A Graph Convolutional Network with Localized Convolution and Readout Operations for Diagnosing Chest X-Rays Using Radiologist Gaze Data

Antonio Nascimento Lutfi
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo
Email: lutfi@usp.br

João do Espírito Santo Batista Neto
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo
Email: jbatista@icmc.usp.br

***Abstract*—In this work-in-progress report we propose a Graph Convolutional Network (GCN) capable of diagnosing chest x-rays using radiologist captured data for training. While other neural networks are capable of making inference on medical image exams with gaze data, the examples found in literature use architectures that combine this data with traditional CNNs that learn from the whole image. Our model, on the other hand, learns from a graph of gaze fixations as nodes, each accompanied by a feature vector describing only their region of observation. Such graph is, naturally, euclidean. Traditional convolution and readout operations in GCNs are not conceived to leverage local features and attributes of euclidean graphs, usually aggregating nodes and edges into a whole-graph representation. Our approach divides the graph in a grid, performing such operations in small regions as to preserve local features. With this we aim to prove two hypotheses: 1) a model can learn from specialist gaze data over an image without being paired with the image in its original structure and 2) it is possible to take advantage of euclidean graphs by not aggregating local features in graph convolution and readout layers.**

I. Introduction

Convolutional Neural Networks (CNNs) have been a standard ML tool for learning and inference in domains with euclidean data, such as images and sound. To leverage the power of the convolution operation into non-euclidean domains, the Graph Neural Network (GNN) [1] and the Graph Convolutional Network (GCN) [2] have been conceived. Instead of using a fixed sized kernel, convolution in a graph is a weighted aggregation of the features of a node and its neighbors'. GCNs have been adopted in a wide range of applications, such as recommendation systems; molecular biology; urban traffic optimization; social network analysis and prediction; among many others [3].

However, some of the data that can be intuitively thought as graphs show euclidean properties when modelled as such. This is the case of gaze tracked data. If one models each eye fixation as a node, these nodes are distributed in a 2D euclidean space that represent the observed area. GCNs typically aggregate all nodes into a graph wide representation (readout) before feeding this aggregation into an MLP for inference. In doing so, nodes of different regions of this 2D space are treated uniformly, possibly losing local particularities.

It is not uncommon for Machine Learning (ML) medical imaging applications to learn from images combined with eye tracking data captured from specialists while examining such images [4]. The present research is one of such instances. In this work-in-progress report, we present a strategy to diagnose chest x-rays (CXRs) using radiologist gaze data, using the REFLACX [5] dataset. We do so by modelling the gaze fixations as nodes in a graph, where the features of each node correspond to the local image features for that fixation. Contrary to existing approaches, however, our model only learns from the fixation graph and these localized features, not performing any learning on the whole medical image.

Hence, the aim of this ongoing research is two-fold: 1) investigate whether it is possible to generate correct diagnoses for CXRs using only local image features and eye fixation data; and 2) while doing so, develop a GNN that takes into consideration the euclidean nature of the fixation graph.

The following sections describe the current state of this research, and its possible future directions. Section II presents the recent works related to our research domain. Section III describes the structure of the adopted dataset, as well as its transformation into a graph of gaze fixations. In section IV we present the proposed model architecture, detailing its components and how they could attain the preservation of euclidean local features during the feed forward process. The current state of the research, as well as its next steps are presented in section V. We end this paper with a brief conclusion in VI.

II. Related Work

The references for this work are compiled according to three research fields we find most relevant: state of the art graph convolutional network literature (II-A); works that use gaze-tracking as insight about medical data (II-B); and specific datasets that can be used to test our hypothesis (II-C).

A. Graph Convolutional Networks

A survey containing industrial applications of graph neural networks was published in Ref. [3]. Most of the applications

of graphs for inference do not have input data in an euclidean space..

In computer vision applications, however, data modelled as euclidean graphs as inputs to GNNs is a common feature. Ref. [6] extracts 3D meshes from 2D pose images using a GNN.

The seminal works on GNN form the theoretical basis for the model proposed in this article. Graph Convolutional Networks (GCN) [7] provide a manner to perform the convolution operation, successful in learning image features, over nodes of a graph. Similarly to the original convolution operation, a node n' in a graph g' will be calculated as some aggregation of the original node n with its neighbors in graph g. After this operation was adopted, GCN and GNN became interchangeable terms. A further development, Graph Attention Networks (GAT) were first proposed in [2], providing a learnable weight matrix for the edges, effectively implementing a self-attention mechanism in GNNs.

Visual GNNs research adapts a lot from both in visual and NLP transformers [8]–[11]. In the next steps of our work – detailed in V-A– we intend to apply some of this knowledge as well.

B. Gaze Data for Inference in Medical Imaging

The term gaze-tracking has two main application in medical research: 1) understand eye movement behaviour in patients and 2) extract medical specialists –usually radiologists– gaze data to elucidate how trained professionals observe medical exams. Our research is concerned only with the second instance. Precisely in this scope, [4] provides an comprehensive survey of recent studies where medical gaze is combined with machine learning models.

In Ref. [12], a GNN is proposed to solve a problem very similar to ours, diagnosing chest x-rays using gaze data. However, the devised graph deploys a multi-modal node structure, using the total amount of gaze time spent in a region as an extra information to the original image.

Ref. [13] presents a study where eye-tracking was employed to differentiate between novice and experienced ophthalmologists while they screened exams for glaucoma.

Ref. [14] examines how image areas observed by experts for a long time, but not marked as relevant, often contribute to false positives in NNs that classify vulvovaginal candidiasis. It also presents a model that takes this multi-modal data and improves the accuracy in detecting the disease.

C. Relevant Datasets

MIMIC-CXR [15] is a dataset of chest x-ray images, containing 377,110 de-identified data points. It is one of the largest and most popular medical imaging datasets.

REFLACX [5] and EYE-Gaze [16] are both datasets that capture radiologists gaze data while examining a subset of MIMIC x-rays. The former contains 3032 data points and, the latter, 1038. In the current stage of the research, we are using REFLACX to train and test our model.

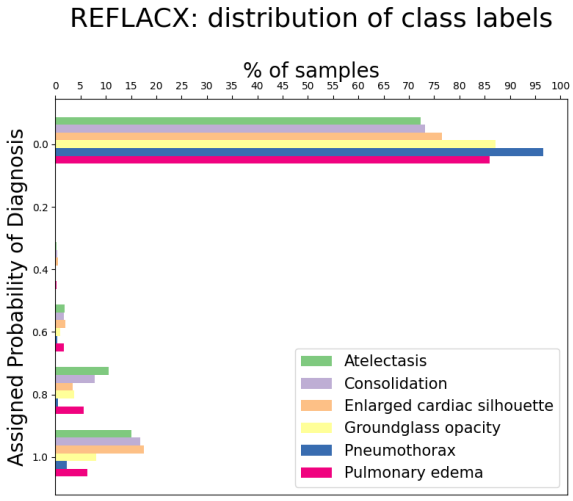

Fig. 1. REFLACX [5] dataset label distribution for 6 abnormalities. For each abnormality studied, a 0 to 5 likelihood is assigned by the observing radiologists. In this work, these values were normalized to a [0, 1] interval and treated as a probability. The significant imbalance between positive and negative cases is an issue to be tackled during the next steps of the research. There are more than six abnormalities, but these are the ones shared by all data-points. Therefore, at this stage of the research, they are the only ones being considered, as they maximize the available data volume for training.

III. Dataset Structure

This section provides details of the data structures used in the project. Section III-A briefly describes the REFLACX [5] data and its relationship to the original MIMIC-CXR [15] dataset. Additionally, it also explores REFLACX data distribution. Section III-B details the fixation graph dataset that is the actual input to the proposed model and how it is compiled from REFLACX.

A. REFLACX Structure and Data Distribution

Each REFLACX data-point is a chest x-ray image from MIMIC-CXR, accompanied by an observation made by a radiologist. This observation is comprised of this radiologist's captured gaze fixations, timed transcript of their recorded voice, and their final diagnoses (figure 3). Each original MIMIC image can be the subject of more than one REFLACX data-point. The dataset is divided into 3 separate phases, differing from one another only by the types of abnormalities being screened for. Only the abnormalities present in all three phases are being considered in this research (figure 1), so that all of the 3052 data-points can be leveraged together. For each x-ray image, a bounding box of the chest area is also provided by the dataset.

As illustrated in figure 1, REFLACX data is significantly imbalanced in favor of x-rays without abnormalities. This asymmetry poses a problem for training the model that will have to be addressed in the next stages of the research.

As for the structure of the fixation data, each data-point contains a list of gaze fixations. Each fixation contains:

- Start and end timestamps;
- (x, y) gaze position in image coordinates;

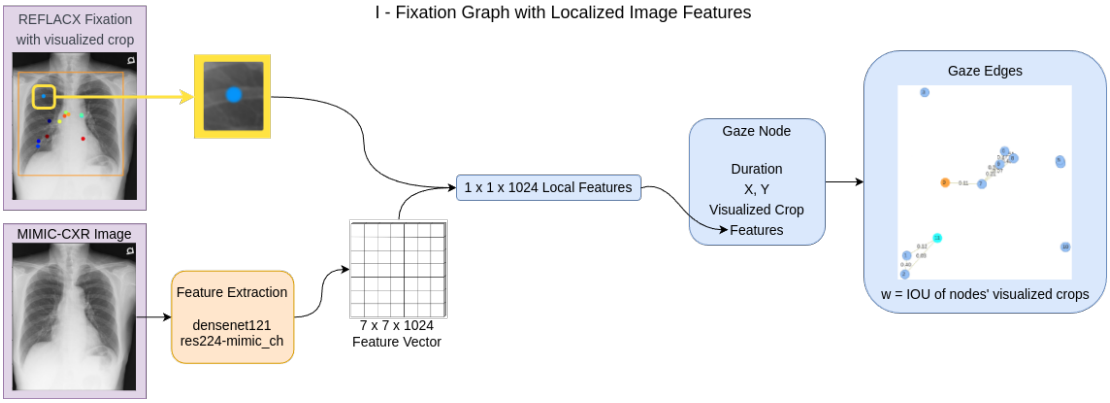

Fig. 2. Transformation of a REFLACX [5] data-point into a graph. DenseNet [17] generates a 7x7x1024 feature map for the corresponding MIMIC-CXR [15] x-ray. Since each fixation contains the visualized image area around the fixated point itself, it's possible to crop a 1x1x1024 feature vector from the DenseNet feature map to represent the fixation region. This feature vector, together with the fixation's duration time, (X, Y) position, and visualized area limits make up for a node's features. An edge exists between nodes if their visualized regions overlap. Its weight is the intersection over union (IOU) between these regions. There is a self edge for all nodes of weight 1 that is omitted.

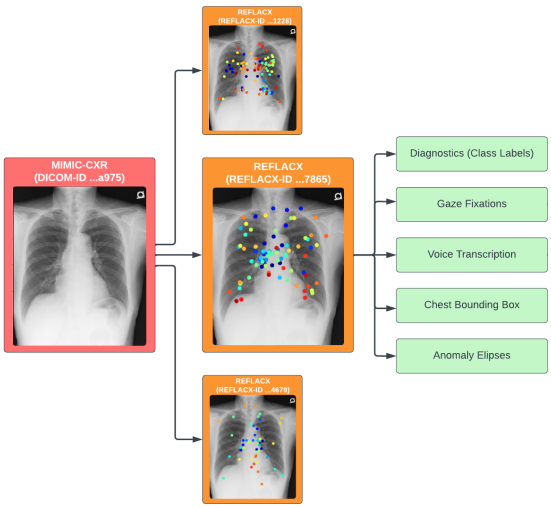

Fig. 3. REFLACX [5] dataset structure summary. Each original MIMIC-CXR [15] image is examined by one or more radiologists. Each pair of image and new examination is a REFLACX data-point, containing the diagnoses; captured gaze fixations; the radiologist's timed voice transcription; manually drawn ellipses around found anomalies; and a bounding box around the chest.

- captured pupil area;
- the angular resolution: calculated pixels per degree of vision, vertically and horizontally;
- the crop of the image that was being displayed on screen (zoom level);
- the location of this last crop in screen coordinates.

B. Fixation Graph Dataset

Each graph node of our GNN architecture corresponds to a gaze fixation in a REFLACX data-point, together with information about the region observed by that fixation in the image.

The dimensions of this region are modelled as a normal distribution. The (x, y) position is the distribution mean point and the standard deviation is one degree of vision, which can be converted into pixels using the available angular resolution. This is how it is implemented in REFLACX's original code [5] while processing the gaze heatmaps, and we adopted the same method. To select the area of the image corresponding to a specific fixation, one needs to select the number of standard deviations and crop the image around the fixation's position. In the experiments performed so far, the crops contain one standard deviation around the fixations.

Hence, a fixation node structure is as follows:

- (x, y) position inside chest bounding box, normalized to [0, 1];
- duration of the fixation, in seconds;
- coordinates of the region visualized by the fixation;
- localized image features corresponding to the visualized region.

For the localized image features, a 7x7x1024 feature map is generated by feeding the entire x-ray through densenet121 [17]. These features are then cropped in a region corresponding to the fixation crop, generating a 1024-sized local feature vector (figure 2).

As for the edges of the graph, many are the possibilities. Three of them have been considered so far: a) connect fixations in the order they were observed; b) inverse euclidean distance, connecting all nodes; and c) intersection over union of fixation crops. In the current stage of the research, we deemed the last option to be the most promising. Two nodes have a connecting edge if their fixation crops overlap, and the edge weight is the intersection over union (IOU) between these crops. All nodes have an IOU of 1 with themselves, so a self edge was added. This prevents silent regression of lone nodes during the graph convolution, and facilitates all convolution calculations.

IV. Methods

This section describes in details the approach used to address the research questions proposed in Section I. Subsection

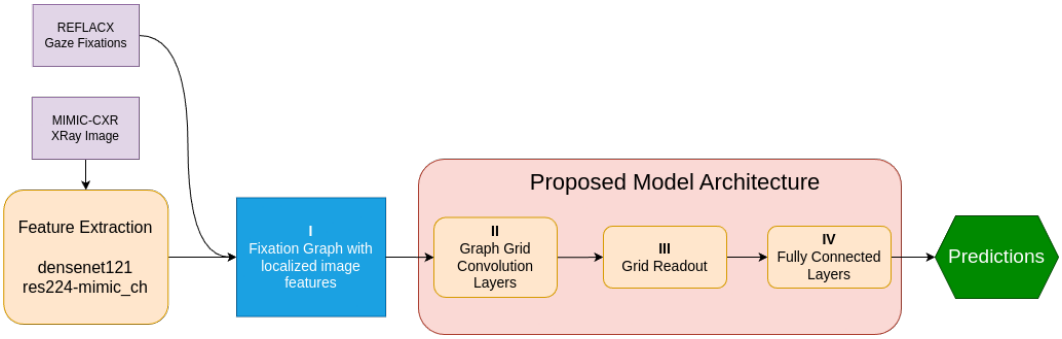

Fig. 4. Summary of model architecture and REFLACX [5] data-flow. Feature extraction is performed on the original MIMIC-CXR [15] x-ray. Then, for each of REFLACX data-point gaze fixations, localized features are extracted from the feature map. A graph with fixations as nodes is then fed forward to the proposed GNN for predictions. Each numbered sub-process is explained in more details in figures 2, 5, and 6.

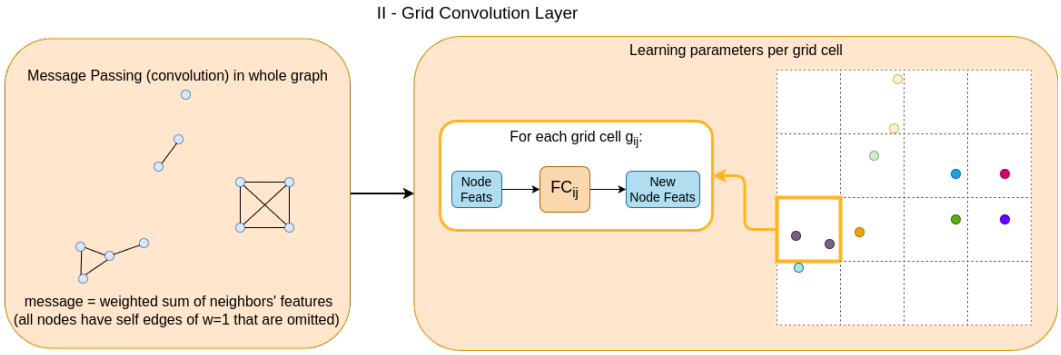

Fig. 5. Convolution operation using a grid. Message passing is performed on the graph considering all its edges. Each node feature is updated according to their neighbors'. The nodes are then divided by their spatial position in a grid. Each of the grid's cells contains a separate linear layer that learns only from that cell's nodes. The nodes are then updated with the output of this linear layer to be used either by a new convolution, or by the readout layer.

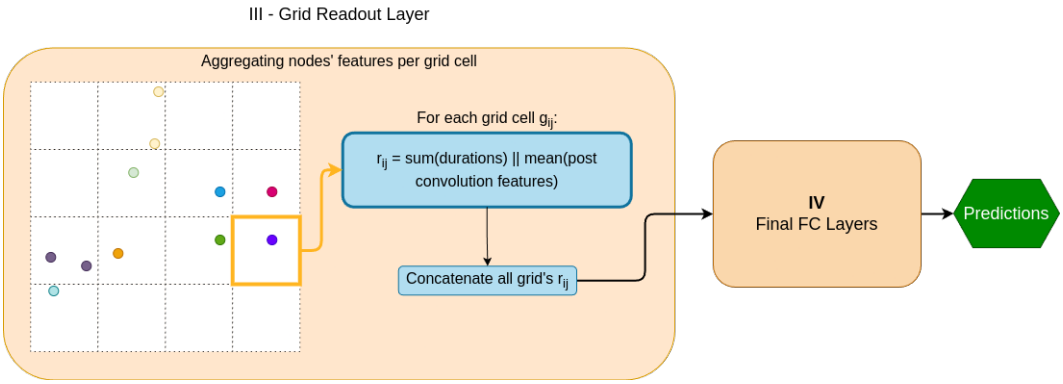

Fig. 6. Graph readout using a grid and final fully connected layers. After all convolution operations, a readout operation is performed on each grid cell. All nodes in a cell are aggregated: their gaze durations summed and their convolved features averaged. The grid is then flattened and forwarded to a series of fully connected layers for predictions.

IV-A provides the general architecture of the proposed model (figure 4) and the following subsections dive into further details of each individual component. The implementation is available at [18]. A detailed explanation on the adjustable parameters particular to this problem is presented at the end of this section.

Our model differs from other GNN implementations in the sense that our data generates, not only an euclidean graph, but a very particular kind of euclidean graph. Since all chest x-rays are oriented in the same way and the morphology of individual chests are similar compared to other image-related ML problems, we can assume that fixations in the same spatial regions are observing similar physiological features. The center is the heart, below is the diaphragm, etc.

General graph ML architectures frequently address problems in non-euclidean spaces. As a result, they are not

concerned with local spatial information. In the convolution stage, this translates to all nodes features being fed into the same NN for feature learning. If we approach REFLACX in the same way, one NN would be fed both shoulder, lung, and heart features, for instance. In parallel, during the readout phase of a typical GNN, all nodes and edges are aggregated into a single graph summary, being it a sum, mean, or max, for example.

For these reasons, our proposed model attempts to preserve local features by performing convolution layer activation and readout using a square grid that divides the graph nodes into smaller, more similar regions (figures 5 and 6). These grid dimensions are user-defined, but are the same for both operations.

A. Model Architecture

The main data-flow of the proposed model (figure 4) consists in assembling the graph using the REFLACX fixations and the localized image features obtained by cropping the x-ray's feature map from densenet121 [17]. This graph is then fed through two convolution layers. The convolved graph is then summarized by performing a readout operation using a grid, and the resulting features are fed through a regular MLP for predictions. As of now, this problem is being treated as a regression, the predictions being the probabilities of the presence of each possible abnormality listed in figure 1.

At the current stage of this research, the whole image is not being examined by the model. The model's only access to the image is through the fixation nodes' cropped densenet121 features. This is being done to verify whether the eye gaze information is enough to predict dataset labels.

B. Grid Convolution

Convolution in GNNs is a two-stage process: 1) assign a node's features to be an aggregation of its neighbors' –and itself– correspondent features and 2) run each resulting new feature through a NN that learns node-level information.

As for stage (1), our proposed model does not differ from traditional GNNs. The first step in our convolution is the concatenation of positions, time duration, and 1024 densenet121 features into a single new feature vector. For each node, this new feature vector is set to be an average of its neighbors' corresponding features, weighted by their respective edges. It is important to recall that each node is adjacent to itself.

Stage (2), however, differs from traditional GNNs (figure 5). Instead of running every convolved feature through the same NN, our module divides the nodes into a grid. Each grid cell contain one NN of its own, so it learns features only from nodes in close proximity. This is due to an intuition that features considered important for the heart may not be so significant to the lungs, for example. We believe that this approach, should it prove itself successful, could be adopted in other problems that deal with euclidean graphs with fixed orientation.

The resulting features, after being fed through the NN, are the input of any subsequent convolution layers.

It is important to distinguish our approach from the one adopted by GazeGNN [12], referenced in section II-B, since their domain and keywords are similar. The main difference of our model is that the nodes are the fixations themselves, enhanced with local features. GazeGNN divides the image in a grid and uses each patch, enhanced by the gaze data, as a node in their graph. Our approach does not use the whole image as an input, precisely because the hypothesis we are trying to prove is that it is possible to make inference about an x-ray using only the regions observed by the radiologist.

C. Grid Readout

Similar to the second stage of the convolution, our readout operation differs from traditional GNNs, as illustrated in Figure 6. While usual readout operations aggregate all nodes – and possibly edges– into a single representation for the whole graph, we perform this aggregation by grid cells. For each cell, the process takes the sum of the cell's nodes time and the average of the features that went through the convolution NNs. This generates a matrix of aggregations, which is then flattened and fed to a regular MLP for predictions.

D. Degrees of Freedom: Adjustable Parameters

The problem at hand, and the implementation chosen, have particularities when compared to traditional neural networks. This section presents the parameters that can be changed and, therefore, require experimentation. Determining the sequence of layers and their dimensions are common issues in all ML problems, so they will not be explored in this article.

As stated in III-B, the area of the image actually observed by a fixation is given by a normal distribution centered around its position with standard deviation of one degree of viewing angle. In the current stage of the research, we are cropping the image with one standard deviation.

Both the convolution and readout operations (IV-B and IV-C) employ a 4x4 grid to divide the fixation graphs. The higher this dimension, more FC layers will be created to deal with smaller, more particular, image regions. This, in theory, generates a more detailed representation. However, smaller grid cells mean less nodes per cell, and even a larger number of empty cells. Smaller grid dimensions, on the other hand, mean larger cells and the loss of more localized features. So, finding the right number for this parameter is crucial in generating a model useful for solving the problem at hand.

The fixation graph edges, as detailed in III-B, are being considered as the IOU between a pair of nodes. Not only would this IOU differ as the number of standard deviations in a fixation crop change, but there would be a myriad of other edge options to try should this approach proves insufficient. An interesting edge type to be tried is an attention layer, making the adjacency matrix learnable. This would make this implementation resemble a GAT [2].

V. Current State of the Research

At the present moment, the implementation of the proposed model is complete and the training process runs through all

data with low loss. However, given the skewness of the dataset (figure 1), and possibly other factors, the model favors predictions of value zero, meaning no abnormalities for each possible type. Dealing with this imbalance is crucial to improve the training process to a usable state.

Moreover, while separate NNs for each grid cell provide local feature learning, they also divide the dataset, making each NN learn from less graph nodes.

A. Next Steps

The adjustment of the distribution of data mentioned in the previous section is the most urgent step to perform. It poses a challenge, since eliminating negative samples to a more evenly balanced subset would yield far too few data-points. This could be reasonable if more of our learnable parameters could benefit from transfer learning. However, apart from the feature extraction, that is not the case.

Experimenting with different values for the model parameters mentioned in IV-D is also a necessary step. Multiple models with different parameters will have to be benchmarked against each other, to empirically determine the best configuration.

Parallelly, an approach that preserves local features using positional embeddings is being studied. Opposite to the learnable embeddings in [12], ours would need to be static, as gaze graphs have no fixed structure. Not only that, positional embeddings for such graphs need to be two-dimensional, as one of the approaches in [8]. This is being studied as an alternative to the grid implementation, specifically during graph convolution.

Lastly, this implementation is the first experiment in this research and is not, by all means, the ultimate. Other approaches are being formulated simultaneously, each bringing new bibliographic references and insights into the problem. One continuous aspect of our work is searching for datasets of images which contain gaze data and even proposing methods of assembling such datasets.

VI. Conclusion

This paper detailed the current state of a piece of research that proposes a GCN architecture to diagnose REFLACX CXRs using gaze data with localized image features.

Our GCN model performs graph convolution and readout operations while preserving particularities of different image regions. This is achieved by having distinct learnable parameters representing each of these regions.

If proven successful, this approach would demonstrate two hypotheses: 1) using a specialist gaze data with localized image features is be enough for inference, without the need of a neural network that learns from the whole image; 2) this GCN architecture is also efficient for other euclidean graph domains.

Acknowledgment

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001

References

[1] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61–80, 2009.

[2] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=rJXMpikCZ

[3] H. Lu, L. Wang, X. Ma, J. Cheng, and M. Zhou, "A survey of graph neural networks and their industrial applications," *Available at SSRN 4822242*, 2024.

[4] S. Moradizeyveh, M. Tabassum, S. Liu, R. A. Newport, A. Beheshti, and A. D. Ieva, "When eye-tracking meets machine learning: A systematic review on applications in medical image analysis," 2024. [Online]. Available: https://arxiv.org/abs/2403.07834

[5] R. B. Lanfredi, M. Zhang, W. F. Auffermann, J. Chan, P.-A. T. Duong, V. Srikumar, T. Drew, J. D. Schroeder, and T. Tasdizen, "REFLACX, a dataset of reports and eye-tracking data for localization of abnormalities in chest x-rays," *Scientific Data*, vol. 9, no. 1, jun 2022. [Online]. Available: https://doi.org/10.1038%2Fs41597-022-01441-z

[6] L. Wang, X. Liu, X. Ma, J. Wu, J. Cheng, and M. Zhou, "A progressive quadric graph convolutional network for 3d human mesh recovery," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 1, pp. 104–117, 2022.

[7] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2017. [Online]. Available: https://arxiv.org/abs/1609.02907

[8] 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. [Online]. Available: https://arxiv.org/abs/2010.11929

[9] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pvt v2: Improved baselines with pyramid vision transformer," *Computational Visual Media*, vol. 8, no. 3, p. 415–424, Mar. 2022. [Online]. Available: http://dx.doi.org/10.1007/s41095-022-0274-8

[10] S. Amir, Y. Gandelsman, S. Bagon, and T. Dekel, "Deep vit features as dense visual descriptors," 2022. [Online]. Available: https://arxiv.org/abs/2112.05814

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. [Online]. Available: https://arxiv.org/abs/1706.03762

[12] B. Wang, H. Pan, A. Aboah, Z. Zhang, E. Keles, D. Torigian, B. Turkbey, E. Krupinski, J. Udupa, and U. Bagci, "Gazegnn: A gaze-guided graph neural network for chest x-ray classification," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2024, pp. 2194–2203.

[13] M. Akerman, S. Choudhary, J. Liebmann, G. Cioffi, R. Chen, and K. Thakoor, "Extracting decision-making features from the unstructured eye movements of clinicians on glaucoma oct reports and developing ai models to classify expertise," *Frontiers in Medicine*, vol. 10, 09 2023.

[14] Y. Kong, S. Wang, J. Cai, Z. Zhao, Z. Shen, Y. Li, M. Fei, and Q. Wang, "Gaze-detr: Using expert gaze to reduce false positives in vulvovaginal candidiasis screening," 2024. [Online]. Available: https://arxiv.org/abs/2405.09463

[15] A. E. W. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng, "Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs," 2019. [Online]. Available: https://arxiv.org/abs/1901.07042

[16] A. Karargyris, S. Kashyap, I. Lourentzou, J. Wu, M. Tong, A. Sharma, S. Abedin, D. Beymer, V. Mukherjee, E. Krupinski *et al.*, "Eye gaze data for chest x-rays," *PhysioNet https://doi. org/10.13026/QFDZ-ZR67*, 2020.

[17] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," *CoRR*, vol. abs/1608.06993, 2016. [Online]. Available: http://arxiv.org/abs/1608.06993

[18] A. Nascimento Lutfi and J. do Espirito Santo Batista Neto, "Lutfi_REFLACX_graph_classification." [Online]. Available: https://github.com/anlutfi/reflacx_graph_classification