Classification of Chest X-ray Images and Patient Metadata Using Multi-Modal Models

Antoine Bonnet, Silvia Romanato & Alexander Sternfeld

I. Introduction

Chest radiography is a crucial diagnostic tool for a variety of conditions, including heart failure and bone fractures. However, the accurate analysis of radiographic images is constrained by the limited availability of radiologists. Research has shown that deep learning models can be applied with success to determine medical conditions from imaging sources such as X-rays. However, these visual models do not factor contextual information on the patient for their diagnosis. Our work expands on these findings by building a multi-modal model that combines patient data with the corresponding chest X-rays. We aim to show that a model that incorporates contextual information into its visual diagnosis outperforms a vision-only model.

II. Data

Two datasets are used for this project: MIMIC-III and MIMIC-CXR-JPG. Both are part of an extensive database of electronic health records of around 60,000 patients from the intensive care unit at Beth Israel Deaconess Medical Center in Boston between 2001 and 2012 [1]. MIMIC-CXR-JPG contains 377,110 chest X-ray JPEG images corresponding to frontal views, in either posteroanterior (PA) or anteroposterior (AP) beam configurations, as well as lateral views. As PA images are the gold standard of chest X-ray images [2], we will use these images as frontal views and filter out AP images. These imaging studies are provided along with 227,827 associated radiology reports from which were extracted 14 different labels indicating the condition of the patient, e.g. lung lesion, fracture or pneumonia. On the other hand, MIMIC-III contains a variety of patient characteristics, such as the age, gender, type of insurance, marital status and ethnicity. Both databases are linked by patient IDs, allowing us to combine the chest X-rays with the features of the corresponding patient. In our research, we will use chest X-ray images and patient characteristics as features, and the condition of the patient as the target label.

III. Model Architecture

A. Multimodality

The aim of our model will be to combine vision (chest X-ray images) and tabular (structured patient information) inputs to produce a comprehensive diagnosis. This approach enables the model to rely on contextual information about the patient such as their age or gender to inform its clinical diagnosis. As proposed by [3], a multi-modal model consists of multiple encoders tailored for different data modalities. All encoders produce a latent representation of their respective inputs, which are then aggregated by the model by concatenating their embedding vectors. This combined representation is then used by a classification head to produce a label — in our case the condition of the patient.

B. Vision encoders

The MIMIC dataset contains both frontal and lateral chest X-Ray images. To use both views simulatenously, we will leverage the dual encoder architecture suggested by [4] and employ two visual encoders, in addition to the tabular encoder. One visual encoder will be trained on frontal images, while the other will be trained on lateral images.

We will experiment with multiple architectures to embed chest X-rays using visual encoders. Convolutional Neural Networks (CNNs) architectures have been applied with success to the task of chest X-ray classification [5]. More recently, the Vision Transformer (ViT) architecture proved to be a viable alternative to CNN-based architectures, leading to state-of-the-art performance on chest X-Rays classification benchmarks [6]. To illustrate, [7] show that Vision Transformers obtain an F1 score of over 0.95 over three different related benchmarks.

Specifically, we will compare the performance of three different models; the CNN-based `ResNet50` [8] and `Densenet` [9] as well as a Vision Transformer (ViT) model (`Google/vit-base-patch16-224`)[10]. All 3 models were pre-trained on millions of images from ImageNet and will be fine-tuned on the MIMIC data.

To encode tabular patient information, we will use a simple fully-connected network. The embeddings for the front view, the lateral view and the tabular data will be concatenated and passed as input for the classification head. For the classification head, we again use a fully connected neural network.

IV. Experimental setting

To compare the performance of our multi-modal model, we will train 7 different models. As a baseline we will train a full-connected network to identify conditions using only the patient features. Since we compare two CNN's and one ViT for the vision data, we will train three models on X-ray images only and three models using both image and tabular data.

To allow for a fair comparison between the different settings, we split the data into a train, validation and test sets using 70%, 10%, 20% splits. For each of the models, we will tune the hyperparameters on the validation set. All computations will be performed using Google Cloud compute.

V. Conclusion

This work addresses the inherent limitations of vision-only models for chest X-ray classification by combining radiographic images with patient characteristics in a multi-modal architecture. By considering factors such as age, gender, and medical history alongside chest X-ray images, the multi-modal architecture aims to provide a more holistic understanding of each patient's health profile and enhance its classification performance by providing a contextualized diagnosis. The findings may have a significant impact in areas where there are insufficient radiologists by providing a high-quality automated analysis of chest X-rays.

REFERENCES

[1] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific Data*, vol. 3, no. 1, 2016.

[2] Y. Akhter, R. Singh, and M. Vatsa, "Ai-based radiodiagnosis using chest x-rays: A review," *Frontiers in Big Data*, vol. 6, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fdata.2023.1120989

[3] P. Hager, M. J. Menten, and D. Rueckert, "Best of both worlds: Multimodal contrastive learning with tabular and imaging data," 2023.

[4] J. Rubin, D. Sanghavi, C. Zhao, K. Lee, A. Qadir, and M. Xu-Wilson, "Large scale automated reading of frontal and lateral chest x-rays using dual convolutional neural networks," 2018.

[5] K. Almezhghwi, S. Serte, and F. Al-Turjman, "Convolutional neural networks for the classification of chest x-rays in the iot era," *Multimedia Tools and Applications*, vol. 80, no. 19, p. 29051–29065, 2021.

[6] G. I. Okolo, S. Katsigiannis, and N. Ramzan, "Ievit: An enhanced vision transformer architecture for chest x-ray image classification," *Computer Methods and Programs in Biomedicine*, vol. 226, p. 107141, 2022.

[7] S. Regmi, A. Subedi, U. Bagci, and D. Jha, "Vision transformer for efficient chest x-ray and gastrointestinal image classification," 2023.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[9] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2018.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.