Accurate Medical Image Classification using Contrastive Graph Cross-View Learning with Multimodal Fusion

Jun-En Ding[1], Chien-Chin Hsu[2], Feng Liu[1]

[1]School of Systems and Enterprises, Stevens Institute of Technology, Hoboken, NJ 07030, USA

[2]Dept. Nuclear Medicine, Kaohsiung Chang Gung Memorial Hospital, Kaohsiung Chang, Taiwan

May 6, 2024

Abstract

Traditional medical image computing research has increasingly utilized deep learning for disease prediction, primarily focusing on medical images, neglecting the data's underlying manifold structure. This work proposes a multimodal approach encompassing both image and non-image features, leveraging contrastive cross-view graph fusion for disease classification. We introduce a novel multimodal co-attention module, integrating embeddings from separate graph views derived from low-dimensional representations of images and clinical features. This enables the extraction of more robust and structured features for improved multi-view data analysis. Additionally, a simplified contrastive loss-based fusion method is devised to enhance cross-view fusion learning. Our graph-view multimodal approach achieves outperformance in two multimodal medical datasets. It also demonstrates superior predictive capabilities on non-image data compared to solely machine learning-based methods.

1. Introduction

In recent years, deep learning models have shown great successes in biomedical applications, including computer-aided detection/diagnosis, image segmentation, image generation, disease staging and prediction.

Graph representation learning that can incorporate node features as well as utilize the connectivity/similarity information among all the nodes achieved good success for machine learning tasks on the graph structured data, such as graph convolutional neural network (GCN) [6]. GCN's capability of leveraging both information of nodes (representing features of entities) and edges (representing connections or relationships between nodes), allowing for feature aggregation in the network level. Advanced versions of GCN include graph attention networks (GATs) [11], graph transformer networks [13] etc. In this paper, we propose a multimodal framework integrating both image features and clinical features by building a contrastive graph cross-view learning approach where the graph represents the similarity of individuals in the embedded space for detecting the PD.

2. Supervised Graph Contrastive Learning

Supervised contrastive learning, as demonstrated by its significant successes in image applications such as visual representations [1, 5], primarily aims to enhance the similarity among positive pairs while simultaneously augmenting the dissimilarity between negative pairs [8]. Contrastive learning with two graph views was proven effective in fMRI-based neuroimaging classification in a previous medical study that sought to improve the diagnosis of neurological disorders in autism and dementia [9]. Investigations have been conducted on to the design of contrastive losses, such as the InfoNCE loss [10], which maximizes the consistency of positive pairs and uses negative sampling to increase the number of negative pairs from different batches for k classes.

3. METHODOLOGY

3.1. Features Extraction and Graph Construction

We can denote a dataset of N patients with the i-th patient's medical image denoted as X_i^m and non-image features denoted as X_i^f, where $X_i^m \in \mathbb{R}^{\mu \times \nu}$ and $X_i^f \in \mathbb{R}^F$ with F features, and the label matrix $Y \in \mathbb{R}^{N \times C}$ with C classes. The multimodal dataset can be described as $\{X_i^m, X_i^f, Y_i\}_{i=1}^N$, where Y_i is the i-th row of Y. In the first stage, we use a convolutional neural network (CNN) based autoencoder $h(\cdot)$ for image feature extraction and flatten the output image feature matrix Q^m in the final layer. Then, we constructed two adjacency matrices A^m, $A^f \in \mathbb{R}^{N \times N}$ using the K nearest neighbors (KNN) algorithm based on the features obtained from the CNN model encoder and the non-image clinical features (such as patient age, biomarkers, symptoms, etc.) for each images encoder and clinical

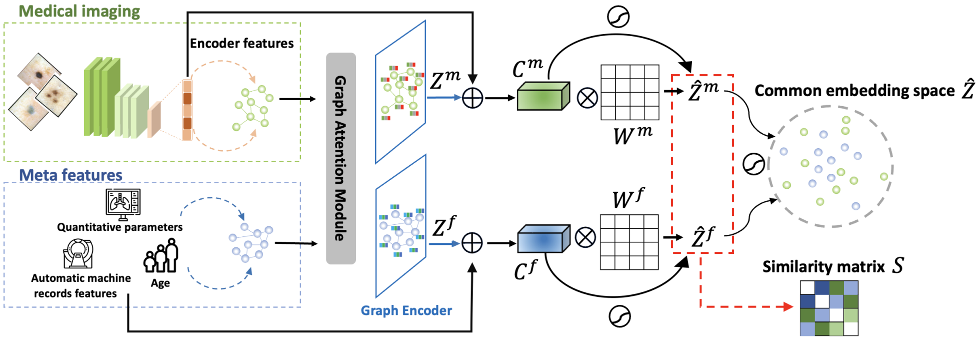

Figure 1. The workflow of multimodal contrastive cross-view graph learning framework.

features subject. Our proposed framework considers two graphs, $\mathcal{G}^m(X^m, A^m)$ and $\mathcal{G}^f(X^f, A^f)$, as different domain inputs. We also considered a self-loop adjacency matrix $\hat{A} = A + I$ between the patients. In a K-neighborhood, two data points i and j are connected by an edge $e_{(i,j)}$ if i is amongst the K nearest neighbors of j, or vice versa.

3.2. Graph Encoder and Cross-View Fusion

In this stage, we constructed two GATs to learn the graph structures of $\mathcal{G}^m(A^m, X^m)$ and $\mathcal{G}^f(A^f, X^f)$ as depicted in (Fig. 1). We utilize encoded image features X^m and clinical features X^f as node attributes for the GAT inputs. Moreover, we introduced a GAT architecture that incorporates dual perspectives, enabling the generation of embeddings for neighboring nodes. The most common expression for attention coefficients as applied to our two cross-views is as follows:

$$\alpha_{ij} = \frac{exp\left(LeakyReLU(\vec{a}_{ij}^T\left[W\vec{h}_i \parallel W\vec{h}_j\right])\right)}{\sum_{k=1}^N exp\left(LeakyReLU(\vec{a}_{ij}^T\left[W\vec{h}_i \parallel W\vec{h}_k\right])\right)} \tag{1}$$

Finally, we can obtain each feature $\vec{h}'$ for two cross-views feature representations as shown in Equation (2):

$$\vec{h}' = \sigma\left(\frac{1}{K}\sum_{k=1}^N \sum_{j\in\mathcal{N}_i} \alpha_{ij}^k W^k \vec{h}_j\right), \tag{2}$$

where α_{ij}^k and W^k are the attention mechanism and linear transformation weight matrix, and $\mathcal{N}_i$ denotes the set of neighborhood nodes of i.

The extracted nodes representation from the GAT output is denoted as $Z^m = f(\mathcal{G}^m)$, and the non-image feature embeddings are represented by $Z^f = f(\mathcal{G}^f)$, where Z^m and Z^f are the embeddings in a low-dimensional space $\mathbb{R}^{F'}$. Afterward, we concatenated the encoder matrix Q^m with Z^m to form C^m, and the clinical features X^f with Z^f to create C^f as shown below:

$$C^m = [Q^m \parallel Z^m] \tag{3}$$

$$C^f = \left[X^f \parallel Z^f\right] \tag{4}$$

where the $C^m \in \mathbb{R}^{N\times(F+F')}$ and $C^f \in \mathbb{R}^{N\times(F+F')}$ represent the two concatenated matrices from cross-views. The improved fusion embedding $\hat{Z}^m$ and $\hat{Z}^f$, can be obtained by $\sigma(C^m W^m)$ and $\sigma(C^f W^f)$ respectively, where W^m and W^f are trainable weight matrices, and $\sigma(\cdot)$ is non-linear activation function.

3.3. Contrastive Cross-View Loss

In order to learn the common embedding $\hat{Z}$, we fuse the two cross-views of node embeddings between $\hat{Z}^m$ and $\hat{Z}^f$ as

$$\hat{Z} = \hat{Z}^m + \hat{Z}^f \tag{5}$$

To better integrate the feature spaces of image and non-image data in the same embedded space, we constructed a similarity matrix $S \in \mathbb{R}^{N\times N}$ for each pair of similar patients using the final embedding $\hat{Z}$ learned from the model. We can define the similarity between the i-th and j-th patients as follows:

$$S_{ij} = \hat{Z}_i \cdot (\hat{Z}_j)^T, \forall i, j \in [1, N] \tag{6}$$

In order to enhance the effectiveness of fusing two types of view embeddings in contrastive learning, we have designed positive and negative losses to capture the differences in distance between positive and negative pairs in terms of the similarity and dissimilarity of our samples. The definitions of positive pair $D_{pos} = S \odot (\hat{A}^m \odot \hat{A}^f)$, while negative pair $D_{neg} = (\mathbb{I} - S) \odot \left[(\mathbb{I} - \hat{A}^m) \odot (\mathbb{I} - \hat{A}^f)\right]$, where $\mathbb{I}$ denotes the matrix with all elements being 1 with the related dimension, and the two adjacency matrices with self-looped is denoted as $\hat{A}^m$ and $\hat{A}^f$ [7]. Then, we can present the loss function of positive and negative pairs as shown below:

$$\begin{cases} \mathcal{L}_{pos} = -\|D_{pos} \cdot Y\|_2^2 \\ \mathcal{L}_{neg} = -\|\max\{D_{neg} - \delta\mathbb{I}, 0\}(\mathbb{I} - Y)\|_2^2 \end{cases} \tag{7}$$

Table 1. Comparison of the accuracy of multi-class classification performance between baseline and proposed models on a melanoma dataset.

Model	Backbone	Image	Non-image	BWV	DaG	PIG	PN	RS	STR	VS	DIAG
Baseline											
Logistic Model	-	✔	✗	0.83±0.00	0.58±0.20	0.51±0.15	0.68±0.10	0.77±0.00	0.73±0.10	0.83±0.03	0.75±0.19
Xgboost	-	✔	✗	0.82±0.00	0.49±0.14	0.68±0.12	0.67±0.09	0.76±0.00	0.74±0.10	0.84±0.03	0.74±0.19
ResNet18	CNN	✔	✗	0.84±0.01	0.48±0.14	0.59±0.11	0.65±0.09	0.74±0.01	0.64±0.09	0.82±0.03	0.75±0.20
2-layer	CNN	✔	✗	0.80±0.01	0.43±0.12	0.60±0.12	0.56±0.08	0.71±0.01	0.69±0.1	0.80±0.03	0.72±0.20
7-point† [4]	CNN	✔	✔	0.85	0.60	0.63	0.69	0.77	0.74	0.82	0.73
AMFAM† [12]	GAN	✔	✔	0.88	**0.69**	0.71	0.71	**0.81**	0.75	0.83	0.75
Proposed Model											
GCN+GCN	2-layer CNN	✔	✔	0.81±0.01	0.60±0.06	0.73±0.08	0.74±0.20	0.73±0.01	**0.80±0.04**	0.81±0.03	0.74±0.20
GAT+GAT		✔	✔	0.79±0.02	0.46±0.13	0.60±0.12	0.69±0.20	0.71±0.03	0.69±0.09	0.80±0.03	0.72±0.20
GCN+GCN	ResNet18	✔	✔	**0.87±0.01**	0.64±0.02	**0.76±0.03**	0.75±0.05	0.78±0.02	0.76±0.02	0.86±0.02	**0.76±0.20**
GAT+GAT		✔	✔	0.86±0.01	0.67±0.03	0.74±0.03	**0.75±0.03**	0.78±0.02	0.77±0.02	**0.86±0.01**	0.75±0.20

†Denotes the average of accuracy.

where the $\delta > 0$ is the controllable margin and Y is the label matrix. By using Eq. 7, we can ultimately obtain the combined losses, incorporating both positive and negative loss, written as: $\mathcal{L}_{contrastive} = \mathcal{L}_{pos} + \mathcal{L}_{neg}$. By minimizing $\mathcal{L}_{contrastive}$, the similarity intra-class and the dissimilarity inter-class can be maximized.

3.4. Optimization Objective Function

To optimize the loss function and predict final disease probability, we considered embedding both $\hat{Z}^m$ and $\hat{Z}^f$ in the supervised classification loss using the softmax function. The cross-entropy loss function can be written as:

$$\mathcal{L}_m = -\sum_{i=1}^{N} y_i^T \ln(\text{softmax}(\hat{\mathrm{y}}_\mathrm{i}^\mathrm{m})) \tag{8}$$

$$\mathcal{L}_f = -\sum_{j=1}^{N} y_j^T \ln(\text{softmax}(\hat{\mathrm{y}}_\mathrm{j}^\mathrm{f})) \tag{9}$$

During the optimization process, we also designed the overall loss function to combine cross-entropy and contrastive loss from the two cross views. To effectively improve the cross-graph view module, we also took into account the mean square error loss between the similarity matrix S and the diagonal matrix $D_{ii} = \sum_i A_{ii}$ when computing the clustering of the view structure of the two modules.

$$\mathcal{L}_{diag} = \frac{1}{N} \sum_{i,j}^{N} (S_{ij} - D_{ii})^2 \tag{10}$$

We use the β coefficient to control the optimization weight of the overall loss defined as follows:

$$\mathcal{L} = (1-\beta)(\mathcal{L}_m + \mathcal{L}_f) + \beta\mathcal{L}_{contrastive} + \mathcal{L}_{diag} \tag{11}$$

where β can be set between 0 and 1. The contribution level of different losses is controlled through the coefficient of β.

4. EXPERIMENTS

4.1. Dataset

- **Melanoma dataset [4]** The melanoma open dataset consists of a 7-point multimodal dataset comprising dermoscopic images and patient clinical data, with 413 training samples and 395 testing samples. The classification task involves seven labels: 1) Pigment Network (PN), 2) Blue Whitish Veil (BWV), 3) Vascular Structures (VS), 4) Pigmentation (PIG), 5) Streaks (STR), 6) Dots and Globules (DaG), and 7) Regression Structures (RS), along with five categories of DIAGNOSIS: 1) Basal Cell Carcinoma (BCC), 2) blue nevus (NEV), 3) melanoma (MEL), 4) miscellaneous (MISC), and 5) seborrheic keratosis (SK). The dermoscopic images have dimensions of 512×768, and the clinical data includes patient gender and lesion location.
- **Parkinson's Disease (PD)** Our data was collected at Kaohsiung Chang Gung Memorial Hospital in Taiwan from January 2017 to Jun 2019 with 416 patients [2]. The data was annotated by four expert physicians to provide a label either as normal or abnormal PD. Tc99m TRODAT single-photon emission computed tomography (SPECT) images were acquired using a hybrid SPECT/CT system (Symbia T, Siemens Medical Solution). SPECT images were obtained with 30s per step, acquiring 120 projections over a circular 360-degree rotation using low-energy, high-resolution parallel-hole collimators. To facilitate model building, we resized the original SPECT images (800 × 1132) to a fixed size of 128 × 128. We then used the remaining 412 preprocessed images and their corresponding quantitative DaTQUANT data for model training (n=300) and testing (n=112).

In this study, we conducted a comparative analysis using popular machine learning algorithms (logistic regression and XGBoost) as baseline methods [3] for melanoma

Table 2. Evaluation of AUC and ACC for the proposed model and machine learning methods in PD classification using image and non-image data.

Model	Backbone	Image	Non-image	ACC	AUC
Baseline					
Logistic	-	✗	✔	0.80±0.00	0.80±0.00
Xgboost	-	✗	✔	0.78±0.00	0.78±0.00
ResNet18	CNN	✔	✗	0.85±0.01	0.85±0.02
2-layer CNN	CNN	✔	✗	0.76±0.14	0.77±0.13
Proposed Model					
GCN+GCN	2-layer CNN	✔	✔	0.88±0.03	0.88±0.02
GAT+GAT		✔	✔	0.83±0.01	0.84±0.01
GCN+GCN	ResNet18	✔	✔	0.89± 0.01	0.89± 0.01
GAT+GAT		✔	✔	**0.90±0.02**	**0.90±0.01**

and PD image classification. In our research, we began by extracting image features using a CNN model and non-imaging variables, followed by constructing two cross-views of the graph representation. We subsequently concatenated these different modalities to improve the predictive capability after model fusion. We first perform multi-class classification results on the modality of skin lesions as described in Table 1. While fundamental CNN-based models achieve reasonable performance, their accuracy suffers when they neglect non-image data. Yang et al. [12] proposed an adversarial multimodal fusion with an attention mechanism to address this limitation. Their approach effectively concentrates on fusing features, leading to improved model performance. However, our experiment showcases results from utilizing two methods (GCN and GAT) for learning graph structures and generating embeddings to enhance both classification accuracy and stability across all classes.

For the classification tasks on PD, we also employed a two-layered CNN model with a ResNet18 backbone, specifically utilizing ResNet18 for the classification of SPECT images. As shown in Table 2, relying solely on the CNN model for prediction did not lead to superior performance. However, our graph-based fusion method offers an advantage for two-domain fusion. Our proposed multimodal revealed that, when a cross-view approach was employed, the GAT method achieved a macro average accuracy rate of 90% along with the area under the receiver operating characteristic curve (AUC) in normal and abnormal classes.

5. RESULTS AND DISCUSSION

In summary, we successfully integrate two-domain features from image and non-image data to enhance prediction accuracy in the medical image classification task. Our research findings indicate that models based solely on CNNs can have certain limitations in interpreting image features. However, those limitations can be overcome to some extent through the integration of non-image data and the application of contrastive loss learning, which significantly enhanced the overall performance and predictive capacity of our model.

References

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709, 2020. 1

[2] Jun-En Ding, Chi-Hsiang Chu, Mong-Na Lo Huang, and Chien-Ching Hsu. Dopamine transporter spect image classification for neurodegenerative parkinsonism via diffusion maps and machine learning classifiers. *Annual Conference on Medical Image Understanding and Analysis*, 2021. 3

[3] Shih-Yen Hsu, Hsin-Chieh Lin, Tai-Been Chen, Wei-Chang Du, Yun-Hsuan Hsu, Yichen Wu, Yi-Chen Wu, Po-Wei Tu, Yung-Hui Huang, and Huei-Yung Chen. Feasible classified models for parkinson disease from 99mtc-trodat-1 spect imaging. *Sensors*, 2019. 3

[4] Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE journal of biomedical and health informatics*, 23(2): 538–546, 2018. 3

[5] Prannay Khosla, Piotr Teterwak, Chen Wang, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, A. Maschinot, Aaron Maschinot, Ce Liu, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. 2020. 1

[6] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017. 1

[7] Yue Liu, Xihong Yang, Sihang Zhou, and Xinwang Liu. Simple contrastive graph clustering. *arXiv.org*, 2022. 2

[8] Yue Liu, Xihong Yang, Sihang Zhou, and Xinwang Liu. Simple contrastive graph clustering. *arXiv.org*, 2022. 1

[9] Liang Peng, Nan Wang, Jie Xu, Xiao lan Zhu, and Xiaoxiao Li. Gate: Graph cca for temporal self-supervised learning for label-efficient fmri analysis. *IEEE Transactions on Medical Imaging*, 2022. 1

[10] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv: Learning*, 2018. 1

[11] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. *arXiv: Machine Learning*, 2017. 1

[12] Yan Wang, Yangqin Feng, Lei Zhang, Joey Tianyi Zhou, Yong Liu, Rick Siow Mong Goh, and Liangli Zhen. Adversarial multimodal fusion with attention mechanism for skin lesion classification using clinical and dermoscopic images. *Medical Image Analysis*, 81:102535, 2022. 3, 4

[13] Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, and Hyunwoo J Kim. Graph transformer networks. *Advances in neural information processing systems*, 32, 2019. 1