Comparison of Pre-trained vs Domain-specific Convolutional Neural Networks for Classification of Interstitial Lung Disease

Jorge Bautista Martinez
Department of Computer Science
Sonoma State University, Rohnert Park, CA
bautisjo@sonoma.edu

Gurman Gill (Contact Author)
Department of Computer Science
Sonoma State University, Rohnert Park, CA
gillg@sonoma.edu

***Abstract*—Interstitial Lung Disease (ILD) is an umbrella term used to describe different variations of lung diseases that affect humans. Due to the difficulty of classifying ILD, the time span taken to have a patients CT scans analyzed by a radiologist is rather long. To speed up the process, Computer-Aided Diagnostic (CAD) systems have been built. In this paper, one such approach based on Convolutional Neural Network (CNN) is proposed to classify ILD from CT scans. We investigate how a generalized pre-trained neural network compares to a domain-specific neural network. In addition, we propose different methods to help improve a CNNs performance in classifying ILD and assess how our data impacts performance for both models. The InceptionV3 model produced by Google Inc. yielded the best F1 score of 0.80 ± 0.11 rivaling a domain-specific model. When using an Ensemble Network composed of InceptionV3, VGG16, and ResNet50, the accuracy increased to 0.827 ± 0.08.**

***Index Terms*—Convolutional Neural Networks, Computed Tomography, Interstitial Lung Disease, Medical Imaging**

Type of submission: Short paper for CSCI-ISHI

I. Introduction

Interstitial Lung Disease (ILD) describes a group of different diseases that affect the interstitium and space around the alveoli. In a recent study, it was reported that approximately 80 per 100,00 men suffered from ILD while 67 per 100,000 women suffered from ILD in the US. Current new cases rates of 31.5% for men and 26.1% per woman are made per year [1]. Medical professionals have provided the MedGIFT dataset, that contains marked regions in computed tomography (CT) scans [5]. Our on-going research explores how to use a Convolutional Neural Network (CNN) to help classify ILD. Due to the recent popularity of CNN's, they have become the method of choice for computer aided detection of ILD in CT scans [3].

A. Convolutions Neural Networks

A CNN is a type of deep learning algorithm that is used to help recognize and analyze patterns in images. A CNN takes in an image as input, which is passed through different convolutional layers to help filter the image into different distinct textures. The convolutional layer outputs are then passed into a series of fully connected (FC) layers which will eventually output prediction scores using an activation function.

A way to intuitively understand CNNs is as follows: The very early layers help extract edges and colors in the input image. Once the image goes through the early convolutional layers, pooling and convolutions layers help extract textures and shapes. The final FC layers give weights to all the extracted textures and edges to determine features that are critical for image classification. The final layer is then the prediction layer that outputs the predicted category along with their confidence scores (Fig. 1).

Due to the increased popularity of CNNs in the last decade, major tech companies have funded and invested time into creating a generalized CNN to help classify a multitude of image variations. These generalized CNNs are targeted to classify natural images [4]. They contain millions of parameters and are trained on state of the art GPU's and datasets such as ImageNet that contains 1.2 million images belonging to 1000 non-medical categories [10]. The CNN architectures used in this research are InceptionV3 (Google Inc.), VGG19 [8], VGG16 [8] and ResNet50 [9]. All networks are provided by Keras [6] that were pre-trained on ImageNet.

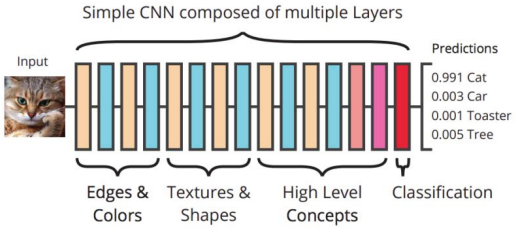

Fig. 1. Generalized CNN that takes in an Image and outputs a probability score for different categories.

II. Methods: Transfer learning

A large amount of data and GPU power to train on large batches of data is not readily available. A technique called transfer learning can be used as an alternative to train a CNN model from scratch [2]. It uses a pre-trained network

978-1-7281-5584-5/19/$31.00 ©2019 IEEE
DOI 10.1109/CSCI49370.2019.00188

and makes it more domain-specific by training on a limited amount of data. Transfer learning is done by removing the classification layer in the trained networks, then attaching new untrained layers into the model. These layers consisted of two fully connected (FC) layers and two dropout layers, followed by a final soft-max activation layer (Fig. 2).

In addition, our ongoing research assesses the efficacy of various other techniques that will help boost the classification performance of a transfer learning model to show how it can compete with more domain-specific models. In order to improve transfer learning, we propose using an ensemble of pre-trained networks.

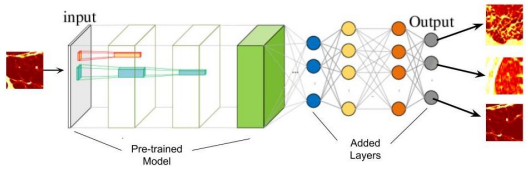

Fig. 2. Schematic diagram showing the process of adapting an existing CNN to classify images from a new domain.

A. Ensemble Networks

A technique we propose to help improve classification rates of transfer learning models is using an ensemble of networks. In an ensemble network, we run and train 3 independent models. Once all predictions for the dataset have been made with these 3 models, we sum all the predictions score and takes the average of the results. This will produce new scores that takes into consideration what each model learns.

III. Dataset

The MedGIFT dataset is used to generate the image dataset that will be used to perform experiments. The database contains CT scans from 128 patients diagnosed as healthy or having 1 or more of 10 different ILD's [5]. Each CT scan contains labeled regions of interest (ROI).

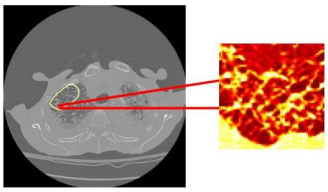

Fig. 3. Patch extraction from a CT scan. ROI are marked by the yellow curve. The yellow region was then converted from HU to RGB.

A. Patch Extraction

All models in this paper required three-channel (RGB) input images. Since MedGIFT contains raw CT volumes of a patient, a MATLAB framework was developed to extract patches from the CT volumes. The framework traversed through all the ROI's and extracted patches of a fixed size and converted them to Red, Green, Blue (RGB) images (Fig. 3). Each patch has at least 50% of its pixel images inside the designated ROI. The number of patches extracted can be changed by changing the patch frequency, as well as patch size. Patch frequency designates how often patches are sampled from within ROI. Currently, there are three patch size presets: 16x16, 32x32, and 64x64 pixels, with CNN's performing better when trained and classifying on the larger 64x64 pixel images.

CT scans are displayed in Hounsfield Units (HU), where each pixel value represents underlying tissue density (Air: -1000, Lung: -900, Bone: +1000). The HU were linearly mapped as follows: HU range (-1000, -600) to the Red channel, HU range (-601, -200) to the Green channel, and HU range (-201, 200) to the Blue channel.

The final dataset generated from the CT scans consists of 9 classes, 122 patients, and 20138 image patches (Fig. 4). The classes were as followed: Bronchiectasis(BR), Consolidation (CN), Emphysema (EM), Fibrosis (FI), Ground Glass (GG), Healthy (H), MacroNodules (MaN), MicroNodules (MiN), Reticulation (R).

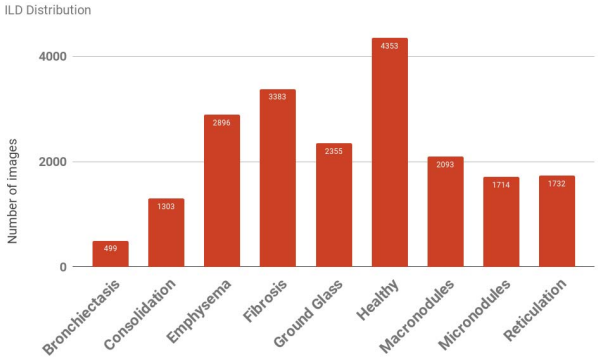

Fig. 4. Class distribution of image patches.

IV. Evaluation

The performance of CNN models can be adversely affected by data imbalance, wherein a number of images of different classes hugely varies. To address data imbalance, oversampling was used to even out class distributions during training. This was done by finding the highest group count and sampling all other subsequent classes to the same class count of the higher class.

Cross-validation was used to assess the model's average performance. The cross-validation technique Leave One Group Out (LOGO) was used to train the transfer learning and domain-specific CNN models. In Leave One Group Out, each patient, not disease category, is considered a Group. This ensures that only information specific to the disease is used during training. By removing certain patients images from the training set, we can ensure that the model is not influenced by patient patterns, but by disease textures. Once training is complete, then we only use the removed patient images as the test set. This allows us to make an estimate of the models performance for any CT scan, regardless of the patient.

Each models performance was evaluated by calculating their F1-Score, which is defined as:

$$F_1 = \frac{2*precision*recall}{precision+recall} \quad (1)$$

$$Precision = \frac{TruePositives}{TruePositives+FalsePositives} \quad (2)$$

$$Recall = \frac{TruePositives}{TruePositives+FalseNegatives} \quad (3)$$

A Confusion Matrix (CM) was also used to help visualize the actual class versus predicted class for all images. The vertical axis represents actual classification while the horizontal axis represents predicted classification.

A. Transfer Learning Results

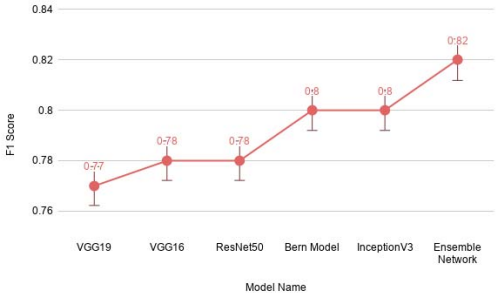

Fig. 5. Final Average F1 Score for each Model after running through all 122 LOGO runs.

Four different pre-trained models were used to evaluate the performance of transfer learning. These models were: VGG16, VGG19, ResNet50, and InceptionV3. The transfer learning layers began with a FC layer containing 1020 neurons. Followed by a dropout layer with a 0.5 dropout rate to prevent overfitting. Following the dropout, one more FC layer, with 520 neurons, and dropout layer was added. The final layer following the dropout was a soft-max activation output prediction layer. All models were found to reach the best performance with the parameters in Table 1.

The best notable performing model was InceptionV3, while our worst performing model was VGG19 (Fig. 5). InceptionV3 quickly rivaling the domain-specific Bern Model shows that it had better performance at identifying the key textural features in the lung patches, while ResNet50 and VGG16 both struggled to capture the same amount of detail (Fig. 5).

TABLE I
TRANSFER LEARNING PARAMETERS

Parameter Values			
Epochs	***Batch Size***	***Learning Rate***	***Optimizer***
8	32	0.0001	Adam

B. Using an Ensemble Network

We used InceptionV3, ResNet50, and VGG19 to create the ensemble network. The ensemble network yielded the best results at an F1 score of .827, showing us that combining the models lead to a higher classification rate on average. From the Ensembles confusion matrix, shown in Fig. 6, we can note that the worst-performing class is Bronchiectasis at an F1 score of 0.64, while the best performing class was Consolidation at 0.92.

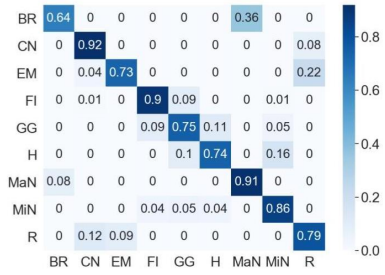

Fig. 6. Final CM for the Ensemble Network.

C. Using a Domain Specific CNN

We wanted to compare how well a transfer learning model can compete with a CNN that is domain-specific to the problem. In order to test this idea, we created our own LeNet5 inspired model along with another model created by the Bern team [7]. Our in-house CNN used a LeNet5 variation that consists of 2 convolutions layers follows by a max pooling, activation, convolutional layer, max-pooling layer, activation layer and then passed through three FC nodes. Our in-house CNN did not perform well producing a sub-optimal F1 score of 0.38 while the domain-specific Bern model proved to yield high classification results, with an F1 score of 0.80. One key thing to note is ILD textures are categorized by its textural features and having a max-pooling layer early on could result in removing these key features in the images [7]. Future models will be tested where the max-pooling occurs at the end, similar to the Bern model.

The Bern performed notably well, with a classification score very similar to the InceptionV3 transfer learning model. From the confusion matrix in Fig. 7, the best performing class classification is MacroNodules with an average 0.91 F1 score. The worst performing class was within Bronchiectasis, at an F1 score of 0.46 on average. Overall, Ensembles based model performed better than the Bern model.

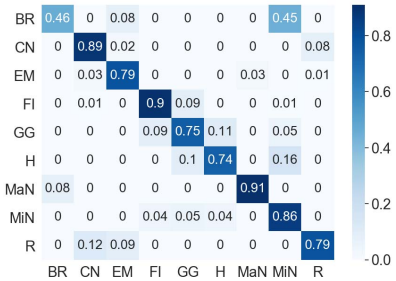

Fig. 7. Bern Model Confusion Matrix

V. Discussion

From these preliminary results, we can now see that we are able to take a pre-trained generalized CNN and make it more domain-specific. Performing transfer learning allows us to achieve near similar results to a specialized CNN such as the Bern Model. This can prove to be very useful as we do not have to train from scratch and dedicate GPU usage and spend days. When using transfer learning, we can optimize the time required it takes to train the network.

Using our best performing network, Ensemble Networks, we were able to increase the classification performance of two different classes. In the Bern model, the Bronchiectasis class only scored an F1 score of 0.46 on average, while in the ensemble networks, this class performance increased to 0.64. Similarly, a slight increase was also seen in Consolidation with its F1 score increasing from 0.89 to 0.92. Performance of all other classes, excluding Emphysema, remained fairly consistent, which is discussed next.

The accuracy of Emphysema class lowered by a notable margin dropping from an average F1 score of 0.79 in the Bern model to 0.73 in the Ensembles model. However, we can note from the Ensembles Confusion matrix (Fig. 6) that 22% of Emphysema patches were being classified as Reticulation. This is dramatically different from the 1% of such misclassifications in the Bern model (Fig. 7). Not much can be commented as to why this has happened, as the Reticulation disease did not have any notable increase or decrease in its performance, but this occurrence will be studied in future research.

Overall, using a single well trained pre-trained model with transfer learning would be enough to compete with a domain-specific CNN. However, by combining multiple pre-trained models that were trained using transfer learning, we can begin to increase the classification performance. As shown in Fig. 8, the Bern model has high confidence in predicting if a patient scan had fibrosis or consolidation. However, it has lower confidence in determining MicroNodules. Likewise, Inception performed well at having high confidence scores for Consolidation and MicroNodules, but did not perform notably well with Fibrosis. But by combining InceptionV3, ResNet50, and VGG16, all three models were able to, on average, figure out all three diseases with high confidence. This could be due to each pre-trained model learning different key features in the patient's CT image.

The convergence of all models happens at around 8 epochs. This gives us insight that given enough data, not many epochs are needed to quickly identify key features in ILD CT scans. An advantage that pre-trained models have is that training time is nearly half of what is required from training a CNN from scratch. On average, a pre-trained CNN model with transfer learning takes an hour for training. As opposed to a CNN from scratch, such as our LeNet5 inspired model, and Bern Model, which on average take about 2 hours for training. However, due to the nature of Ensembles having to run three different networks, the run time for Ensembles was on average greater than a single model, at a run time of 4 hours.

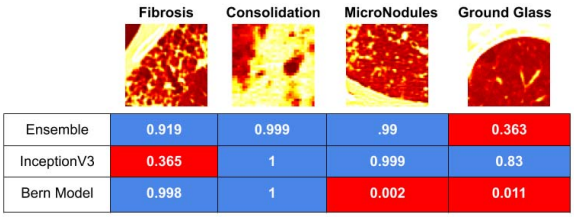

Fig. 8. Confidence Score of three different diseases. Scores in blue were correctly classified, while scores in red were incorrectly classified.

VI. On going Work

This on-going research project is funded by Research, Scholarship, Creative Activities Program (RSCAP) at Sonoma State University. Ongoing work will involve creating our own model to categorize critical features to improve the model classification. In addition, the MedGIFT dataset will be used to generate 2.5D images and create a 2.5D model to classify ILD. By using 2.5D images, the model will have information regarding textures of the disease by looking at the nearby slices. This will add more depth to train and test for each patient's data. Using a 2.5D dataset may also allow for better classification results since the disease may grow vertically along with the patient, not just horizontally [11]. Using full 3D images is also being taken into consideration to have a full volume of CT information be given to the CNN model, as this will provide more data to the CNN, not just an area of the CT image information.

References

[1] Interstitial Lung Disease (ILD). American Lung Association, https://www.lung.org/lung-health-and-diseases/lung-disease-lookup/interstitial-lung-disease/.

[2] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks?. Neural Information Processing Systems 27 (NIPS '14), pages 3320 - 3328

[3] H. Shin, H. Roth, M. Chen, Le Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, R. Summers, "Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning", IEEE Trans. on Medical Imaging, May 2016

[4] A. Krizhevsky, I. Sutskever, and G. Hinton, ImageNet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst., p. 9, 2012.

[5] Depeursinge A. et al., Building a reference multimedia database for interstitial lung diseases. Computerized Medical Imaging and Graphics 36(3) pp 227-38, July 2012

[6] Francois Chollet et al. Keras.https://keras.io, 2015

[7] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou, Lung pattern classification for interstitial lung diseases using a deep convolutional neural network, IEEE Trans. Med. Imag., vol. 35, no. 5, pp. 1207-1216, May 2016.

[8] Karen Simonyan and Andrew Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, 2014, arXiv:1409.1556

[9] Kaiming He and Xiangyu Zhang and Shaoqing Ren and Jian Sun,Deep Residual Learning for Image Recognition,2015, arXiv:1512.03385,

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in IEEE CVPR, 2009.

[11] H. Roth, L. Lu, J. Liu, J. Yao, A. Seff, K. M. Cherry, E. Turkbey, and R. Summers, Improving computer-aided detection using convolutional neural networks and random view aggregation, in IEEE Trans. on Medical Imaging, May 2016.