From Retinal Disease to COVID-19, Deep Learning CNNs for Biomedical Image Classification: Achieving Accuracy and Interpretability Despite Small Sample Sizes

Thomas Carr, Saber Sami, Julie Sanderson

Introduction

Convolutional Neural Networks (CNNs) are deep learning algorithms specialised for image interpretation and classification. Rather than being explicitly programmed, deep learning algorithms learn by example from a training dataset. When trained with sufficient data, CNNs can achieve human-like accuracies identifying natural images and clinician-like accuracies with medical images.

One of the greatest drawbacks of deep learning is the large training data required, usually tens or hundreds of thousands of images. Sample size represents a huge barrier, both financially and practically, for those seeking to build CNN based systems.

Our goal was to demonstrate that accurate medical image classifiers can be trained with the small datasets you would be able to collect in a typical clinical environment. By using existing data and increasing the value of our data, we provide sufficient information for our algorithm to learn from.

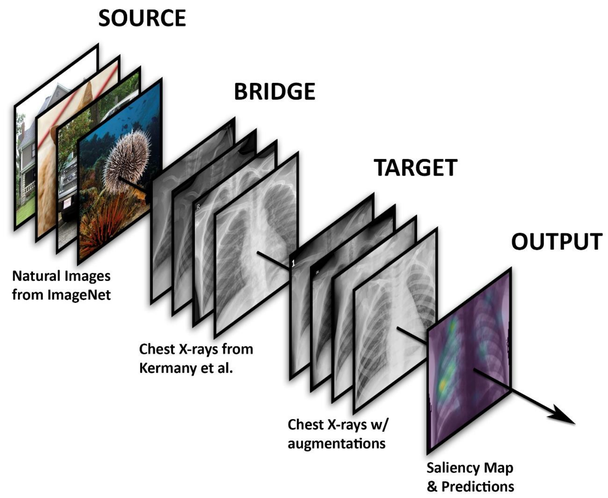

Figure 1. Training and Assessment workflow. The algorithm is initialised with Natural images, fine-tunes with same-modality images, and final training on the target dataset expanded with augmented images. The algorithm outputs saliency maps and predictions for any inputted images.

Method

To maximise our baseline performance, we used Inception V4 as our CNN architecture, one of the best performing open-source CNNs [5]. Figure 1 outlines our training and assessment workflow, with the algorithm training over three steps.

1. Initialised with parameters learned from a source dataset.
2. Fine-tune parameters on a large bridge dataset of the same image domain as the target dataset.
3. Trained on the target dataset.

Each step tunes the algorithm by first learning general image recognition, followed by domain specific features, before final training with the target [4].

We artificially boosted the size of the target dataset through data augmentation, increasing the available data without capturing additional samples. The algorithm was assessed against unseen data, producing accuracy metrics and attention visualisations to help explain how these predictions are made [6,7].

COVID-19

Chest X-rays are among the widest available biomedical imaging techniques worldwide. They are used to triage patients presenting with breathing difficulties. Currently, most patients presenting with breathing difficulties or fever are assumed to have COVID-19. In many situations there is no rapid diagnostic capability due to overwhelmed lab capacity.

Our goal was to train an algorithm to accurately differentiate those with COVID-19 pneumonia from those with Bacterial or Viral pneumonia and those without pneumonia. Our dataset was limited to 142 subjects, 102 of which were used in the target dataset, with augmentation expanding this to 1122 images per class [2].

Our bridge dataset was made up of open-source Chest X-rays depicting Clinically Normal, Viral Pneumonia, and Bacterial Pneumonia. 102 subjects were removed from the bridge dataset for target training [3].

Results: Our algorithm achieved 0.944 total accuracy (Figure 3), 0.945 class-averaged sensitivity (+/- 0.095), and 0.943 class-averaged f1-score (+/- 0.0525). Class-averaged Receiver Operating Characteristics (ROC) Area Under Curve (AUC) was 0.979. Saliency maps [7] indicate a clear right-sided bias across all classes, with the algorithms attention predominantly located within the thoracic cavity (Figure 2).

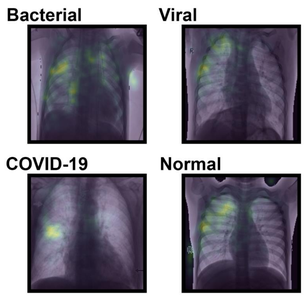

Figure 2. Chest X-ray saliency maps indicating localised importance for the correct classification. The algorithm heavily biases towards using features from the subjects' right side for classification in almost all cases.

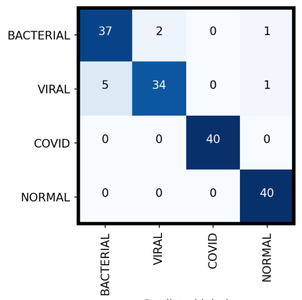

Figure 3. Confusion matrix showing performance of the Chest X-ray algorithm against a held-back assessment dataset.

Retinal Disease

Optical Coherence Tomographic (OCT) is a laser based imagine modality used frequently in ophthalmology. In many clinical settings imaging capacity is greater than the ability of clinicians to review and action.

Our goal was to train an algorithm with limited data that can accurately classify images belonging to one of four possible diagnosis's; Normal, Macula Hole (MH), Diabetic Retinopathy (DR), or Central Serous Retinopathy (CSR). We limited our dataset to 50 samples per diagnosis, with augmentation expanding this to 550 images per class (10 augmentations per sample) [1].

Our bridge dataset was made up of 108,309 open-source OCT images of either Clinically Normal, Retinal Drusen, Diabetic Macular Oedema (DME), and Choroidal Neovascularisation (CNV) [3].

Results: Our algorithm achieved 0.94 total accuracy (Figure 5), 0.93 class-averaged sensitivity (+/- 0.06), and 0.93 class-averaged f1-score (+/- 0.03). Class-averaged Receiver Operating Characteristics (ROC) Area Under Curve (AUC) was 0.983. SHAP maps [6] indicate that the algorithm successfully identified the location of diagnostically important pathology (Figure 4).

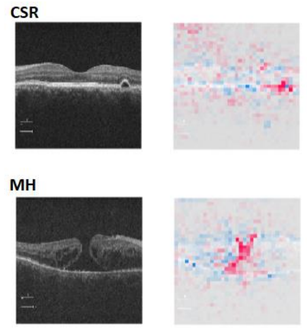

Figure 4. OCT SHAP maps indicating positive importance in red and negative importance in blue. In both cases the algorithm correctly associates areas of pathology with the correct diagnosis.

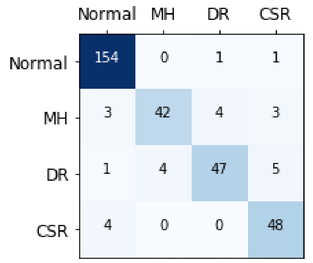

Figure 5. Confusion matrix showing performance of the OCT algorithm against a held-back assessment dataset.

References: [1] Gholami, P., Roy, P., Parthasarathy, M. K., & Lakshminarayanan, V. (2020). OCTID: Optical coherence tomography image database. *Computers and Electrical Engineering*, *81*. https://doi.org/10.1016/j.compeleceng.2019.106532
[2] Cohen, J. P., Morrison, P., & Dao, L. (2020). COVID-19 Image Data Collection. Retrieved from http://arxiv.org/abs/2003.11597
[3] Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C. S., Liang, H., Baxter, S. L., … Zhang, K. (2018). Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell*, *172*(5), 1122-1131.e9. https://doi.org/10.1016/j.cell.2018.02.010
[4] Kim, H. G., Choi, Y., & Ro, Y. M. (2017). Modality-bridge Transfer Learning for Medical Image Classification. *Proceedings - 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics, CISP-BMEI 2017*, *2018-January*, 1–5. Retrieved from http://arxiv.org/abs/1708.03111
[5] Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017). Inception-v4, inception-ResNet and the impact of residual connections on learning. In *31st AAAI Conference on Artificial Intelligence, AAAI 2017* (pp. 4278–4284). AAAI press.
[6] Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, *2017-December*, 4766–4775. Retrieved from http://arxiv.org/abs/1705.07874
[7] Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *2nd International Conference on Learning Representations, ICLR 2014 - Workshop Track Proceedings*. Retrieved from http://arxiv.org/abs/1312.6034