Generalization of Deep Convolutional Neural Networks - A Case-study on Open-source Chest Radiographs

Nazanin Mashhaditafreshi, BSc1, Amara Tariq, Ph.D.3, Judy Wawira Gichoya, MD MS.2,3, Imon Banerjee, Ph.D.2,3

1Department of Computer Engineering, K. N. Toosi University of Technology, Tehran, Iran; 2Department of Radiology, Emory School of Medicine, Atlanta, GA, USA; 3Department of Biomedical Informatics, Emory School of Medicine, Atlanta, GA, USA

Abstract

Deep Convolutional Neural Networks (DCNNs) have attracted extensive attention and been applied in many areas, including medical image analysis and clinical diagnosis. One major challenge is to conceive a DCNN model with remarkable performance on both internal and external data. We demonstrate that DCNNs may not generalize to new data, but increasing the quality and heterogeneity of the training data helps to improve the generalizibility factor. We use InceptionResNetV2 and DenseNet121 architectures to predict the risk of 5 common chest pathologies. The experiments were conducted on three publicly available databases: CheXpert, ChestX-ray14, and MIMIC Chest X-ray JPG. The results show the internal performance of each of the 5 pathologies outperformed external performance on both of the models. Moreover, our strategy of exposing the models to a mix of different datasets during the training phase helps to improve model performance on the external dataset.

Introduction

The proliferation of big data coupled with non-linear data abstraction (filters) and high performance computing has spurred rapid advancement in deep learning applications including speech recognition, sentiment analysis, computer vision, and machine translation. These areas were previously thought to be extremely hard for computers to analyze and required hundreds of hours of manual feature engineering yet deep learning techniques deliver state-of-the-art performance with minimal human intervention. Medicine is witnessing rapid adoption and application of deep learning. For example, a large volume of radiology studies are performed daily in most centers yet the number of available trained radiologists remains constant. The opportunity to standardize the clinical workflow is thus seen as a low hanging fruit for automation using deep learning, with lots of efforts deemed as hype that try to replace radiologists using deep learning.

Deep Convolutional Neural Networks (DCNNs) apply multiple layers of convolution operations to extract translation and scale invariant features from images, and are widely used to analyze radiology image content to assist in diagnosis. DCNNs have achieved expert-level performance for various chest pathologies. Beyond classification tasks on radiology images, researchers have attempted to rebuild the imaging workflow, assessing DCNNs performance after non-image data fusion for classification of multi-label chest X-ray images. Despite a plethora of multiple publications improving on the state-of-the-art, validation and scalability of deep learning in medicine remains limited, since model development and validation is frequently performed on a single institutional dataset. A review of studies published in 2018 found that only 6% (31 of 516) of published studies performed external validation (i.e. studies had a diagnostic cohort design, included data from multiple institutions, and performed prospective data collection).

Overfitting is a well-known limitation of complex DCNN models which may produce an overly optimistic performance. Therefore, it is important for an optimized DCNN model to have sustained performance on unseen external datasets to promote model generalizibility and translation of models to real life clinical work. Despite the expensive cost of labeling medical datasets, there are several publicly available datasets for chest radiographs that can be used for testing the model generalization. These datasets include the MIMIC Chest X-ray JPG (MIMIC-CXR-JPG) Database v2.0.0 from Beth Israel Deaconess Medical Center in Boston, the CheXpert dataset released from the Stanford Hospital, performed between October 2002 and July 2017 coded with 14 common radiographic diseases, and the ChestX-ray14 dataset from the U.S. National Institutes of Health. The labels for these three datasets were derived from radiology text reports using natural language processing algorithms. A large number of publications have been published from these three datasets focusing on novel DCNN design and development, and present state-of-the-art...
performance. However, there are only a limited number of studies on the generalizability of a DCNN model trained on chest X-ray images, specifically assessing whether the model retains its performance and generalizes well on unseen datasets. 

In this study, we perform thorough experiments to understand the generalizability of state-of-the-art DCNN models using data from three publicly available chest radiograph datasets. We selected 5 common pathologies (Cardiomegaly, Edema, Atelectasis, Consolidation, Pleural Effusion) and trained two state-of-the-art DCNN models (DenseNet121, InceptionResNetV2). To evaluate the performance, we adopted performance metrics that have been previously published for disease recognition tasks. We compared the external and internal performance of the models by training them on different partitions of data from the three datasets, and subsequently tested each model with various combination of test sets of these datasets. We report the test AUC of each experiment which shows that the performance of the DCNNs on internal data outperforms the external performance for the test sets.

Materials and Methods

In this section, we present details of the public datasets and architecture of the two DCNN models that were used for our experiments. We also describe the experimentation details to test the generalization capability of the DCNN models.

Datasets

The CheXpert dataset comprises of 224,316 frontal and lateral chest radiographs of 65,240 patients. The ChestX-ray14 dataset contains 112,120 frontal-view X-ray images of 30,805 patients. The MIMIC-CXR-JPG dataset consists of 377,110 images of 65,379 patients. We selected 5 diseases (Atelectasis, Edema, Pleural Effusion, Consolidation, Cardiomegaly) which were common among the above datasets. We randomly split the CheXpert dataset into training (161,035 images), validation (18,097 images), and test (44,282 images) sets. ChestX-ray14 dataset was also randomly divided into training (80,657 images), validation (9,169 images), and test (22,294 images) sets. There were no overlapping patients between the training, validation and test sets for the CheXpert and ChestX-ray14 datasets. In addition, we preserved the original test set of the MIMIC-CXR-JPG dataset (5,159 images) and considered it as an external test set for every configuration.

The ChestX-ray14 and CheXpert datasets provide some non-image features such as age, gender, and radiographic positioning. Figure 1a shows the distribution of patients’ age and gender for the ChestX-ray14 dataset. The average age is 46.13 years with a standard deviation of 17.01 years for this dataset. Patients’ age and gender distributions for CheXpert dataset are shown in Figure 1b. The average age and the standard deviation are 60.31 and 18.56 years, respectively. The distribution of gender is quite similar among the two datasets, but the ChestX-ray14 dataset has a larger proportion of younger patients as compared to the CheXpert dataset.

Data Preprocessing

For the ground truth labels, we used the binary mapping approach for handling uncertainty in labels, in which the uncertain labels were replaced with 1 (U-Ones model), or 0 (U-Zeroes model). We used U-Ones model for uncertain
labels for Atelectasis, Edema, and Pleural Effusion and U-Zeroes model for Cardiomegaly and Consolidation based on the CheXpert results[10].

For image preprocessing, we applied contrast-limited adaptive histogram equalization (CLAHE) technique[15] for contrast enhancement on all training and validation images before feeding them into the network. Thereafter, we normalized all images based on the mean and standard deviation of images in the ImageNet[17] training set. All the images were resized to 299 × 299 pixels for InceptionResNetV2 and 224 × 224 pixels for DenseNet121 architecture. The scikit-image transform module, which applies first order spline interpolation for image downscaling and Gaussian filter to eliminate aliasing artifacts, was used to resize the images. A constant value of 0 was used to fill the points outside the input boundaries. 50% of the training data was augmented with random horizontal flipping.

**Model Architecture and Implementation**

We compared two well-known architectures; i) DenseNet121[13] and ii) InceptionResNetV2[14]. InceptionResNetV2 combines the Inception architecture (which is a very deep convolutional neural network) with residual connections while DenseNets are used to simplify the connectivity pattern between layers by connecting all layers directly with each other. While residual connections (used in Inception networks) sum up outputs of multiple layers, DenseNets concatenate outputs of multiple connected layers. DenseNets are known to avoid learning redundant feature maps and have much better feature reuse than traditional convolutional neural networks.

InceptionResNetV2 consists of 782 layers and 54,283,877 trainable parameters (Figure 2) and is more complex than DenseNet121 which has 429 layers and 6,958,981 trainable parameters (Figure 3). It intuitively follows that DenseNet121 requires less memory than InceptionResNetV2 and is less susceptible to the vanishing-gradient problem. InceptionResNetV2 achieves better top-1 accuracy on ImageNet-1k validation set for image classification task[15].

![Figure 2: InceptionResNetV2 Architecture. (I) Schematic of the InceptionResNetV2 model; (II) A, B, C are Inception modules which comprise of several convolutional layers; (III) A and B are reduction modules which reduce the size of the output[13]](image)

We selected these two architectures to assess the impact of model complexity on external test performance. We largely preserved each architecture while adapting it to our classification task. We removed the top layer and replaced it with a global average 2D pooling layer and a dense layer with sigmoid activation as the last fully-connected layer. Parameter n was set to 5 for the last layer to match the 5 labels of our classification problem.

We used binary cross-entropy loss and Adam accumulate optimizer for training each network starting with pre-trained ImageNet weights. For model training, we set learning rate = 1 × 10^{-3}, decay = 1 × 10^{-5}, and batch size = 32. We used GTX 1080 Ti GPU and Keras Python library.
Experiments

We used three distinct ways to assess the generalization capabilities of state-of-the-art neural networks. For ChestX-ray14 and CheXpert, we randomly split the patients to train, validation, and test sets (Figure 4, 5). For the MIMIC-CXR-JPG, we kept the original test set of 5,159 images from 293 patients (Figure 6). The details of the three evaluation configurations are described in detail below.

1) In the first configuration, we trained our models on the train set of CheXpert. We tested the models on the test set of the CheXpert dataset for the in-sample data and the test set of the MIMIC-CXR-JPG dataset as the external test set. 
2) In the second configuration, we trained our models on the train set of ChestX-ray14. Models were tested on the test set of the ChestX-ray14 dataset as the in-sample test and the test set of the MIMIC-CXR-JPG dataset as the external test. 
3) In the third configuration, to increase the variation of training samples, we trained our models on the combination of train sets of CheXpert and ChestX-ray14 datasets. Models were tested on combined test sets of CheXpert and ChestX-ray14 datasets as the in-sample and the MIMIC-CXR-JPG test set as the external test.

![Figure 3: DenseNet121 Architecture (Dx: Dense Block x, Tx: Transition Block x, DLx: Dense Layer x)](image-url)

**ChestX-ray14 Dataset**

| Dataset       | # Images   | # Patients | Male (%) | Female (%) |
|---------------|------------|------------|----------|------------|
| **Train Set** | 80,657     | 22,179     | 54%      | 46%        |
| Positive      |            |            |          |            |
| Cardiomegaly  | 1,944      |            |          |            |
| Edema         | 1,622      |            |          |            |
| Consolidation | 3,357      |            |          |            |
| Atelectasis   | 8,321      |            |          |            |
| Pleural Effusion | 9,739   |            |          |            |
| **Test Set**  | 22,294     | 6,161      | 54%      | 46%        |
| Positive      |            |            |          |            |
| Cardiomegaly  | 618        |            |          |            |
| Edema         | 454        |            |          |            |
| Consolidation | 951        |            |          |            |
| Atelectasis   | 2,207      |            |          |            |
| Pleural Effusion | 2,553    |            |          |            |
| **Validation Set** | 9,169 | 2,465 | 54% | 46% |
| Positive | | | | |
| Cardiomegaly | 214 | | | |
| Edema | 227 | | | |
| Consolidation | 359 | | | |
| Atelectasis | 1,031 | | | |
| Pleural Effusion | 1,025 | | | |

**Figure 4: Flowchart of images used in this study from the ChestX-ray14 dataset for train, validation, and test sets.**
Results

In this section, we present a discussion on the results of our experiments. We also include observations made by a radiologist on the performance of the two models.

Model Performance

We performed thorough experimentation to evaluate the robustness and generalization capabilities of two very popular deep learning classification architectures, i.e., InceptionResNetV2 and DenseNet121, for the classification task of common chest diseases diagnosis. As explained in the previous section, we used three different training schemes and tested each trained model over internal and external test sets. Our diagnosis task includes 5 labels (Cardiomegaly, Edema, Atelectasis, Consolidation, Pleural Effusion) common to the three datasets that we used. We employed AUC (area under the curve of receiver operating characteristics curve) as an evaluation measure. The three public datasets have severe class imbalance for the five selected pathologies. The overall performance on both models in terms of AUC values is reported in Table 1.

The two architectures have widely different complexity in terms of the depth of the network as well as the number of trainable parameters. Still, both architectures have the same performance on internal and external test sets. In general, the performance of each model on the internal test set is better than its performance on the external test set for every configuration. As demonstrated in Table 1, both architectures have better performance for the diagnosis of Cardiomegaly and Edema on the internal test set compared to the external test set under evaluation configuration 1 (training over CheXpert dataset). On the other hand, performance for diagnosis of Atelectasis, Consolidation, and...
Pleural Effusion is quite similar for internal and external test sets under the same configuration for both architectures. Both architectures have much better performance on internal test sets for all labels under configuration 2 (training over ChestX-ray14 dataset) as compared to corresponding performance values under configuration 1. On the other hand, performance on external test sets is worse than the corresponding internal test performance for both architectures for all labels under this configuration.

**Table 1:** Performance of InceptionResNetV2 and DenseNet121 models for all evaluation configurations for internal and external test sets in terms of AUC-ROC value

| Model         | Evaluation Configuration (Training/Validation set) | Comparison Type | Test Set                    | Cardiomegaly | Edema | Consolidation | Atelectasis | Pleural Effusion |
|---------------|-----------------------------------------------------|-----------------|-----------------------------|--------------|-------|---------------|-------------|------------------|
| DenseNet121   | Configuration 1 (CheXpert)                          | Internal        | CheXpert                    | 0.853        | 0.852 | 0.714         | 0.711       | 0.864            |
|               |                                                     | External        | MIMIC-CXR-JPG               | 0.735        | 0.808 | 0.706         | 0.704       | 0.854            |
|               | Configuration 2 (ChestX-ray14)                      | Internal        | ChestX-ray14                | 0.905        | 0.88  | 0.797         | 0.803       | 0.873            |
|               |                                                     | External        | MIMIC-CXR-JPG               | 0.674        | 0.732 | 0.614         | 0.679       | 0.792            |
|               | Configuration 3 (CheXpert and ChestX-ray14)         | Internal        | CheXpert and ChestX-ray14   | 0.878        | 0.900 | 0.747         | 0.775       | 0.893            |
|               |                                                     | External        | MIMIC-CXR-JPG               | 0.744        | 0.817 | 0.694         | 0.729       | 0.859            |
| InceptionResNetV2 | Configuration 1 (CheXpert)                        | Internal        | CheXpert                    | 0.854        | 0.857 | 0.719         | 0.716       | 0.866            |
|                |                                                     | External        | MIMIC-CXR-JPG               | 0.729        | 0.815 | 0.694         | 0.716       | 0.856            |
|                | Configuration 2 (ChestX-ray14)                      | Internal        | ChestX-ray14                | 0.910        | 0.871 | 0.796         | 0.805       | 0.870            |
|                |                                                     | External        | MIMIC-CXR-JPG               | 0.651        | 0.702 | 0.626         | 0.682       | 0.773            |
|                | Configuration 3 (CheXpert and ChestX-ray14)         | Internal        | CheXpert and ChestX-ray14   | 0.881        | 0.902 | 0.752         | 0.781       | 0.894            |
|                |                                                     | External        | MIMIC-CXR-JPG               | 0.734        | 0.808 | 0.706         | 0.724       | 0.862            |

Configuration 3 involves training over a larger and more generalized set by combining training sets of ChestX-ray14 and CheXpert datasets. We observed similar performance comparisons between internal and external test sets for both architectures as was observed in the other two configurations. Performance for internal sets is generally better than performance for external sets. There is a mixed trend in terms of performance improvement for internal sets as compared to corresponding performance values for the other two configurations. For both architectures, AUC for Edema and Pleural Effusion are better than corresponding values of other configurations for the same models. On the other hand, external test set performance improves for almost all labels under this configuration for both models. Hence, generalized training sets seem critical in improving the generalization capabilities of trained models.
Figure 7: ROC curves for DenseNet121 for all evaluation configurations

Figure 8: ROC curves for InceptionResNetV2 for all evaluation configurations
Increasing the difference in the prevalence of all the 5 diseases changes model performance with improved AUCs on the external sets when compared to corresponding values of other configurations except for two cases. First, DenseNet121 trained at CheXpert to detect Consolidation has better external test AUC (0.706) compared to joint CheXpert-ChestX-ray14 (AUC = 0.694). Second, InceptionResNetV2 trained at CheXpert to detect Edema has better external test AUC (0.815) compared to joint CheXpert-ChestX-ray14 (AUC = 0.808).

ROC curves for all evaluation configurations are displayed in Figures 7 and 8 for DenseNet121 and InceptionResNetV2 respectively. It is evident that ROC curves have larger areas under the curve for internal test sets for all configurations for both models than corresponding curves for external test sets. For the joint training set (CheXpert-ChestX-ray14), curves have larger areas under the curve for the external test than the other two configurations for both models. This trend indicates better generalization capacity of the trained model under joint training configuration.

Radiologist Evaluation

We observed that the models tend to perform very differently for different labels. For example, the performance of both models is quite poor for Cardiomegaly under all evaluation configurations. We consulted a board-certified radiologist to further investigate this trend. The radiologist reviewed randomly chosen MIMIC-CXR-JPG images (external test set) of Cardiomegaly label that were incorrectly labeled by trained models under configuration 3, i.e., trained over joint CheXpert and ChestX-ray14 datasets. Table 2 shows a few samples of such images with their groundtruth label and predicted probability of that label for each model.

Radiologists rely on the cardiothoracic ratio to diagnose Cardiomegaly. It is the ratio of maximal horizontal cardiac diameter and maximal horizontal thoracic diameter (inner edge of ribs/edge of pleura) and is measured on a Posteroanterior (PA) chest X-ray. A normal measurement should be less than 0.5. The first and third images shown in Table 2 were incorrectly predicted to have Cardiomegaly by DenseNet121. However, InceptionResNetV2 was able to correctly predict the absence of Cardiomegaly for these images. Also, the DenseNet121 model predicted incorrectly that the second image shown in Table 2 does not have Cardiomegaly. InceptionResNetV2 correctly predicted Cardiomegaly for this image. On the other hand, InceptionResNetV2 incorrectly predicted the fourth image shown in Table 2 to have Cardiomegaly. Radiologist assessed that these are all borderline Cardiomegaly cases based on the cardiothoracic ratio. The AI models performed poorly for borderline cases. Another observation is that the pacemaker is closely tied with Cardiomegaly for both models. The presence of this foreign object may bias the decision of the DCNN models.

Table 2: Incorrectly classified radiographic images of borderline Cardiomegaly pathology; Groundtruth label for each image as well as probabilities of that label predicted by both models are included. The maximum diameter of the heart and the thoracic diameter were measured by the radiologist (shown by red lines) to estimate cardiothoracic ratio to diagnose Cardiomegaly. All images were classified as Borderline Cardiomegaly by the radiologist.
Conclusion

In this paper, we thoroughly studied the robustness and generalization capabilities of complex deep learning classification models. We used three publicly available chest X-ray datasets (CheXpert, ChestX-ray14, MIMIC-CXR-JPG) and experimented with two different DCNN models (InceptionResNetV2, DenseNet121). Our experiments indicate that these models have limited generalization capacity when tested over images outside of the training dataset, i.e., external test set. For all class labels under every evaluation configuration, the performance of each model is better for the internal test set than it is for the external test. We also worked on improving generalization capabilities of these models. Our technique relies on improving the quality of the training data by combining images from different datasets, thus increasing the data variation the models are exposed to during the training phase. This technique proved effective as performance of models was significantly improved for external test sets. In this case, even incorrectly predicted labels tend to be borderline cases of their corresponding pathologies. Interestingly, the performance of each model for internal test sets remains approximately the same level. Therefore, we can conclude that generalization of deep learning classification models to a larger variety of items is heavily dependent on the quality and heterogeneity of the training dataset. Exposing the model to multiple datasets with wide variation during the training phase is an effective technique for improvement in the generalization capabilities of the trained model.

References

1. van Assen M, Banerjee I, De Cecco CN. Beyond the artificial intelligence hype: what lies behind the algorithms and what we can achieve. Journal of Thoracic Imaging. 2020;.
2. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJ. Artificial intelligence in radiology. Nature Reviews Cancer. 2018;18(8):500–510.
3. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. CheXNet: radiologist-level pneumonia detection on chest x-rays with deep learning. CoRR. 2017;abs/1711.05225. Available from: [http://arxiv.org/abs/1711.05225](http://arxiv.org/abs/1711.05225).
4. Majkowska A, Mittal S, Steiner DF, Reicher JJ, McKinney SM, Duggan GE, et al. Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. Radiology, 2020 Feb;294(2):421–431. Available from: [https://doi.org/10.1148/radiol.2019191293](https://doi.org/10.1148/radiol.2019191293).
5. Baltruschat IM, Nickisch H, Grass M, Knopp T, Saalbach A. Comparison of deep learning approaches for multi-label chest x-ray classification. Scientific Reports. 2019 Apr;9(1). Available from: [https://doi.org/10.1038/s41598-019-42294-8](https://doi.org/10.1038/s41598-019-42294-8).
6. Kim DW, Jang HY , Kim KW, Shin Y, Park SH. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. Korean journal of radiology. 2019;20(3):405–410.
7. Johnson A, Peng Y, Lu Z, Mark R, Berkowitz S, Horng S. MIMIC-CXR-JPG - chest radiographs with structured labels; 2019. Available from: [https://doi.org/10.13026/8360-t248](https://doi.org/10.13026/8360-t248).
8. Johnson AEW, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng C, et al. MIMIC-CXR: a large publicly available database of labeled chest radiographs. CoRR. 2019;abs/1901.07042. Available from: [http://arxiv.org/abs/1901.07042](http://arxiv.org/abs/1901.07042).
9. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet. Circulation. 2000 Jun;101(23). Available from: [https://doi.org/10.1161/01.cir.101.23.e215](https://doi.org/10.1161/01.cir.101.23.e215).
10. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, et al. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. CoRR. 2019;abs/1901.07031. Available from: [http://arxiv.org/abs/1901.07031](http://arxiv.org/abs/1901.07031).
11. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. CoRR. 2017;abs/1705.02315. Available from: [http://arxiv.org/abs/1705.02315](http://arxiv.org/abs/1705.02315).
12. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLOS Medicine. 2018 Nov;15(11):e1002683. Available from: https://doi.org/10.1371/journal.pmed.1002683.

13. Huang G, Liu Z, Weinberger KQ. Densely connected convolutional networks. CoRR. 2016;abs/1608.06993. Available from: http://arxiv.org/abs/1608.06993.

14. Szegedy C, Ioffe S, Vanhoucke V. Inception-v4, Inception-ResNet and the impact of residual connections on learning. CoRR. 2016;abs/1602.07261. Available from: http://arxiv.org/abs/1602.07261.

15. Bianco S, Cadène R, Celona L, Napoletano P. Benchmark analysis of representative deep neural network architectures. CoRR. 2018;abs/1810.00736. Available from: http://arxiv.org/abs/1810.00736.

16. Pizer SM, Amburn EP, Austin JD, Cromartie R, Geselowitz A, Greer T, et al. Adaptive histogram equalization and its variations. Computer vision, graphics, and image processing. 1987;39(3):355–368.

17. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. Ieee; 2009. p. 248–255.

18. Wang Y, Guan Q, Lao I, Wang L, Wu Y, Li D, et al. Using deep convolutional neural networks for multi-classification of thyroid tumor by histopathology: a large-scale pilot study. Annals of Translational Medicine. 2019;7(18). Available from: http://atm.amegroups.com/article/view/28771.

19. Ruiz P. Understanding and visualizing DenseNets. Towards Data Science; 2018. Available from: https://towardsdatascience.com/understanding-and-visualizing-densenets-7f688092391a.