CheXNet Model Card

Christian Garbin

November 2020

1 CheXNet

This paper converts the description of the CheXNet model from the prose of its original paper into the structured format of a model card [1].

In late 2017, a team from Stanford announced CheXNet, "radiologist-level pneumonia detection on chest x-rays with deep learning" [5]. It was developed based on the ChestX-ray8 dataset [6], with an important enhancement: a team of four radiologists labeled the test set (as opposed to relying on the NLP-extracted labels from ChestX-ray8).

Information for the model card was compiled from:

- The latest (third) version of the paper [5].
- CheXNet: an in-depth review [2].

The model card is written from the first-person point of view, as if the authors had created it, to make it more realistic. Whenever applicable, the source for the information used in the model card is cited.

2 Model card

Model Card - CheXNet

Model Details

- Developed by researchers at Stanford University Department of Computer Science, Department of Medicine, and Department of Radiology.
- The latest version of the model was released in December 2017.
- Dense Convolutional Neural Net (DenseNet) with 121 layers, initialized "with weights from a model pretrained on ImageNet" [5].
- See the description of the work in arXiv.

Intended Use

- Determine the "probability of pneumonia along with a heatmap localizing the areas of the image most indicative of" [5] "pneumonia-like features" [2] in frontal chest X-ray images.
- Not intended to be used with lateral chest x-rays.
- It is intended to assist health care professionals, not to perform the final diagnosis.

Factors

- Male and females instances are balanced in the training and test sets.
- The average patient age is 46.9 years, with a standard deviation of 16.6 years. There are few images of infants, toddlers, and preschoolers in the training and test sets. It is not known if the model performs well for those age groups.
- Training and test images are from one institution. It is known that "performance on chest X-rays from outside hospitals [may be] lower than on held-out X-rays from the original hospital system." [8] [3] [7]. Further evaluations with images from other institutions is an open investigation item.

Metrics

- The metrics include the F1 score of the model and four radiologists (the same radiologists who labeled the test set). The threshold for the decision is 0.5 [2].
- The metrics also include the 95% confidence interval (2.5$^{\text{th}}$ and 97.5$^{\text{th}}$ percentiles of the F1 scores), computed with bootstrap (10,000 samples, with replacement).
- "We find that the difference in F1 scores — 0.051 (95% CI 0.005, 0.084) — does not contain 0, and therefore conclude that the performance of CheXNet is statistically significantly higher than radiologist performance." [5]

Training Data

- ChestX-ray14 (frontal X-ray images), with images that have pneumonia are labeled "positive" examples and all other images are labeled as "negative" examples.
- The ChestX-ray14 images were downscaled to 224 × 224 pixels and normalized based on the mean and standard deviation of the ImageNet dataset.
- Out of the 112,120 images and 30,805 patients in ChestX-ray14, we selected 98,637 images and 28,744 patients for the training set and 6,351 images of 1,672 patients for the validation set.
- Augmentation with random horizontal flipping.

Evaluation Data

- Test set with 420 images of 389 patients. The images were not randomly selected from the full set. They were "sampled to contain at least 50 cases of each of the original [14] pathology labels." [4] [a]

- The test set was annotated independently by “four practicing radiologists at Stanford University, who were asked to label all 14 pathologies in [ChestX-ray14].” [5]
- There is no overlap of patients between the training, validation, and test sets.

Ethical Considerations

- No personally identifiable information was used to train and test the model.
- The output of the model must not be used to make automated healthcare decisions. The model is intended to provide information to assist healthcare professionals.

Caveats and Recommendations

- The images in the ChestX-ray14 dataset are downscaled from the original DICOM images to 1024 × 1024 pixels and 256-level gray scale. The DICOM images have 2-3× as many pixels, and 3,000 gray levels [6] [2].
- The images are downsized again when training the model to 224 × 224 pixels.
- The radiologists annotating the test data used the downscaled ChestX-ray14 images, not the original DICOM images.
- The radiologists did not have access to the patient records. Not knowing the patient history “decrease[s] radiologist diagnostic performance in interpreting chest radiographs” [5].
- “Detecting pneumonia in chest radiography can be difficult for radiologists. The appearance of pneumonia in X-ray images is often vague, can overlap with other diagnoses, and can mimic many other benign abnormalities. These discrepancies cause considerable variability among radiologists in the diagnosis of pneumonia”. [5]

Quantitative Analyses

	F1 Score	95% CI	
Radiologist 1	0.383	0.309	0.453
Radiologist 2	0.356	0.282	0.428
Radiologist 3	0.365	0.291	0.435
Radiologist 4	0.442	0.390	0.492
Radiologist average	0.387	0.330	0.442
CheXNet	**0.435**	**0.387**	**0.481**

Table 1: “We compare radiologists and our model on the F1 metric. . . CheXNet achieves an F1 score of 0.435 . . . , higher than the radiologist average of 0.387 . . . . We use the bootstrap to find that the difference in performance is statistically significant.” [5]

[a] This was unclear in the CheXNet paper. It was clarified in [2]. The quote is from a follow-up paper, CheXNeXt [4].

3 Appendix

3.1 Model cards

Model cards are “short documents accompanying trained machine learning models that provide benchmarked evaluation in a variety of conditions, such as across different cultural, demographic, or phenotypic groups . . . and intersectional groups . . . that are relevant to the intended application domains. Model cards also disclose the context in which models are intended to be used, details of the performance evaluation procedures, and other relevant information."

Model cards were motivated by systematic bias in commercial applications that were discovered only after the models were released. To counter that, the authors "advocate for measures of model performance that contain quantitative evaluation results to be broken down by individual cultural, demographic, or phenotypic groups, domain-relevant conditions, and intersectional analysis combining two (or more) groups and conditions." The emphasis on ethic aspects of the measurements is a distinguishing feature of model cards, compared to other proposals to document models.

References

[1] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. Model Cards for Model Reporting v2. 2018.

[2] L. Oakden-Rayner. CheXNet: an in-depth review. Link to publication 2020-07-27, 2018.

[3] E. H. P. Pooch, P. L. Ballester, and R. C. Barros. Can we trust deep learning models diagnosis? The impact of domain shift in chest radiograph classification. 2019.

[4] P. Rajpurkar, J. Irvin, R. L. Ball, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. P. Langlotz, B. N. Patel, K. W. Yeom, K. Shpanskaya, F. G. Blankenberg, J. Seekins, T. J. Amrhein, D. A. Mong, S. S. Halabi, E. J. Zucker, A. Y. Ng, and M. P. Lungren. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLOS Medicine*, 15(11):e1002686, nov 2018.

[5] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning v3. 2017.

[6] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. ChestX-ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases v5. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jul 2017.

[7] L. Yao, J. Prosky, B. Covington, and K. Lyman. A Strong Baseline for Domain Adaptation and Generalization in Medical Imaging. *arXiv e-prints*, page arXiv:1904.01638, Apr. 2019.

[8] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, 15(11):e1002683, nov 2018.