CS230

Exploring Different Deep Learning Architectures on a Large Chest Radiograph Dataset

Sean Afshar
Department of Electrical Engineering
Stanford University
safshar@stanford.edu

Abstract

In this study, multi-class convolutional neural networks are trained to diagnose 14 types of observations from chest radiographs. The radiographs come from a large public chest X-ray dataset known as CheXpert. The three models, VGG-16, ResNet-50, and DenseNet-121 are trained using mini-batch gradient descent, a binary cross entropy loss function, and the adam optimizer. The DenseNet-121 model is trained using both transfer learning and traditional learning. The DenseNet-121 model performed the best out of all three models in every metric, but received low F1-scores. This was a result of class imbalance, which was remedied with an upsampling of underrepresented classes. The DenseNet-121 model was then trained again on the upsampled data, improving in accuracy on the test set and F1-scores.

1 Introduction

As of the time this manuscript is being written, the COVID-19 pandemic is making a massive surge in the United States, which has now experienced 10,000,000 cases [1]. In addition to the coronavirus, many patients are also suffering from diseases that can trigger similar symptoms, such as pneumonia and influenza. This ambiguity in diagnosis has played a role in the early spread of the virus in the United States. It is of the utmost importance that medical professionals have tools in order to discern between these diseases.

One such tool is the chest X-ray. Accurate chest radiograph interpretation would not only be a valuable aide to diagnosis and clinical decision making, but it would also improve workflow prioritization in a time where medical resources are stretched thin. In this work we analyze CheXpert(**Che**st e**Xpert**), a large chest radiograph dataset with 224,316 chest radiographs of 65,240 patients labeled for 14 attributes common in chest radiographs. The deep learning task was formulated as a multi-classification problem. The radiograph images were fed into each of the three models used for this study, which would then predict probabilities for each of the 14 observations. A probability would then become a positive (1) or negative classification (0) if it passed a threshold of 0.5, as seen in Figure 1. The three models selected for the study were VGG-16, ResNet-50, and DenseNet-121. The performance of each model was evaluated based upon accuracy, recall, precision, and F1 scores.

2 Related work

Deep learning applied to chest radiographs became popular in 2017 with the public release of the ChestX-ray14 dataset. Many different teams used novel architectures for computer aided diagnosis,

CS230: Deep Learning, Winter 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

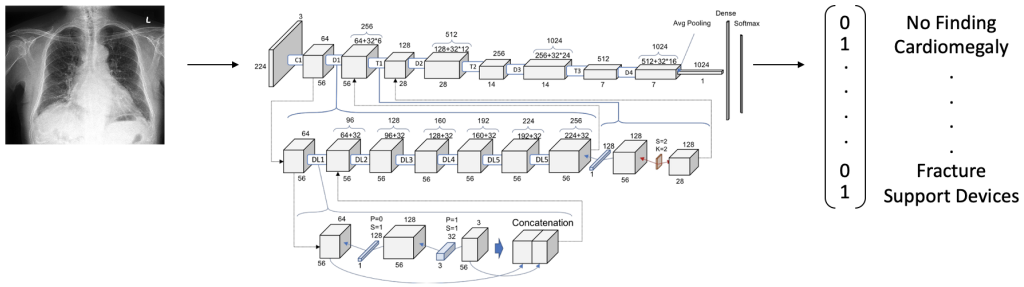


Figure 1: An example of the workflow of this paper: a chest radiograph is fed into a DenseNet-121 which outputs the vector of predictions.

such as [2,3]. However, the most popular architecture was the CheXnet, a 121-layer convolutional network (CNN) which was touted as being even more effective than doctors in diagnosing pneumonia [4].

The authors of CheXnet then expanded upon their work with their analysis of the CheXpert dataset in 2019 [5]. Both CheXnet and CheXpert are DenseNet-121s that predict probabilities for the same 14 observations, however, CheXpert is trained on the more robust CheXpert dataset. A DenseNet-121 is a CNN that differs from normal convolutional networks by connecting the layers in a feed forward fashion as outlined in [6]. DenseNet-121 is the current state of the art for this task. This paper will validate this consensus and also explore the potential of other architectures for this task.

3 Dataset and Features

The CheXpert dataset is a large public dataset for chest radiograph interpretation, consisting of 224,316 chest radiographs of 65,240 patients labeled for the presence of 14 observations as positive, negative, or uncertain [2]. The observations and their statistics are reported in Table 1. The dataset was obtained from Stanford ML, the authors of the original CheXpert paper. Once the data was loaded, each image was reshaped into a $224 \times 224 \times 3$ array which could be fed into each model.

Pathology	Positive (%)	Uncertain (%)	Negative (%)
No Finding	16627 (8.86)	0 (0.0)	171014 (91.14)
Enlarged Cardiom.	9020 (4.81)	10148 (5.41)	168473 (89.78)
Cardiomegaly	23002 (12.26)	6597 (3.52)	158042 (84.23)
Lung Lesion	6856 (3.65)	1071 (0.57)	179714 (95.78)
Lung Opacity	92669 (49.39)	4341 (2.31)	90631 (48.3)
Edema	48905 (26.06)	11571 (6.17)	127165 (67.77)
Consolidation	12730 (6.78)	23976 (12.78)	150935 (80.44)
Pneumonia	4576 (2.44)	15658 (8.34)	167407 (89.22)
Atelectasis	29333 (15.63)	29377 (15.66)	128931 (68.71)
Pneumothorax	17313 (9.23)	2663 (1.42)	167665 (89.35)
Pleural Effusion	75696 (40.34)	9419 (5.02)	102526 (54.64)
Pleural Other	2441 (1.3)	1771 (0.94)	183429 (97.76)
Fracture	7270 (3.87)	484 (0.26)	179887 (95.87)
Support Devices	105831 (56.4)	898 (0.48)	80912 (43.12)

Table 1: The number of samples in the CheXpert dataset for each observation.

3.1 Data Processing

The CheXpert dataset consists of both frontal and lateral chest radiographs. However, there are only 33,087 lateral radiographs compared to the 191,229 frontal cardiographs. In order to make the dataset more uniform, all lateral radiographs were removed. The dataset originally came with separate training and validation sets, but since the validation set was so small compared to the training set, the two were combined. After recombining the two data sets the overall data was shuffled, and then cut in a 80/10/10 ratio corresponding to the training, validation, and test sets respectively.

3.2 Uncertainty Handling

For some samples, certain observations were labeled as uncertain. After reviewing different attempts at analyzing the CheXpert data, I decided to replace all the uncertain labels with a positive label. Time constraints rendered more sophisticated models infeasible and this simple policy is somewhat excusable in a real life setting. If a patient receives a false negative, he or she is more likely to accept the result than in the case of a false positive. In this case, the patient is more likely to get a second opinion which can clear up a classification.

4 Methods

The three models used for this study were VGG-16, ResNet-50, and DenseNet-121 [7,8]. Each model was created using Keras and loaded with weights pre-trained on ImageNet. Furthermore each model was augmented by adding the following layers at the end in the following order: a global average pooling layer, a fully connected layer with 14 outputs and a ReLu activation function, and finally a logistic layer with a sigmoid activation function. The model confirms the existence of an observation when the predicted probability is above a threshold.

Each model was trained to minimize the binary cross entropy loss for each vector of observations, Y, which is given below.

$$L(Y, \hat{Y}) = \sum_{i=1}^{14} -[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

The overall loss is summed over each observation as seen above, and then summed over each image. I used the Adam optimizer with default β parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ and learning rate $\alpha = 1 \times 10^{-4}$ with no learning rate decay. Batches were sampled using a fixed batch size of 32 images, with training conducted over 3 epochs.

5 Experiments/Results/Discussion

5.1 DenseNet-121

The DenseNet-121 was trained via two methods. The first method was a transfer learning method where the weights learned from training the network on ImageNet were preserved, leaving the model with $\approx 10^6$ trainable parameters. In the second method the model was trained from the ground up, leaving the model with $\approx 8 \times 10^8$ trainable parameters. The number of training epochs, batch size, and optimizer constants were chosen to be consistent with the CheXpert paper. The results are shown below.

Pathology	Precision (TL)	Recall (TL)	F1-Score (TL)
No Finding	0.81 (0.41)	0.09 (0.24)	0.16 (0.30)
Enlarged Cardiom.	0.75 (0.00)	0.00 (0.00)	0.00 (0.00)
Cardiomegaly	0.59 (0.25)	0.52 (0.24)	0.55 (0.24)
Lung Lesion	0.75 (0.00)	0.10 (0.00)	0.18 (0.00)
Lung Opacity	0.59 (0.62)	0.82 (0.51)	0.69 (0.56)
Edema	0.67 (0.60)	0.71 (0.09)	0.69 (.16)
Consolidation	0.55 (0.00)	0.07 (0.00)	0.12 (0.00)
Pneumonia	0.57 (0.00)	0.21 (0.00)	0.31 (0.00)
Atelectasis	0.62 (0.00)	0.29 (0.00)	0.40 (0.00)
Pneumothorax	0.77 (0.00)	0.30 (0.00)	0.43 (0.00)
Pleural Effusion	0.63 (0.59)	0.88 (0.51)	0.73 (0.55)
Pleural Other	0.75 (0.00)	0.05 (0.00)	0.09 (0.00)
Fracture	0.49 (0.00)	0.01 (0.00)	0.02 (0.01)
Support Devices	0.92 (0.70)	0.79 (0.88)	0.85 (0.78)
Average	0.68 (0.23)	0.35 (.18)	0.37 (0.17)

Table 2: The precision, accuracy, and F1-Scores of the DenseNet-121 under each training regime. The values in parentheses correspond to the transfer learning (TL) method.

The transfer learning trained DenseNet-121 performed worse than its fully trained counterpart for nearly every observation and every metric. This trend is also seen in the test set performance. The transfer learning trained DenseNet-121 achieved an accuracy of 0.76 on the test set while the fully trained DenseNet-121 achieved an accuracy of 0.83 on the test set.

The improvement in F1-score between the two models implies that the transfer learning trained DenseNet-121 was overfitting to the training data. Even though both models had >75% accuracy on the test set, they both suffered from relatively low F1-scores on certain observations such as Fracture. It was later seen during error analysis that the disparity in F1-scores was a result of the distribution of the training data.

5.2 VGG-16 and ResNet-50

The results of training the ResNet-50 and VGG-16 models are shown below:

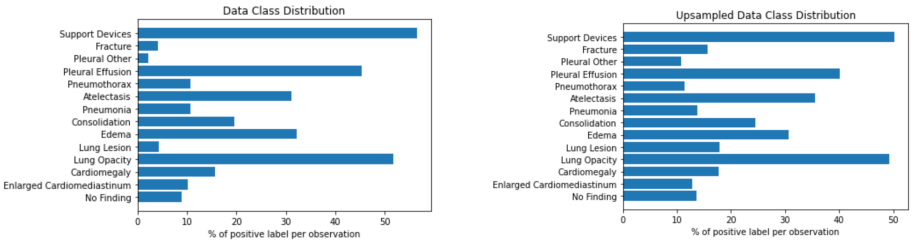
Architecture	Average F1-Score	Test Set Accuracy
DenseNet-121	0.37	0.83
ResNet-50	0.33	0.78
VGG-16	0.31	0.77

Table 3: Performances of the ResNet-50, VGG-16, and fully trained DenseNet-121 models.

The DenseNet-121 did the best in all metrics, giving credence to its status as the state of the art architecture for this task. It is interesting to note that while both the ResNet-50 and VGG-16 had more trainable parameters ($\approx 26 \times 10^6$ and $\approx 15 \times 10^6$ respectively) than the DenseNet-121, they both did worse than the DenseNet-121. One possible explanation could be that the feed forward connections in the DenseNet-121 help avoid vanishing and exploding gradients. This is supported by the fact that the ResNet-50, which connects layers with additional skip connections, performed better than the VGG-16.

5.3 Class Balancing

Upon further analysis of the data, I deduced that the low F1-scores in classification were a byproduct of class imbalance in the training dataset. Classes with a bigger representation tended to have higher F1-scores as seen in Tables 1 and 2. In order to alleviate this problem, the smaller classes were upsampled, creating a more balanced distribution. Figure 2 supports this claim, by showing the class distribution of the original data side by side the class distribution for the upsampled data.



(a) The original class distribution.

(b) The upsampled class distribution.

Figure 2: Original vs upsampled class distributions for the CheXpert training set.

There were alternatives to this method. I could have used a weighted loss function which penalized more for classification mistakes made in the minority classes or used other algorithms such as the synthetic mirror oversampling technique (SMOTE) to generate more minority class data. The upsampled data was then shuffled and cut again in an identical manner to the original experiments. The DenseNet-121 model was then trained on the upsampled data and produced the following results:

Pathology	Precision	Recall	F1-Score
No Finding	0.75	0.81	0.78
Enlarged Cardiom.	0.66	0.34	0.45
Cardiomegaly	0.57	0.72	0.64
Lung Lesion	0.76	0.69	0.72
Lung Opacity	0.72	0.93	0.81
Edema	0.72	0.77	0.74
Consolidation	0.48	0.74	0.58
Pneumonia	0.62	0.43	0.51
Atelectasis	0.68	0.66	0.67
Pneumothorax	0.61	0.53	0.57
Pleural Effusion	0.81	0.79	0.80
Pleural Other	0.56	0.98	0.71
Fracture	0.54	0.99	0.70
Support Devices	0.88	0.82	0.85
Average	0.67	0.73	0.68

Table 2: The precision, accuracy, and F1-Scores of the DenseNet-121 trained with the upsampled data.

The DenseNet-121 achieved an accuracy of 0.86 on the test set. We also see improvement in the F1-scores of each class.

6 Conclusion/Future Work

We examined the efficacy of three different CNNs as they attempted to diagnose different diseases from chest radiographs. The best architecture was the DenseNet-121 model, which achieved an accuracy of 0.83 on the test set. However, the F1-scores were low for all three models. This was discovered to be caused by class imbalance in the data. Once the class imbalance issue was fixed via upsampling, the DenseNet-121 model went from an average F1-score of 0.37 to 0.68 and a testing accuracy of 0.83 to 0.86.

In the future, weighted loss functions can be explored to better deal with class imbalance. More sophisticated uncertainty handling policies could also be adopted. It would also be helpful to implement weighted gradient class activation maps to have a better visualization of the data and the predictions being made.

7 Contributions

All the work was done solely by the author. I would like to thank the instructors of CS230 for their one on one guidance throughout this project.

References

[1] "CDC COVID Data Tracker." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, covid.cdc.gov/covid-data-tracker/.

[2] Salehinejad, H., Valaee, S., Dowdell, T., Colak, E. and Barfett, J., 2020. Generalization Of Deep Neural Networks For Chest Pathology Classification In X-Rays Using Generative Adversarial Networks. [online] arXiv.org. Available at: <https://arxiv.org/abs/1712.01636> [Accessed 17 November 2020].

[3] Li, Zhe, et al. "Thoracic Disease Identification and Localization with Limited Supervision." ArXiv:1711.06373 [Cs, Stat], 20 June 2018, arxiv.org/abs/1711.06373. Accessed 17 Nov. 2020.

[4] Rajpurkar, Pranav, et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. 25 Dec. 2017.

[5] Irvin, Jeremy, et al. "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison." Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 17 July 2019, pp. 590–597, 10.1609/aaai.v33i01.3301590.

[6] Huang, Gao, et al. "Densely Connected Convolutional Networks." ArXiv.org, 2016, arxiv.org/abs/1608.06993.

[7] Simonyan, Karen, and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." ArXiv.org, 2014, arxiv.org/abs/1409.1556.

[8] He, Kaiming, et al. "Deep Residual Learning for Image Recognition." ArXiv.org, 2015, arxiv.org/abs/1512.03385.

[9] Pytorch, Keras, Tensorflow, sklearn, panda, numpy, matplotlib libraries.