THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY | LIBRARY

HKUST SPD - INSTITUTIONAL REPOSITORY

Title	Semi-supervised multimodality learning with Graph Convolutional Neural Networks for Disease Diagnosis
Authors	Huang, Yongxiang; Chung, Albert Chi Shing
Source	Proceedings - International Conference on Image Processing, ICIP, v. 2020-October, October 2020, article number 9191172, p. 2451-2455
Version	Accepted Version
DOI	10.1109/ICIP40778.2020.9191172
Publisher	IEEE
Copyright	© 2020 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creatin

This version is available at HKUST SPD - Institutional Repository (https://repository.hkust.edu.hk/ir)

If it is the author's pre-published version, changes introduced as a result of publishing processes such as copy-editing and formatting may not be reflected in this document. For a definitive version of this work, please refer to the published version.

SEMI-SUPERVISED MULTIMODALITY LEARNING WITH GRAPH CONVOLUTIONAL NEURAL NETWORKS FOR DISEASE DIAGNOSIS

Yongxiang Huang and Albert C. S. Chung

Lo Kwee-Seong Medical Image Analysis Laboratory, Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong

ABSTRACT

There is a trend that digitalized clinical data increases dramatically every year. Part of data is multi-modal with imaging and non-imaging data such as phenotypic and genetic information. Though the success of CNNs has empowered a wide range of applications in learning from the imaging data, incorporating both the imaging and non-imaging data complementarily to improve the diagnostic quality is still challenging. To tackle this challenge, we propose a novel graph-convolutional model which is based on the proposed concept of edge adapter for learning an adaptive population graph from a multi-modal database. The edge adapter can be jointly optimized with the proposed graph convolutional neural network for semi-supervised node classification. Experimental results on two challenging multimodal medical databases demonstrate the potential of our method in learning from multi-modal data for disease diagnosis.

***Index Terms*—** Multimodality learning, Graph convolution, Computer-aided diagnosis

1. INTRODUCTION

With the fast increasing volume of digitalized clinical data, there is a growing need in processing and learning from multi-modal data for computer-aided medical analysis[1, 2]. Non-imaging data, such as genetic information and phenotypic data (e.g., age, gender), which are complementary to the imaging data such as CT and MRI scans, are usually collected to derive more comprehensive diagnostic assessments. Besides, non-imaging data can potentially reveal the correlation between subjects and explain the representation variance in the imaging features. Although off-the-shelf CNNs [3, 4] are powerful in representing visual features for both natural and biomedical images [5, 6], learning from both the imaging data and non-imaging data complementarily in a unified model to achieve better task performance is not straightforward. The method proposed in [7] applied a joint fully connected layer on the concatenation of the extracted imaging features and non-image features of a subject, followed by a feed-forward neural network for the final prediction. However, this approach fails to model the interaction and correlation between subjects, leading to a weak generalization when the multi-modal data of the subjects are heterogeneous due to the acquisition variance. Simply concatenating the multi-modal features may bring no benefits in improving the performance of using a single modality (which is confirmed experimentally). Graph models provide another perspective to model this problem. Graph convolutional neural networks (GCNs) [8, 9] have shown great potential in learning from non-structural data and allowing semi-supervised learning with less labels [10].

This paper presents novel graph-convolutional modeling by introducing a new trainable module, called edge adapter, to encode the non-imaging data into the connectivity of a population. In this setting, a population graph represents multiple subjects in a database, where each node represents the main features of a subject and each edge is defined to capture the association between a pair of subjects. The proposed edge adapter allows to learn the pairwise associations between subjects based on the non-imaging data such as phenotypic information (e.g., age, gender and site) during the training of the followed GCN model. While the goal is to predict the disease state of the unlabelled subjects under the supervision of the labeled ones in a population graph, the learned graph can be easily applied on clustering analysis by thresholding, which is new. We mathematically show that the edge adapter in conjunction with spectral graph convolutions [8] is differentiable for gradient descent optimizations. Thus, the overall model can be trained end-to-end from different modalities to improve diagnostic performance. Experiments on two challenging multimodal databases confirmed the superiority of the proposed method.

2. METHODOLOGY

In this section, we further present the proposed model, called Edge Adaptable GCN (EA-GCN), for incorporating the multimodal data in a database for disease prediction. The overview of the pipeline is shown in Fig. 1. The EA-GCN model accepts the imaging and non-imaging data of N subjects and represent them as a population graph (partially labeled) via the edge adapter (EA), followed by our GCN architecture for semi-supervised node classification [10], yielding a fully labeled diagnostic graph.

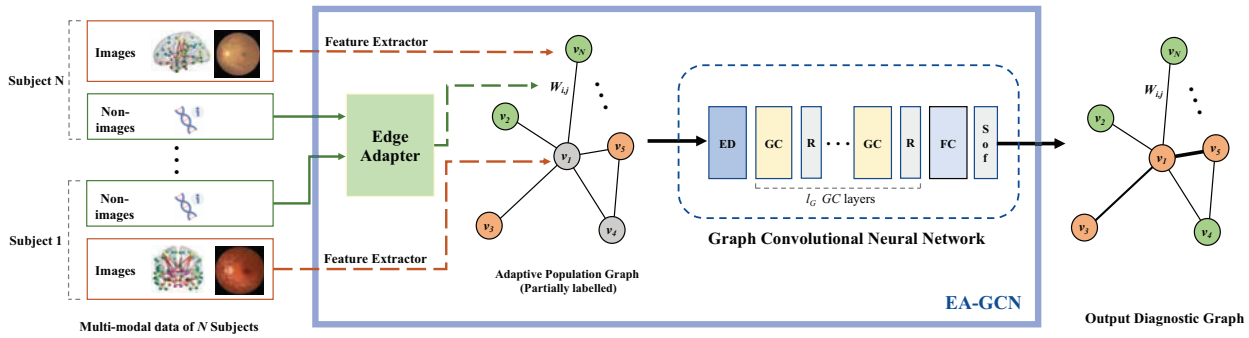

Fig. 1. Overview of the proposed EA-GCN model. ED: edge dropout, GC: graph convolution, R: ReLU, FC: fully connected layer, Sof: softmax activation. Colors in the graphs: green and orange - labelled disease states, grey - unlabelled.

2.1. Learning a Graph Representation for Multimodal Data with EA

Given the observation of N subjects composed of imaging and non-imaging data, where only a subset of subjects are labeled (e.g., healthy or diseased), we aim to construct a population graph $G = (V, E, W)$ with $|V| = N$ vertices and W being the edge weights, such that each vertex $\mathbf{v} \in V$ represents the diagnostic features of a subject and each edge captures the association between two subjects concerning their non-image measurements (e.g., age and gender) and disease state. We define $\mathbf{v}_i \in \mathbf{R}^C$ as a C-dimensional feature vector extracted from the imaging data of subject i, under the observation that the imaging data (e.g., microscopic image, functional Magnetic Resonance Images) usually provide the most important evidence for diagnosis. The modeling of the edge weights is critical for representing the population graph properly and substantially influences the performance of a graph-based learning model (e.g., GNNs or GCNs). Previous methods [11, 12] used hand-crafted affinity metrics for computing the edge weights to construct a static similarity graph, which require to fine-tune different thresholds for different modalities. The threshold tuning can be overwhelming when the number of modalities increases and the constructed static graph can be inappropriate for the task (seen in our experiments). We proposed to define the edge weight between the i-th and j-th vertex w_{ij} as a learnable function of their non-imaging measurements which provide complementary information, via the proposed Edge Adapter module. This results in a non-static population graph with adaptable edge weights. Hence, the EA-GCN model is able to learn the graph representation for multimodal data without using hand-crafted similarity metrics.

Edge Adapter The edge adapter is a crucial module to model the associations between subjects and builds the connectivity in a population graph accordingly. As depicted in Fig. 2, the edge adapter consists of a twin network (before L1), which accepts distinct inputs $\mathbf{x}_i$ and $\mathbf{x}_j$, and a metric

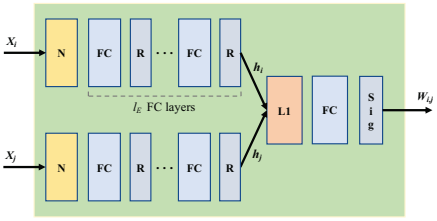

Fig. 2. The neural network architecture of Edge Adapter. N: normalization, L1: L_1 distance layer. Sig: sigmoid activation.

network to score the association between them in the latent space [13]. The non-imaging inputs $\mathbf{x}_i$ and $\mathbf{x}_j$ are first normalized by rescaling to $[0, 1]$ interval and subtracting mean element-wise, to avoid the vanishing gradient problem in back-propagation, which is important in this problem as data from different modalities has varying statistical properties. The twin network parallelly encodes the normalized inputs into two feature vectors $\mathbf{h}_i$ and $\mathbf{h}_j$ by two multi-layer perceptrons (MLP) with sharing weights using l_E hidden layers ($l_E = 1$ in experiments). Formally, the metric network scores the association between vertex i and j as

$$W_{i,j} = \sigma(\sum_d \alpha_d |\mathbf{h}_i^{(d)} - \mathbf{h}_j^{(d)}|)$$

, where σ is the sigmoid activation function and α_d is the d-th dimension trainable parameter defined in the fully connected layer after the L_1-distance layer. This computes the weighted L_1 distance of the two non-imaging feature vectors combined with sigmoid activation to map onto the $[0, 1]$ interval. The trainable parameters in the edge adapter are initialized using K. He initialization [14].

Notice that the population graph is constructed on both the labeled and unlabeled subjects, which allows to perform semi-supervised learning with graph convolutional neural networks [10]. The population graph acts as a regularizer in training to force the GCN model to aggregate both the labeled

and unlabelled nodes for prediction to minimize the cross-entropy loss on the labeled set. Moreover, during testing, the learned associations between an unlabelled subject and its neighboring nodes should provide additional references for disease prediction via graph convolutions, especially when the imaging data of the centered subject is imperfect or is obtained from a diverse distribution.

2.2. Differentiable EA-GCN on Adaptive Graphs

In this subsection, we discuss our GCN model on the adaptive population graph. We employ Chebyshev graph convolution (ChebGConv) [8] in EA-GCN for its weighted graphs. This operation is a spectral approach which exploits the fact that spatial graph convolutions can be computed in the Fourier domain as multiplications using the tool of graph Fourier transform (GFT) [15].

Background A spectral convolution of a graph signal $x \in \mathbf{R}^N$ with a filter g_θ is defined as $g_\theta \star x = U g_\theta U^T x$, where $U^T x$ is the graph Fourier transformation of x, U is the matrix of eigenvectors of the normalized Laplacian $L = I_N - D^{-1/2} W D^{1/2}$ with D being the diagonal degree matrix. As performing the above spectral convolution is computationally expensive ($\mathcal{O}(N^2)$), [8] proposed to approximate g_θ with a truncated expansion of Cheybyshev Polynomials $T_k(x)$ in K orders. The Chebyshev Polynomials are defined recursively as $T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x)$, with $T_0(x) = 1$ and $T_1(x) = x$. Accordingly, in ChebGConv, the convolution of a graph signal x with a filter $g_{\theta'}$ parameterized by $\theta' \in \mathbf{R}^K$ is given as $g_{\theta'} \star x \approx \sum_{k=0}^{K} T_k(\tilde{L}) \theta'_k x$, where $\tilde{L} = \frac{2}{\lambda_{max}} L - I_N$ is the rescaled Laplacian and λ_{max} is the largest eigenvalue of L.

Let us consider a convolution layer $l+1$ in the network and the corresponding input graph $G^l = (V^l, E^l, W^l)$ with $|V^l| = N$ nodes represented by feature vectors $\mathbf{H}^l \in \mathbf{R}^{N \times C^l}$ where C^l is the dimensionality of each node feature vector in layer l. Based on the above discussion, the convolution operation can be derived as

$$\mathbf{H}^{l+1} = \sum_{k=0}^{K} T_k(\tilde{L}) \mathbf{H}^l \mathbf{\Theta}^l_k \tag{1}$$

, where $\mathbf{\Theta}^l_k \in \mathbf{R}^{C^l \times C^{l+1}}$ are the weights in the k-th order Chebyshev Polynomial filter. While $\mathbf{\Theta}^l_k$ works as a node feature transformer, the polynomials $T_k(\tilde{L})$ acts as a k-localized aggregator, i.e. it combines the neighboring nodes that are k-step away from the central node.

Regularized GCN The architecture of our graph convolutional neural network is shown in Fig. 1. We propose to apply an edge dropout layer on the input graph, which randomly zeroes out a fraction of edges, acts as a data augmenter and increases the sparsity of the graph to reduce overfitting. We keep the architecture relatively shallow to avoid the over-smoothing problem in deepening GCNs [16]. The rest of the architecture consists of l_G K-order Chehbyshev graph convolutional layers equipped with ReLU non-linearity and ends with a fully connected layer with softmax activation ($K = 3$, $l_G = 2$ for ABIDE and $l_G = 1$ for ODIR in experiments). Cross-entropy loss computed on the labeled nodes is used to train the GCN and Edge Adapter model concurrently.

After training with SGD, the output diagnostic graph (Fig. 1) are fully labeled with each node representing the predicted disease state (i.e., class) of a subject and the edge weights capturing the pairwise associations between subjects.

Differentiability Let us consider the overall model, where spectral graph convolutions are performed on an adaptive graph of which the edge weights depend on the trainable module EA. It requires new formulations to prove it is optimizable with SGD. To optimize the parameters α in EA, we need to guarantee the final loss $\mathcal{L}$ is differentiable w.r.t. α. By chain rule, we can derive $\frac{\partial \mathcal{L}}{\partial \alpha} = \frac{\partial \mathcal{L}}{\partial \mathbf{H}^l} \frac{\partial \mathbf{H}^l}{\partial W} \frac{\partial W}{\partial \alpha}$. Both $\frac{\partial \mathcal{L}}{\partial \mathbf{H}^l}$ and $\frac{\partial W}{\partial \alpha}$ are differentiable independently as they corresponds to the back-propagated gradients in the GCN model and EA model alone respectively. The key is the derivative of the feature vectors of a node w.r.t. the input edge weights $\frac{\partial \mathbf{H}^l}{\partial W}$. For a $K = 1$ order ChebGConv, we can derive $\frac{\partial \mathbf{H}^l}{\partial W}\big|_{K=1} = \frac{\partial (I_N - D^{-1/2} W D^{1/2})}{\partial W} = \frac{\partial (D^{-1/2} W D^{1/2})}{\partial W}$ based on Eq. 1. Since the polynomial term $T_k(\tilde{L})$ for higher order ChebGConv is defined recursively, after expanding $T_k(\tilde{L})$ we can prove that $\frac{\partial \mathbf{H}^l}{\partial W}$ is still differentiable for $K > 1$ by induction and is not always zero. Thus, the EA-GCN is differentiable and can be optimized end-to-end.

3. EXPERIMENTAL RESULTS

We evaluate the proposed method on two challenging medical databases, i.e., the Autism Brain Imaging Data Exchange (ABIDE) database and the Ocular Disease Intelligent Recognition (ODIR) database.

3.1. Autism Disease Diagnosis on the ABIDE Database

Dataset and Experimental Settings The ABIDE database [1] collects data from 20 different acquisition sites and shares functional magnetic resonance imaging (fMRI) data of 1112 subjects with corresponding phenotypic data (e.g., age, gender and acquisition site) for identifying Autism Spectrum Disorder (ASD) from normal. For a fair comparison with the ABIDE state of the art [17, 11], we choose the same 871 subjects composing of 403 normal and 468 ASD individuals and perform the same preprocessing and feature extraction steps. Hereby a $C = 2000$ dimensional feature vector is derived from the fMRI data to represent the brain functional connectivity of a subject, as a node in the graph where the phenotypic data are the input non-imaging data. 10-fold stratified cross-

Table 1. Comparison with the baselines and recent state-of-the-art (SoTA) methods on the ABIDE database. INI: both imaging (I) and non-imaging data (NI) are used. $\times$ means only imaging data is used. JFC: joint fully connected layer to concatenate features of I and NI. θ is a threshold tuned for computing a static similarity graph in [11]. P: Parameters (K).

Methods	INI	ACC(%)	AUC(%)	P.
Ridge Classifier	$\times$	65.30	70.5	**2**
DNN	$\times$	71.99	74.16	550
DNN-JFC [7]	✓	72.10	73.48	635
Abraham et al.[17]	$\times$	66.80	-	-
Parisot et al.[11] $\theta = 2$	✓	75.50	81.05	96
Parisot et al.[11] $\theta = 3$	✓	71.87	82.02	96
Kazi et al. [12]	✓	75.66	79.10	290
EA-GCN	✓	**80.17**	**83.70**	97

validation is employed for evaluation as in [11], from which we report the mean accuracy and AUC.

Quantitative Results In Table. 1, we can see that the proposed approach outperforms the competing methods substantially. Simply concatenating the features of multi-modal data into a deep neural network (DNN-JFC) only brings marginal improvement (0.1%) compared to using one modality (DNN). Comparatively, graph model-based methods ([11, 12] and ours) show more promising results in leveraging the multi-modal data on ABIDE. The recent SoTA methods [11] and [12] require to finetune the affinity thresholds to construct a static similarity population graph. We can see that the performance of [11] varies a lot under different thresholds ($\theta = 2$ and $\theta = 3$). While Kazi et al. [12] using InceptionGCN can reach 75.66% accuracy after finetuning, our EA-GCN model delivered significantly more promising diagnostic performance with 80.17% accuracy and 0.837 AUC with less parameters. Ablation results in Table. 1 highlight the importance of using the proposed Edge Adapter to learn an adaptive graph from multi-modal data. Eliminating EA from our model and replacing the adaptive graph with a random graph (i.e. edges are connected uniformly at random between subjects) or the hand-crafted similarity graph [11, 12] instead highly deteriorates the original performance. Meanwhile, on the same hand-crafted similarity graph, our GCN architecture yields improved performance (77.26% over 75.50%), which demonstrates the effectiveness of our regularized GCN.

Table 2. Ablation study for the proposed EA-GCN approach on the ABIDE database. Rnd: random, Sim: similarity.

Methods	Accuracy(%)	AUC(%)
w/o EA, RndGraph-GCN	75.43	78.60
w/o EA, SimGraph-GCN	77.26	80.30
w/ EA, AdaptiveGraph-GCN	**80.17**	**83.70**

3.2. Ocular Disease Diagnosis on the ODIR Dataset

The ODIR dataset [18] contains fundus photographs of left and right eyes and non-imaging data including age, gender and diagnostic words collected from 5000 patients in different medical centers. We select a set of 1500 annotated subjects with fair imaging quality in the experiments. Each patient is labeled for 7 types of ocular diseases including diabetes, glaucoma, etc. We compared our method with two SoTA CNNs and the graph-based approach [11], using 5-fold cross-validation. To construct the graph, we employed a CNN pre-trained on ImageNet without classifier to extract a C dimensional feature vector ($C = 3072$ for InceptionV4 [19] and $C = 2048$ for EfficientNet-B0 [4]) from the images of a patient. Diagnostic words are not used to avoid label leaking.

Table 3 compares the proposed method with recent state-of-the-art approaches. We can clearly see that the EA-GCN is able to boost the classification performance for EfficientNet [4] from 0.8431 AUC to 0.8726 and for InceptionV4 from 0.84 to 0.8639. On average, the proposed semi-supervised multimodality learning method can improve the performance of a trained CNN model by 2.67% on ODIR, by properly learning to incorporate the complementary non-imaging data encoded in the graph. It is interesting to note that the static graph-based method [11], where the required thresholds are already finetuned [11], downgrades the performance of EfficientNet, which further implies the merits of constructing a learnable population graph compared to a hand-crafted one.

Table 3. Comparison results on ODIR reporting mean AUC from 5-fold CV. (I) or (E): InceptionV4 or EfficientNet as the adopted feature extractor. D: Diabetic, G: Glaucoma, M: Myopia, All: All 8 classes including normal.

Methods	D	G	M	All
InceptionV4 [3]	64.26	69.89	96.65	84.00
EfficientNet [4]	66.90	71.91	96.99	84.31
EA-GCN (I)	69.41	65.32	96.90	86.39
EA-GCN (E)	**70.78**	**73.24**	**97.49**	**87.26**
Parisot et al. [11] (E)	56.75	68.17	64.48	77.63

4. CONCLUSIONS AND DISCUSSIONS

In this paper, we have proposed a novel graph-convolutional framework to tackle the challenges in learning from multi-modal data for disease diagnosis, which leverages the proposed edge adapter for population graph construction and the designed GCN architecture for weighted adaptive graphs. Experimental results show the improved performance in two different domains, i.e., Austim disease prediction and ocular disease diagnosis, where we can clearly see the superiority of the proposed concept of edge adapter in representing a population graph for multimodality learning. This method provides a new potential to unlock a better usage of imaging and non-imaging data for computer-aided diagnosis in clinics.

5. REFERENCES

[1] Adriana Di Martino, Chao-Gan Yan, Qingyang Li, Erin Denio, Francisco X Castellanos, Kaat Alaerts, Jeffrey S Anderson, Michal Assaf, Susan Y Bookheimer, Mirella Dapretto, et al., "The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism," *Molecular psychiatry*, vol. 19, no. 6, pp. 659–667, 2014.

[2] Paul M Thompson, Jason L Stein, Sarah E Medland, Derrek P Hibar, Alejandro Arias Vasquez, Miguel E Renteria, Roberto Toro, Neda Jahanshad, Gunter Schumann, Barbara Franke, et al., "The enigma consortium: large-scale collaborative analyses of neuroimaging and genetic data," *Brain imaging and behavior*, vol. 8, no. 2, pp. 153–182, 2014.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[4] Mingxing Tan and Quoc V Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *arXiv preprint arXiv:1905.11946*, 2019.

[5] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[6] Yongxiang Huang and Albert CS Chung, "Evidence localization for pathology images using weakly supervised learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 613–621.

[7] Tao Xu, Han Zhang, Xiaolei Huang, Shaoting Zhang, and Dimitris N Metaxas, "Multimodal deep learning for cervical dysplasia diagnosis," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2016, pp. 115–123.

[8] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Advances in neural information processing systems*, 2016, pp. 3844–3852.

[9] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu, "A comprehensive survey on graph neural networks," *arXiv preprint arXiv:1901.00596*, 2019.

[10] Thomas N Kipf and Max Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[11] Sarah Parisot, Sofia Ira Ktena, Enzo Ferrante, Matthew Lee, Ricardo Guerrerro Moreno, Ben Glocker, and Daniel Rueckert, "Spectral graph convolutions for population-based disease prediction," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2017, pp. 177–185.

[12] Anees Kazi, Shayan Shekarforoush, S Arvind Krishna, Hendrik Burwinkel, Gerome Vivar, Karsten Kortüm, Seyed-Ahmad Ahmadi, Shadi Albarqouni, and Nassir Navab, "Inceptiongcn: receptive field aware graph convolutional network for disease prediction," in *International Conference on Information Processing in Medical Imaging*. Springer, 2019, pp. 73–85.

[13] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*. Lille, 2015, vol. 2.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[15] David I Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE signal processing magazine*, vol. 30, no. 3, pp. 83–98, 2013.

[16] Qimai Li, Zhichao Han, and Xiao-Ming Wu, "Deeper insights into graph convolutional networks for semi-supervised learning," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[17] Alexandre Abraham, Michael P Milham, Adriana Di Martino, R Cameron Craddock, Dimitris Samaras, Bertrand Thirion, and Gael Varoquaux, "Deriving reproducible biomarkers from multi-site resting-state data: An autism-based example," *NeuroImage*, vol. 147, pp. 736–745, 2017.

[18] "Ocular disease intelligent recognition," https://odir2019.grand-challenge.org/dataset/, 2018.

[19] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-first AAAI conference on artificial intelligence*, 2017.