CoViD X-Ray Image Classification using Deep Learning

Robert Wilcox, Sr.
U. of Illinois - Champaign
Monee, IL
rwilcox3@illinois.edu

Emil Varghese
U. of Illinois - Champaign
Burbank, CA
emilv2@illinois.edu

Tyler Bybee
U. of Illinois - Champaign
Des Moines, IA
tbybee2@illinois.edu

Charles Beyer
U. of Illinois - Champaign
Glenview, IL
ccbeyer2@illinois.edu

1. ABSTRACT

Background and Objective: Novel Coronavirus also known as COVID-19 originated in Wuhan, China around December 2019 and has now spread across the world. The study tries to classify COVID-19 chest x-ray images against pneumonia and healthy patients using convolutional neural networks (CNNs)

Methods: In this study, we try to develop a prediction model as suggested in the paper, FLANNEL (Focal Loss bAsed Neural Network EnsembLe) for COVID-19 detection [2], which uses several standard convolutional neural network models and fuse them together for better accuracy using Ensemble methods. These publicly available datasets has been used for preliminary analysis. Kaggle Chest X-Ray images and IEEE Covid-19 data set. The following base learners have been used. AlexNet, ResNeXt101, Densenet161, Inception V3, VGG19, and Resnet152. In addition to these we also plan to benchmark using SqueezeNet, ShuffleNet and MNASNet algorithms. The idea is to take the best 5 base learners and use them to train the ensemble model for better prediction. In addition to the transforms mentioned in the original paper other image augmentation methods like Color jitter and random rotation are done to increase accuracy.

Preliminary Results: Preliminary training was done using the following pretrained pytorch models. AlexNet, ResNeXt101, Densenet161, Inception V3, VGG19, and Resnet152. Based on the initial results, it appears that some models perform better with the modified transforms (e.g.: - ResNeXt101), while some models perform better with the base image transformations. The results are also better than the base model evaluation in the original paper. The number of images in the dataset has increased from the time of publishing the original paper and this could be one of the reasons for the better performance in this approach. The number of publicly available images for Covid-19 is still, however, limited.

Discussion: Based on the initial results, it appears that some base models are performing better with the modified image transforms and some are not. We are currently looking at the cause for the poor performance of those models. One of the challenges with training has been the time it takes for training images. On average we have seen around 5 mins per epoch and this significantly slows down the process. Based on the limited resources available to students, it might not be possible to efficiently run and compare the various models in a timely manner.

Preliminary Conclusion: From the initial training, it appears that the base models are performing better than the original FLANNEL [2] paper. This could be due to the fact that 1) The number of Covid-19 images in the dataset have increased from the time of the original paper publication, 2) the models are running into an unbalanced data problem that hinders classification accuracy and limit model generalization. We plan to use resampling techniques to overcome this issue.

Keywords
COVID-19 detection, Convolutional neural networks (CNNs), computer-assisted radiographic image interpretation, Class imbalance

2. INTRODUCTION

2.1 Objective

As of March, 2021, there have been over 100 million cases of the novel coronavirus disease 2019 (COVID-19) around the world, and almost 3 million deaths [1]. Due to the recency of the disease, there are relatively fewer x-ray images related to the disease compared to non-COVID-19 related x-ray images, which makes developing effective deep learning models difficult [2].

There have been developments in ensemble neural networks, specifically, FLANNEL (Focal Loss bAsed Neural Network EnsembLe) to classify x-rays of COVID-related pneumonia, and we hope to increase the accuracy of a FLANNEL model by introducing preprocessing in the form of image augmentation and by attempting to improve the accuracy of the worst performing basis model of the ensemble by replacing it with a better performing model. Traditional data augmentation methods have demonstrated promise for increasing the accuracy of Convolutional Neural Networks and we hope that by increasing the accuracy of a basis model of the ensemble, the overall accuracy would increase [4]. This would allow for faster and more accurate diagnosis of patients suspected of having COVID-19, to assess appropriate treatment options faster.

2.2 Background and Significance

2.2.1 FLANNEL Approach (CNNs, Ensemble, Class Imbalance

This paper leverages the work in the FLANNEL [2] paper and attempts to improve upon that work. Within that document is given an excellent background on the elementary Deep learning techniques that were employed to develop the FLANNEL approach. The authors of this document defer to the expertise of the FLANNEL paper for their well-thought out discussion on these items. The main sections referenced from FLANNEL are on Convolutional Neural Networks, Ensemble Methods, and the Class imbalance challenge, which are covered well in these papers. As a brief recap, Convolutional Neural Networks are the choice solution for Image processing and this papers objective is image classification. Ensemble methods are a technique to combine multiple approaches into a single result with the goal of achieving a better overall performance than any one of the individual techniques can achieve. Class imbalance challenge refers to the limited amount of CoViD imagery as compared to the other classes. Having a class imbalance in a dataset is a problem in many ways. First, it is difficult to train a model with limited information as the model may learn features incorrectly with a limited dataset as opposed to a larger dataset. Second, when there's a class imbalance it means that are far fewer examples of a given type of class as compared to other classes. This is particularly a problem when the class is the main class of interest, as it is in the datasets being referenced in this paper. The problem occurs when the overall model performance doesn't reflect the performance of the imbalanced class type. Thus, someone reviewing the model could draw in an invalid conclusion on the model overall when the classes are imbalanced.

2.2.2 Preprocessing

Preprocessing of images is a technique to improve the overall performance of the model. During the review of a dataset, a key strategy is to look for data leakage. An example of this would be to have an arrow on an X-ray image that points to a particular disease. That arrow "leaks" the fact that the image does not contain a normal condition and the models will learn that the presence of an arrow indicates an unhealthy image. Strategies for this type of leakage would be to attempt to remove either the leakage or to remove the image. The existing FLANNEL paper did crop images to remove data leakage on the borders of the image. In our work, we are adding in more datasets which has other data leakage considerations, such as the arrows.

Other pre-processing techniques include resizing, cropping, random horizontal flipping, and noise introduction, which were employed within the existing FLANNEL work. Random image rotation and contrast variation are additional techniques added by this paper in order to improve model performance.

2.2.3 Alternate Models

In ensemble, the model selection is crucial to the overall performance. There are a number of CNNs to choose from. The existing FLANNEL work used vgg19, inception_v3, ResNext101, Densenet161, and Resnet152. This paper will investigate the use of Alexnet, Squeezenet, Shufflenet, and MNASNet as alternative models.

3. RELATED WORK

FLANNEL paper produced by Prof. Sun, et al and is the basis of our project. FLANNEL uses an ensemble technique combining 5 state-of-the-art convolutional neural network (CNN) classifiers as based models [2]. Data used in the paper is available publicly as well as the basis of the code for the project. The model compared with the baseline shows a higher macro-F1 score with 6% relative increase on the COVID-19 identification task [2].

"COVID-19 detection from scarce chest x-ray image data using few-shot deep learning approach" [5] uses a few-shot approach which is well-suited to scarcity in data, which still seems to be the case for our problem domain, to the author's best knowledge. Like the FLANNEL paper, the approach employs transfer learning, but the use of unsupervised learning and few-shot seem to be a potentially novel aspect of their approach. The work produced a 96.4% accuracy over the 83% baseline. The work also classifies normal, covid, and pneumonia in the models. The few-shot approach employs a Siamese network, which there are reference implementations for, though not included in pytorch. Given the capabilities of few-shot learning that are exhibited by GPT-3, the authors feel that this could be a potentially useful avenue to integrate into the FLANNEL approach.

The COVD-Net paper which was used in the FLANNEL paper to compare performance and is also referenced by a number of other papers in this area of research. The models achieved a 93.3% accuracy and employs CNNs in what the authors titled PEPIX to represent their Projection-Expansion-Projection-Expansion architecture [6]. In order to add explainability as well as audit the model, the authors used GSInquire. This tool highlights areas on the image that the model believes are features. This was then evaluated by radiologists to verify that the model was using valid features for classification.

The AICOVID paper is another related work that was also used for comparison within the FLANNEL paper [7]. In this approach, deep learning is used in the detection of COVID, normal, and pneumonia data. What is unique about this paper is that CT images were used instead of X-rays. The paper is a good resource for us to see another approach, though CT images are not as readily available to the authors best knowledge and so would not be a source for a potential solution. This work had very good performance, reporting a 96% accuracy. However, in comparison of the F1 scores done in the FLANNEL paper [2] and considering the higher quality of CT imagery, the authors find the results surprising.

"Deep-COVID: Predicting COVID-19 from chest X-ray images using deep transfer learning" [8]. In this paper, the authors use ResNet18, ResNet50, SqueezeNet, and DenseNet-161 models to do transfer learning on covid-19 classification. The authors used a fine tuning technique on these existing models that is similar to the approach in FLANNEL. For a sensitivity rate of 98%, these models achieved a specificity rate of around 90% on average [8]. The authors also applied a heatmap solution similar to the one used in FLANNEL. In reviewing the heatmap against the radiologist's analysis, it appears that although the classification was correct, the features used in making that decision were off at times.

"Classification of COVID-19 chest X-rays with deep learning: new models or fine tuning" [9] is the final relates work reviewed by the authors. It is very similar in its approach to the Deep-COVID paper. This time the author compared the performance of AlexNet, GoogleNet, and SqueezeNet. The approach utilized CNN in their models. The author's approach was to gather 6 separate datasets, 2 of which used 3 class classification (normal, pneumonia, Covid) and focused again on adjusting the parameters of the pre-trained models. Accuracy measured on the range of 95.9 – 99.2%. The related works section in the paper is a very good resource for additional research on this topic.

4. DATA

The following data sets will be utilized for this project.

COVID Chest X-ray Dataset. This is a public dataset containing the chest X-rays of patients who are positive or suspected of COVID-19 or other viral and bacterial pneumonias. This data is collected from public sources. This data set has 542 images from 262 people. The data set is available at
https://github.com/ieee8023/covid-chestxray-dataset

COVID-19 X-ray Dataset This dataset contains X-rays of patients with COVID-19, pneumonia, and no disease. This dataset is a combination of the data from multiple sources. This contains 127 images collected from COVID dataset and Normal and pneumonia images are from ChestX-ray8 database. This data set is available at
https://github.com/muhammedtalo/COVID-19

Chest X-Ray Images (Pneumonia) This dataset contains many non-covid images and was used have enough non-covid and "normal" images to compare to COVID-19 images. This dataset is available at:
https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia

As part of the project, we combined the COVID-19 datasets to set up the experimental data universe. Since COVID X-ray images are rare, we have combined all three datasets to increase the number of positive COVID examples. Manual inspection of the COVID images showed many possible signs of data leakage, such as markings on the edge of the images or wires (such as those connected to patient monitors) crossing the image. There were no arrows found in any of the non-COVID images, but there were wires visible in some of the non-COVID images. Because most images had markings on the edges, we used a center-crop to remove those markings, in the same method as the original FLANNEL paper. We also manually removed any images with arrows that wouldn't be removed by the crop. We did not remove any images due to visible wires though, or our dataset would have been significantly reduced in number, and it is unlikely that future data to predict on will not have any wires, i.e. removing images with wires in them could reduce real-world performance of the model.

The data sets used in the original FLANNEL paper had fewer COVID images and fewer types of "other" diagnoses. We chose to separate the images into "Normal" i.e. healthy, "Pneumonia", and "COVID-19" classes, as our main focus was COVID identification. We selected images that fell into those categories from the datasets above and removed images with leakage as mentioned previously. This resulted in 440 COVID-19 examples, 1,591 normal examples, and 4,290 pneumonia examples for training and evaluating from (compared to 100 COVID-19 examples, 1,118 normal examples, and 4,290 pneumonia examples in the original FLANNEL). Additionally, resizing and scaling was done as part of preprocessing, so the images are consistent across all sources.

5. APPROACH / METRICS

We took two approaches to try to improve the original FLANNEL basis models. The first was to attempt to improve the performance of the pretrained models by performing traditional image augmentation of the training data. The original preprocessing consisted of resizing all images to the same size, doing a center crop of that to get a 224x224 image, randomly horizontally flipping the image, and introducing noise to the images. We hoped to increase the generalizability of the model by also adding a random rotation, random brightness change, and random contrast change [3][4]. We added rotation up to +/-30 degrees as not all of the images were perfectly vertical. Images also had varying amounts of brightness and contrast, so those were also added at +/- 20% of the original values as well. We compared performance of the models trained on the original image modifications to models trained on the more-modified images using F1-Score. Additionally, the train/evaluation split was performed prior to image modifications, so the data splits for each type of model are on the same base images.

We are also attempting to improve the accuracy of the FLANNEL model by replacing the lowest performing models of the original basis models, with models we find with better performance according to F1-score. There are various CNNs we will evaluate for this, but pretrained weights (parameters) will be used in the same method as FLANNEL, where the pretrained model first has its final layer replaced with 3 outputs pertaining to the x-ray labels, at which point it is trained with all of the weights frozen besides the output layer. So far, we have evaluated Alexnet with the original data and the modified data, and we are planning on evaluating more models as well.

Once we are done training more models (at least 10) on both datasets, we will choose the 5 highest scoring models to use as the basis of the FLANNEL architecture **[see figure below]**. Depending on their performance, some models could be trained on the highly preprocessed data, and the rest of the models could be trained on the less-preprocessed data.

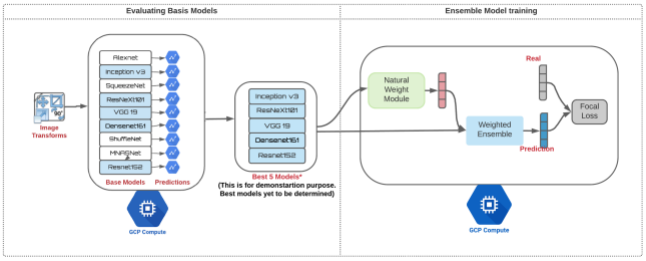

Figure. 1: FLANNEL Reference Architecture [2]

6. SETUP

The following software will be used for this project.

Software – Python, PyTorch, Torchvision. Jupyter Notebooks

Hardware
Google cloud VM (2 vCPUs, 13 GB memory), 1 x NVIDIA Tesla T4

Google cloud VM (2 vCPUs, 13 GB memory), 1 x NVIDIA Tesla K80

7. EXPERIMENTAL RESULTS

As our experiment has multiple experimental components, the experimental results are provided in sections covering each specific experiment.

Base Learner F1 Score Performance

As outlined in previous sections of the document, our first task is to identify the "Best 5" base learner models which will be subsequently pushed through the Ensemble Model Learning process. Each model will be evaluated against the source data using only basic preprocessing as well as our RED Flannel image pre-processing.

Table 1 *(below)* contains the performance, based on F1 Score, for the base learner models utilizing the source data with only basic preprocessing.

Table 1. Performance comparison on F1 score for standard image pre-processing

	COVID-19	Pneumonia	Normal	Macro-F1	Avg. Train Time (per Epoch)
Base Leaners					
Inception V3	0.87	0.93	0.85	0.88	478 seconds
VGG19	0.69	0.84	0.40	0.65	6.2 seconds
ResNeXt101	0.96	0.98	0.95	0.98	774 seconds
Resnet152	0.96	0.98	0.94	0.96	428 seconds
Densenet161	0.97	0.98	0.95	0.97	392 seconds
Alexnet	0.91	0.84	0.64	0.82	140 seconds
ShuffleNet*	--	--	--	--	--
MNRSNet*	--	--	--	--	--
Flexnet*	--	--	--	--	--
SqueezeNet*	--	--	--	--	--

* These base learners will be evaluated before final submission

Table 2 *(below)* contains the performance, based on F1 Score, for the base learner models utilizing our "RED Flannel" image pre-processing.

Table 2. Performance comparison on F1 score for RED Flannel image pre-processing

	COVID-19	Pneumonia	Normal	Macro-F1	Time (per
Base Leaners					
Inception V3	0.88	0.93	0.84	0.88	518 seconds
VGG19*	--	--	--	--	--
ResNeXt101	0.96	0.98	0.96	0.98	769 seconds
Resnet152**	0.09	0.21	0.25	0.19	412 seconds
Densenet161**	0.19	0.04	0.40	0.26	408 seconds
Alexnet**	0.02	0.79	0.63	0.48	147 seconds
ShuffleNet*	--	--	--	--	--
MNRSNet*	--	--	--	--	--
Flexnet*	--	--	--	--	--
SqueezeNet*	--	--	--	--	--

* These base learners will be evaluated before final submission

** Results for these base learners are being reviewed due to large variances against base image pre-processing results

Please note the following while interpreting the tables:

- All data provided in the tables are preliminary and we continue to test / evaluate our models
- Not all models were tested prior to submission of our draft document. Models without data are indicated as such.
- We are investigating some unexpected variances in our model performance data. (E.g. Resnet152 performs significantly worse than expected when using our "RED Flannel" image pre-processing

In addition to the F1 table results, we have also generated a confusion matrix to graphically represent the predictive capability of the model for COVID-19, Pneumonia, and Normal images. While we have generated matrices for all the base learner models, we are only including one such matrix in the document for the sake of brevity.

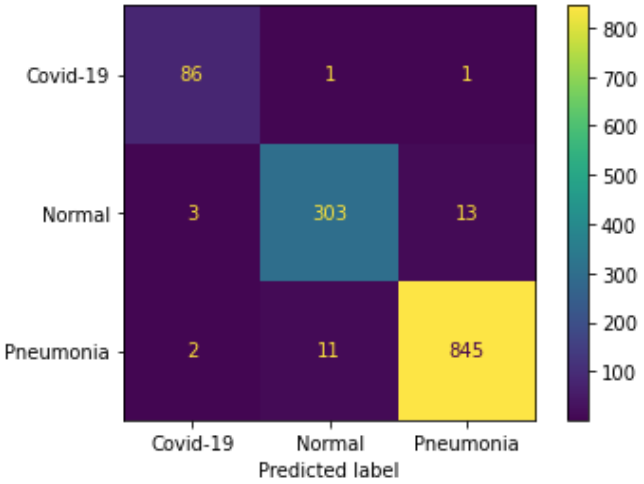

Figure 2 - ResNeXt101 RED Flannel

Ensemble Model Learning

Our second experimental component consists of benchmarking our "Best 5" base learners against the original FLANNEL base learners. While previous experiment's F1 scores will give us insight into which 5 models perform the best individually, we still need to verify that the better individual performances result in a better Ensemble Model Learning result.

Please note that at this point, we do not have results as we are still evaluating the base learners and we are actively working on this. For the purpose of the draft submission, please consider this section **TBD**.

8. DISCUSSION

From our initial testing, we have noted multiple notable findings:

- FLANNEL Baseline Challenges
- High COVID-19 Classification Rate
- "RED Flannel" Performance Issues *(for certain base learners)*

FLANNEL Baseline Challenges

Given that the FLANNEL research paper included the original source code and data sources, we did not expect much trouble in recreating the original baseline results. Unfortunately, we encountered a few surprises which complicated this effort.

The first issue was that the provided source code was not complete and was lacking sections of code. (E.g. Model Training) Additionally, we encountered some code defects as well as sections that required updates due to changes in the data. *(that will be further clarified below)* We believe that we have remedied most of these issues at this point and have been in discussion with the Github project maintainer to ensure updates are committed to the FLANNEL code base.

A second issue that we encountered was the data itself. Even though we went to the exact same data sources, the data has changed since the original FLANNEL implementation. Variances were encountered with image quality, quantity, and even classification labeling! While the original labels simply consisted of Bacterial Pneumonia / Viral Pneumonia / COVID-19 / Normal, far more specific labels were assigned to images in the newer dataset requiring updates to the original code.

The final issue we have encountered is in the equipment for processing our models. We have tried Colab, Azure, AWS, and Google Cloud Platform(GCP). In each of the platforms, we have run into issues. Colab terminated our training mid-run. In GCP, we ran into issues getting access to GPUs. We believe are past these issues now.

High COVID-19 Classification Rate

Compared to Pneumonia and Normal X-Ray images, our positive classification rate for COVID-19 was consistently higher across our testing. While we would like to think that our models are simply exceptional, we suspect that there may be another explanation.

Our suspicion is that COVID-19 patients, who are sick enough to be in a hospital setting and receiving an X-Ray, are likely to be experiencing significant medical issues. Because of this, they may have a higher incidence of other items appearing in the X-Ray images. (e.g. wires for monitors, tubes, etc.)

We have performed an initial review of the COVID-19 images and have removed multiple images where there were obvious markers. As we are still seeing unexpectly high results, we will have another review of the images to see if there are other images that we should remove from training.

"RED Flannel" Performance Issues *(for certain base learners)*

F1 Scores for certain base learners: Resnet152, Densenet161, and Alexnet were significantly worse when training against our "RED Flannel" modified images. *(see below table)*

Table 3. Performance variance on F1 score for RED Flannel vs. normal image processing

	COVID-19	Pneumonia	Normal	Macro-F1	Avg. Train Time (per Epoch)
Base Leaners					
Inception V3	0.01	0	-0.01	0	40 seconds
VGG19*	--	--	--	--	--
ResNeXt101	0	0	0.01	0	-5 seconds
Resnet152**	-0.87	-0.77	-0.69	-0.77	-16 seconds
Densenet161**	-0.78	-0.94	-0.55	-0.71	16 seconds
Alexnet**	-0.89	-0.05	-0.01	-0.34	7 seconds
ShuffleNet*	--	--	--	--	--
MNRSNet*	--	--	--	--	--
Flexnet*	--	--	--	--	--
SqueezeNet*	--	--	--	--	--

* These base learners will be evaluated before final submission

** Results for these base learners are being reviewed due to large variances against

The variances are significant and unexpected. We are currently still investigating to determine why we are seeing this result.

9. CONCLUSION/OPTIMIZATION

Based on initial results the base models are performing better than the original FLANNEL [2] paper. This could be because 1) The number of Covid-19 images in the dataset has increased from the time of the original paper publication and 2) the models are running into unbalanced data problem with a bias towards Covid-19 images. Another interesting observation is that the addition of additional image transformations like Color Jitter and random rotation seems to adversely affect some pre trained models while improving accuracy for some others. We plan to take a best of breed approach to improve accuracy.

Our objective is to improve the base model prediction performance in identifying COVID-19 from other pneumonia images and healthy images by feeding the output of top performing 5 models to the ensemble training step. This step will be able to automatically combine and use the outputs of individual base learners as features to create a more accurate global model.

10. REFERENCES

[1] COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). https://coronavirus.jhu.edu/map.html. Access: 2021-03-29.

[2] Z. Qiao, A. Bae, L. Glass, C. Xiao, J. Sun. FLANNEL (Focal Loss bAsed Neural Network EnsembLe) for COVID-19 detection *Journal of the American Medical Informatics Association*, 28(3), 2021, 444–452. (https://github.com/qxiaobu/FLANNEL)

[3] L.F. Rodrigues, M.C. Naldi, J.F. Mari. Comparing convolutional neural networks and preprocessing techniques for HEp-2 cell classification in immunofluorescence images. *Computers in Biology and Medicine*, 116, 2020.

[4] J. Wang, L. Perez. The Effectiveness of Data Augmentation in Image Classification using Deep Learning. *arXiv*:1712.04621, 2017.

[5] S. Jadon. COVID-19 detection from scarce chest x-ray image data using few-shot deep learning approach. Medical Imaging 2021: *Imaging Informatics for Healthcare, Research, and Applications*, Proceedings Vol. 116010, 2021.

[6] L. Wang, Z.Q. Lin, & A.Wong. COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. *Scientific Reports*, 10, 2020.

[7] X. Bai, R. Wang, Z. Xiong, et al. Artificial intelligence augmentation of radiologist performance in distinguishing COVID-19 from pneumonia of other origin at chest CT. *Radiology*, 296, 2020.

[8] S. Minaee, R. Kafieh, M. Sonka, S. Yazdani, G.J. Soufi. Deep-COVID: Predicting COVID-19 from chest X-ray images using deep transfer learning, *Medical Image Analysis*, Volume 65, 2020.

[9] T.D. Pham. Classification of COVID-19 chest X-rays with deep learning: new models or fine tuning? *Health Information Science and Systems*, 9, 2021.