Exploring Transfer Learning, Fine-tuning of Thyroid Ultrasound Images

K.V. Sai Sundar, S. Siva Sankara Sai
Sri Sathya Sai Institute of Higher Learning
Prasanthi Nilayam, Puttaparthi, Anantapur-515134
saisundarkandarpa@gmail.com

Abstract

Deep Learning for medical imaging has been on the forefront of its numerous applications, thanks to its versatility and robustness in deployment. In this paper we explore various classification methodologies that are employed for datasets of relatively small in size to actually train a deep learning algorithm from scratch. Thyroid ultrasound images are classified using a small CNN from scratch, transfer learning and fine-tuning of Inception-v3, VGG-16. We present a comparison of the aforementioned methods through accuracy, sensitivity and specificity.

1 Introduction

1.1 Background

Image Classification task of a Machine Learning/Deep Learning algorithm is widely found to be useful in the medical imaging domain wherein the algorithm is trained to classify medical images into benign or malignant in case of cancer detection and various other ailments based on the symptoms in the images. Arriving at automated diagnosis seems to be the dream of every researcher associated with computer vision coupled with biomedical imaging. Contrary to famous datasets like the Imagenet dataset, Cifar-10, Cifar-100 which have thousands of images in each category, the availability of medical data is limited. When running a deep learning algorithm with millions of parameters, less data hurdles the training with overfitting. The model eventually tends to fail at generalization of learning giving low accuracy on the test dataset. Regularization can reduce the high variance to some extent but training a deep learning framework from scratch remains out of bounds. Data augmentation can be employed to further boost the dataset size. Therefore in order to arrive at ac-curacies which can be of deployment standard, we resort to training a smaller CNN from scratch, transfer learning-using bottleneck features from deep CNNs to train a new FC layer or a different classifier and finally fine-tuning deep architectures to classify the custom dataset. We also explored deployment of Thyroid Cancer Classification by training the mobilenet-224 with Thyroid images and integrating a mobile application to do the classification. The code in this work is written in TensorFlow Abadi et al. [2016] and Keras Chollet et al. [2015].

1.2 Motivation

The Thyroid Ultrasound domain was chosen following consultation with local doctors who brought to our attention the high prevalence of thyroid ailments in the region. Thyroid Ultrasound is predominantly used to detect thyroid nodules, classify them as benign or malignant and also to identify goiter, thyroiditis. The problem consists of binary classification initially identifying images of patients who probably require a biopsy in order confirm malignancy of nodule and eventually multi-class classification identifying various other ailments apart from cancer. The aim of this work is to de-

Submitted to 1st Conference on Medical Imaging with Deep Learning (MIDL 2018). Do not distribute.

velop an automated thyroid diagnosis system that could aid the radiologist and fasten the diagnosis. In this paper we only discuss our implementations of binary classification.

2 Dataset Description

We have used two datasets in this work. The first dataset is a publically available one consisting of 298 images and their corresponding biopsy verified reports in .xml format.Pedraza et al. [2015] The TIRADS scores are given for each of the images ranging from 2 to 5 on the scale of increasing probability for malignancy. Since our task in this work dealt only with probably benign or malignant test scenario we considered scores 2 and 3 as benign and all the scores above these as malignant. The second dataset used in this work was the local database of images from GE LOGIQ P9 which were labelled by an experienced doctor and the reports were written in word format. The various cases of cancerous nodule, thyroiditis, simple goiter, multinodular goiter, toxic goiter and normal were present in this database. Again we considered only the relevant images as mentioned for the previous dataset. This dataset consisted of thyroid images of 127 patients.

3 Classification Methodology

3.1 Training a small CNN from scratch

The first method constituted training a CNN from scratch using the medical data. The layer architecture is 3 Convolutional Layers with 3x3 kernels of numbers 32, 32 and 64 respectively. The features from the FC layer were classified using a regular sigmoid function in-to the two classes benign or malignant. The training of the CNN was done on the GPU and since the model is relatively shallow and the dataset small, the training completed in an hour's time.

3.2 Transfer Learning

Bottleneck features and the new FC Layer - The bottleneck features from the VGG-16 and Inception-v3 were obtained and then trained on the CNN from the previous method. In VGG-16 the last three FC layers were discarded and the CNN model which we used in the first method was fed these features. The Imagenet pre-trained weights were loaded into the models and after the forward pass of the image through the network bottleneck features were saved.

Bottleneck features and SVM- The CNN was replaced with the popular linear classifier Support Vector Machine which was fed the bottleneck features for classification. The simple default parameters of the SVM implementation provided by the scikit-learn were used in this meth-od. Deep CNNs are known to be excellent feature extractors and using linear classifier to use these features proves to be an excellent way of tackling smaller datasets. Razavian et al. [2014]

3.3 Fine-tuning Inception-v3 and Vgg-16

The Inception-v3 model was imported with the help of tf-slim high level API pro-vided by TensorFlow. With the help of checkpoints provided for each of the models, pre-trained models could be availed and fine-tuned. The Inception-v3 net provided in the slim API returns the list 'end points' and 'logits' which can be fed to a classifier to predict the class. We obtained the end points['pre-logits'] which is a layer prior to the last layer in the architecture and customized the FC layer, to give output as a binary classifier. Softmax classifier was used for the classification. The last three FC layers of the VGG-16 which contribute to huge computations were discarded and new FC layer was attached after the 'pool-5' layer. For fine-tuning all layers above the conv 5_2 were frozen. So essentially the last three layers (excluding the FC layers) and the new custom FC layer were trained in this approach.

4 Results

The metrics used for evaluating the aforementioned methods are accuracy i.e. ratio of the total number of correct predictions to the total number of images predicted, sensitivity, which gave an indication of true positive rate and specificity for true negative rate. The classification was done on both the datasets separately and on the combined dataset. It was observed that the first public dataset gave high sensitivity and low specificity while the second dataset gave just the opposite. This is due to the nature of the data wherein the first public dataset consisted of biased data with number of cancerous samples on the higher side. The local dataset on the other hand had the bias towards normal samples. This problem was handled by combing the datasets and also data augmentation achieved by flipping, rotating and adding noise to the existing images. The table below summarizes the results obtained on the combined dataset alone. This combined dataset consisted of 2525 training samples and 613 test samples. The metrics are tabulated as percentages.

Sl No	Model	Accuracy	Sens.	Spec.
1	CNN from Scratch	0.82	0.89	0.82
2	VGG_16-Bottleneck+SVM	0.99	1	0.985
3	Inc_v3-Bottleneck+SVM	0.967	0.985	0.95
4	VGG_16-Bottleneck+New FC	0.94	0.96	0.92
5	Inc_v3-Bottleneck+New FC	0.98	0.99	0.97
6	VGG_16 Finetuning	0.76	0.84	0.68
7	Inc_v3 Finetuning	0.79	0.8	0.77

Table 1: Summary of Classification Results on Combined Dataset

5 Discussion

The delicate balance between the various evaluation metrics is important for performance analysis of deep learning algorithm in biomedical imaging. Owing to the fact that the dataset size is small Fine-tuning the Deep Architectures resulted was bettered by the other approaches. The amalgamation of linear classifier like the SVM to the architecture (especially the VGG-16) turned out to be the best model with stable metrics. Linear classifiers are considered to be useful when the deep neural networks is trained on datasets which are very different from domain in question. In order to take

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

François Chollet et al. Keras, 2015.

Lina Pedraza, Carlos Vargas, Fabian Narvaez, Oscar Duran, Emma Munoz, and Eduardo Romero. An open access thyroid ultrasound image database. In *10th International Symposium on Medical Information Processing and Analysis*, volume 9287, page 92870W. International Society for Optics and Photonics, 2015.

Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*, pages 512–519. IEEE, 2014.