Multimedia Tools and Applications (2020) 79:25023–25050
https://doi.org/10.1007/s11042-020-09227-4

A deep multimodal generative and fusion framework for class-imbalanced multimodal data

Qing Li[1] · **Guanyuan Yu**[1] · **Jun Wang**[1] · **Yuehao Liu**[1]

Received: 27 May 2019 / Revised: 12 June 2020 / Accepted: 15 June 2020 /
Published online: 28 June 2020

Abstract
The purpose of multimodal classification is to integrate features from diverse information sources to make decisions. The interactions between different modalities are crucial to this task. However, common strategies in previous studies have been to either concatenate features from various sources into a single compound vector or input them separately into several different classifiers that are then assembled into a single robust classifier to generate the final prediction. Both of these approaches weaken or even ignore the interactions among different feature modalities. In addition, in the case of class-imbalanced data, multimodal classification becomes troublesome. In this study, we propose a deep multimodal generative and fusion framework for multimodal classification with class-imbalanced data. This framework consists of two modules: a deep multimodal generative adversarial network (DMGAN) and a deep multimodal hybrid fusion network (DMHFN). The DMGAN is used to handle the class imbalance problem. The DMHFN identifies fine-grained interactions and integrates different information sources for multimodal classification. Experiments on a faculty homepage dataset show the superiority of our framework compared to several start-of-the-art methods.

Keywords Multimodal classification · Class-imbalanced data · Deep multimodal generative adversarial network · Deep multimodal hybrid fusion network

1 Introduction

Multimodal data consist of several feature modalities, where each modality is represented by a group of similar data sharing the same attributes. The aim of multimodal classification is to process and integrate information from multiple modalities to make decisions. In the era of big data, many applications of interest involve multimodal classification problems, including audio-visual speech recognition (AVSR) [40], affective computing [39], human emotion recognition [32], medical image analysis [22], user profiling [13], and stock

✉ Guanyuan Yu
kennis.yu@smail.swufe.edu.cn

[1] Fintech Innovation Center and School of Economic Information Engineering, Southwestern University of Finance and Economics, Chendu, China

movement prediction [29]. However, two challenging problems usually arise when fusing information from multiple interactive modalities for multimodal classification.

The first major challenge is multimodal representation. The heterogeneity in the statistical properties of multimodal data makes it more difficult to learn a joint representation using information from multiple sources [3, 17, 24]. A good example is the joint processing of images (which are real-valued and dense) and texts (which are discrete and sparse), which typically have different dimensions and structures [52]. In previous studies, a common strategy has been to separately map each modality into a common latent space [39, 49], for example, using a Gaussian probability distribution [52, 56]. However, in practice, samples usually appear to come from a distribution that is skewed, very peaked, or very flat or shows some other discrepancy relative to a Gaussian distribution [47]. Information provided by data from such distribution would be distorted if estimated using a Gaussian probability distribution.

The second major challenge is multimodal alignment, which involves identifying the direct relationships between components from different modalities [12, 54]. These interactions are essential to consider when developing a model for decision-making. For example, considering the interactions between stock data, news articles, and discussion boards can boost the performance of a model for predicting stock movements [27]. However, due to the heterogeneity in the statistical properties of multimodal data, it is difficult to find direct relationships and correspondences between the various components of multimodal features [40]. Traditional approaches, such as feature vector concatenation, will disconnect the links between such components [41]. As illustrated in Fig. 1, the layout feature set of a webpage consists of four types of tags, that is, $\langle h1 \rangle$, $\langle p \rangle$, $\langle div \rangle$, and $\langle footer \rangle$. Each tag contains different textual information. Intuitively, the words embedded in an $\langle h1 \rangle$ tag are more important than those in a $\langle footer \rangle$ tag. Such interrelationships are ignored, however, if the tag and text features are concatenated into a compound vector.

In addition, in many applications, such as abnormal brain tumor recognition [44] and credit scoring classification [31], the class imbalance problem is encountered. Class imbalance often arises when the samples of the majority class outnumber those of the minority class [53]. Once a dataset becomes imbalanced, model performance declines because the

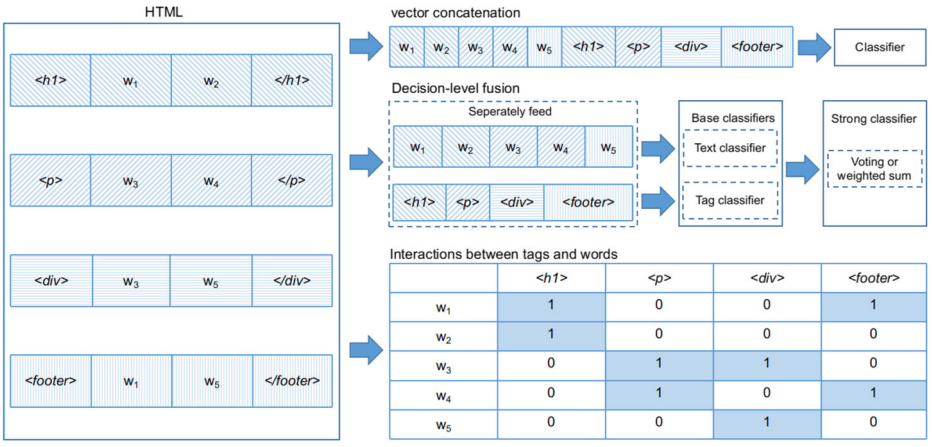


Fig. 1 Modeling multimodal data

characteristics of the minority class cannot be learned effectively [19]. Multimodal classification becomes more troublesome in the presence of a class imbalance because an auxiliary mechanism needs to be designed to rebalance the interdependent multimodal data to maintain multimodal classification performance. However, to date, there have been few works concerning multimodal classification with class-imbalanced datasets.

To address the three challenges mentioned above, we propose a deep multimodal generative and fusion framework. This framework is composed of a deep multimodal generative adversarial network (DMGAN) and a deep multimodal hybrid fusion network (DMHFN) and presents three unique contributions to the literature:

- The DMGAN synthesizes samples to address the problem of imbalanced multimodal data.
- A gate mechanism in the DMHFN is used to find the direct relationships among fine-grained elements across different feature modalities.
- An interaction mechanism in the DMHFN is proposed to solve the fusion problem for heterogeneous data by means of a co-occurrence matrix (a probability matrix). Through this interaction mechanism, every data modality is mapped to its own finite discrete distribution space.

The remainder of this article is structured as follows. Section 2 briefly describes the previous work related to our research. Section 3 presents the design details of the proposed framework. Section 4 examines the effectiveness of our approach. Finally, Section 5 concludes this article and suggests future work.

2 Related work

2.1 Class imbalance

The class imbalance problem refers to the situation in which the members of the majority class greatly outnumber the members of the minority class [53]. Most classical learning algorithms are best suited for balanced datasets [11]. Once the classes of a dataset become imbalanced, model performance declines because the characteristics of the minority class cannot be learned effectively [19]. A classifier trained on imbalanced data will tend to overrepresent the majority class compared to the minority class. The minority class may be so small that members of this class may be easily ignored, treated as noise, or misidentified as the majority class [20, 23, 26, 38]. Louzada et al. reported that class-imbalanced data lead to severe deterioration in model performance [31].

The traditional techniques for dealing with the class imbalance include random oversampling (OS), random undersampling (US), the synthetic minority oversampling technique (SMOTE) [5], and adaptive synthetic sampling (ADASYN) [18].

In the US approach, observations from the majority class are randomly dropped until the remaining number of majority-class samples matches the number of samples in the minority class. This approach results in a loss of valuable information, which inevitably reduces the classification performance [8].

In contrast, the OS approach involves randomly duplicating observations from the minority class until the total number of minority-class samples matches the number of samples in the majority class. However, this approach tends to cause overfitting due to the simple replication of samples from the minority class [10, 19, 50]. In essence, this approach cannot provide additional valuable information for use in classification.

To overcome the weaknesses of the OS approach, Chawla et al. proposed SMOTE [5]. In this approach, the number of samples in the minority class is increased by creating virtual samples, each of which is a linear combination of two real samples from the minority class that are located near each other, to rebalance the data. However, these virtual samples are merely linear combinations of local information instead of being drawn from the overall minority-class distribution [10]. In addition, SMOTE may produce noisy samples when the boundary between the majority and minority classes is not sufficiently clear [19].

In ADASYN [18], different numbers of samples are generated for each minority class in accordance with their data distributions. The synthetic samples are generated through the linear combination of a data point with a randomly chosen minority data sample from among its K nearest neighbors. However, this technique seems to have the same problems as SMOTE.

Consequently, generating synthetic samples based on the real minority-class distribution is a critical challenge. Some researchers have taken the further step of utilizing generative adversarial networks (GANs). Such a framework involves training a generator network and a discriminator network, which compete with each other in a zero-sum game [16]. A well-trained generator can estimate the latent distribution of the real data and then produce fake samples based on the global distribution instead of using only local information. For example, Shin et al. utilized a GAN to rebalance medical data by synthesizing abnormal brain tumor MRI images because of the limited availability of data from patients with cancer [44].

Previous generative approaches for multimodal data include multimodal stochastic recurrent neural networks (MS-RNNs) [45], hierarchical long short-term memory with adaptive attention (hLSTMat) [15], attention-based long short-term memory with semantic consistency (aLSTMs) [14], and dual conditional GANs (Dual cGANs) [46]. However, these approaches are aimed at multimodal translation or domain transformation and cannot be used to generate completely new samples, i.e., when all multimodal features are missing. In this study, we propose a DMGAN for synthesizing samples, even when all multimodal features are missing, to address the problem of imbalanced multimodal data. Specifically, the DMGAN generates fake samples by simultaneously using features from different feature modalities in accordance with the global feature distributions and their relationships to improve the performance of multimodal classification with imbalanced data.

2.2 Multimodal classification

Multimodal data consist of several data modalities, where each modality is represented by a group of related data sharing the same attributes. The aim of multimodal classification is to process and integrate information from multiple modalities to make decisions based on multimodal fusion. Multimodal fusion is defined as a category of techniques that integrate information from different sources [61]. These techniques usually face two challenging problems:

- **Multimodal Representation:** The heterogeneity in the statistical properties of multimodal data makes it more challenging to learn a joint representation using information from multiple sources [3, 17, 24]. For example, images (which are real-valued and dense) and texts (which are discrete and sparse) typically have different dimensions and structures [52].
- **Multimodal Alignment:** Multimodal alignment involves identifying direct relationships between fine-grained components from different modalities [12, 54]. These

interactions are essential to consider when developing a model for decision-making. For example, considering the interactions between stock data, news articles, and discussion boards can boost the performance of a model for predicting stock movements [27]. However, due to the heterogeneity in the statistical properties of multimodal data, it is difficult to find direct relationships and correspondences between the various components of multimodal features [40].

To address these two problems, many traditional and commonly applied techniques have been developed in previous studies, including vector concatenation and the decision-level fusion approach [61]. These techniques can solve the heterogeneity problem, but they weaken or even ignore the interactions between multiple modalities [3].

Vector concatenation involves concatenating features from various information sources into a single compound vector. For example, [49] used two deep Boltzmann machines (DBMs) to separately map image features and text features to higher-level features and finally concatenated them into a joint representation. [36] adopted similar procedures to obtain multimodal joint representations. Such approaches learn global relationships between low-level features from multiple modalities.

Decision-level fusion approaches input information from different sources separately into different classifiers, which are then assembled into a single strong classifier for final prediction using a weighted sum or voting scheme. For instance, [35] reduced the unimodal model error and improved the variety of multimodal models to enhance the effect of a voting scheme. [40] aggregated the results of an audio hidden Markov model (HMM) and a visual HMM using a weighted sum to improve the performance of speech recognition [40]. Intuitively, these approaches fuse information sources at the decision level and ignore the fine-grained interactions among features.

Some researchers have taken a further step by estimating the interactions across multiple modalities through a joint probability distribution. For example, in the joint multimodal variational autoencoder (JMVAE) [52] and multimodal variational autoencoder (MVAE) [56] approaches, every modality is mapped to a probability space to obtain a multimodal representation, and the joint probability distributions are used to identify the relationships among different modalities. However, in these two techniques, the prior distribution of data is assumed to be a Gaussian distribution, which may be inconsistent with the real situation. In practice, samples usually seem to come from a distribution that is skewed, peaked, flat, or shows some other discrepancy relative to a Gaussian distribution [47]. Therefore, to estimate the natural distribution of every modality, we adopt a flexible finite discrete distribution space [47].

In this study, a DMHFN is proposed to integrate information from diverse sources by means of feature fusion at multiple levels for multimodal classification. In particular, to solve the problems of multimodal representation and alignment, a gate mechanism is used to find direct relationships among fine-grained elements from different feature modalities. In addition, an interaction mechanism is used to map each modality to its own finite discrete distribution space, and a co-occurrence matrix is then used to integrate these distribution spaces (Table 1).

3 System design

In this study, we propose a deep multimodal generative and fusion framework for multimodal classification with class-imbalanced data. Figure 2 presents an overview of this

Table 1 Representative research on class imbalance and multimodal classification

Category	Model	Weakness
Class imbalance	US	Loses valuable information.
	OS	Overrepresents samples through replication and cannot provide additional valuable information.
	SMOTE [5] ADASYN [18]	Are limited to local information. Introduce noise when the boundary between the majority and minority classes is not sufficiently clear.
	MS-RNNs [45] hLSTMat [15] aLSTMs [14] Dual cGANs [46]	Are aimed at multimodal translation or domain transformation and cannot be used when all multimodal features are missing.
	GANs [44]	Are aimed at unimodal data.
Multimodal classification	Dual DBMs [49] MDBM [36]	Ignore the interactions between multiple modalities.
	MinCq [35] HMMs [40]	Ignore the interactions between multiple modalities.
	JMVAE [52] MVAE [56]	Are limited to spherical Gaussian distributions.

framework. Multimodal features are preprocessed to form a multimodal dataset with n feature modalities. The DMGAN first rebalance the dataset by generating pseudofeatures for each modality and combining them to form fake samples. Then, the DMHFN finds fine-grained interactions among features and integrates information from diverse sources at different fusion levels for multimodal classification.

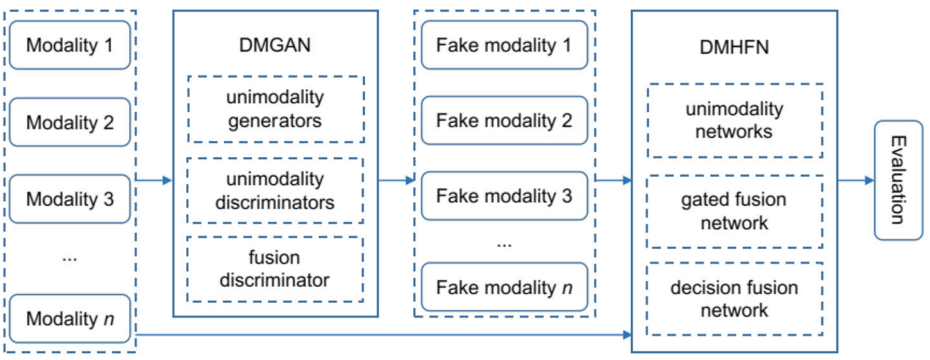


Fig. 2 The deep multimodal generative and fusion framework

3.1 Deep multimodal fusion generative adversarial network (DMGAN)

To overcome the class imbalance problem for multimodal data, we have designed a novel architecture called a DMGAN to generate artificial samples of the minority class. The DMGAN first creates fake samples for each feature modality via an iterative adversarial training process. In this way, the fused distribution of the counterfeit features can be made to approach the joint distribution of the real features while preserving the individual characteristics of each feature set and the relationships among them.

Figure 3 shows an overview of the DMGAN. It consists of n generators, where G_i denotes the i-th generator; n unimodal discriminators, where D_i denotes the i-th unimodal discriminator; and a modality-fused discriminator D_f. The n generators create fake samples $G_1(z), G_2(z), \cdots, G_n(z)$ and their combination $(G_1(z), G_2(z), \cdots, G_n(z))$ using random samples z. During training, these outputs are further fed into the corresponding discriminators along with the real sample data $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$ and $(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n)$. Then, the $n+1$ discriminators attempt to determine whether the inputs come from the real distribution or a fake one. Essentially, the n generators aim to fool the discriminators, which act as anti-fraud agents and provide feedback that is used to adjust the weights of the generators to improve their fraudulent capabilities. The objective function of the DMGAN is defined as follows:

$$L_{mgan} = \sum_{i=1}^{n} (\alpha_i \times LG_i) + \alpha_f \times LG_f. \tag{1}$$

Here, L_{mgan} represents the overall loss of the DMGAN. LG_i represents the loss for the i-th modality; these loss functions ensure the preservation of the individual characteristics of each feature modality. LG_f is the fusion loss function, which ensures that the fused distribution of the fake features approaches the joint distribution of the real features. α_i and α_f are parameters that control the importance of the corresponding loss functions. The loss for the i-th modality, LG_i, is defined as follows:

$$LG_i = \mathbb{E}_{\boldsymbol{x}_i \sim p_{x_i}}[\log D_i(\boldsymbol{x}_i)] + \mathbb{E}_{G_i(z) \sim p_{g_i}}[\log 1 - D_i(G_i(z))], \tag{2}$$

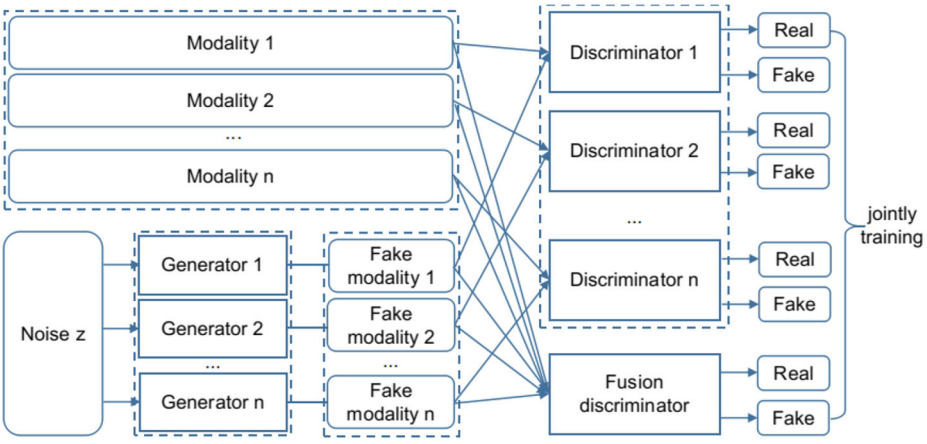


Fig. 3 Deep multimodal fusion generative adversarial network (DMGAN)

where p_{x_i} and p_{g_i} represent the real and fake distributions, respectively, for the i-th modality.

The fusion loss function is defined as follows:

$$LG_f = \mathbb{E}_{f(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n)\sim p_{x_f}}[log D_f(f(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n))] + \mathbb{E}_{f(G_1(z),G_2(z),\cdots,G_n(z))\sim p_{g_f}}[log(1-D_f(f(G_1(z),G_2(z),\cdots,G_n(z)))], \quad (3)$$

where p_{x_f} is the joint distribution of the real samples $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$; p_{g_f} is the fused distribution of the generated samples $G_1(z), G_2(z), \cdots, G_n(z)$; and $f(\cdot)$ is a fusion function embedded in the modality-fused discriminator to preserve the fine-grained interactions among different feature modalities, as illustrated in Fig. 4.

Intuitively, the interactions among the modalities are determined by the joint effects of the relevant sources. For example, the impact of an independent variable on a dependent variable can be measured in terms of the magnitudes of other associated independent variables [1]. Therefore, such interactions can be expressed in the form of a product, if the information sources are scalars, or a Kronecker product if the information sources are vectors [4, 55]. Rendle and Steffen adopted the product approach to capture the interactions among features in factorization machines [42]. The feature space that captures the interactions among different information modalities can be formally expressed as follows:

$$\begin{aligned} \boldsymbol{i}_{1\cdot 2} &= \boldsymbol{h}_1 \otimes \boldsymbol{h}_2, \\ \boldsymbol{i}_{1\cdot 3} &= \boldsymbol{h}_1 \otimes \boldsymbol{h}_3, \\ &\cdots \\ \boldsymbol{i}_{i\cdot j} &= \boldsymbol{h}_i \otimes \boldsymbol{h}_j, \\ &\cdots \\ \boldsymbol{i}_{(n-1)\cdot n} &= \boldsymbol{h}_{n-1} \otimes \boldsymbol{h}_n, \end{aligned} \quad (4)$$

where $\otimes$ denotes the Kronecker product, which is used to represent all possible interactions between elements in every pair of feature vectors (note that there are $n(n-1)/2$ possible

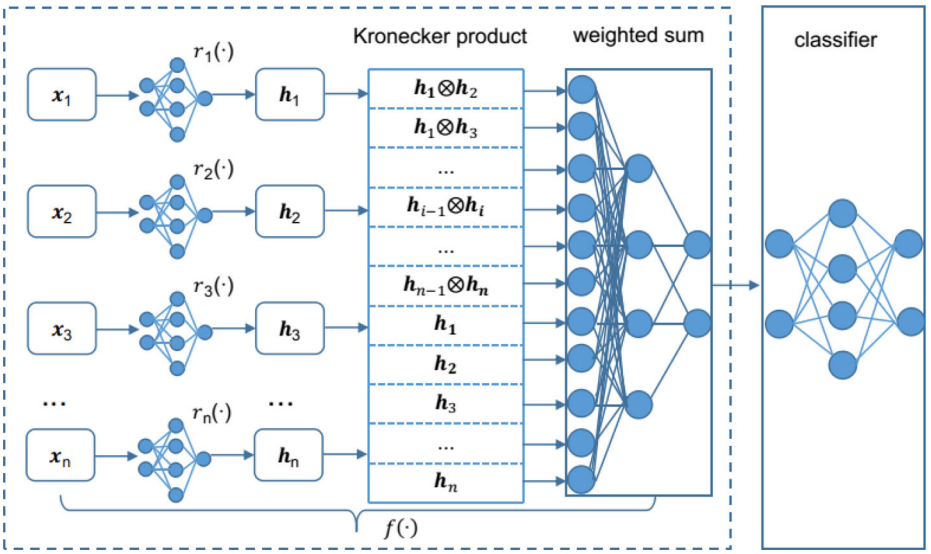


Fig. 4 Modality-fused discriminator

fusions for n modalities), and $\boldsymbol{i}_{i\cdot j}$ represents the fusion of features from the i-th and j-th modalities via the Kronecker product.

The fusion function $f(\cdot)$ is defined as follows:

$$f(\boldsymbol{x}_1, \boldsymbol{x}_2, , \cdots, \boldsymbol{x}_n) = \sum_{i=1}^{n}(\boldsymbol{h}_i \boldsymbol{w}_i) + \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(\boldsymbol{i}_{i\cdot j}\boldsymbol{w}_{i\cdot j}) + b. \tag{5}$$

In (5), $\boldsymbol{h}_i = r_i(\boldsymbol{x}_i)$. $\boldsymbol{h}_i$ is a high-level mapping of $\boldsymbol{x}_i$ obtained through a sub-network $r_i(\cdot)$. Here, $r_i(\cdot)$ and $f(\cdot)$ are both sub-networks of the modality-fused discriminator $D_f(\cdot)$, and $\boldsymbol{w}_i$, $\boldsymbol{w}_{i\cdot j}$, and b are network parameters.

In (3), the vector $f(\boldsymbol{x}_1, \boldsymbol{x}_2, , \cdots, \boldsymbol{x}_n)$ is the result of fusing the real feature sets $(\boldsymbol{x}_1, \boldsymbol{x}_2, , \cdots, \boldsymbol{x}_n)$. Similarly, the vector $f(G_1(z), G_2(z), \cdots, G_n(z))$ is the result of fusing the n fake feature sets. In the fraud and anti-fraud game between the generators and discriminators, the goal of the n generators is to confuse the discriminators (make $D_f(f(G_1(z), G_2(z), \cdots, G_n(z))$ close to 0) while making the discriminators believe that the generated features are real ones (make $D_f(f(G_1(z), G_2(z), \cdots, G_n(z))$ close to 1) for fixed states of the discriminators. In contrast, the purpose of the discriminators is to distinguish real features from generated ones. That is, the discriminators are trained to make $D_f(f(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n))$ close to 1 and make $D_f(f(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n))$ close to 0 for fixed states of the n generators. By solving the above objective function, we obtain $p_{g_f} = p_{x_f}$, which indicates that the fused distribution of the features generated by the n generators can converge to the joint distribution of the real features. The pseudocode for the proposed DMGAN is shown in Algorithm 1.

Algorithm 1 DMGAN training with mini-batch gradient descent.

inputs: $\boldsymbol{x}_i$ with positive label $\boldsymbol{y}_{pos}$
outputs: n generators $G_1, G_2, \cdots, G_n$

1: **while** k steps **do**
2: Draw m samples $\boldsymbol{x}_i$ from the training set
3: Draw m noise samples z from a real distribution
4: $G_i(z) \leftarrow$ output of G_i given z as input
5: Assign the negative label $\boldsymbol{y}_{neg}$ to the generated samples $G_i(z)$, and use them and the real samples to train the $n + 1$ discriminators
6: Repeat Step 5; assign the positive label $\boldsymbol{y}_{pos}$ to the generated samples $G_i(z)$ and use them to train the n generators
7: **end while**

3.2 Deep multimodal hybrid fusion network (DMHFN)

In this study, we propose a DMHFN for multimodal classification with interdependent feature modalities. Figure 5 presents an overview of this network, which mainly includes three types of sub-networks, namely, n unimodal networks, a feature-level fusion network, and a decision-level fusion network. As described in Table 2, the modality i network is the unimodal network designed for the i-th modality. The gated fusion network is the feature-level fusion network, which is designed to integrate information from the n modalities via gate and interaction mechanisms. The decision-level fusion network aggregates the decisions made by the n unimodal networks and the gated fusion network.

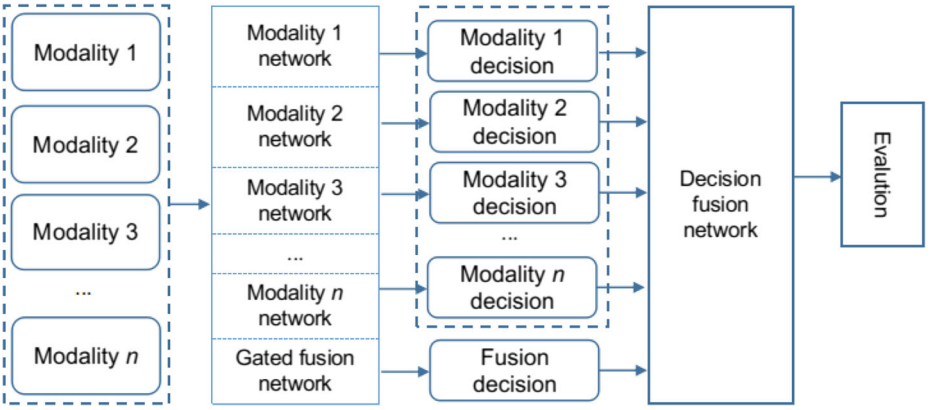


Fig. 5 Deep multimodal hybrid fusion network (DMHFN)

3.2.1 Modality *i* networks

The modality i network is designed based on the statistical properties of the i-th modality. Thus, such a network can be a feedforward neural network (FNN), a recurrent neural network (RNN), or a convolutional neural network (CNN). The loss function of this network, which is denoted by LU_i, can be designed for a specific task. For example, for a binary classification task, this loss function is defined as follows:

$$\begin{aligned} LU_i &= -\frac{1}{m}\sum_{j=1}^{m} y_{ij}\log\hat{y}_i(\boldsymbol{x}_{ij}) + (1-y_{ij})\log(1-\hat{y}_i(\boldsymbol{x}_{ij})), \\ LU &= \sum_{i=1}^{n} LU_i, \end{aligned} \tag{6}$$

where y_{ij} is the true label of sample $\boldsymbol{x}_{ij}$, $\hat{y}_i(\boldsymbol{x}_{ij})$ is the predicted probability generated by the modality i network when given sample $\boldsymbol{x}_{ij}$ as input, and m denotes the mini-batch size.

Table 2 Descriptions of the sub-networks of the DMHFN

Network	Category	Function
Modality i network	Unimodal network	Makes a decision based on the i-th modality
Gated fusion network	Feature-level fusion network	Finds the direct relationships among the fine-grained elements from the n modalities via a gate mechanism and fuses these relationships based on a co-occurrence matrix via an interaction mechanism
Decision fusion network	Decision-level fusion network	Aggregates the decisions made by the n unimodal networks and the gated fusion network through a stacking mechanism

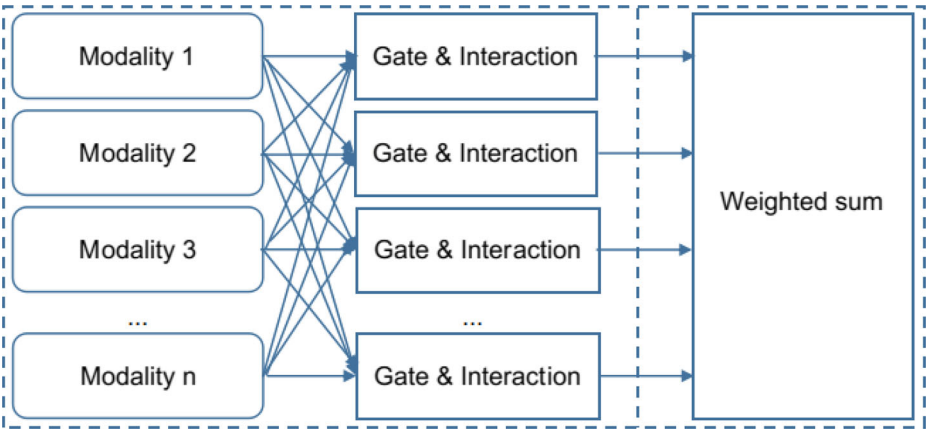


Fig. 6 Gated fusion network

3.2.2 Gated fusion network

The gated fusion network mainly includes $n(n-1)/2$ gate and interaction blocks, one for each pair that can be selected from among the n modalities, where each such block applies an attention-based gate mechanism and a co-occurrence matrix-based interaction mechanism, as illustrated in Fig. 6. The gate mechanism attempts to extract the direct relationships among the fine-grained components from every pair of modalities. The purpose of the interaction mechanism is to fuse the related information based on a co-occurrence matrix.

Gate Mechanism Autoencoders are a type of self-supervised learning model that can learn a compressed representation of input data. Intuitively, it is intractable to construct the large sparse co-occurrence matrix, as illustrated in Fig. 8, by using the original information from multiple modalities. Instead, it is reasonable to construct this matrix using a compressed vector of a fixed size learned by an autoencoder, the components of which still represent the original information [37, 48, 51]. For example, Li and Mandt [59] proposed a variational autoencoder (VAE)-based deep generative model for mapping high-dimensional sequential data to a latent representation that is split into a static part and a dynamic part. Mathieu et al. [33] disentangled the hidden factors of variation within a set of labeled observations. They separated such factors into complementary codes by combining deep convolutional autoencoders with a form of adversarial training. Inspired by these two network architectures, we use an autoencoder to compress and decompose the i-th modality before this modality is subjected to the gate mechanism, as illustrated in Fig. 7. The reconstruction loss of the autoencoder for the i-th modality is defined as follows:

$$\begin{aligned} LE_i &= \frac{1}{m}\sum_{j=1}^{m}(\boldsymbol{x}_{ij} - \hat{\boldsymbol{x}}_{ij})^2, \\ LE &= \sum_{i=1}^{n} LE_i, \end{aligned} \tag{7}$$

where this reconstruction loss is based on the mean square error (MSE). $\boldsymbol{x}_{ij}$ represents the true value, while $\hat{\boldsymbol{x}}_{ij}$ represents the prediction of the encoder. m denotes the mini-batch size.

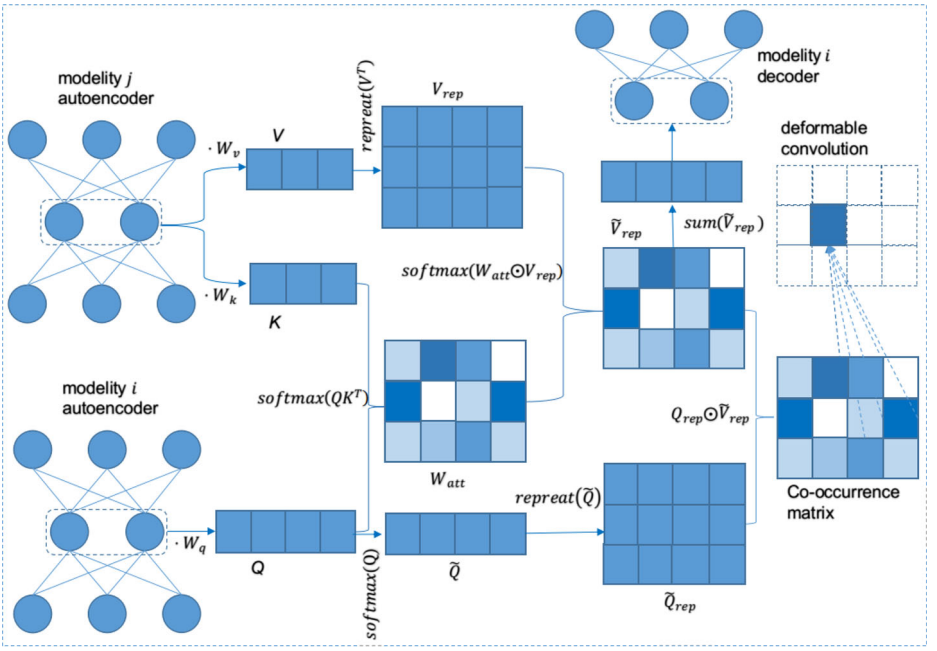


Fig. 7 Gate & interaction mechanisms

If firm fine-grained interactions exist between components from different modalities describing a common phenomenon, then these interactions are essential to consider in the corresponding decision-making model. For example, considering the interactions between stock data, news articles, and discussion boards can boost the performance of a model for predicting stock movements [27]. Concatenation-based approaches, which depend only on the globally encoded information from every modality, neglect such important fine-grained information [57]. This problem becomes severe in the case of complex modality information. To find the direct relations and correspondences between components from multiple modalities and allow this component-level information to be utilized, an attention-based gate mechanism is proposed to enable a classifier to extract the components from one modality that are most relevant to components from another modality. Attention [2] can be described as a mechanism for mapping a value to an output in accordance with weights based on the correspondence between a query and a key. An attention-based gate mechanism can assign higher weights to related components from different feature modalities.

For simplicity, we consider modalities $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ as an example. As shown in Fig. 7, in the proposed attention-based gate mechanism, the set of compressed representations from the modality $\boldsymbol{x}_i$ autoencoder is denoted by $\boldsymbol{H}_i = [\boldsymbol{h}_{i1}, \boldsymbol{h}_{i2}, \cdots, \boldsymbol{h}_{im}] \in \mathbb{R}^{d_i \times m}$, whereas the set of compressed representations from the modality $\boldsymbol{x}_j$ autoencoder is denoted by $\boldsymbol{H}_j = [\boldsymbol{h}_{j1}, \boldsymbol{h}_{j2}, \cdots, \boldsymbol{h}_{jm}] \in \mathbb{R}^{d_j \times m}$. Here, d_i and d_j represent the hidden layer sizes.

$$\begin{aligned} \boldsymbol{q} &= \boldsymbol{H}_i \boldsymbol{w}_q \in \mathbb{R}^{d_i}, \\ \boldsymbol{k} &= \boldsymbol{H}_j \boldsymbol{w}_k \in \mathbb{R}^{d_j}, \\ \boldsymbol{v} &= \boldsymbol{H}_j \boldsymbol{w}_v \in \mathbb{R}^{d_j}, \end{aligned} \tag{8}$$

where $\boldsymbol{w}_q$, $\boldsymbol{w}_k$, and $\boldsymbol{w}_v \in \mathbb{R}^m$ are learnable parameters. $\boldsymbol{q}$, $\boldsymbol{k}$, and $\boldsymbol{v}$ are referred to as the query vector, the key vector, and the value vector, respectively.

$$\boldsymbol{W}_{att} = softmax\left(\frac{\boldsymbol{q}\boldsymbol{k}^T}{\sqrt{d_2}}\right) \in \mathbb{R}^{d_i \times d_j}, \tag{9}$$

where $\boldsymbol{W}_{att}$ is a scaled attention weight matrix. The (s,t)-th value in the attention matrix $\boldsymbol{W}_{att}$ measures the relationship between the s-th value from modality $\boldsymbol{x}_i$ and the t-th value from modality $\boldsymbol{x}_j$. The value range of the elements in every row of this matrix is between 0 and 1, and the sum of all elements in the same row is equal to 1; these elements represent the weights assigned to every column of the value matrix. The vector $\boldsymbol{v}^T \in \mathbb{R}^{d_j}$ is expanded to the matrix $\boldsymbol{V}_{rep} \in \mathbb{R}^{d_i \times d_j}$. $\boldsymbol{W}_{att} \circ \boldsymbol{V}_{rep}$, which denotes the element-wise product between $\boldsymbol{W}_{att}$ and $\boldsymbol{V}_{rep}$, represents the operation of filtering $\boldsymbol{V}_{rep}$ by $\boldsymbol{W}_{att}$. In this operation, every element in the same column of $\boldsymbol{V}_{rep}$ is fine-tuned by means of the corresponding weight in $\boldsymbol{W}_{att}$. If the s-th value of $\boldsymbol{x}_i$ is strongly related to the t-th value of $\boldsymbol{x}_j$, the latter will be amplified by the corresponding weight in $\boldsymbol{W}_{att}$. Otherwise, the t-th value of $\boldsymbol{x}_j$ will be reduced.

$$\begin{aligned} \boldsymbol{V}_{rep} &= repeat(\boldsymbol{v}^T) \in \mathbb{R}^{d_i \times d_j}, \\ \widetilde{\boldsymbol{V}} &= \boldsymbol{W}_{att} \circ \boldsymbol{V}_{rep} \in \mathbb{R}^{d_i \times d_j}, \\ \boldsymbol{h}_s &= \sum_{k=1}^{d_2} \widetilde{\boldsymbol{V}}_{sk} \in \mathbb{R}^{d_i}. \end{aligned} \tag{10}$$

Here, $\boldsymbol{h}_s$ represents the weighted sum of the elements along each column of $\boldsymbol{x}_j$. Through an FNN, the latent vector $\boldsymbol{h}_s$ is transformed into the outputs $\hat{\boldsymbol{x}}_i$.

$$\begin{aligned} LT_l &= \mathbb{E}(\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i)^2, \\ LT &= \sum_{l=1}^{n(n-1)/2} LT_l. \end{aligned} \tag{11}$$

Equation (11) is the transformation loss function defined for the l-th gate and interaction block. Minimizing this loss function is equivalent to making the outputs $\hat{\boldsymbol{x}}_i$ generated from modality $\boldsymbol{x}_j$ approximate modality $\boldsymbol{x}_i$. The significant benefit of this procedure is that the weights used in the attention-based gate mechanism can be sufficiently trained based on the gradient of such a loss function. After this pre-training process, the internal alignment between the components from modality $\boldsymbol{x}_i$ and the corresponding components from modality $\boldsymbol{x}_j$ can be found.

Interaction Mechanism The heterogeneity in the statistical properties of multimodal data makes it more challenging to learn a joint representation using information from multiple sources [3, 24]. For example, images (which are real-valued and dense) and texts (which are discrete and sparse) typically have different dimensions and structures [52]. In previous studies, a typical strategy has been to separately map each modality to a common latent space [36, 39, 49], for example, using a Gaussian probability distribution [52, 56]. However, in practice, samples often appear to come from a distribution that is skewed, very peaked, or very flat or shows some other discrepancy relative to a Gaussian distribution [47].

Therefore, for estimating real-world data distributions, we abandon the Gaussian assumption and instead use the activation function $softmax(\cdot)$ in the neural network to transform the latent representation $\widetilde{\boldsymbol{V}}$ into a probability distribution $\boldsymbol{P}_v$ and the latent representation $\boldsymbol{q}$ into a probability distribution $\boldsymbol{p}_q$. Both distributions may have various shapes. The probability distribution $\boldsymbol{P}_v$ is adjusted by means of the attention-based gate mechanism mentioned above. In Eq. (12), the output $\boldsymbol{C}$ is the co-occurrence matrix, which measures the

likelihood that components from modality $\boldsymbol{x}_i$ and components from modality $\boldsymbol{x}_j$ co-occur. As shown in Fig. 8, in the case of equal weights, a higher overall probability of occurrence of one component of a single modality may also result in a higher co-occurrence probability.

$$\begin{aligned} \boldsymbol{P}_v &= softmax(\widetilde{\boldsymbol{V}}) \in \mathbb{R}^{d_i \times d_j}, \\ \boldsymbol{p}_q &= softmax(\boldsymbol{q}) \in \mathbb{R}^{d_i}, \\ \boldsymbol{P}_q &= repeat(\boldsymbol{p}_q) \in \mathbb{R}^{d_i \times d_j}, \\ \boldsymbol{C} &= \boldsymbol{P}_v \circ \boldsymbol{P}_q \in \mathbb{R}^{d_i \times d_j}. \end{aligned} \tag{12}$$

As illustrated in Fig. 8, one component of modality $\boldsymbol{x}_i$ is usually associated with other elements and their neighbors, and vice versa. Therefore, transforming such a matrix into a vector will compromise these inherent spatial features. A convolutional kernel is an effective approach for extracting spatial features. Additionally, the co-occurrence matrix is a relatively large sparse matrix; only a small fraction of its elements are non-zero, and the dense areas in such a matrix may have various geometrical shapes. However, traditional CNN modules sample the input feature map at fixed locations; such approaches have difficulty handling these geometric variations.

Therefore, we introduce a deformable convolutional kernel [7, 62] in our interaction mechanism. Such a kernel is suitable for accommodating the sparsity and geometric variability of the co-occurrence matrix. Here, an autoencoder with a deformable convolutional kernel is used to reconstruct the co-occurrence matrix. The reconstruction loss is defined as follows:

$$\begin{aligned} LD_l &= \mathbb{E}(\boldsymbol{C}_l - \hat{\boldsymbol{C}}_l)^2, \\ LD &= \sum_{l=1}^{n(n-1)/2} LD_l, \end{aligned} \tag{13}$$

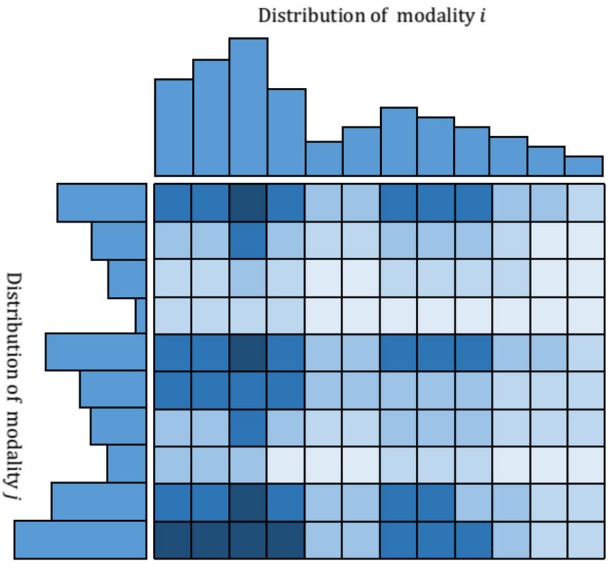


Fig. 8 Co-occurrence matrix under the condition of equal weights

where $\hat{\boldsymbol{C}}_l$ is the reconstructed result obtained from the l-th gate and interaction block. Because the dimensionality of the hidden layer is lower than the dimensionality of the visual layer, the sparse co-occurrence matrix is compressed into a dense vector.

Classification Loss The l-th set of compressed co-occurrence information is further fed into a classifier. We define the following cross-entropy-based binary classification loss for the gated fusion network:

$$\begin{aligned} LC_l &= \mathbb{E}[y \log(\hat{y}) + (1-y)\log(1-\hat{y})], \\ LC &= \sum_{l=1}^{n(n-1)/2} \beta_l LC_l, \end{aligned} \tag{14}$$

where y is the true label, $\hat{y}$ is the predicted probability, and β_l is a parameter that controls the importance of the l-th classifier.

Overall Loss We must simultaneously minimize four loss functions, namely, LE, LT, and LD, and LC:

$$LG = LE + LT + LD + LC. \tag{15}$$

The gradient of LG is backpropagated to train the initial feature compression, the gate mechanism, the interaction mechanism, the co-occurrence matrix compression, and the classifier.

3.2.3 Decision fusion network

The decision fusion network aggregates the $n+1$ decisions from the various sub-networks. The loss function of the decision fusion network is defined as follows:

$$LDF = \mathbb{E}[y \log(\hat{y}) + (1-y)\log(1-\hat{y})], \tag{16}$$

where y is the true label and $\hat{y}$ is the predicted probability.

To evaluate the total loss of the DMHFN, we define the following total loss function:

$$L_{mhfn} = \gamma_1 LU + \gamma_2 LG + \gamma_3 LDF. \tag{17}$$

In Eq. (17), L_{mhfn} is the total loss of the DMHFN, LU is the loss of the unimodal classifiers calculated in Eq. (6), LG is the loss of the gated fusion network calculated in Eq. (15), LDF is the loss of the decision fusion network calculated in Eq. (16), and $\gamma_1 \sim \gamma_3$ are parameters that control the importance of each loss function.

Algorithm 2 DMHFN training with mini-batch gradient descent.

inputs: Feature modalities $\boldsymbol{x}_i$ and labels y
outputs: Well-trained DMHFN

1: **while** k steps **do**
2: Draw m examples $\boldsymbol{x}_i$ from the training set
3: Feed $\boldsymbol{x}_i$ to the modality i network and calculate the loss according to Eq. (6)
4: Feed the n modalities into the $n(n-1)/2$ gate and interaction blocks to calculate LG
5: Aggregate the results from the $n+1$ sub-networks utilizing the decision fusion network and calculate the classification loss LDF according to Eq. (16)
6: Calculate the total loss of the DMHFN according to Eq. (17)
7: **end while**

Table 3 Four data subsets with different levels of class imbalance

Subset	Samples	Imbalance Level
FW_1	7,943 faculty homepages among a total of 15,886 webpages	1 : 1
FW_2	7,943 faculty homepages among a total of 23,829 webpages	1 : 2
FW_3	7,943 faculty homepages among a total of 31,645 webpages	1 : 3
FW_4	7,943 faculty homepages among a total of 39,715 webpages	1 : 4

4 Experimental evaluation

This section presents a series of experiments conducted to gauge the effectiveness of the proposed deep multimodal generative and fusion framework. In particular, this evaluation targets the framework's internal functions for processing imbalanced multimodal data.

The standard *accuracy* metric is applied to evaluate the model performance [34]. However, due to the class imbalance in the multimodal data, this *accuracy* metric alone is unable to provide a comprehensive evaluation of model performance. Consequently, we also adopt additional assessment metrics, such as $specificity$ ($precision$), $sensitivity$ ($recall$), and $F1$, to evaluate the model performance [19].

The experimental platform is a Linux server with 80 CPU cores (Intel(R) Xeon(R) CPU E5-2698 v4 @ 2.20GHz), 500 GB of RAM, and 4 GPUs (NVIDIA Tesla M10).

4.1 Data

The task of recognizing faculty homepages is a typical binary classification problem involving multimodal features, namely, text, image, and HTML layout features, all of which are related in a complicated fashion. In addition, recognizing faculty webpages is also a class-imbalanced problem, as the total number of samples belonging to the minority class (faculty homepages) is far smaller than the total number of samples belonging to the majority class (non-faculty webpages).

In this study, we use a faculty homepage dataset to evaluate the performance of our proposed framework. This dataset includes 39,715 samples collected by [60] from several universities in the United States. The dataset is divided into four subsets to evaluate the system performance for various levels of data imbalance, as described in Table 3. In addition, our source code can be accessed at GitHub[1].

The text, image, and layout features of a webpage are described in detail below.

- **Text features**: The text of a webpage can be represented as a word list with a dimensionality of 400. To enhance the semantic and contextual information it contains, the text can be further represented by word embeddings. In this study, we use the Google word vector model [2] to convert this word list into an embedding matrix with an embedding size of 300. Then, we apply convolutional kernels to extract the semantic information from the word embeddings, as suggested by [58].
- **Image features**: Instead of representing images as pixels, we abstract the image features of a webpage as a four-dimensional vector. The elements of this vector include the

[1] https://github.com/kennis-coder/multimodal_generative_fusion_framework.git

[2] This model comprises 3 million 300-dimensional English word vectors and is accessible at https://code.google.com/archive/p/word2vec/

number of zero-face images, the number of one-face images, the number of multiple-face images, and the total number of images. Whether an image includes one or more faces is determined using a Histograms of Oriented Gradients (HOG)-based face recognition algorithm [9].

- **Layout features**: The layout features of a webpage are represented by a tag vector with a dimensionality of 300. Each element in this vector reflects the number of HTML tags of a particular type, such as $\langle a \rangle$, $\langle p \rangle$, or $\langle span \rangle$.

4.2 Model parameters

In this study, we propose a deep multimodal generative and fusion framework for multimodal classification with class-imbalanced data. This framework is composed of a DMGAN and a DMHFN. To evaluate the performance of this framework, we compare it with several benchmark models.

4.2.1 Deep multimodal generative and fusion framework

DMGAN We conducted a series of preliminary experiments to find the optimal settings for the four discriminators and three generators. The final chosen settings are as follows. The text discriminator has a convolutional layer with 250 1D filters, each with a size of 3, and four fully-connected layers with 250 units each. The activation function is the sigmoid function. The image and layout discriminators each comprise five fully-connected layers, but with different numbers of units; the image discriminator has 100 units per layer, and the layout discriminator has 500 units per layer. The ReLU activation function is used in the discriminators. The modality-fused discriminator has three convolutional layers, each with 300 1D filters with a size of 5, and five fully-connected layers with 300 units each. The activation function is the sigmoid function. Among the generators, the text generator consists of five convolutional layers with 300 1D filters, each with a size of 3. The image and layout generators each comprise five fully-connected layers but with different numbers of units; the image generator first has 100 units per layer, and the layout generator has 500 units per layer. The ReLU activation function is used in the generators.

We implemented three tricks to alleviate the challenges of non-convergence, mode collapse, and slow training when training the DMGAN. Specifically, we added batch normalization layers only to the four discriminators [21] to accelerate and stabilize the training process, chose the Adam optimizer as the top-priority solver to accelerate the training process [25], and added random noise to both the real and fake samples [43] to alleviate mode collapse.

DMHFN To address the problem of multimodal data with interdependencies, our proposed DMHFN consists of five sub-networks, namely, a text-based network, an image-based network, a layout-based network, a gated fusion network, and a decision fusion network. The image-based network, layout-based network, and decision fusion network each have three fully-connected layers with different numbers of units; the first layer has 10 units, the second has 200 units, and the last has 4 units. The ReLU activation function is applied in these networks. The gated fusion network is composed of three gate and interaction blocks. Each block contains two feature autoencoders, a gate mechanism, and an interaction mechanism. The text autoencoder is based on a sequence-to-sequence framework with long short-term memory (LSTM). The image and layout autoencoders are each an FNN with a single hidden layer. The activation function for the gate mechanism is the softmax function. A deformable

convolutional kernel is applied to extract features from the co-occurrence matrix in the interaction mechanism. Here, we adopt the Adam optimizer to train the DMHFN. More detailed information can be found by referring to our source code.

4.2.2 Benchmark models

To gauge the overall performance of the proposed framework and its internal functions, we compare this framework with several state-of-the-art algorithms and frameworks, namely, SMOTE [5], the unimodal GAN framework, the extreme gradient boosting (XGBoost) algorithm [6], the SMOTE-XGBoost framework, the random forest (RF) algorithm, the SMOTE-RF framework, the CNN algorithm, the decision tree ensemble based on SMOTE and bagging with differentiated sampling rates (DTE-SBD) algorithm [50], and the GAN-CNN framework. Table 4 gives detailed descriptions of the other model configurations.

4.3 Comparison

To gauge the overall performance of the proposed framework, we compare it with several state-of-the-art models, including the DTE-SBD, XGBoost, RF, SMOTE-XGBoost, SMOTE-RF, and GAN-CNN models aforementioned, on the four subsets of the experimental dataset, namely, $FW_1 \sim FW_4$. Figure 9 compares the results obtained on these subsets in terms of the four selected assessment metrics.

It can be observed that the proposed approach outperforms the other methods on all four subsets. As the amount of training data increases, the proposed approach becomes more accurate, thus demonstrating the scalability of the proposed framework. As the level of imbalance increases, the DTE-SBD, XGBoost, RF, SMOTE-XGBoost, and SMOTE-RF models become more vulnerable to imbalance effects, while the GAN-CNN model and the proposed framework both show robust performance despite the imbalance. Moreover, the proposed framework performs better than GAN-CNN. Overall, the DMGAN generates better samples for dataset rebalancing than the unimodal GAN does.

4.4 Internal functions

In this study, we propose a deep multimodal generative and fusion framework consisting of two unique modules to address the challenges that arise in learning from imbalanced multimodal data. To our knowledge, this paper is the first time that a DMGAN has been introduced to rebalance a dataset by generating pseudofeatures for each modality and then combining them to form fake samples. In addition, a DMHFN with gate and interaction mechanisms is presented to capture the fine-grained relationships among different feature modalities and integrate these relationships into the model in the form of a co-occurrence matrix.

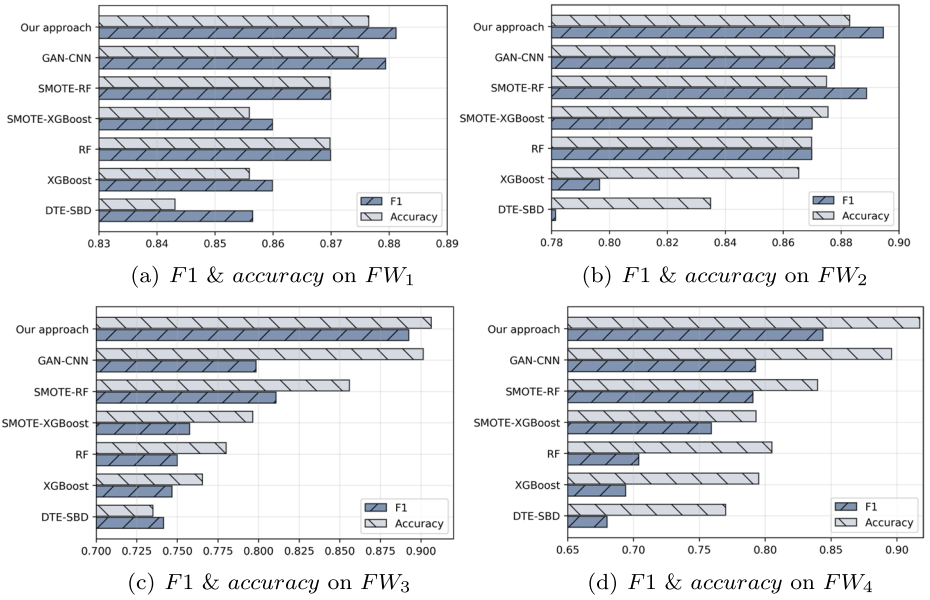
To gauge the effectiveness of the DMGAN and DMHFN modules, in the rest of this section, we first examine the performance of the DMGAN when faced with class imbalance, and then validate the ability of the DMHFN to capture the relationships among different feature modalities. Finally, we explore the robustness of the DMHFN for different data sizes.

4.4.1 DMGAN for class imbalance

We carry out a series of experiments using our DMHFN on imbalanced data, data augmented with classic SMOTE, data augmented with a state-of-the-art GAN, and

Table 4 Descriptions of the other model configurations

Category	Model	Configuration
Class imbalance	SMOTE	The number of nearest neighbors to be used to construct synthetic samples is set to 5. The number of nearest neighbors to be used to determine whether a minority sample is in danger is set to 10. The estimator is chosen to be a support vector machine (SVM).
	Unimodal GAN	Consists of three GANs, namely, a text GAN, an image GAN, and a layout GAN. The generators in the unimodal GAN framework have the same configurations as the generators in the DMGAN. In contrast, the text discriminator in the unimodal GAN framework has a convolutional layer with 300 1D filters with a size of 3 and one fully-connected layer with 500 units. The layout discriminator has two fully-connected layers with 500 units each. The image discriminator has two fully-connected layers with 100 units each.
Multimodal classification	XGBoost	The number of gradient-boosted trees is 50. The maximum tree depth for the base learners is 10. The boosting learning rate is 0.005. The booster is a gradient-boosted tree. The minimum loss reduction is 0.05. The $L2$ regularization term for the weights is 0.3. The sub-sampling ratio for the training instances is 0.5. The sub-sampling ratio for the columns when constructing each tree is 0.5.
	RF	Includes 50 trees and an MSE-based split equality estimator, and the minimum number of leaves is 5.
	CNN	Has the same settings as the gated fusion network except for the gate and interaction mechanisms.
Combined	SMOTE-XGBoost	Combines SMOTE and XGBoost.
	SMOTE-RF	Combines SMOTE and RF.
	DTE-SBD	Described in [50].
	GAN-CNN	Combines the unimodal GAN and CNN approaches.



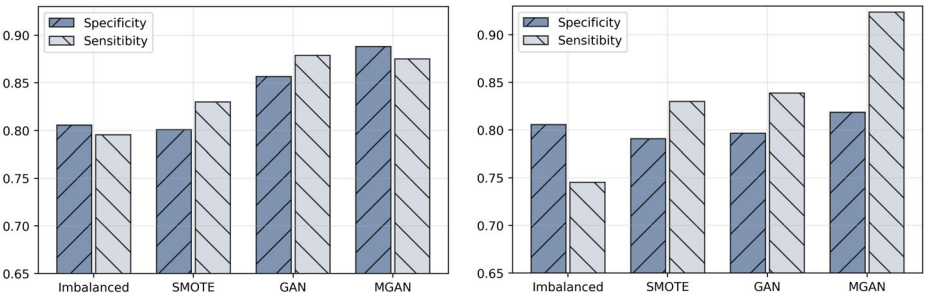
(a) $F1$ & $accuracy$ on FW_1 (b) $F1$ & $accuracy$ on FW_2

(c) $F1$ & $accuracy$ on FW_3 (d) $F1$ & $accuracy$ on FW_4

Fig. 9 $F1$ & $accuracy$ results for the different models on the four subsets

data augmented with our proposed DMGAN. The original imbalanced data used in these experiments consist of the previously introduced subsets FW_3 and FW_4, which represent two different class imbalance situations. Figure 10 shows the results of the different methods in terms of $specificity$ and $sensitivity$. As the class imbalance problem worsens, the $specificity$ and $sensitivity$ of the DMHFN degrade. Accordingly, the $specificity$ and $sensitivity$ of the DMHFN are worse on the more imbalanced subset.

In the SMOTE approach, the number of samples in the minority class is increased by creating fake samples, each of which is a linear combination of two real samples from the minority class that are located near each other, to rebalance the data. However, these fake samples are merely linear combinations of local information instead of being drawn



(a) $Specificity$ & $sensitivity$ on FW_3 (b) $Specificity$ & $sensitivity$ on FW_4

Fig. 10 $Specificity$ & $sensitivity$ achieved with different data augmentation methods

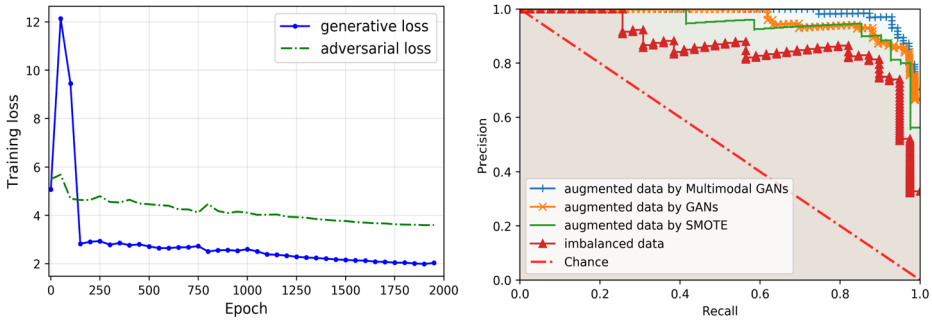
from the overall minority-class distribution [10]. As Fig. 10 demonstrates, the performance of SMOTE is not very good. We argue that the capabilities of this method are limited.

Previous studies on the application of GANs to imbalanced data have focused on unimodal data rather than multimodal data. Figure 10 shows the results obtained when using the GAN model in the proposed framework to generate fake features. The $specificity$ and $sensitivity$ of identifying faculty homepages are generally improved compared to those in the cases of no data augmentation and SMOTE augmentation.

In the proposed framework, to overcome the class imbalance problem in the case of multimodal data, we use the DMGAN to generate fake samples to augment the minority class (faculty homepages). The DMGAN consists of three generators and four discriminators. The generators create features for each feature modality through iterative interaction with the four discriminators, thus causing the fused distribution of the generated data to gradually approach the real distribution over multiple iterations while preserving the individual characteristics of each feature modality.

Figure 11a shows the adversarial loss (of the discriminators) and the generative loss (of the generators) during each iteration of the network learning process. Both the adversarial and generative losses show an initial sharp decrease and then gradually converge to lower values. The generative loss declines very smoothly throughout the entire learning process, while the adversarial loss fluctuates over a relatively broad range at the beginning of the learning process. Both losses converge when the number of iterations exceeds 2,000, indicating that the fake fused distribution can effectively imitate the real distribution after the competitive fraud/anti-fraud game between the generators and discriminators. Then, by applying the proposed DMGAN, the imbalanced dataset is transformed into an augmented dataset with balanced classes.

As seen in Fig. 10a and b, compared with the other three cases, the $specificity$ and $sensitivity$ for identifying faculty homepages are further improved with DMGAN augmentation, even in the case of high class imbalance. Figure 11b presents the precision-recall (PR) curves of the above three methods. Here, a curve that is closer to the upper right corner represents a model with better performance. The results illustrate that DMGAN outperforms the other two approaches. A good explanation of the superior performance of the DMGAN over the classical GAN and SMOTE methods is that its ability to generate fake features for each feature modality allows it to preserve both the individual characteristics of each feature set and the relationships among the multiple modalities.



(a) Loss comparison (b) PR comparison

Fig. 11 Loss & PR comparisons

4.4.2 Gate and interaction mechanisms in the DMHFN

The task of recognizing faculty homepages is essentially a multimodal classification problem, in which a target faculty homepage is identified based on three different types of features, namely, text, image, and layout features. Our proposed DMHFN can well address this type of problem. Specifically, the DMHFN contains gate and interaction mechanisms. The gate mechanism identifies direct and fine-grained relationships between feature modalities, and the interaction mechanism integrates these relationships into the model in the form of a co-occurrence matrix. This section reports a series of experiments carried out on subset FW_4 to evaluate the effectiveness of these two proposed mechanisms. Specifically, experimental evaluations are conducted using two model variants, as follows:

- $G\&I$: The proposed DMHFN with both the gate and interaction mechanisms enabled.
- $NG\&NI$: The proposed DMHFN with both the gate and interaction mechanisms disabled.

Figure 12 shows the performance achieved by these two variants in terms of $F1$ and $accuracy$. As seen in Fig. 12, $G\&I$ outperforms $NG\&NI$ in terms of $F1$, which indicates that $G\&I$ can achieve excellent $precision$ and $recall$. In addition, $G\&I$ shows improved $accuracy$. These findings provide evidence that the gate and interaction mechanisms are essential for improving model performance.

Next, we will provide insight into the ability of the gate mechanism to find direct and fine-grained relationships among different feature modalities based on the attention mechanism and the ability of the interaction mechanism to integrate these relationships into the model in the form of a co-occurrence matrix. We consider the text and layout features as an example. For simplicity, we remove the text autoencoder and the layout autoencoder from the DMHFN. For illustration, a segment of HTML source code is displayed in Fig. 13.

After such a segment of HTML source code is processed by the gate and interaction mechanisms, an attention matrix (Fig. 14a) and a co-occurrence matrix (Fig. 14b) are obtained. Based on the input code segment, the gate mechanism can find the direct and fine-grained relationships between the text and layout features. The interaction mechanism can then further integrate the frequencies of occurrence of various text and layout features on the basis of the attention matrix.

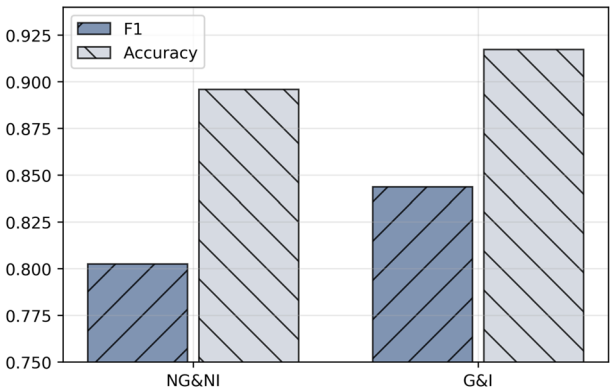


Fig. 12 $F1$ & $accuracy$ results for two model variants

```
<div class="field__label">Title</div>
<div class="field__item">Professor of the Graduate School</div>
<div class="field__label">Department</div>
<div class="field__item">Dept of Economics</div>
<div class="field__label">Research Expertise and Interest</div>
<div class="field__items">
<span class="field__item">finance</span>,
<span class="field__item">probability theory</span>,
<span class="field__item">mathematical economics</span>,
<span class="field__item">nonstandard analysis</span>
</div>
<div class="field__label">Research Description</div>
<p>His research interests include mathematical economics,
finance, nonstandard analysis, and probability theory.</p>
```

Fig. 13 An example of HTML source code

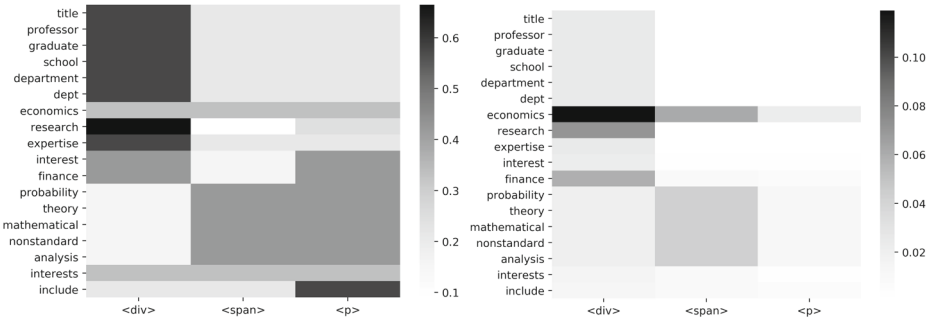
In brief, the gate mechanism serves to find direct and fine-grained relationships between the two feature modalities. In addition, the interaction mechanism integrates these relationships while considering the frequencies of the various components of the two feature modalities. The combination of both mechanisms enables a trade-off that improves the multimodal classification performance in the case of interdependent feature modalities.

4.4.3 Autoencoders in the DMHFN

This section reports a series of experiments carried out on subset FW_4 to evaluate the effectiveness of the autoencoders in the DMHFN. Specifically, experimental evaluations are conducted using two model variants, as follows:

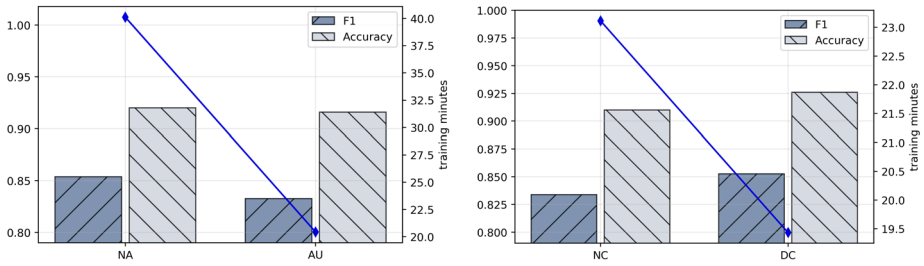
- NA: The proposed DMHFN without autoencoders.
- AU: The proposed DMHFN with autoencoders.

According to our understanding, autoencoders can learn compressed representations of input data. As seen in Fig. 15a, after the removal of the text, image, and layout autoencoders, the $F1$ and $accuracy$ values of the DMHFN are only slightly increased. However, the training time is nearly doubled. These findings provide evidence that the use of autoencoders can effectively save training time while maintaining the performance of the DMHFN.



(a) Attention matrix (b) Co-occurrence matrix

Fig. 14 Heat maps of the attention matrix and co-occurrence matrix



(a) Performance with and without autoencoders (b) Performance with and without a deformable CNN

Fig. 15 Performance with and without autoencoders & a deformable CNN

4.4.4 Deformable CNN in the DMHFN

The co-occurrence matrix in the DMHFN is a relatively large sparse matrix; only a small fraction of the elements are non-zero, and the dense regions have various geometrical shapes. Traditional CNN modules sample the input feature maps at fixed locations; therefore, such approaches have difficulty handling geometric variations. To handle this issue, we introduce a deformable convolutional kernel [7, 62]. Such a kernel is suitable for accommodating the sparsity and geometric variability of the co-occurrence matrix. This section reports a series of experiments carried out on subset FW_4 to evaluate the effectiveness of using a deformable CNN in the DMHFN. Specifically, experimental evaluations are conducted using two model variants, as follows:

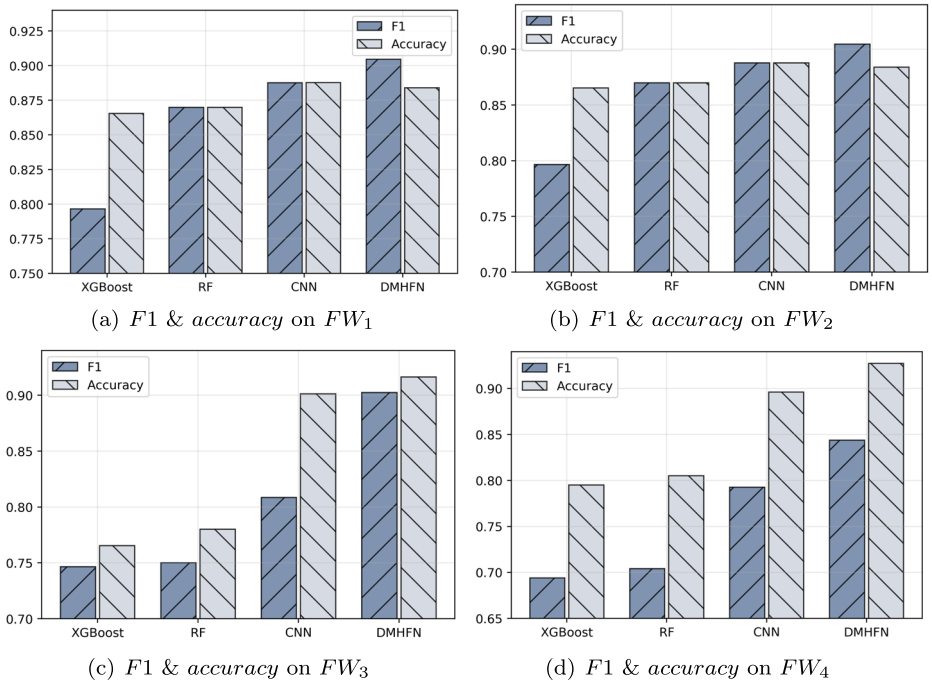
- NC: The DMHFN with a normal CNN.
- DC: The DMHFN with a deformable CNN.

As displayed in Fig. 15b, compared to the DMHFN with a normal CNN, the DMHFN with a deformable CNN shows slight increases in both $F1$ and $accuracy$ and a significant decrease in training time.

4.4.5 Comparison of the DMHFN with other models

To gauge the overall performance of the proposed DMHFN, we compare it with three state-of-the-art algorithms: XGBoost, RF, and CNN. In addition, to make the performance comparison more convincing, we compare the DMHFN with the other models on all four subsets $FW_1 \sim FW_4$ aforementioned. Figure 16 shows the detailed experimental results in terms of $F1$ and $accuracy$. The proposed framework achieves the best performance, generally followed (in approximate order of decreasing performance) by the CNN, RF, and XGBoost models, as the class imbalance becomes more severe. However, our model does not have an overwhelming advantage over the CNN model when evaluated on the two smaller subsets of data.

Our model exhibits more obvious advantages when evaluated on FW_3 and FW_4, as shown in Fig. 16c and d. The values of $F1$ and $accuracy$ further increase. Figure 16 shows that the DMHFN has the highest true positive rate and the lowest false positive rate among all tested models, thus indicating that the proposed approach ensures the highest probability of correctly and successfully identifying faculty homepages.



(a) $F1$ & $accuracy$ on FW_1 (b) $F1$ & $accuracy$ on FW_2

(c) $F1$ & $accuracy$ on FW_3 (d) $F1$ & $accuracy$ on FW_4

Fig. 16 $F1$ & $accuracy$ results for different models on the four subsets of data

5 Conclusions and future work

In this study, we propose a deep multimodal generative and fusion framework for multimodal classification with class-imbalanced data. This framework consists of two modules, namely, a deep multimodal generative adversarial network (DMGAN) and a deep multimodal hybrid fusion network (DMHFN). The DMGAN handles the class imbalance problem by generating fake features for each feature modality through iterative adversarial training. As a result of this procedure, the fused distribution of the counterfeit features approaches the joint distribution of the real features. At the same time, the individual characteristics of each modality and the relationships among different information sources are preserved. The DMHFN integrates information from diverse sources at different fusion levels for multimodal classification. Specifically, in the DMHFN, a gate mechanism is introduced to find direct interactions among fine-grained components from multiple modalities, while an interaction mechanism is used to aggregate these relationships based on the co-occurrence matrix. The task of recognizing faculty homepages is a typical binary classification problem involving multimodal features, namely, text, image, and HTML layout features, all of which are related in a complicated fashion. In addition, recognizing faculty webpages is a class-imbalanced problem, as the total number of samples in the minority class (faculty homepages) is far smaller than the total number of samples in the majority class (non-faculty webpages). Experiments on a faculty homepage dataset we collected show the effectiveness of the internal functions of the proposed framework and its advantages over other state-of-the-art models.

The proposed deep multimodal generative and fusion framework can be generalized to many other multimodal classification problems involving class-imbalanced data and interdependent feature modalities. One good example is media-aware stock movement prediction, in which the market information space consists of several modalities, including transaction data, news articles, and investors' moods in bear markets [27, 28, 30]. However, the effectiveness of the proposed deep multimodal generative and fusion framework has yet to be explored in other related fields. We plan to perform such explorations in the near feature.

Acknowledgements This work was supported by the National Natural Science Foundation of China (NSFC) (71671141 and 71873108), the National Social Science Foundation of China (NSSFC) (Grant No. 19BFX120), the Fundamental Research Funds for the Central Universities (JBK 171113, JBK 170505, JBK 1806003, and JBK 2002030), the Science and Technology Department of Sichuan Province (2019YJ0250), the Fintech Innovation Center of Southwestern University of Finance and Economics, and the Financial Intelligence and Financial Engineering Key Laboratory of Sichuan Province.

References

1. Ai C, Norton EC (2003) Interaction terms in logit and probit models. Econ Lett 80(1):123–129
2. Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: International conference on learning representations (ICLR)
3. Baltrušaitis T, Ahuja C, Morency L-P (2018) Multimodal machine learning: a survey and taxonomy. IEEE Trans Pattern Anal Mach Intell (TPAMI) 41(2):423–443
4. Basu A (1976) Elementary statistical theory in sociology. Brill Archive, 12
5. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res (JAIR) 16:321–357
6. Chen T, Guestrin C (2016) XGBOOST: a scalable tree boosting system. In: Proceedings of the 22nd ACM international conference on knowledge discovery and data mining (SIGKDD), pp 785–794
7. Dai J, Qi H, Xiong Y, Li Y, Zhang G, Hu H, Wei Y (2017) Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision (ICCV), pp 764–773
8. Dal Pozzolo A, Caelen O, Johnson RA, Bontempi G (2015) Calibrating probability with undersampling for unbalanced classification. In: IEEE Computational intelligence, 2015 IEEE symposium series, pp 159–166
9. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition (CVPR), IEEE, vol 1, pp 886–893
10. Douzas G, Bacao F (2018) Effective data generation for imbalanced learning using conditional generative adversarial networks. Expert Syst Appl (ESWA) 91:464–471
11. Douzas G, Bacao F (2018) Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. Inf Sci 465:1–20
12. Dwibedi D, Aytar Y, Tompson J, Sermanet P, Zisserman A (2019) Temporal cycle-consistency learning. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 1801–1810
13. Farnadi G, Tang J, De Cock M, Moens M-F (2018) User profiling through deep multimodal fusion. In: Proceedings of the eleventh ACM international conference on web search and data mining, ACM, pp 171–179
14. Gao L, Guo Z, Zhang H, Xu X, Shen HT (2017) Video captioning with attention-based LSTM and semantic consistency. IEEE Trans Multimed (TMM) 19(9):2045–2055
15. Gao L, Li X, Song J, Shen HT (2019) Hierarchical LSTMs with adaptive attention for visual captioning. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)
16. Goodfellow I, Pouge Abadie J, Mirza M, Xu B, Warde Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Advances in Neural Information Processing Systems (NIPS), pp 2672–2680
17. Guo W, Wang J, Wang S (2019) Deep multimodal representation learning: a survey. IEEE Access 7:63373–63394
18. He H, Bai Y, Garcia EA, Li S (2008) ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence), IEEE, pp 1322–1328

19. He H, Garcia EA (2008) Learning from imbalanced data. IEEE Trans Knowl Data Eng (TKDE), (9)1263–1284
20. He H, Shen X (2007) A ranked subspace learning method for gene expression data classification. In: International conference on artificial intelligence (ICAI), pp 358–364
21. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning (ICML)
22. James AP, Dasarathy BV (2014) Medical image fusion: a survey of the state of the art. Inform Fusion 19:4–19
23. Japkowicz N, Stephen S (2002) The class imbalance problem: a systematic study. Intell Data Anal (IDA) 6(5):429–449
24. Khaleghi B, Khamis A, Karray FO, Razavi SN (2013) Multisensor data fusion: a review of the state-of-the-art. Inform Fusion 14(1):28–44
25. Kingma DP, Ba JL (2015) Adam: a method for stochastic optimization. In: International conference on learning representations (ICLR)
26. Kubat M, Holte RC, Matwin S (1998) Machine learning for the detection of oil spills in satellite radar images. Mach Learn 30(2-3):195–215
27. Li Q, Chen Y, Jiang LL, Li P, Chen H (2016) A tensor-based information framework for predicting the stock market. ACM Trans Inf Syst (TOIS) 34(2):11
28. Li Q, Tan J, Wang J, Chen H (2020) A multimodal event-driven LSTM model for stock prediction using online news. IEEE Transactions on Knowledge and Data Engineering (TKDE). https://doi.org/10.1109/TKDE.2020.2968894
29. Li Q, Wang J, Wang F, Li P, Liu L, Chen Y (2017) The role of social sentiment in stock markets: a view from joint effects of multiple information sources. Multimed Tools Appl (MTAP) 76(10):12315–12345
30. Li Q, Wang T, Gong Q, Chen Y, Lin Z, Song S (2014) Media-aware quantitative trading based on public web information. Decision Support Systems (DSS) 61:93–105
31. Louzada F, Ferreira Silva PH, Diniz CarlosAR (2012) On the impact of disproportional samples in credit scoring models: an application to a brazilian bank data. Expert Syst Appl (ESWA) 39(9):8071–8078
32. Mansoorizadeh M, Charkari NM (2010) Multimodal information fusion application to human emotion recognition from face and speech. Multimed Tools Appl (MTAP) 49(2):277–297
33. Mathieu MF, Zhao JJ, Zhao J, Ramesh A, Sprechmann P, LeCun Y (2016) Disentangling factors of variation in deep representation using adversarial training. In: Advances in Neural Information Processing Systems (NIPS), pp 5040–5048
34. Metz CE (1978) Basic principles of roc analysis. In: Seminars in Nuclear Medicine, vol 8. Elsevier, Amsterdam, pp 283–298
35. Morvant E, Habrard A, Ayache S (2014) Majority vote of diverse classifiers for late fusion. In: Joint IAPR international workshops on statistical techniques in pattern recognition (SPR) and structural and syntactic pattern recognition (SSPR). Springer, Berlin, pp 153–162
36. Ngiam J, Khosla A, Kim M, Nam J, Lee H, Ng AY (2011) Multimodal deep learning. In: Proceedings of the 28th international conference on machine learning (ICML), pp 689–696
37. Oshri B, Khandwala N (2015) There and back again: autoencoders for textual reconstruction
38. Pearson R, Goney G, Shwaber J (2003) Imbalanced clustering for microarray time-series. In: Proceedings of the international conference on machine learning (ICML), vol. 3
39. Poria S, Cambria E, Bajpai R, Hussain A (2017) A review of affective computing: from unimodal analysis to multimodal fusion. Inform Fusion 37:98–125
40. Potamianos G, Neti C, Gravier G, Garg A, Senior AW (2003) Recent advances in the automatic recognition of audiovisual speech. Proc IEEE 91(9):1306–1326
41. Qi X, Davison BD (2009) Web page classification: features and algorithms. ACM Comput Surv (CSUR) 41(2):12
42. Rendle S (2010) Factorization machines. In: 2010 IEEE international conference on data mining (ICDM), pp 995–1000
43. Roth K, Lucchi A, Nowozin S, Hofmann T (2017) Stabilizing training of generative adversarial networks through regularization. In: Advances in neural information processing systems (NIPS), pp 2018–2028
44. Shin HC, Tenenholtz NA, Rogers JK, Schwarz CG, Senjem ML, Gunter JL, Andriole KP, Michalski M (2018) Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In: International workshop on simulation and synthesis in medical imaging (SASHIMI). Springer, Berlin, pp 1–11
45. Song J, Guo Y, Gao L, Li X, Hanjalic A, Shen HT (2018) From deterministic to generative: multimodal stochastic rnns for video captioning. IEEE Trans Neural Netw Learn Syst (TNNLS) 30(10):3047–3058
46. Song J, Zhang J, Gao L, Liu X, Shen HT (2018) Dual conditional GANs for face aging and rejuvenation. In: International joint conference on artificial intelligence (IJCAI), pp 899–905

47. Sprent P, Smeeton NC (2000) Applied nonparametric statistical methods. Chapman and Hall/CRC
48. Srivastava N, Mansimov E, Salakhudinov R (2015) Unsupervised learning of video representations using LSTMs. In: International conference on machine learning (ICML), pp 843–852
49. Srivastava N, Salakhutdinov RR (2012) Multimodal learning with deep boltzmann machines. In: Advances in neural information processing systems (NIPS), pp 2222–2230
50. Sun J, Lang J, Fujita H, Li H (2018) Imbalanced enterprise credit evaluation with DTE-SBD: decision tree ensemble based on SMOTE and bagging with differentiated sampling rates. Inf Sci 425:76–91
51. Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. In: Advances in neural information processing systems (NIPS), pp 3104–3112
52. Suzuki M, Nakayama K, Matsuo Y (2017) Joint multimodal learning with deep generative models. In: International conference on learning representations (ICLR) (Workshop)
53. Tsai C, Lin W, Hu Y, Yao G (2019) Under-sampling class imbalanced datasets by combining clustering analysis and instance selection. Inf Sci 477:47–54
54. Tsai Y-HH, Bai S, Liang PP, Kolter JZ, Morency L-P, Salakhutdinov R (2019) Multimodal transformer for unaligned multimodal language sequences. In: The 57th annual meeting of the association for computational linguistics (ACL 2019), pp 6558–6569
55. Vartak MN, et al. (1955) On an application of kronecker product of matrices to statistical designs. Ann Math Stat 26(3):420–438
56. Wu M, Goodman N (2018) Multimodal generative models for scalable weakly-supervised learning. In: Advances in neural information processing systems (NIPS), pp 5575–5585
57. Xu T, Zhang P, Huang Q, Zhang H, Gan Z, Huang X, He X (2018) AttnGAN: fine-grained text to image generation with attentional generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 1316–1324
58. Yih Wt, He X, Meek C (2014) Semantic parsing for single-relation question answering. In: Proceedings of the 52nd annual meeting of the association for computational linguistics (NAACL) (Short Papers), vol 2, pp 643–648
59. Yingzhen L, Mandt S (2018) Disentangled sequential autoencoder. In: International conference on machine learning (ICML), pp 5670–5679
60. Yu G, Li Q, Wang J, Zhang D, Liu Y (2020) A multimodal generative and fusion framework for recognizing faculty homepages. Inf Sci 525:205–220
61. Zhang C, Yang Z, He X, Deng L (2020) Multimodal intelligence: representation learning, information fusion, and applications. IEEE Journal of Selected Topics in Signal Processing, 1–1
62. Zhu X, Hu H, Lin S, Dai J (2019) Deformable convnets v2: more deformable, better results. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 9308–9316