function to judge, as shown below -

| Problem Type | Output Type               | Final Activation Function | Loss Function              |
|--------------|---------------------------|---------------------------|----------------------------|
| Regression   | Numerical value           | Linear                    | Mean Squared Error         |
| Classification | Binary outcome           | Sigmoid                   | Binary Cross Entropy       |
| Classification | Single label, multiple classes | Softmax               | Cross Entropy              |
| Classification | Multiple labels, multiple classes | Sigmoid               | Binary Cross Entropy       |

4. Experimental results
The experiment in this paper is carried out on Windows 10 Enterprise, Intel (R) Core (TM) i9-10900K @ 3.70 GHz with Nvidia GeForce RTX 3080, based on the platform of Anaconda with Spyder Python 3.7.3, it is programming through the Python language and Python library function. In the Endoscopic image databases, a total of 6 different kinds, each image has 1000 different samples, a total of 6000 endoscopy images, each time 2 types are selected as a training subject, for example, here a total of 500 training samples, then a total of 200 validation samples, with a total of 300 testing samples.
Figure 1. Datasets on the first experiment.
(The multiple images scenario of esophagus and no-esophagus)

Figure 2. Graphs showing training and validation accuracies/losses.
(First experiment)
Figure 3. Datasets on the second experiment.
(The multiple images scenario of esophagitis and normal pylorus)

Figure 4. Graphs showing training and validation accuracies/losses.
(Second experiment)
Figure 5. Datasets on the third experiment.
(The multiple images scenario of normal z-line and ulcerative colitis)

Figure 6. Graphs showing training and validation accuracies/losses.
(Third experiment)
Table 2. The processing item and classification rate of three experiments.

| Item Analysis       | Number of Training Samples | Percentage Rate (%) |
|---------------------|-----------------------------|----------------------|
| First Experiment    | 1000 (500 x 2)              | 91.33                |
| Second Experiment   | 1000 (500 x 2)              | 94.50                |
| Third Experiment    | 1000 (500 x 2)              | 99.1                 |

4.1. Image identification analysis and review

This essay proposed an algorithm for automatically determining the category of endoscopic images, which is mainly divided into two parts - feature extraction and output classification. The procedure of this method is as follows: firstly, the ReLU function is used in CNN, followed by the sigmoid function; secondly, image features are extracted; finally, the last layer of CNN will be treated as a linear classifier for identifying these images. This method has the traits of high recognition rate and short operation time. CNN still highly maintains immutability in extracting complex features of images, regardless of its deformation, rotation, scaling or other forms of deformity. Thus the model has learned rich feature representations for a variety of images. [6]

Image classification is a significant step in the broad domain of artificial intelligence and computer vision. By expanding the number of convolutional layers and hidden neurons, the results can be made more precise. The model can be used to identify objects from blurred images. Image categorisation is a superior modelling topic for the study of neural networks, which provides a good method for exploiting more progressed deep learning technologies.

Sigmoid-although the nonlinearity will be increased to a certain extent, it is prone to supersaturation, that is to say the edge interval is longer, the effective range is shorter, it will cause a lot of data to fall in the edge interval, the gradient line will be lying flat.

CNNs have strong feature extraction capabilities, which been used to extract features from endoscopy images, the proposed method can achieve good performance.

5. Summary

Here, this study will use deep learning that is different from previous fixed algorithms, with using deep learning way to implement image classification, by through the training of neural network weights, due to the size of its training model is mainly changed according to the structure of neural network, under the condition of large amount of training data, compared with the local binary method, can effectively decrease the size of the training model, and reduce the model loading time of the identification system.

The outcome of this research will be used in further studies to develop analysis and experimentation for vision processing improvements which the deep learning algorithms will help to advance the image science. To meet a steadily growing trend with complex and diversity, the science technology concerning the medical image processing requires perpetual exploration and research.

References

[1] Agrawal, Carreira, Jitendra, Joao, Malik and Pulkit 2015 Learning to see by moving. In Proceedings of the IEEE International Conference on Computer Vision, pp 37 - 45
[2] Hinton G, Bengio Y and LeCun Y, Deep learning’ May 2015 Nature, vol 521 no 7553 pp 436 -444
[3] Eigen D, Mathieu M, Sermanet P, Fergus R, Zhang X and Le C Y 2014 Overfeat: Integrated recognition, localization and detection using convolutional networks. In ICLR
[4] Zisserman A and Simonyan K 2015 Very deep convolutional networks for large-scale image recognition, in International Conference on Learning Representations
[5] Nguyen A, Yosinski J and Clune J 2016 Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. arXiv preprint arXiv:1602.03616
[6] Alvarez J M, Porikli F and Zhou H 2016 Less Is More: Towards Compact CNNs pp 662-677. Springer International Publishing