DEGREE PROJECT IN COMPUTER ENGINEERING,
FIRST CYCLE, 15 CREDITS
STOCKHOLM, SWEDEN 2017

Training Set Size for Skin Cancer Classification Using Google's Inception v3

PATRIC RIDELL

HENNING SPETT

KTH ROYAL INSTITUTE OF TECHNOLOGY
SCHOOL OF COMPUTER SCIENCE AND COMMUNICATION

Training Set Size for Skin Cancer Classification Using Google's Inception v3

PATRIC RIDELL AND HENNING SPETT

Bachelor in Computer Science
Date: June 5, 2017
Supervisor: Pawel Herman
Examiner: Örjan Ekeberg
Swedish title: Storlek på träningsdata vid klassificering av hudcancer med Googles Inception v3
School of Computer Science and Communication

Abstract

Today, *computer aided diagnosis* (CADx) is a common occurrence in hospitals. With image recognition, computers are able to detect signs of breast cancer and different kinds of lung diseases. For a *convolutional neural network* (CNN) that classifies images, the accuracy depends on the amount of data it is trained on and performs better as the amount of training data increase. This introduces a need for relevant images for the classes the classifier is supposed to differentiate between. However, when input data is increased, so does the computational cost, leading to a trade-off between accuracy and computational time. In a study by Cho et al. the accuracy improvement stagnates, when comparing the accuracy with different amounts of training data. This creates interests in finding that point of stagnation, since further increase of input data would lead to longer computational time but little effect on the accuracy. In this study, the pre-trained CNN Google Inception v3 is retrained with various amounts of skin lesion images. The objective is to detect whether the image represents a benign nevus or malignant melanoma. When comparing the accuracy for these different training sessions it is concluded that the accuracy increases when trained with more data. However, a stagnation point for the accuracy is not found.

Sammanfattning

Datorstödd diagnostik (CADx) är idag vanligt förekommande inom sjukvården. Med datorseende är det möjligt att undersöka huruvida bilder påvisar tecken för till exempel bröstcancer och lungsjukdomar. Träffsäkerheten för *convolutional neural networks* (CNN) klassificering av sjukdomar beror till viss del på hur mycket data det tränats på. Stora datamängder är en förutsättning för att CNN ska kunna ge pålitliga diagnoser. En stor mängd indata innebär dock att bräkningstiden ökar. Detta medför att det kan behöva göras en avvägning mellan träffsäkerhet och beräkningstid. Cho et al. har i en studie visat att träffsäkerhetens förbättring stagnerar när mängden indata ökar. Det finns därför ett intresse i att hitta den punkt där träffsäkerheten stagnerar, eftersom ytterligare ökning av indata skulle innebära längre beräkningstid men med liten förbättring i träffsäkerhet. I denna uppsats tränas Googles förtränade CNN om på varierade mängder bilder på hudfläckar, i syfte att avgöra om en bild föreställande en hudfläck visar tecken på malignt melanom eller om den bedöms vara godartad. Studiens resultat ger indikationer på att träffsäkerheten för klassificeraren förbättras när mängden träningsdata ökar. Däremot finns inte underlag för att fastställa en punkt då träffsäkerheten stagnerar.

Contents

Chapter 1

Introduction

Today, *computer aided diagnosis* (CADx) has become a common occurrence in the clinical work of medical professionals. The availability of CADx has helped both patients and medical staff to detect and treat diseases at an earlier stage. In radiology departments, computers are used to detect breast cancer and different kinds of lung diseases [8] and recently, *convolutional neural networks* (CNN) has proven to be as accurate at classifying cancerous/non-cancerous skin lesions as trained dermatologists [9]. Most of the available CADx systems require medical grade instruments and personnel. Skin cancer is one of few classes of serious diseases that does not necessarily need such resources for a first preliminary diagnosis. Since the symptoms are visible on the surface of the skin, hence the prospective applications are promising. A mobile application could make skin cancer classification available for a wide range of patients that may otherwise suffer or even die without early detection of the disease.

The potential for computerized visual examination of skin lesions has been made possible by the recent advances in image recognition. In the ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC 2012), Krizhevsky et al. displayed a significant breakthrough with the use of a convolutional neural network (known as AlexNet[19]) reaching a 15.4% top-5 accuracy. When classifying several different classes, a common practice is to measure the accuracy of a network with top-5 accuracy, which states how often a correct label is not within the network's 5 top predictions. The runner up in ILSVRC 2012 had a top-5 accuracy of 26.2% which is 10.8 percentage points less accurate than AlexNet. Since 2012, CNN have made great strides and been the state of the art when it comes to a variety of computer vision tasks [30].

Two common attributes of the current state of the art CNN are that their training is a computationally intensive task and they require large sets of training data. Transfer learning offers an interesting solution to this problem. Instead of training a network from scratch the outer layers are fine-tuned (trained on a specific task) while keeping the pre-trained inner layers fixed. This approach can significantly lower the training time needed. The fine-tuning can be used to train the network on a specific task with a smaller training set. However, if the restrictions on available training data are severe, how much training data is enough?

1.1 Problem statement

This thesis intends to investigate how the amount of training images affects the accuracy on a specific CNN, namely Google's Inception v3 [13], which is pre-trained on approximately 1.26 million images of generic object classes. Furthermore, the aim is to study if there is a point where increased data does not further improve the classifier's accuracy. This point is valuable since additional training would lead to unnecessary computational cost and time waste. It can also give insight into how regulations for future implementations in clinical settings are asserted. This leads to a two part research question:

- How does the accuracy of the Google Inception v3 vary with increased amount of training data?
- Is there a point of stagnation in accuracy for Google Inception v3?

1.2 Scope

To examine the research questions the analysis is made with respect to a retrained model of the Inception v3. Furthermore, hyperparameters for the training algorithm is kept the same while varying the image amount. There is little preprocessing of the data fed to the network by oversampling a small part of input data and flipping them horizontally.

The data set is assembled from the ISIC-archive [18], which is an open source archive of labeled skin lesion images. In total, 2679 images of benign nevi and malignant melanoma were available to use for training, validation and testing. The CNN classifies whether a skin lesion displays signs of malignant melanoma or benign nevus (i.e. a binary classifier), since only a fraction of available images represented other diagnoses. No other diagnostic labels are considered.

Chapter 2

Background

This chapter introduces a technical background and related work. The first section gives a brief description about skin cancer, followed by clinical terms used to measure the accuracy of clinical tests. The second section presents artificial neural networks, with a focus on convolutional neural networks. Furthermore, the concept of transfer learning and the architecture of Google's Inception v3 are described. In the final section, Related work, studies which used machine learning to detect skin cancer are presented, along with studies that pre-process images to enhance performance of classifiers.

2.1 Skin cancer

Cancer is a collective name used for a group of diseases where cells start to divide in an uncontrolled manner, creating a tumor (in most cases) [17]. A tumor can be benign, where it can grow large but does not spread into the surrounding tissues, or malignant, where it does. Most benign tumors are not life threatening and seldom grow back once surgically removed, whereas malignant tumors are more dangerous and can spread not only to connected tissues, but can also break off and be transported via the bloodstream or the lymphatic system to develop in other places of the body. In total, there are more than 100 types of cancer, among which, skin cancer is one of the most common ones [17].

Skin cancer, the abnormal growth of skin cells, mostly affects skin that is often exposed to sunlight. It affects people of all ethnicities, however, people with skin that is easily sunburnt have a higher risk of developing the disease [7]. There are several different types of skin cancer, that can be developed in the outer layer of the skin (called epidermis). The layer consist of three types of cells: basal cells, squamous cells, and melanocytes [16].

The most common type of skin cancer is *basal cell carcinoma* [7]. It is developed in the basal cells and is recognized as a small and slowly growing, white (or skin colored), lump that may bleed [15]. The second most common type of skin cancer is called *squamous cell carcinoma* [7]. The cancer is found in cells called squamous cells that make up the surface of the skin and can also be found in the lining of the digestive tract, the respiratory system, as well as other hollow organs. [15]. The deadliest form of skin cancer is *melanoma* [7]. It is a more aggressive form of cancer than the two previously mentioned and is developed in the melanocytes (the cells that produce melanin, which gives skin its color). It can begin as a nevus (a skin mole made up of a cluster of melanocytes) that changes shape, color, starts to itch or even bleed. Together, the basal cell carcinoma and squamous cell carcinoma, are usually referred to as nonmelanoma skin cancer, as they are usually not spread to other parts of the body and are easily treatable. Melanoma makes up only 2% of the total amount of skin cancer cases, but it is the most common cause of skin cancer related deaths [16].

The following characteristics are some of the factors that a CNN will use to differentiate between the two classes (benign nevus and malignant melanoma). There are different types of nevi; the common nevus (common mole) and the dysplastic nevus. The distinguishing characteristics of a common mole is that it is usually small (<5mm diameter) with distinct edges and a smooth surface, the color (usually brown, tan, or pink) is even across the mole and the shape is round or oval [14]. A dysplastic nevus, sometimes referred to as an atypical mole, can be larger than the common mole, the color of the nevus does not have to be uniform, it can have a mixture of colors [14]. Both a common and dysplastic nevus may develop melanoma, but most often they remain stable [14] and will then be classified as benign nevus.

2.1.1 Clinical tests and terms

To test for a certain disease a clinical test is used. Within clinical testing there are four fundamental terms (see Table 1) used to describe the result produced by a test. These, in turn, are used to define sensitivity and specificity, two performance metrics used to evaluate a test's performance. Within the context of this study the test will consist of the CNN receiving an image as input and returning a label, either benign or malignant. The result can then be described as one of the four terms in Table 1.

True positive	a patient with the disease receives a positive diagnosis (i.e. the test confirms the presence of the disease)
False positive	a patient without the disease receives a positive diagnosis
True negative	a patient without the disease receives a negative diagnosis (i.e. the test refutes the presence of the disease).
False negative	a patient with the disease receives a negative diagnosis

Table 1. Fundamental clinical terms [21].

Sensitivity is a measure of a test's ability to identify patients that have the disease the test is screening for. Likewise, specificity measures how well the test can identify those without the disease. A sensitivity or specificity of 100% means that the test can identify all the patients with or without the disease.

$$Sensitivity = \frac{True\ positives}{True\ positives + False\ negatives} \tag{2.1}$$

$$Specificity = \frac{True\ negatives}{True\ negatives + False\ positives} \tag{2.2}$$

2.2 Artificial neural network

An *artificial neural network* (ANN) is a set of nodes connected through edges. In a feed-forward network, the nodes form layers as they connect with neighboring layers as a bipartite graph (see Figure 1). Activations are sent from the first layer (input layer) to the last layer (output layer). The output for each hidden node is manipulated with a transfer function, this introduces nonlinearity which stabilizes the network by suppressing the impact of divergent neurons [31].

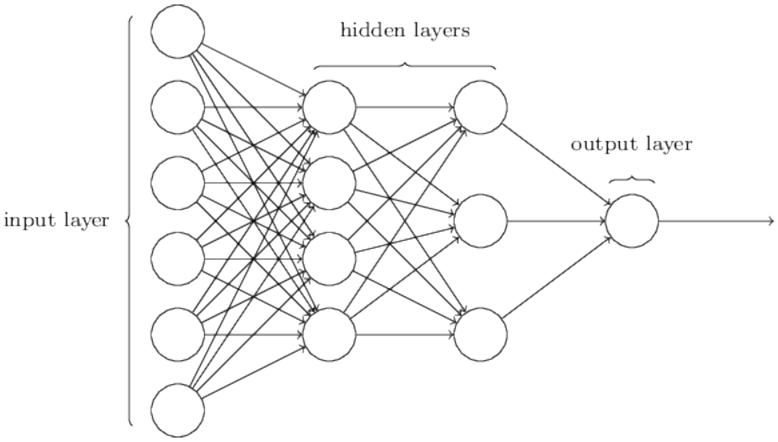

Figure 1. Example of a feed forward ANN, with six nodes in the input layer, four nodes in the first hidden layer, three nodes in the second hidden layer and one node in the output layer.
Source: Neural Networks and Deep Learning [24].

2.2.1 Convolutional neural networks

Convolutional neural networks or CNN are ANN which processes data with known topology [10]. When applying data of images (often matrices of pixels) the spatiality of each pixel has to be considered. In traditional neural networks every output interacts with every input. Therefore, backpropagation learning leads to a gradient that either vanishes or explodes when propagated through many layers making deeper architectures of nets impossible to train. To overcome this problem, CNN typically have sparse connections by using filters with lower dimensions than the input data [10]. Furthermore, algorithms using CNN often implement a greedy approach meaning that the net is trained layer-wise and preserves the complexity of each layer by freezing its weights once trained and sending its output as input to the next layer. Thus, the convolutional layers (called feature extractors) in early layers serves to detect simple shapes (for example lines and circles) of given image while the later convolutional layers detect more complex features. Figure 2 briefly explains how an image is fed to a CNN with several layers. The image of $MxNxC$ dimensions is translated to $1xK$ vector with probability values for which ever of the K classes the image represents.

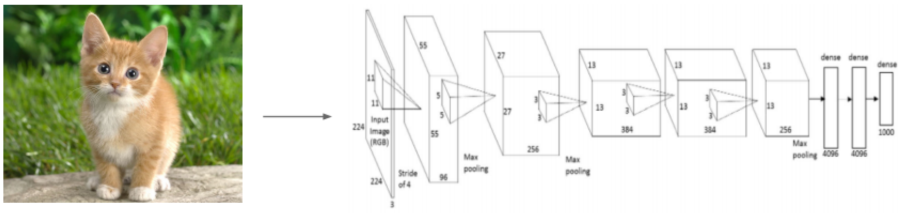

Figure 2. Schematic design of a image fed into a CNN. It consists of several convolutional and max-pooling layers finished by fully connected layers.
Source: Visualizing, Training & Designing ConvNets [28].

CNN consist of three different kinds of layers; convolutional layers, pooling layers and fully-connected layers. *Convolutional layers* are sets of learnable filters [5]. When scanning through its input, a 2-dimensional activation map is produced describing which filter is active at each spatial location of the input [5]. The purpose of *pooling layers* is to reduce the spatial size of the data representation, hence reducing the amount of computations and the risk of overfitting [5]. *Fully-connected layers* utilize the nodes in the previous layer to translate data to a non spatially dependent one dimensional data structure [5].

2.2.2 Transfer learning

Training a CNN with many layers from scratch can take a considerable amount of time because of the large computational cost. To combat this issue transfer learning has been developed. The method decreases the time it takes to train a network for a specific set of object classes. To start with, a fully trained network is used and retrained on a new set of objects, with the main purpose of having a general network that can be adapted to suit different needs without major restructuring and training. To achieve this a common practice is to use ImageNet's [20] training set, which is an academic benchmark set specifically for computer vision, consisting of 1.26 million images and 1000 generic object classes.

Traditional machine learning assumes that the data for testing has the same distribution as the data the net were trained on, in many cases this is not true [6]. However, the knowledge gained from the source task (the pre-training) may be transferred to serve as valuable knowledge for a domain the network has not previously encountered. Depending on the amount of data available in the new domain and the variation between the new and old domain, the net can either serve as a feature extractor by replacing the last fully-connected layer with something more appropriate for the new domain or by fine-tuning the network by adjusting weights of suitable amount of layers [27].

2.2.3 Google's Inception v3 architecture

Google's Inception v3 [30] is an architecture that provides a high performance network with a relatively low computational cost. To measure the accuracy there are two main measurements used in the literature [30]; top-5 error rate and top-1 error rate. These measure the rate at which the network fails to include the correct class in the top-5 and the top-1 output respectively. Inception v3 achieves a 5.6% top-5 error rate and a 21.2% top-1 error rate. To put these rates into context; the AlexNet [19] achieves a top-5 error rate of 15.3% and the original Inception (GoogLeNet) [29] achieves 6.67% in the same category.

The architecture of Google's Inception v3 emphasizes the importance of memory management and restrictions on computational capacity. This could mean that Inception v3 is more suitable for mobile applications than other more computationally expensive architectures.

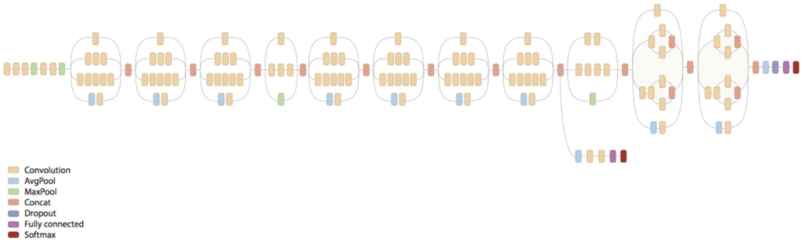

Figure 3. Schematic diagram of the Inception v3 architecture.
Source: Inception in Tensorflow [13].

2.3 Related work

In a recent study, Esteva et al. [9] compared the accuracy of deep learning to several dermatologists when classifying images of skin lesions. A training set of 129 450 images was used to train the network specifically on skin lesions, after being pre-trained on 1.28 million images from the ImageNet dataset. The network reached an accuracy of 72.1%, that is at least as good as the average of 23 board certified dermatologists [9]. Another study published by Nylund [25] also concluded that CNN can perform on the same level as dermatologists. Similar to the work of Esteva et al. [9], Nylund [25] used a network that was pre-trained on the ImageNet dataset, achieving an accuracy of 89.3%. In the study several different datasets, with a total of over 20 000 images, were used to retrain the network. In contrast, Majtner et al. [22] used significantly less training images (900) to train a skin lesion classifier which reached an accuracy of 82.6%. The amount of images used varies in the different studies, which reflects the difficulty to assemble a sufficient data set. Furthermore, while an increase in input data implies increased computations it is shown [3] that the accuracy of a classifier converges when adding more input data.

To increase the amount of training data, recent studies investigate techniques to both augment data and to utilize unlabeled data when pre-training the network [23]. Masood et al. [23] proposed a model which benefits from unlabeled data as well as labeled data. It is a semi-supervised model, using both self-advised support vector machines (SVM) and deep belief neural networks. This model seems to outperform previously state-of-the-art methods such as SVM and expectation maximization (EM) in every tested set of training data. However, Masood et al. emphasize that the proposed semi-supervised model only performs better when the training data is both labeled and unlabeled [23]. The value in investigating models for semi-supervised learning is due to the lack of labeled data, according to the article [23]. This study provides new insights into the dependency between accuracy and the data used for training. The issue with restricted access to data is attempted to be solved by manipulating input data and extracting valuable features by hand (called handcrafted features) before applying a deep learning method. In studies by Majtner et al. [22] and Permaladha et al. [26] images of skin lesions are first translated from a colorized image to a thresholded image, only describing the border of the skin lesion. Parameters like asymmetry index and circularity index can then be derived through certain formulas and fed to the learning algorithm, instead of the actual image [26]. In comparison to other tested methods, the study concludes that their current method is more accurate with less computational time [26].

Classifiers have been shown to perform worse [11] when training is performed with an imbalance in the image amount for each class. Oversampling the training data can therefore result in a performance (accuracy) that is as good as if it was trained on balanced data [11].

Chapter 3

Method

In this chapter the datasets used to train and test the network are outlined. This is followed by a brief description of the training process and a motivation for the parameters used. Lastly, the measurements used to evaluate the network's performance are presented.

3.1 Dataset

The images used to train, validate and test the network comes from the open-source ISIC Dermoscopic Archive [18]. When assembling the data, only images that were labeled as either malignant melanoma or benign nevus was selected. This ruled out a minority of images that had several different diagnoses. The amassed dataset consists of 2679 images (1800 benign nevi and 879 malignant melanomas). Every image is labeled through histopathology, which refers to biopsy examination or examination of surgically extracted specimen by a pathologist [18]. No distinction based on background information about the person with the skin lesion in question was made.

To balance unevenly distributed training data, which may negatively impact the performance of the CNN, it has been shown that oversampling an underrepresented object class can be an effective solution [11]. Oversampling is a technique where the number of images is increased by slightly modifying an original image, for example by mirroring or rotating the image. Therefore, the dataset of malignant melanoma is oversampled by adding random images from the dataset but mirrored horizontally. This gives a wider range of possible images for training, thus easier to illustrate the difference in performance when varying the amount. As the risk for overfitting increases if oversampling is used excessively, the set of benign nevi is kept the same.

Image set	Malignant	Benign	Total
Original	879	1800	2679
Oversampled	1800	1800	3600

Table 2. The amount of images in each class, both when oversampled and not.

3.2 Network architecture

A Google's Inception v3 network [30], pre-trained on the ImageNet [12] training set was used (see section 2.2.3).

3.2.1 Training the CNN

The retraining of the CNN was performed on a Intel i7 3.0 GHz and was done in two separate stages. In the first stage, 25 images per class were added in separate training steps until a total of 200 images per class was used for training (resulting in 8 training steps). In the second stage, 200 images per class were added until all images of the training set, 1600 per class, had been added (also resulting in 8 different training steps). When performing the first stage of training, only original images were used. In the second stage the oversampled images were added to the data set. The separation into two different stages allows for comparison between the results from Cho et al. [3], where they saw a quick convergence in accuracy after approximately 200 images per class.

In addition to the training images, approximately 400 (adjusted as a percentage of the total amount of input images) were used as the test set, regardless of training set size. The test set was only used at the end of each training session to evaluate the final accuracy of the network. All images, both for training and testing, were randomly sampled from the data set.

The Inception v3 model along with the retrainer script are collected from the Tensorflow github repository [13]. The retraining script is built with Bazel [1] and run ten times for each image group with random images from the data set, which is divided into groups of benign or melanoma. Furthermore, the model is configured to make predictions whether the given input images are benign or melanoma.

3.2.2 Hyperparameters

Most of the hyperparameters were set to their default values. The exceptions were: the learning rate, testing percentage, training batch size and validation batch size. The learning rate is probably the most important hyperparameter to change if there is a time constraint to the training (i.e. when exhaustive parameter-testing is not an option) [2]. When fine-tuning a network the learning rate should be decreased [30]. Hence, it was changed from the default 0.01 to 0.001. The testing percentage (i.e. the percentage of input images that were separated from training images to form the test set) was varied to keep the test set size as close to 400 images as possible. This gives more consistent readings, especially in the first phase of stage 1. A 10% testing percentage at this point would give a test set size of 5 images. The training batch size (i.e. how many pictures used in each training iteration) was the default 100 in the second stage of training, however, in the first stage it was set to 10 since there were only a few training images. Lastly, the validation batch size (i.e. the amount of images used of validation in each epoch) was set to include all images in the validation set. This leads to more stable training results across the different training iterations [30]. Table 3 displays some of the main hyperparameters used for the experiments.

Learning rate	0.001
Number of epochs	4000
Training batch size	10/100
Number of test images	~400
Validation batch size	100%
Validation percentage	10%

Table 3. The parameters selected when training and testing.

3.3 Validation of results

An average of the *test accuracy* for every 10 iterations with a fixed image amount was calculated, as well as standard deviation to study the fluctuation between sessions. *Sensitivity* and *specificity* was also calculated as the average of the 10 training iterations. These measurements are further described in section 2.1.1.

Chapter 4

Results

This chapter presents the results from training stage 1 followed by stage 2.

4.1 Training stage 1

As seen in Table 4, the accuracy had a steady increase (with minor dips at 125 and 175 training images per class) over stage 1. The standard deviation decreases quickly during the first 3 steps of the training, as the results become more stable. The accuracy increases steadily when using 25-100 training images, and stabilizes on around 68% when 100-200 images are used. This is further illustrated in Figure 4.

# of images	Accuracy (S.D) [%]	Sensitivity (S.D.) [%]	Specificity (S.D.) [%]
25	55.6 (±4.55)	73.2 (±20.82)	39.4 (±27.83)
50	58.9 (±3.93)	65.5 (±9.70)	52.3 (±11.50)
75	64.2 (±2.23)	72.1 (±4.88)	56.5 (±6.64)
100	67.1 (±2.24)	72.6 (±6.05)	62.0 (±7.22)
125	66.8 (±2.82)	71.3 (±5.34)	62.3 (±3.80)
150	68.5 (±2.00)	69.7 (±6.41)	67.2 (±6.06)
175	67.4 (±1.34)	71.0 (±3.81)	64.0 (±4.06)
200	69.0 (±1.75)	68.6 (±3.04)	69.2 (±5.04)

Table 4. Average test accuracy, sensitivity and specificity at each training step during stage 1 of the training (training up to 200 images per class).

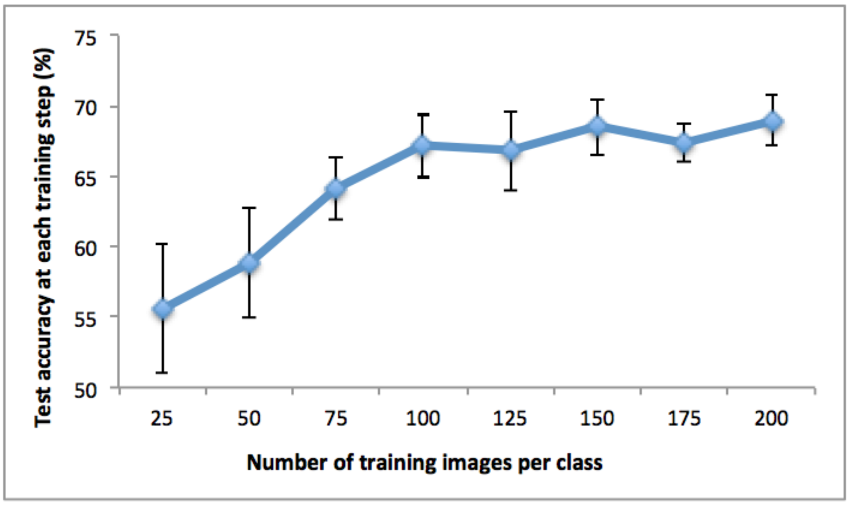

Figure 4. Average accuracy at training step 1-8, image amount used varies from 25 to 200. Error bars displaying the standard deviation within 10 iterations of each training step are shown.

Notice how the specificity increases steadily while the sensitivity varies little for the classifier trained on 25-200 images (Figure 5). The standard deviation quickly decreases after the first couple of training steps, this behavior can also be observed in Figure 4.

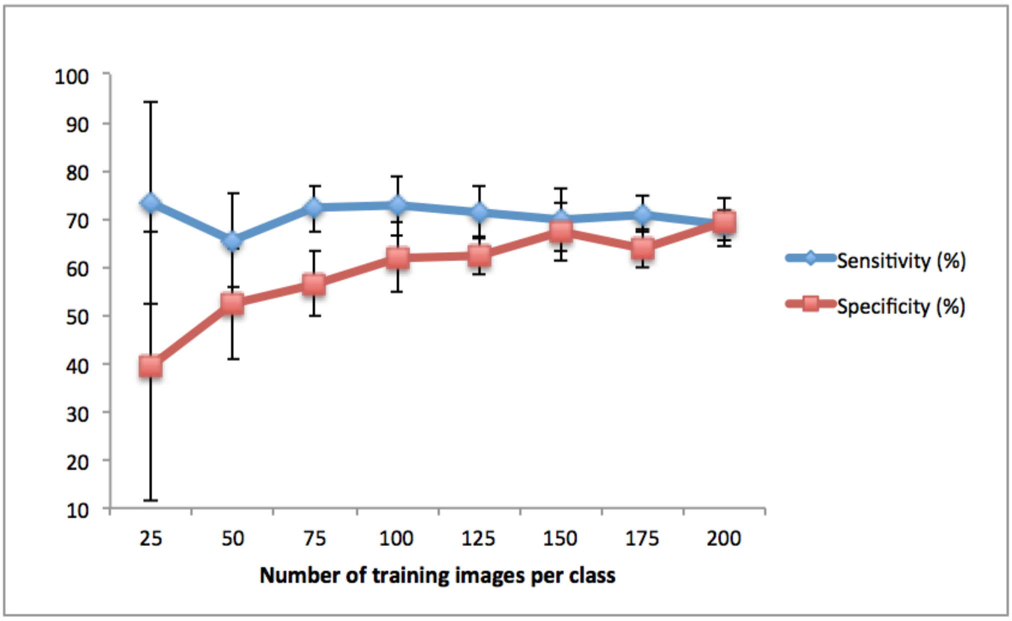

Figure 5. The average sensitivity and specificity for the classifier, trained with dataset sizes from 25 to 200 images. Error bars

displaying the standard deviation within 10 iterations of each training step are shown.

4.2 Training stage 2

Table 5 and Figure 6 displays the average accuracy and standard deviation when the classifier was trained on a large amount of images (200 to 1600). Just like Figure 4 (see section 4.1), the accuracy is improved when increasing input data, while the standard deviation is decreasing. A noticeable difference is that in Figure 6 the increase of accuracy does not seem to stagnate.

# of images	Accuracy (S.D) [%]	Sensitivity (S.D) [%]	Specificity (S.D) [%]
200	70.8 (±2.48)	69.3 (±2.93)	72.2 (±3.83)
400	71.9 (±1.76)	65.3 (±3.71)	74.7 (±2.82)
600	73.7 (±1.55)	71.9 (±2.22)	75.5 (±3.05)
800	75.3 (±1.38)	73.8 (±2.24)	77.7 (±4.99)
1000	74.3 (±1.59)	73.0 (±2.58)	76.3 (±3.19)
1200	75.3 (±1.56)	73.0 (±2.17)	76.7 (±2.91)
1400	76.9 (±1.14)	74.4 (±1.38)	79.1 (±2.26)
1600	77.5 (±0.63)	75.3 (±1.62)	79.4 (±2.22)

Table 5. Average test accuracy, sensitivity and specificity at each training step during stage 2 of the training (up to 1600 images per class).

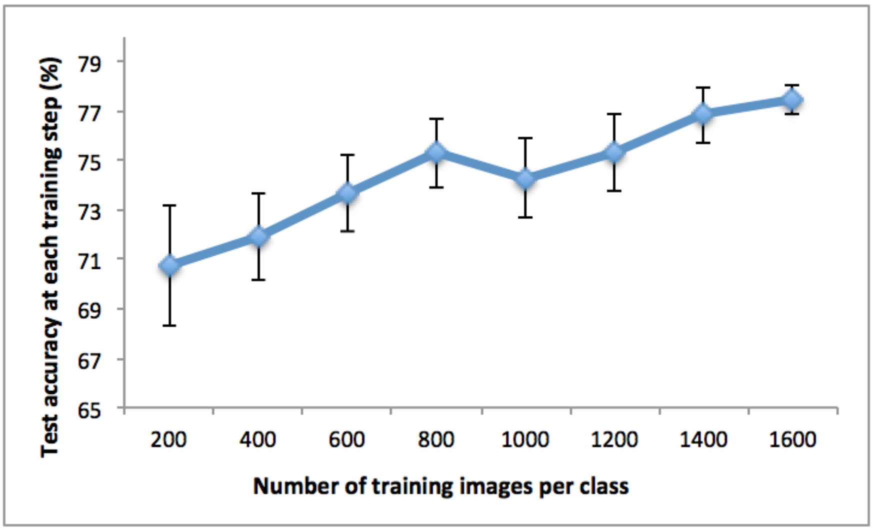

Figure 6. Average accuracy at training step 1-8, image amount used varies from 200 to 1600. Error bars displaying the standard deviation within 10 iterations of each training step are shown.

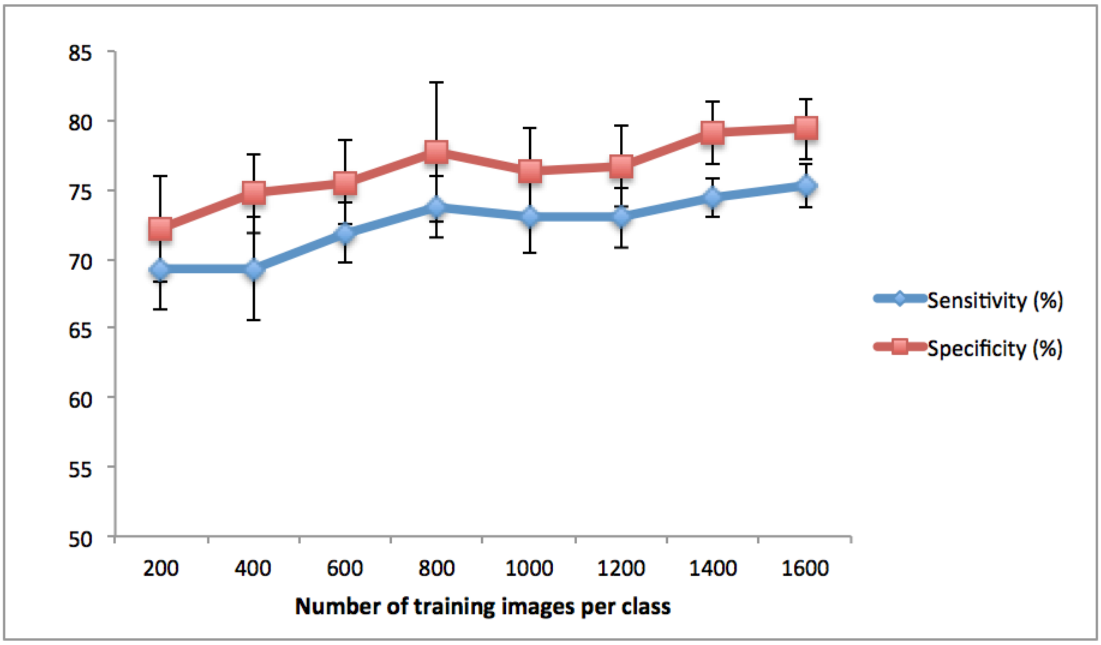

Figure 7. The average sensitivity and specificity for the classifier, trained with dataset sizes from 200 to 1600 images. Error bars displaying the standard deviation within 10 iterations of each training step are shown.

In contrast to Figure 5, the sensitivity and specificity seem to increase at approximately the same rate in Figure 7. Also, the standard deviation for both measurements stabilized after a few training steps in stage 1 (Figure 5), while in stage 2 (Figure 7), they remain consistent, although with some variation.

Chapter 5

Discussion

5.1 Results

As discussed by Cho et al. [4] a clear convergence is expected to be revealed when increasing the amount of input when training a classifier. While Figure 6 indicates an increase in accuracy, any convergence is not present. However, as shown by Figure 4, 5 and 6 there is a decrease in variance when training data is increased, which may be an indicator of improvement in the stability of the classifier.

As for now, no conclusion about a stagnation point can be drawn. However, Esteva et al. [9] used almost 130 000 images to retrain Inception v3 and reached an accuracy of 72.1% compared to 77.5% achieved in this study. An even better accuracy was attained by Nylund [25] who used approximately 20 000 original images along with oversampling methods producing 7 images for each original. The accuracy was 89.3%. Both of these studies used a network pre-trained on the ImageNet data set and the results were notably different. Nonetheless, the dermatologists in [9] achieved an accuracy of approximately 66%, which is lower than the CNN in [9], [25] and the results from this study.

To answer the research question about point of stagnation, the graph would have to clearly show a convergence of accuracy when increasing the dataset size, similar to the results presented in the study by Cho et al. [3]. With the size of the data set used, there is no clear sign of a convergence in accuracy. The main reason convergence is not visible in the result is probably due to the lack of larger datasets to train with. Larger datasets could be extracted from the total accessible database by decreasing the share of images given to the test dataset. This would however decrease the reliability of the given results. Further techniques to augment the data is also a possible way to increase the amount of training data, with a risk of overfitting. In this thesis oversampling was used purely with the intention to balance the training data, which in our experiments boosted the classifiers performance.

A possible reason that more training data is needed to see a convergence in accuracy, is the fact that the visual difference between a benign nevus and malignant melanoma is subtle. In contrast, the image classes studied by Cho et al. [3] were far easier to differentiate between, which means that most people would be able to tell them apart with only minute training. Whereas, even a trained dermatologist has a hard time differentiating between a benign nevus and malignant melanoma.

All hyperparameters for the tests were kept constant with the intention to only observe the impact of image amount. However, there is a possibility that a different set of parameters would have given a result more suitable to answer the research question.

5.2 Limitations

The main limitation in this study is the restricted availability of training data. However, one of the aims of the study was to see how the learning curve progressed through the training. It may still be useful to compare the results to training with a larger data set.

The Inception model was only modified by retraining the last layer, limiting the possibility for the classifier to adapt for our data. Since creating and testing the classifier is time consuming the observations are only measured with 16 different data set sizes, which may hide some relevant information about the accuracy variation.

5.3 Conclusion

This study has shown that with only a small set of training data a pre-trained Inception v3 model can reach an accuracy comparable to that of trained dermatologist. Additionally, it seems that more data is needed to reach the point where the accuracy is not improving significantly with increased input data.

5.4 Future work

Firstly, to see a convergence in the accuracy for the Inception v3, it would be valuable to test it on a larger training set. Secondly, it would be interesting to use the same training set and train the network from scratch and compare the results to our fine-tuned model. The same can be done with a fully pre-trained model where multiple layers are fine-tuned. This is especially interesting for a network if it is to be used in a clinical setting since the extra accuracy may be vital, and therefore worth the extra computational cost.

For clinical implementation of a CNN that classifies skin cancer it would be interesting to study how well dermatologists perform (to a greater extent than that of [9]). This would give health care institutions guidance as to when it is appropriate to use a CNN as a second opinion, or eventually even replace the human factor.

Bibliography

[1] *Bazel*. Accessed: 2017-05-10. URL: https://bazel.build/.

[2] Yoshua Bengio. “Practical recommendations for gradient-based training of deep architectures”. In: *CoRR* abs/1206.5533 (2012). URL: http://arxiv.org/abs/1206.5533.

[3] Junghwan Cho et al. “Medical Image Deep Learning with Hospital PACS Dataset”. In: *CoRR* abs/1511.06348 (2015). URL: http://arxiv.org/abs/1511.06348.

[4] Junghwan Cho et al. “Medical Image Deep Learning with Hospital PACS Dataset”. In: *CoRR* abs/1511.06348 (2015). URL: http://arxiv.org/abs/1511.06348.

[5] *Convolutional Neural Networks (CNNs / ConvNets)*. Accessed: 2017-05-10. URL: http://cs231n.github.io/convolutional-networks/.

[6] Wenyuan Dai et al. “Boosting for Transfer Learning”. In: *Proceedings of the 24th International Conference on Machine Learning*. ICML ’07. Corvalis, Oregon, USA: ACM, 2007, pp. 193–200. ISBN: 978-1-59593-793-3. DOI: 10.1145/1273496.1273521. URL: http://doi.acm.org/10.1145/1273496.1273521.

[7] American Academy of Dermatology. *Types of skin cancer*. Accessed: 2017-05-10. URL: https://www.aad.org/public/spot-skin-cancer/learn-about-skin-cancer/types-of-skin-cancer.

[8] Kunio Doi. “Computer-Aided Diagnosis in Medical Imaging: Historical Review, Current Status and Future Potential”. In: *Computerized medical imaging and graphics: the official journal of the Computerized Medical Imaging Society* 31 (2007), pp. 198–211. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1955762/.

[9] Andre Esteva et al. "Dermatologist-level classification of skin cancer with deep deep neural networks". In: *Nature* 542 (2017), pp. 115–118. URL: http://dx.doi.org/10.1038/nature21056.

[10] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. URL: http://www.deeplearningbook.org.

[11] Paulina Hensman and David Masko. "The impact of Imbalanced Traning Data for Convolutional Nueral Networks". In: (2015). URL: https://www.kth.se/social/files/588617ebf2765401cfcc478c/PHensmanDMasko_dkand15.pdf.

[12] Imagenet. Accessed: 2017-05-10. URL: http://www.image-net.org/.

[13] *Inception in TensorFlow*. Accessed: 2017-05-10. URL: https://github.com/tensorflow/models/tree/master/inception.

[14] National Cancer Institute. *Common Moles, Dysplastic Nevi, and Risk of Melanoma*. Accessed: 2017-05-10. URL: https://www.cancer.gov/types/skin/moles-fact-sheet.

[15] National Cancer Institute. *NCI Dictionary Of Cancer Terms*. Accessed: 2017-05-10. URL: https://www.cancer.gov/about-cancer/understanding/what-is-cancer.

[16] National Cancer Institute. *Skincancer?* Accessed: 2017-05-10. URL: https://www.cancer.gov/about-cancer/understanding/what-is-cancer.

[17] National Cancer Institute. *What is cancer?* Accessed: 2017-05-10. URL: https://www.cancer.gov/about-cancer/understanding/what-is-cancer.

[18] *ISIC-archive*. Accessed: 2017-05-10. URL: https://isic-archive.com/.

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems 25*. Ed. by F. Pereira et al. Curran Associates, Inc., 2012, pp. 1097–1105. URL: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems 25*. Ed. by F. Pereira et al. Curran Associates, Inc., 2012, pp. 1097–1105. URL: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

[21] Abdul Ghaaliq Lalkhen and Anthony McCluskey. "Clinical tests sensitivity and specificity". In: *Continuing Education in Anaesthesia Critical Care & Pain* 8.6 (2008), p. 221. DOI: 10.1093/bjaceaccp/mkn041. eprint: /oup/backfile/content_public/journal/ceaccp/8/6/10.1093/bjaceaccp/mkn041/2/mkn041.pdf. URL: +%20http://dx.doi.org/10.1093/bjaceaccp/mkn041.

[22] T. Majtner, S. Yildirim-Yayilgan, and J. Y. Hardeberg. "Combining deep learning and hand-crafted features for skin lesion classification". In: *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*. Dec. 2016, pp. 1–6.

[23] A. Masood, A. Al- Jumaily, and K. Anam. "Self-supervised learning model for skin cancer diagnosis". In: *2015 7th International IEEE/EMBS Conference on Neural Engineering (NER)*. Apr. 2015, pp. 1012–1015. DOI: 10.1109/NER.2015.7146798.

[24] Michael A. Nielsen. *Neural Networks and Deep Learning*. Accessed: 2017-05-10. 2015. URL: http://neuralnetworksanddeeplearning.com/chap1.html.

[25] Andreas Nylund. "To be, or not to be Melanoma: Convolutional neural networks in skin lesion classification". In: *Dissertation* (2016). URL: http://kth.diva-portal.org/smash/get/diva2:950147/FULLTEXT01.pdf.

[26] J. Premaladha and K. S. Ravichandran. "Novel Approaches for Diagnosing Melanoma Skin Lesions Through Supervised and Deep Learning Algorithms". In: *J. Med. Syst.* 40.4 (Apr. 2016), pp. 1–12. ISSN: 0148-5598. DOI: 10.1007/s10916-016-0460-2. URL: http://dx.doi.org/10.1007/s10916-016-0460-2.

[27] Stanford. *Transfer Learning*. Accessed: 2017-05-10. URL: http://cs231n.github.io/transfer-learning/.

[28] Josephine Sullivan. *Lecture 8 - Visualizing, Training & Designing ConvNets*. Accessed: 2017-05-10. URL: https://www.kth.se/social/files/58fdbdfdf276546e343765e3/Lecture8.pdf.

[29] Christian Szegedy et al. "Going Deeper with Convolutions". In: *CoRR* abs/1409.4842 (2014). URL: http://arxiv.org/abs/1409.4842.

[30] Christian Szegedy et al. "Rethinking the Inception Architecture for Computer Vision". In: *CoRR* abs/1512.00567 (2015). URL: http://arxiv.org/abs/1512.00567.

[31] Sun-Chong Wang. "Artificial Neural Network". In: *Interdisciplinary Computing in Java Programming*. Boston, MA: Springer US, 2003, pp. 81–100. ISBN: 978-1-4615-0377-4. DOI: 10.1007/978-1-4615-0377-4_5. URL: http://dx.doi.org/10.1007/978-1-4615-0377-4_5.