A MULTIMODAL GENERATIVE AND FUSION FRAMEWORK FOR RECOGNIZING FACULTY HOMEPAGES

Anonymous

ABSTRACT

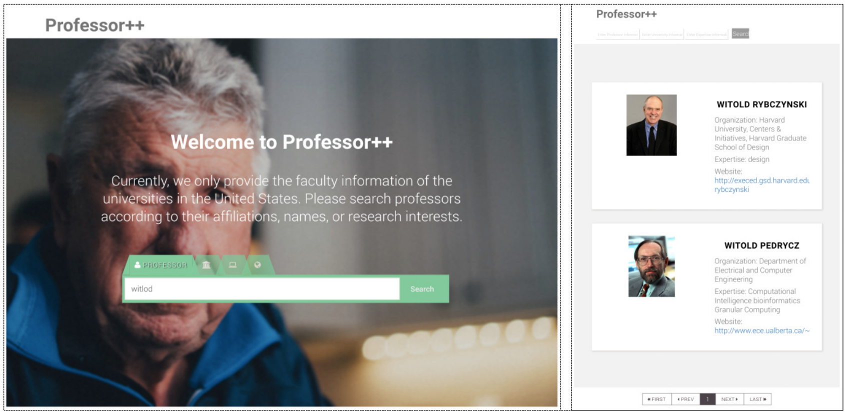
Multimodal data consists of several data modes: each mode is a group of similar data sharing the same attributes. Recognizing faculty homepages is essentially a multimodal classification problem in which a target faculty webpage is determined by three different information sources, namely, textual content, images, and layout. Common strategies in previous studies have been either to concatenate features of various information sources into a compound vector or to feed the features separately into several different classifiers, which are then assembled into a stronger classifier for the final decision. Both approaches ignore the interactions among different feature sets. We argue that such interactions are important to enhance multimodal classification. In addition, recognizing faculty webpages is a class-imbalanced problem in which the total sample number of a minority class is far less than the total sample number of other classes. In this study, we propose a multimodal generative and fusion framework for multimodal learning in the case of imbalanced data and interactive feature modes. In particular, a multimodal generative adversarial net is first introduced to rebalance the dataset by generating pseudo features in terms of each mode and integrating them to create a fake sample. Then, a gated fusion network with interactive and gate mechanisms is presented to capture the links among different feature modes and reduce noise for generalization. Experiments on a faculty homepage dataset show the superiority of the proposed framework.

***K*eywords** faculty homepages · multimodal generative adversarial nets · gated fusion network

1 Introduction

A faculty search engine aims to obtain relevant information about researchers and to trace hot research topics. In this research, we designed and implemented a vertical search engine, Professor++ [1], for this purpose, shown in Figure 1. At present, the search engine covers the Top-100 universities in the United States, and Professor++ will be extended to all universities in the near future. This faculty-oriented search engine provides query functions in terms of name, university and research area. It also supports advanced statistical analyses, including the distribution of faculty members in terms of research interests, ethnicity, and gender.

[1] `http://www.findingprofessor.com`



Welcome to Professor++

Currently, we only provide the faculty information of the universities in the United States. Please search professors according to their affiliations, names, or research interests.

witlod

WITOLD RYBCZYNSKI

Organization: Harvard University, Centers & Initiatives, Harvard Graduate School of Design

Expertise: design

Website: http://execed.gsd.harvard.edu rybczynski

WITOLD PEDRYCZ

Organization: Department of Electrical and Computer Engineering

Expertise: Computational Intelligence bioinformatics Granular Computing

Website: http://www.ece.ualberta.ca/~

Figure 1: Professor++ screen shot.

Automatic recognition of faculty homepages on university websites is critical to building such a faculty-oriented search engine. This task is essentially a multimodal classification problem that consists of several data modes, each of which shares the same attributes. In particular, a target faculty webpage is recognized in terms of three different information sources, namely, textual content, images, and layout. Two critical challenges are considered in this study.

- **Data imbalance**: An imbalanced training sample means that one class is represented by a large number of examples while the other is represented by only a few. Such imbalance may cause deterioration of the classification accuracy because the model pays less attention to the minority classes [1]. Generally, the number of faculty homepages is much less than the total number of university webpages. A good example is the official website of Yale University, which includes fewer than 5,000 faculty and researcher homepages but 50,000+ webpages overall.
- **Multimodality**: Generally, the page source of a webpage consists of three parts, that is, textual content, layout, and images. A targeted homepage and other webpages must be distinguished on the basis of these three features. Common strategies in previous studies have been either to concatenate features of various information sources into a compound vector or to feed the features separately into different classifiers, which are then assembled into a stronger classifier for the final decision. Both approaches inevitably ignore the interactions among different features [2]. For example, as displayed in Figure 2, the layout feature set consists of three tags, that is, $< title >$, $< p >$, and $< footer >$. Each tag contains some textual information. The words embedded in the tag $< title >$ are more important than those within the tag $< footer >$. Such interlinks are ignored, however, if the tag and text features are concatenated into a compound vector.

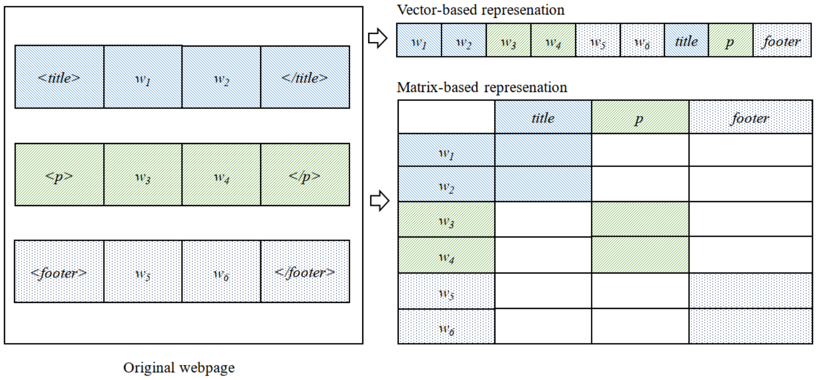


Figure 2: Modeling multimodal data.

In this study, we propose a multimodal generative and fusion framework for recognizing faculty homepages, which is a standard multimodal learning problem with imbalanced data and interactive feature modes. This work makes three unique contributions:

- We propose a fusion neural network for interactive multimodal data. This network is achieved by utilizing the Kronecker product to capture the interactions among different modal data and by applying a gate mechanism to alleviate the noise and redundant features while capturing such interactions.
- To overcome the class-imbalanced problem for multimodal data, we design an advanced framework called multimodal generative adversarial networks, which generate fake multimodal features while preserving interactions among them.
- To the best of our knowledge, Professor++ is the first faculty-oriented search engine. Our source code for recognizing faculty homepages can be accessed via GitHub[2].

The remainder of this article is organized as follows. Section 2 briefly describes the related work. Section 3 presents the design details of the proposed frameworks. Section 4 examines the effectiveness of our approach. The article is concluded with our feature works in Section 5.

2 Related Work

2.1 Multimodal Classification

Recognizing faculty homepages is essentially a classification problem. The typical method of identifying homepages is to train a binary classifier on webpage features with labelled data and then apply this well-trained classifier to recognize new homepages. However, most previous work partitioning data relied on only one type of webpage feature, such as URLs [3, 4, 5], textual content [6] or HTML layout [7]. In fact, recognizing faculty homepages is a multimodal data problem with three different data modes (textual content, layout, and images) that transmit different viewpoints to support supervised learning. Many studies have found that a classifier that utilizes multiple feature sets is able to achieve better performance than one using a single feature set [8, 9, 10, 11].

Two common strategies are considered in previous studies to use multiple feature sets. The first is to feed the different features separately into different classifiers, which are then further assembled into a stronger classifier via a voting or stacking mechanism to reach the final decision [2]. For instance, Joachims et al. trained several support vector machines (SVMs) with different feature sets, including URLs and text, and then combined these SVMs into a strong classifier via a voting mechanism [12]. Glover et al. combined the results from an SVM on extended anchor text and an SVM on full-text to achieve improved performance [8]. Chen and Hsieh first built an SVM with webpage literal words and then trained an SVM on semantic information via latent semantic analysis (LSA). A weighted schema was further utilized to assemble the results to make the final decision [13].

An alternative approach is to concatenate features of various information sources into a super compound vector. For example, Kang and Kim combined the feature sets of textual content, links and URLs with a weighted sum [10].

Both approaches ignored or disconnected the interactions among different feature sets (modes), which are critical for multimodal supervised learning. In this study, three data modes (text, layout, and image) are used to identify faculty homepages. The interactions among these feature modes are important for identifying homepages. A good example is that textual information within different layout tags has varying importance. Separating these interlinks causes a loss of valuable information and reduces the classification accuracy. We argue that such interactions are critical for multimodal classification. Therefore, we propose the gated fusion network (GFN) to incorporate the interlinks among different feature modes for multimodal data classification. This goal is achieved by introducing the Kronecker product to preserve interactions among different modes and implementing a gate mechanism to filter out redundant information and noise.

2.2 Imbalanced Classification

The class-imbalanced problem refers to case where one class is represented by a large number of examples (majority class) while the other is represented by only a few (minority class). Most classic learning algorithms are designed for balanced datasets. The model performance declines once the dataset becomes imbalanced, as the minority class cannot be learned effectively [14]. The classifier tends to over-represent the majority rather than adequately represent the minority. The minority class is so small that it is easily ignored, treated as noise, or identified as the majority class [15, 16, 17, 18]. Louzada et al. showed that class-imbalanced data can severely deteriorate model performance [19].

The traditional techniques for addressing the class-imbalanced problem include random oversampling (OS), random undersampling (US) and the synthetic minority oversampling technique (SMOTE) [20].

[2] `https://github.com/mrspider520/gated_fusion_network.git`

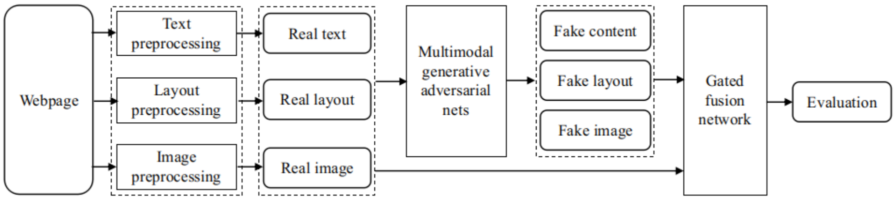


Figure 3: System framework.

In the random US approach, observations from the majority class are randomly dropped to match the number of samples in the minority class. This US results in the loss of valuable information and inevitably reduces the classification performance [21].

The random OS approach randomly duplicates observations from the minority class to match the number of samples in the majority class. The OS approach is prone to overfitting due to the simple replication of samples from the minority class [22, 14, 23]. In essence, OS does not provide additional valuable information for classification.

To overcome the weaknesses of random OS, Chawla et al. proposed SMOTE for imbalanced classification [20]. SMOTE rebalances the data by increasing the number of samples in the minority class by creating virtual samples, each of which is a linear combination of two similar real samples in the minority class. However, these virtual samples are linear combinations of local information instead of the overall minority class distribution [22]. In addition, the linear combination may produce noisy samples if the boundary between the majority and the minority classes is not sufficiently clear [14].

The generation of synthetic samples based on the real minority class distribution is a critical challenge. Some researchers have been taking a further step by utilizing generative adversarial nets (GANs). Such a framework relies on training a generator and a discriminator, which compete with each other in a zero-sum game. The well-trained generator is able to estimate the latent distribution of the real data and produce perfect fake samples according to the global distribution rather than local information [24]. For example, Shin et al. utilized GANs to rebalance medical data due to limited numbers of cancer patients by synthesizing abnormal MRI images with brain tumours [25].

However, previous studies in the application of GANs for imbalanced data focus on unimodal data rather than multimodal data. In this study, we propose multimodal generative adversarial nets to synthesize samples for imbalanced multimodal data. Specifically, our model generates fake faculty homepages by using three modal features (text, layout and image) simultaneously according to the global feature distribution and the interactions of the features to improve the classification of imbalanced webpages.

3 System Design

In this study, we propose a multimodal generative and fusion framework for multimodal learning in the case of imbalanced data and interactive feature modes. Figure 3 presents an overview of the proposed generative and fusion framework. Image features, layout features, and text features are first extracted from original webpages and then preprocessed to form a multimodal dataset. A multimodal GAN is introduced to rebalance the dataset by generating pseudo features for each mode and integrating these features to describe a fake sample. Then, a gated fusion network with interactive and gate mechanisms is presented to capture the links among different feature modes and reduce the noise for generalization.

3.1 Data

The source of a webpage typically consists of three feature sets, namely, text, images, and layout. Detailed descriptions of these features are as follows, and the description of all notations sees Table 1.

- **Text feature**: The text of a webpage can be represented as a word list $\boldsymbol{x}_t \in \mathbb{R}^N$, where N is the number of words on a webpage. To enhance the semantic and context information, text can be further represented by embedded words. In this study, we utilize the Google word vector model [3] to convert $\boldsymbol{x}_t$ into matrix $\boldsymbol{X}_t \in \mathbb{R}^{N \times E}$, where E is the embedding size. We apply convolution kernels to extract the semantic information from embedded words, as suggested by [26]. $\boldsymbol{X}_t$ is further transformed into a more abstract feature $\boldsymbol{h}_t$ via convolution kernels.

[3] The model consists of 3 million 300-dimension English word vectors and is accessible at `https://code.google.com/archive/p/word2vec/`

- **Image feature**: Instead of representing an image via pixels, we extract the image feature of a webpage as a four-dimensional vector $\boldsymbol{x}_i \in \mathbb{R}^4$. The elements of this vector include the number of zero-face images, the number of one-face images, the number of multiple-face images, and the total number of images. The faces in each image are recognized by a HOG face recognition algorithm [27]. $\boldsymbol{h}_i$ is a high-level feature of $\boldsymbol{x}_i$.

- **Layout feature**: The layout feature of a webpage is represented as a tag vector $\boldsymbol{x}_s \in \mathbb{R}^M$, where M is the total number of tags. Each element in this vector is the number of HTML leaf tags, including $< a >$, $< p >$, $< span >$, and so on. $\boldsymbol{h}_s$ is a corresponding high-level feature.

Table 1: Description of notations

Symbol	Description
$\boldsymbol{x}_t$	a vector including N words, $\boldsymbol{x}_t \in \mathbb{R}^N$
$\boldsymbol{X}_t$	an embedding matrix with embedding size E, $\boldsymbol{X}_t \in \mathbb{R}^{N \times E}$
$\boldsymbol{h}_t$	the more abstract text features
$\boldsymbol{g}_t$	the text gate with the same shape as $\boldsymbol{h}_t$
$\boldsymbol{f}_t$	the filtered text after the gate mechanism
$\boldsymbol{x}_s$	the layout features including M HTML leaf tags, $\boldsymbol{x}_s \in \mathbb{R}^M$
$\boldsymbol{h}_s$	the more abstract layout features
$\boldsymbol{g}_s$	the layout gate with the same shape as $\boldsymbol{h}_s$
$\boldsymbol{f}_s$	the filtered layout features after the gate mechanism
$\boldsymbol{x}_i$	an image vector with feature size 4, $\boldsymbol{x}_i \in \mathbb{R}^4$
$\boldsymbol{h}_i$	the more abstract image features
$\boldsymbol{g}_i$	the image gate with the same shape as $\boldsymbol{h}_i$
$\boldsymbol{f}_i$	the filtered image features after the gate mechanism
$\boldsymbol{z}_t$	the content noise, $\boldsymbol{z}_t \in \mathbb{R}^O$, $\boldsymbol{z}_t \sim N(0, 1)$
$\boldsymbol{z}_s$	the layout noise, $\boldsymbol{z}_s \in \mathbb{R}^Q$, $\boldsymbol{z}_s \sim N(0, 1)$
$\boldsymbol{z}_i$	the image noise, $\boldsymbol{z}_i \in \mathbb{R}^P$, $\boldsymbol{z}_i \sim N(0, 1)$
$\boldsymbol{w}$	a weight vector
b	a bias
G_s, G_i, G_t	three generators for layout, image and text features
D	a discriminator in multimodal generative adversarial nets
$\nabla_x y$	the gradient of x given y
p_x	data distribution about x
$\bigodot$	the Hadamard product
$\bigotimes$	the Kronecker product
σ	the sigmoid active function $\sigma(x) = \frac{1}{1+e^{-x}}$

3.2 MGANs

To overcome the class-imbalanced problem for multimodal data, we design a novel architecture called multimodal generative adversarial nets (MGANs) to generate fake samples for the minority class. MGANs creates fake features in terms of each feature mode via iterative adversarial training. This procedure can make the fusion distribution of the fake features approach that of the real features while preserving the independence of each feature set and the interactions among them.

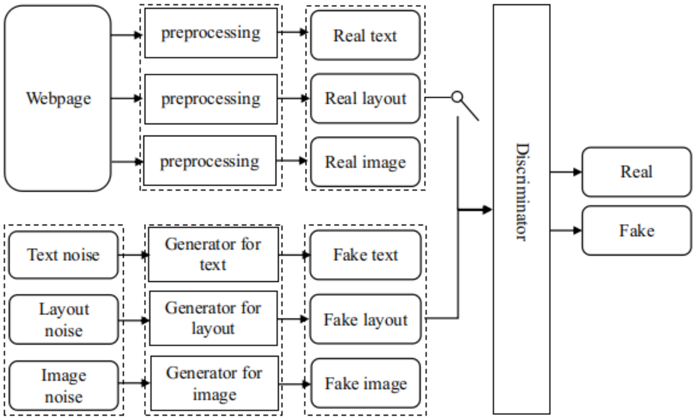


Figure 4: The framework of multimodal generative adversarial nets.

Fig. 4 shows an overview of the MGANs framework. The framework consists of three feature generators, G_t, G_i, and G_s, and one discriminator, *D*. The generators forge fake samples $G_t(\boldsymbol{z}_t)$, $G_i(\boldsymbol{z}_i)$, and $G_s(\boldsymbol{z}_s)$ using random samples $\boldsymbol{z}_t$, $\boldsymbol{z}_i$, and $\boldsymbol{z}_s$. The outputs are further fed into a discriminator along with real sample data ($\boldsymbol{X}_t$, $\boldsymbol{x}_i$, $\boldsymbol{x}_s$) to train and determine whether the inputs come from the real or fake distribution. Essentially, the three generators aim to fool the discriminator, which acts as an anti-fraud agent and provides knowledge to improve the fraudulent ability of the three generators via adjusting their weights. The objective function of the MGANs can be defined as follows:

$$\begin{aligned}&\min_{G_s,G_i,G_t}\max_{D} L(G_s,G_i,G_t,D)\\&=\min_{G_s,G_i,G_t}\max_{D}\{\mathbb{E}_{c(\boldsymbol{X}_t,x_i,x_t)\sim p_{x_c}}[logD(c(\boldsymbol{X}_t,x_i,x_s))]+\\&\mathbb{E}_{c(G_s(\boldsymbol{z}_s),G_i(\boldsymbol{z}_i),G_s(\boldsymbol{z}_t)\sim p_{g_c}}[log(1-D(c(G_s(\boldsymbol{z}_s),G_i(\boldsymbol{z}_i),\\&G_s(\boldsymbol{z}_t)))]\},\end{aligned} \tag{1}$$

where $c(\cdot)$ is a fusion function embedded in the discriminator. For example, $c(\boldsymbol{X}_t,\boldsymbol{x}_i,\boldsymbol{x}_s)$ is the fusion result of real feature sets including $\boldsymbol{X}_t$, $\boldsymbol{x}_i$ and $\boldsymbol{x}_s$. p_{x_c} denotes the distribution. Similarly, $c(G_s(\boldsymbol{z}_s),G_i(\boldsymbol{z}_i),G_s(\boldsymbol{z}_t))$ is the fusion result of the fake feature sets, and p_{g_c} is the distribution. In the fraud and anti-fraud game between the generators and the discriminator, the purpose of the three generators is to confuse the discriminator (make $D(c(\boldsymbol{X}_t,\boldsymbol{x}_i,\boldsymbol{x}_s))$ be close to 0) and to make the discriminator believe that the generated features are real features (make $D(c(G_t(\boldsymbol{z}_t),G_i(\boldsymbol{z}_i),G_s(\boldsymbol{z}_s)))$ be close to 1) with the discriminator fixed. By contrast, the purpose of the discriminator is to distinguish real features from generated features; that is, the discriminator is trained to make $D(c(\boldsymbol{X}_t,\boldsymbol{x}_i,\boldsymbol{x}_s))$ approximate 1 to make $D(c(G_t(\boldsymbol{z}_t),G_i(\boldsymbol{z}_i),G_s(\boldsymbol{z}_s)))$ be close to 0, with the generators fixed. After solving the above objective function, we can obtain $p_{g_c}=p_{x_c}$, which means that the fusion distribution of the features generated by three generators is able to converge to one of the real features. The rest of this section gives a detailed proof.

For simplicity, let

$$\begin{aligned}\boldsymbol{x}_c&=c(\boldsymbol{x}_s,\boldsymbol{x}_i,\boldsymbol{X}_t)\\\boldsymbol{g}_c&=c(G_s(\boldsymbol{z}_s),G_i(\boldsymbol{z}_i),G_t(\boldsymbol{z}_t)).\end{aligned} \tag{2}$$

The following function is maximized to find the optimal discriminator *D* given three generators,

$$\begin{aligned}&L(G_s,G_i,G_t,D)\\&=\int_{\boldsymbol{x}_c}[p_{x_c log(D(\boldsymbol{x}_c))}]d\boldsymbol{x}_c+\int_{\boldsymbol{g}_c}[p_{g_c}log(1-D(\boldsymbol{g}_c))]d\boldsymbol{g}_c\\&=\int_{\boldsymbol{x}_c}[p_{x_c}log(D(\boldsymbol{x}_c))+p_{g_c}log(1-D(\boldsymbol{x}_c))]d\boldsymbol{x}_c.\end{aligned} \tag{3}$$

Thus,

$$\begin{aligned}\frac{\partial L}{\partial \boldsymbol{x}_c} &= \frac{p_{x_c}}{D(\boldsymbol{x}_c)}\frac{\partial D(\boldsymbol{x}_c)}{\partial \boldsymbol{x}_c} - \frac{p_{g_c}}{1-D(\boldsymbol{x}_c)}\frac{\partial D(\boldsymbol{x}_c)}{\partial \boldsymbol{x}_c} \\ \frac{\partial L}{\partial D(\boldsymbol{x}_c)} &= \frac{\frac{\partial L}{\partial \boldsymbol{x}_c}}{\frac{\partial D(\boldsymbol{x}_c)}{\partial \boldsymbol{x}_c}} = \frac{p_{x_c}}{D(\boldsymbol{x}_c)} - \frac{p_{g_c}}{1-D(\boldsymbol{x}_c)} = 0.\end{aligned} \tag{4}$$

Therefore, the best discriminator is

$$D^*(\boldsymbol{x}_c) = \frac{p_{x_c}}{p_{x_c}+p_{g_c}}. \tag{5}$$

Given three generators G_s, G_i and G_t, the maximum of the objective function (Eq. (1)) with the optimal discriminator D^* is

$$\begin{aligned}&\max_D L(G_s, G_i, G_t, D) = L(G_s, G_i, G_t, D^*) \\ &= \mathbb{E}_{\boldsymbol{x}_c\sim p_{x_c}}[logD^*(\boldsymbol{x}_c)] + \mathbb{E}_{\boldsymbol{x}_c\sim p_{g_c}}[(1 - logD^*(\boldsymbol{g}_c))] \\ &= \mathbb{E}_{\boldsymbol{x}_c\sim p_{x_c}} log\frac{p_{x_c}}{p_{x_c}+p_{g_c}} + \mathbb{E}_{\boldsymbol{x}_c\sim p_{g_c}} log\frac{p_{g_c}}{p_{x_c}+p_{g_c}}.\end{aligned} \tag{6}$$

Then, $L(G_s, G_i, G_t, D^*)$ is minimized to obtain the optimal generators (G_s, G_i and G_t),

$$\begin{aligned}&\min_{p_{g_c}} L = \min_{p_{g_c}}[\mathbb{E}_{\boldsymbol{x}_c\sim p_{x_c}} log\frac{p_{x_c}}{p_{x_c}+p_{g_c}} + \mathbb{E}_{\boldsymbol{x}_c\sim p_{g_c}} log\frac{p_{g_c}}{p_{x_c}+p_{g_c}}], \\ &\frac{\partial L}{\partial p_{g_c}} = \frac{1}{p_{g_c}} - \frac{2}{p_{x_c}+p_{g_c}} = 0, \\ &p_{g_c} = p_{x_c}.\end{aligned} \tag{7}$$

Therefore, the minimum value of $L(G_s, G_i, G_t, D^*)$ is $-\log 4$, and we can obtain $p_{g_c} = p_{x_c}$.

$$\begin{aligned}&JS(p_{x_c}||p_{g_c}) = \frac{1}{2}[KL(p_{x_c}||p_{g_c}) + KL(p_{g_c}||p_{x_c})] \\ &= \frac{1}{2}[p_{x_c}log\frac{p_{x_c}}{p_{g_c}} + p_{g_c}log\frac{p_{g_c}}{p_{x_c}}] = 0,\end{aligned} \tag{8}$$

where KL is the Kullback-Leibler divergence and JS is the Jensen-Shannon divergence, both of which measure the similarity between two distributions. Small values indicate a small difference between two distributions. $JS = 0$ indicates that generators G_s, G_i, and G_t are able to learn the real feature distributions. If D^* is certainty, $U = \mathbb{E}_{\boldsymbol{x}_c\sim p_{g_c}}[logD^*(\boldsymbol{x}_c)] + \mathbb{E}_{\boldsymbol{x}_c\sim p_{g_c}}[1 - logD^*(\boldsymbol{g}_c)]$ is a convex function with a unique global optimal solution $p_{g_c} = p_{x_c}$. In other words, during the iteration in network training via gradient descent, p_{g_c} gradually converges to p_{x_c}.

In short, the generators constantly adjust their network parameters through the confrontation with the discriminator so that the fusion distribution of the generated data gradually approaches the real fusion distribution.

Algorithm 1 Multimodal generative adversarial nets training with mini-batch gradient descent.

```
Inputs: $\boldsymbol{x}_t \in \mathbb{R}^N$, $\boldsymbol{x}_i \in \mathbb{R}^4$, and $\boldsymbol{x}_s \in \mathbb{R}^M$ with positive label $y_{pos}$.
Outputs: Three well-trained generators $G_s$, $G_i$, and $G_t$.
1: while $k$ steps do
2:     Draw $n$ batches of size $\boldsymbol{x}_t$, $\boldsymbol{x}_i$, and $\boldsymbol{x}_s$ from the training set
3:     $\boldsymbol{X}_t \in \mathbb{R}^{N\times E} \leftarrow \boldsymbol{x}_t$ is embedded
4:     Draw $n$ batches of size $\boldsymbol{z}_t$, $\boldsymbol{z}_i$, and $\boldsymbol{z}_s$ from the real distribution
5:     $G_t(\boldsymbol{z}_t) \in \mathbb{R}^{N\times E} \leftarrow G_t$ gets $\boldsymbol{z}_t$; $G_i(\boldsymbol{z}_i) \in \mathbb{R}^4 \leftarrow G_i$ gets $\boldsymbol{z}_i$; $G_s(\boldsymbol{z}_s) \in \mathbb{R}^M \leftarrow G_s$ gets $\boldsymbol{z}_s$
6:     Assign a negative label $y_{neg}$ to $G_t(\boldsymbol{z}_t)$, $G_i(\boldsymbol{z}_i)$, and $G_s(\boldsymbol{z}_s)$, and use them and real samples to train the discriminator
7:     Repeat Step 5, assign positive label $y_{pos}$ to $G_t(\boldsymbol{z}_t)$, $G_i(\boldsymbol{z}_i)$, and $G_s(\boldsymbol{z}_s)$, and use those to train the three generators
8: end while
```

3.3 The Gated Fusion Network (GFN)

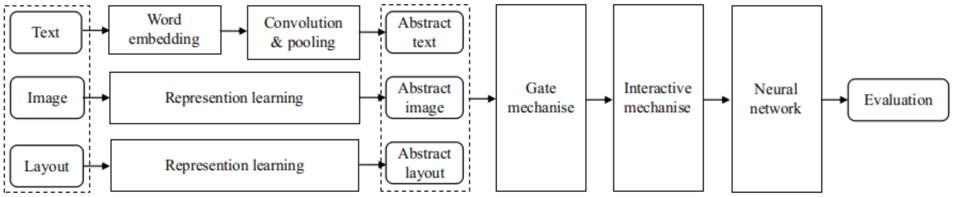
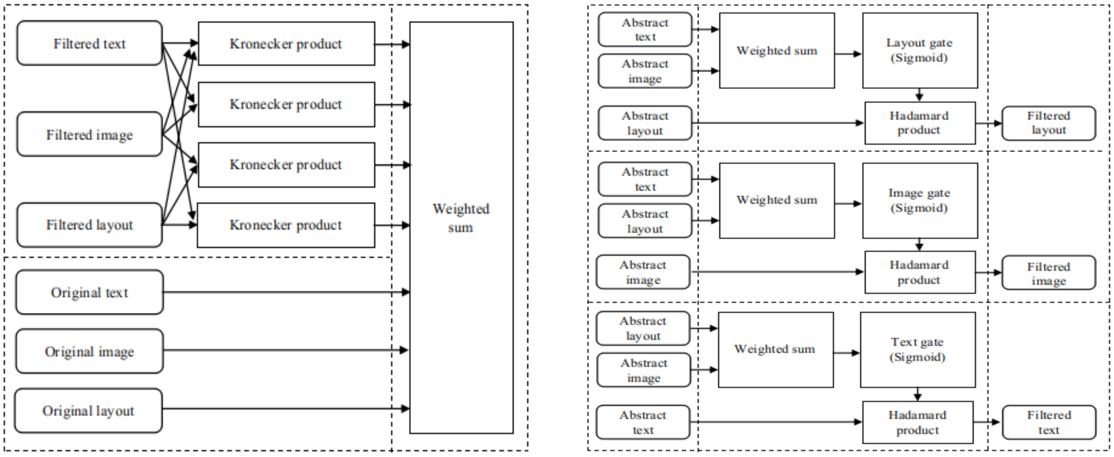


Figure 5: The framework of the gated fusion network.

In this study, we propose a gated fusion network with interactive and gate mechanisms, as shown in Figure 5.



(a) Interactive mechanism (b) Gate mechanism

Figure 6: Two mechanisms.

3.3.1 The Interactive Mechanism

Faculty homepage recognition is essentially a multimodal problem with three different interactive data modes, that is, textual content, layout, and images. Specifically, textual information within different layout tags has varying importance. Separating these interlinks causes a loss of valuable information and reduces the classification accuracy. However, previous studies ignored or disconnected the interactions among feature sets (modes), which are critical for multimodal supervised learning. In the proposed gated fusion network, an interactive mechanism is designed to preserve the interactions among modes. Figure 6(a) describes the strategy of the interactive mechanism. Intuitively, the interactive information is determined by the joint effect of the relevant sources. In particular, an independent variable on a dependent variable can be measured by the magnitude of other associated independent variables [28]. Therefore, interactions can be measured by the product if the information sources are scalars or by the Kronecker product if the information modes are vectors [29, 30]. In fact, Rendle and Steffen adopted the product to capture interactions among features in factorization machines [31].

The feature space for considering the interactions among different information modes can be formally denoted as

$$\begin{aligned} \boldsymbol{o} = \boldsymbol{h}_t\boldsymbol{w}_t + \boldsymbol{h}_s\boldsymbol{w}_s + \boldsymbol{f}_i\boldsymbol{w}_i + (\boldsymbol{f}_t \otimes \boldsymbol{f}_s)\boldsymbol{w}_{ts} + (\boldsymbol{f}_t \otimes \boldsymbol{f}_i)\boldsymbol{w}_{ti} \\ + (\boldsymbol{f}_s \otimes \boldsymbol{f}_i)\boldsymbol{w}_{si} + (\boldsymbol{f}_t \otimes \boldsymbol{f}_s \otimes \boldsymbol{f}_i)\boldsymbol{w}_{tsi} + b, \end{aligned} \tag{9}$$

where $\otimes$ denotes the Kronecker product and $\boldsymbol{h}_t$, $\boldsymbol{h}_s$ and $\boldsymbol{h}_i$ denote the text, layout and image features, respectively. $\boldsymbol{f}_t$, $\boldsymbol{f}_s$ and $\boldsymbol{f}_i$ denote the filtered text, layout and image features, respectively. The reason for applying filtered features instead of the original features is to greatly reduce the information for interactive processing by removing the noise and irrelevant information via a gate mechanism. The gate mechanism is described in detail in Section 3.3.2. $\boldsymbol{f}_t \otimes \boldsymbol{f}_s$, $\boldsymbol{f}_t \otimes \boldsymbol{f}_i$, and $\boldsymbol{f}_s \otimes \boldsymbol{f}_i$ measure the interactions of text and layout features, text and image features, and layout and image features, respectively. $\boldsymbol{f}_t \otimes \boldsymbol{f}_s \otimes \boldsymbol{f}_i$ measures the interactions among text, layout and image features. $\boldsymbol{w}_s$, $\boldsymbol{w}_t$, $\boldsymbol{w}_i$, $\boldsymbol{w}_{ts}$, $\boldsymbol{w}_{ti}$, $\boldsymbol{w}_{si}$, and $\boldsymbol{w}_{tsi}$ are the weight vectors, and b is the bias.

3.3.2 The Gate Mechanism

The gate mechanism was first introduced in long short-term memory to solve the long-term dependency problem of recurrent neural networks [32]. The input gate controls the inputs by adjusting the amount of valuable information to access the hidden states, while the output gate controls the outputs of the hidden states to access the next stage of the model by preventing newly generated noise. In this study, we design three gates, that is, a text gate, a layout gate, and a image gate, to preserve the features of one mode that interact with features in another mode to be utilized in the interactive mechanism. Figure 6(b) shows the framework of the proposed gate mechanism, and the mathematical formulas are as follows:

$$\boldsymbol{g}_t = \sigma(\boldsymbol{h}_s \boldsymbol{w}_s^{(t)} + \boldsymbol{h}_i \boldsymbol{w}_i^{(t)} + b^{(t)}), \tag{10}$$

$$\boldsymbol{g}_s = \sigma(\boldsymbol{h}_t \boldsymbol{w}_t^{(s)} + \boldsymbol{h}_i \boldsymbol{w}_i^{(s)} + b^{(s)}), \tag{11}$$

$$\boldsymbol{g}_i = \sigma(\boldsymbol{h}_t \boldsymbol{w}_t^{(i)} + \boldsymbol{h}_s \boldsymbol{w}_s^{(i)} + b^{(i)}), \tag{12}$$

where $\boldsymbol{h}_t$, $\boldsymbol{h}_s$, and $\boldsymbol{h}_i$ are high-level abstract text, layout and image features, respectively. Eqs. (10) to (12) describe text gate $\boldsymbol{g}_t$, layout gate $\boldsymbol{g}_s$ and image gate $\boldsymbol{g}_i$, respectively. σ denotes the sigmoid activation function. $\boldsymbol{w}_s^{(t)}$, $\boldsymbol{w}_i^{(t)}$, $\boldsymbol{w}_t^{(s)}$, $\boldsymbol{w}_i^{(s)}$, $\boldsymbol{w}_t^{(i)}$, and $\boldsymbol{w}_s^{(i)}$ are weight vectors. $b^{(t)}$, $b^{(s)}$, and $b^{(i)}$ are bias items.

Therefore, the filtered features for the interactive mechanism can be obtained as

$$\boldsymbol{f}_t = f(\boldsymbol{g}_t \odot \boldsymbol{h}_t), \tag{13}$$

$$\boldsymbol{f}_s = f(\boldsymbol{g}_s \odot \boldsymbol{h}_s), \tag{14}$$

$$\boldsymbol{f}_i = f(\boldsymbol{g}_i \odot \boldsymbol{h}_i), \tag{15}$$

where $\odot$ denotes the Hadamard product and $f(\cdot)$ denotes the activation function.

In Eqs. (10) to (12), $\sigma(\cdot)$ is the sigmoid function. Therefore, the value range of the elements of the vectors $\boldsymbol{g}_t$, $\boldsymbol{g}_s$ and $\boldsymbol{g}_i$ is from 0 to 1. When the i-th output value of the sigmoid function approaches 0, the i-th gate turns off. The i-th element in the feature vector is discarded when multiplied with this gate. Therefore, the gate can prevent irrelevant information from entering the next stage of the model. Similarly, the gate can also preserve the valuable information when the i-th output value of the sigmoid function approaches 1. The gate mechanism can control the data flow to obtain filtered features via parameter learning in the network. This gate reduces the amount of irrelevant information processed by the interactive mechanism and amplifies the influence of interactions among different feature modes.

3.3.3 Learning

Algorithm 2 Gated fusion network training with mini-batch gradient descent.

```
Inputs: Feature sets $x_t \in \mathbb{R}^N$, $x_i \in \mathbb{R}^4$, and $x_s \in \mathbb{R}^M$ and label $y$.
Outputs: The well-trained gated fusion network.
while $k$ steps do
    Draw $n$ batches of size $x_t$, $x_i$, and $x_s$ from the training set
    $X_t \in \mathbb{R}^{N \times E} \leftarrow x_t$ is embedded
    Map $X_t$ to $h_t$, $x_i$ to $h_i$, and $x_s$ to $h_s$ after a series of operations
    Calculate the three gates $g_t$, $g_s$, and $g_i$ according to Eqs. (10) to (12)
    Calculate the three filtered feature modes $f_t$, $f_s$, and $f_i$ according to Eqs. (13) to (15)
    Fuse the three filtered feature modes according to Eq. 9
    Feed the fused features to the next stage of the network
end while
```

After the features are preprocessed by the gate and interactive mechanisms, the task becomes a classic binary classification problem. The objective function of the network can be defined as

$$L = \frac{1}{n} \sum_{\boldsymbol{x}_t, \boldsymbol{x}_i, \boldsymbol{x}_s} [y \times log(GFN(\boldsymbol{x}_t, \boldsymbol{x}_i, \boldsymbol{x}_s)) + (1 - y) \times log(1 - GFN(\boldsymbol{x}_t, \boldsymbol{x}_i, \boldsymbol{x}_s))], \tag{16}$$

where $GFN(\boldsymbol{x}_t, \boldsymbol{x}_i, \boldsymbol{x}_s)$ denotes the outputs of the network with respect to $\boldsymbol{x}_t$, $\boldsymbol{x}_i$, and $\boldsymbol{x}_s$. n is the number of samples. To minimize the objective function, we apply adaptive moment estimation (Adam). To implement Adam, the derivative of the prediction is defined as

$$\nabla_{\boldsymbol{w},b}\frac{1}{n}\sum[y \times log(GFN(\boldsymbol{x}_t, \boldsymbol{x}_i, \boldsymbol{x}_s)) + (1-y) \times log(1 - GFN(\boldsymbol{x}_t, \boldsymbol{x}_i, \boldsymbol{x}_s))] \tag{17}$$

The details of the proposed gated fusion network are presented in Algorithm 2.

4 Experimental Evaluation

In this section, a series of experiments are conducted to gauge the effectiveness of the proposed multimodal generative adversarial framework. In particular, the inner functions aimed at interactive multimodal data and imbalanced data are evaluated.

4.1 Measures

The classic *accuracy* is applied to evaluate the model performance [33].

$$accuracy = \frac{TP + TN}{TP + FN + FP + TN}, \tag{18}$$

where *TP* denotes true positives, i.e., the number of faculty homepages that are successfully identified. *TN* denotes true negatives, i.e., the number of non-faculty webpages that are successfully identified. *FP* denotes false positives, i.e., the number of non-faculty webpages that are classified as faculty webpages. *FN* denotes false negatives, i.e., the number of faculty homepages that are incorrectly classified as non-faculty webpages.

Due to the class-imbalanced problem in the multimodal dataset, *accuracy* does not provide an effective comprehensive evaluation of model performance. We adopt alternative assessment metrics, namely, precision, recall, and f1, to evaluate model performance [14].

$$\begin{aligned} precision &= \frac{TP}{TP + FP}, \\ recall &= \frac{TP}{TP + FN}, \\ f1 &= \frac{2 \times recall \times precision}{recall + precision}. \end{aligned} \tag{19}$$

precision represents the proportion of true faculty homepages in the total predicted faculty homepages, *recall* measures how many faculty homepages, out of the total number of true faculty homepages, are correctly identified, and *f1* is a combination of *precision* and *recall*.

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

4.2 Experimental Data

In this study, we implemented a distributed web crawler to collect approximately $31,645$ webpages from the websites of several universities in the United States. Four different datasets that consist of three feature sets (modes) are used in our experimental evaluation.

- FW_1: Includes $7,943$ faculty homepages out of $15,886$ webpages. The ratio of faculty homepages to non-faculty webpages is $1:1$.
- FW_2: Includes $7,943$ faculty homepages out of $23,829$ webpages. The ratio of faculty homepages to non-faculty webpages is $1:2$.
- FW_3: Includes $7,943$ faculty homepages out of $31,645$ webpages. The ratio of faculty homepages to non-faculty webpages is $1:3$.
- FW_4: Includes $7,943$ faculty homepages out of $39,715$ webpages. The ratio of faculty homepages to non-faculty webpages is $1:4$.

4.3 Parameters

4.3.1 Our Framework

Note that we implement three tricks to alleviate the challenges of non-convergence, mode collapse and slow training for MGANs:

- add batch normalization layers into the discriminator [34] to accelerate and stabilize the training;
- choose the Adam optimizer as the top-priority solver to accelerate the training [35];
- add random noise to real and fake samples [36] to alleviate mode collapse.

We conduct a series of preliminary experiments to determine the optimal settings for the discriminator and three generators. Specifically, the discriminator has a convolutional layer with 250 1D filters (or convolutional kernels) in which the size is set to 3 and four fully connected layers with 250 units. The activation function is the sigmoid. For the generators, the text generator consists of five convolutional layers with 300 1D filters, whose size is set to 3. Both the image and the layout generator consist of five fully connected layers with different numbers of units. The first one is 100 units, and the second one is 500 units. The activation function of these generators is ReLU.

We propose a gated fusion network to process multimodal data with interactive feature sets. The network consists of an embedding layer that maps every text to a matrix with dimension 400×300, a convolutional layer and 6 fully connected layers. The activation functions are the sigmoid and hyperbolic tangent for the gate and interactive mechanisms, respectively. More detailed information can be found in our source code.

4.3.2 Other Models

To gauge the overall performance of the proposed framework and the gated fusion network, we compare them with several classic algorithms, namely, support vector machines (SVMs), multilayer perceptron (MLP), decision tree (DT), GANs-based convolutional neural network (CNN), and normal CNN.To gauge the overall performance of the proposed framework, we compare it with several classic algorithms, namely, SVMs, multilayer perceptron (MLP), decision tree (DT), GANs-based convolutional neural network (GANs-based CNN), and normal CNN. We conduct a series of preliminary experiments to determine the optimal settings for the models.

- SVMs: Includes the radial basis kernel function with $\gamma = 0.01$, a shrinking heuristic, a tolerance stopping criterion of $1e - 3$, and a penalty parameter $c = 1$.
- DT: Includes the *gini* split criterion, the *best* strategy used to choose the split at each node, the minimum number of samples to split an internal node is 2, and the minimum number of samples at a leaf node is 1.
- MLP: Includes 2 hidden layers with 2000 units with sigmoid activation functions. The maximum number of iterations 200, and the L2 penalty parameter $1e - 3$. We use the Adam optimizer with the initial learning rate $lr = 0.01$, the first exponential decay rate $\beta_1 = 0.9$, the second exponential decay rate $\beta_2 = 0.999$, and stability $\epsilon = 1e - 8$ as the optimal solver.
- CNN: Has the same settings as the gated fusion network except the gate and interaction mechanisms.
- Unimodal GANs: Consists of three groups of GANs, namely, text GANs, image GANs, and layout GANs. The generators in the unimodal GANs have same settings as the generators in the multimodal GANs, and the text discriminator in the unimodal GANs has a convolutional layer with 300 1D filters (or convolutional kernels) in which the size is set to 3, and one fully connected layers with 500 units, the layout discriminator has two fully connected layers with 500 units, the image discriminator has two fully connected layers with 100 units.
- GANs-based CNN: Combines the unimodal GANs and CNN.

4.4 Model Comparison

Table 2: Comprehensive comparison among different models on different datasets.

Data	Method	Precision	Recall	F1	Accuracy
FW_1	MLP	0.86	0.85	0.85	0.84
	DT	0.85	0.85	0.85	0.86
	SVMs	0.79	0.77	0.78	0.78
	GANs-based CNN	0.88	0.87	0.88	0.87
	Our Model	**0.88**	**0.88**	**0.88**	**0.88**
FW_2	MLP	0.86	0.84	0.85	0.83
	DT	0.76	0.77	0.76	0.77
	SVMs	0.78	0.67	0.72	0.78
	GANs-based CNN	0.88	0.88	0.88	0.88
	Our Model	**0.88**	**0.88**	**0.88**	**0.88**
FW_3	MLP	0.86	0.80	0.83	0.80
	DT	0.88	0.84	0.86	0.85
	SVMs	0.79	0.56	0.65	0.79
	GANs-based CNN	0.90	0.89	0.90	0.89
	Our Model	**0.91**	**0.91**	**0.91**	**0.91**
FW_4	MLP	0.87	0.73	0.79	0.73
	DT	0.89	0.84	0.86	0.84
	SVMs	0.82	0.56	0.67	0.81
	GANs-based CNN	0.91	0.90	0.90	0.90
	Our Model	**0.92**	**0.92**	**0.92**	**0.92**

To gauge the overall performance of the proposed framework, we compare it with several classic algorithms, namely, SVMs, multilayer perceptron (MLP), decision tree (DT), and GANs-based convolutional neural network (GANs-based CNN), on four experimental datasets, that is, FW_1, FW_2, FW_3, and FW_4. Table 2 shows the comparison results in terms of four assessment metrics.

The proposed approach outperforms the other methods on all four datasets. With more training data, the proposed approach becomes more accurate, which indicates the scalability of the proposed approach. As the imbalance level increases, MLP, DT and SVMs become more invulnerable while GANs-based CNN and the proposed approach show robust performance. In addition, the proposed framework is superior to GANs-based CNN. In other words, the multimodal GANs can generate better samples to rebalance the dataset than can the unimodal GANs.

4.5 Inner Functions

In this study, we propose a multimodal generative and fusion framework with two unique modules to address the challenges in imbalanced multimodal data learning. In particular, a multimodal generative adversarial net is introduced to rebalance the dataset by generating pseudo features of each mode and integrating the features to create a fake sample. Then, a gated fusion network with interactive and gate mechanisms is presented to capture the links among different feature modes and reduce noise for generalization. To gauge the effectiveness of these two modules, in the rest of this section, we first examine the performance of MGANs for the class-imbalanced problem, and then assess the capability of the GFN to capture the links among different feature modes. Finally, we explore the robustness of the GFN in terms of different data size.

4.5.1 MGANs for the Class-imbalanced Problem.

Table 3: The classification results of different methods.

(a) Imbalanced dataset

Class	Precision	Recall	F1
Non-faculty webpages	0.96	0.93	0.94
Faculty homepages	0.81	0.89	0.85
Average	0.92	0.92	0.92

(b) GANs

Class	Precision	Recall	F1
Non-faculty webpages	0.82	0.94	0.88
Faculty homepages	0.92	0.80	0.86
Average	0.87	0.87	0.87

(c) MGANs

Class	Precision	Recall	F1
Non-faculty webpages	0.87	0.93	0.9
Faculty homepages	**0.93**	0.86	**0.89**
Average	0.90	0.90	0.90

We first conduct experiments with the proposed multimodal generative and fusion framework by disabling its MGANs function. Table 3(a) shows the results in terms of four measurement metrics when the imbalanced data problem is not addressed. In our dataset, the number of non-faculty homepages (majority class) is approximately three times the number of faculty homepages (minority class). The precision of identifying non-faculty homepages (0.96) is much higher than that of recognizing faculty homepages (0.81).

Previous studies on the application of GANs for imbalanced data focused on unimodal data rather than multimodal data. Table 3(b) shows the results of using the classic GANs, instead of the proposed MGANs, to generate fake features in the proposed framework. The precision of identifying faculty homepage is improved from 0.81 to 0.92, while the recall declines from 0.89 to 0.8. At the same time, the precision of the opposite class is reduced considerably.

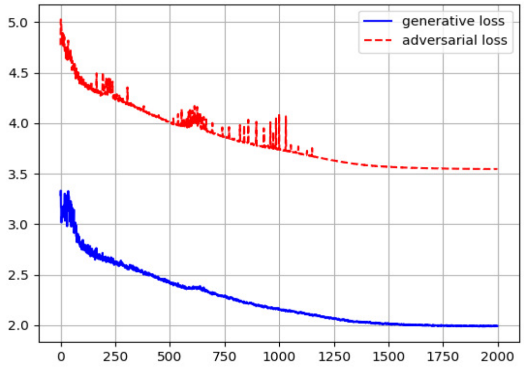


Figure 8: Loss.

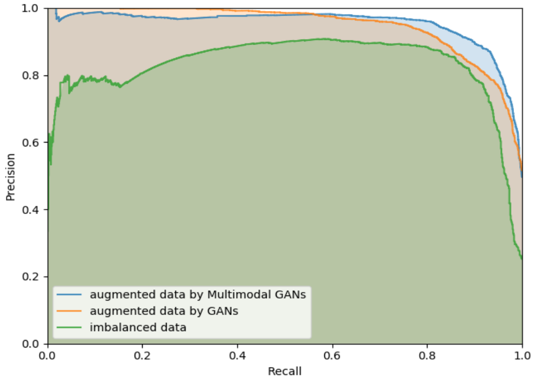


Figure 9: Comparison.

To overcome the class-imbalanced problem for multimodal data, we enable the MGANs function in the proposed framework to generate fake samples to expand the minority class (faculty homepages). MGANs consist of three generators and one discriminator. The generators forge features of each feature mode by iteratively confronting the discriminator to make the fusion distribution of the generated data gradually approach the real distribution. Figure 8 shows the adversarial loss (discriminator) and the generative loss (generators) during each iteration of the learning stage of the network. Both the adversarial and the generative loss decrease sharply and then gradually converge to lower values. The generative loss declines very smoothly throughout the entire learning stage, whereas the adversarial loss fluctuates in a relatively large range at the beginning of the learning stage. Both losses converge when the number of iterations exceeds 1750; that is, the fake fusion distribution sufficiently imitates the real distribution after the fraud and anti-fraud competition between the generators and discriminator. MGANs successfully converts the imbalanced dataset into an augmented dataset with balanced classes.

Table 3(c) shows the results of using MGANs in the proposed framework to address the class-imbalanced problem. Compared with Table 3(b), the precision, recall and $f1$ are improved from 0.92 to 0.93, 0.80 to 0.86, and 0.80 to 0.89, respectively.

Figure 9 presents the precision-recall (PR) curves of the above three methods. Here, the curve closest to the upper right corner represents the model with the best performance. The PR curves show that MGANs outperforms the other two approaches. MGANs outperforms classic GANs because of its ability to generate fake features of each feature mode to preserve the independence of each feature set and the interactions among different features.

4.5.2 Gate and Interactive Mechanisms in GFN

The recognition of faculty homepages is essentially a multimodal classification problem in which a target faculty homepage is determined by three different feature sets, namely, textual content, images, and layout. We propose a GFN with interactive and gate mechanisms. Specifically, the interactive mechanism is utilized to capture the links among different feature modes, and the gate mechanism is introduced to reduce the noise produced by the interactive mechanism. In this section, we conduct a series of experiments to evaluate the effectiveness of the two proposed mechanisms. Specifically, four tracks of experimental evaluation are investigated:

- $G\&I$: The proposed GFN with both gate and interactive mechanisms.
- $G\&NI$: The proposed GFN with the gate mechanism but without the interactive mechanism.
- $NG\&I$: The proposed GFN with the interactive mechanism but without the gate mechanism.
- $NG\&NI$: The proposed GFN without the gate and interactive mechanisms.

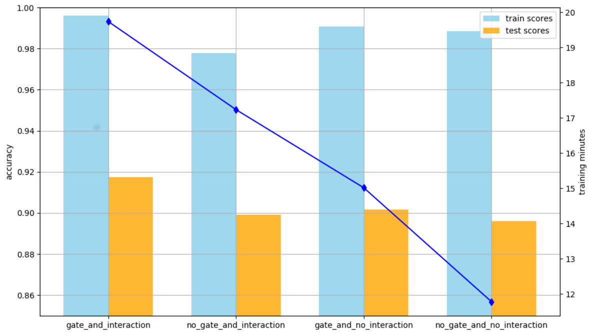
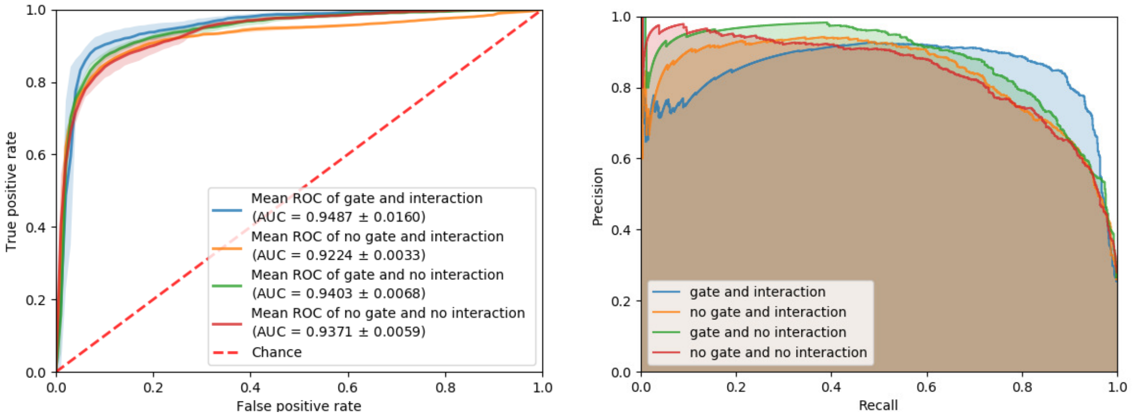


Figure 9: Accuracy of the GFN in terms of four tracks.

Figure 9 shows the performance of these four tracks on the training and testing datasets. We first explore the effectiveness of the gate mechanism. $G\&NI$ outperforms $NG\&NI$, with a significant enhancement on the testing data compared to the training data, and $G\&I$ outperforms $NG\&I$ on both the training and testing datasets by utilizing the gate mechanism. The gate mechanism alleviates overfitting by discarding irrelevant information and, hence, improves the model generalization and accuracy. Next, we examine the efficacy of the interactive mechanism. Compared with $NG\&NI$, $NG\&I$ shows a tiny improvement on the testing data and a small degradation on the training set. One explanation for this difference is that the interactive mechanism captures the interactions of different feature modes while producing noise, which deteriorates the performance if no efficient gate mechanism is available to filter the noise. Figure 10(a) and Figure 10(b) show the ROC and PR curves of the examined tracks, respectively. The superior performance of $G\&I$ confirms the above explanation.

The interactive mechanism captures the interlinks among different information modes but produces noise, and the gate mechanism provides a function to filter out the noise and trivial information. The combination of both methods achieves a tradeoff that results in better performance for multimodal learning with interactive feature modes.



(a) Mean ROC curve

(b) PR curve

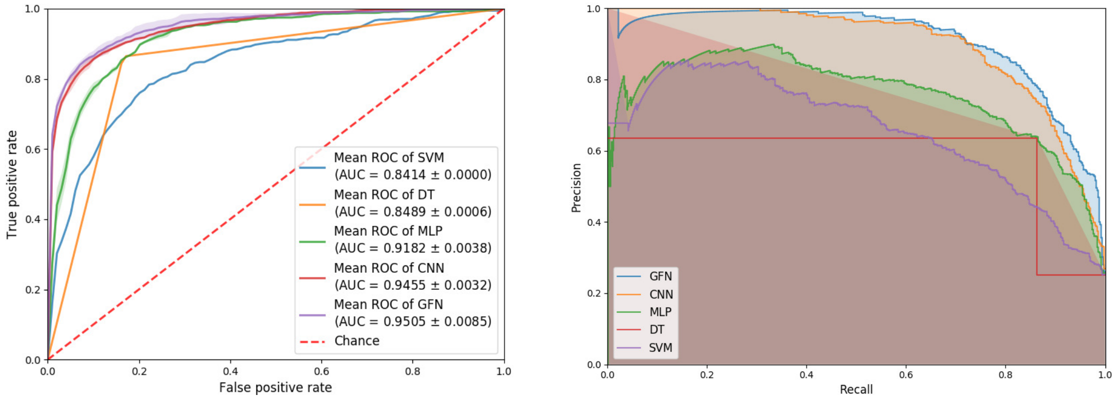
Figure 10: Learning curves of the GFN in terms of four tracks.

4.5.3 Comparison between GFN and Other Models

To gauge the overall performance of the proposed GFN, we compare it with several classic algorithms, namely, SVMs, MLP, DT, and CNN. In addition, to more thoroughly assess the efficacy of the GFN, we compare it with the other models on a small dataset and a large dataset. The large dataset is FW_3 , while the small dataset is FW_1, described in Section 4.2.

Table 4: Comparison among different models on a small dataset.

Model	Precision	Recall	F1	Accuracy
MLP	0.85	0.75	0.80	0.74
DT	0.87	0.84	0.85	0.84
SVMs	0.80	0.63	0.70	0.83
CNN	0.90	0.90	0.90	0.90
GFN	**0.91**	**0.91**	**0.91**	**0.92**



(a) Mean ROC curve

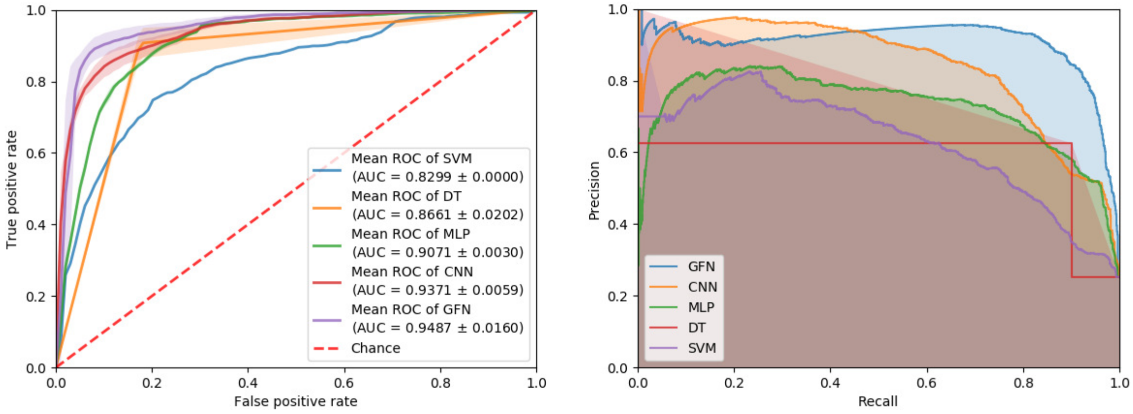
(b) PR curve

Figure 11: Learning curves of different models on a small dataset.

Table 4 shows the details of our experimental results in terms of the four assessment metrics. Figures 11(a) and 11(b) show the ROC and PR curves of these approaches on a small dataset, respectively. The proposed framework achieves the best performance, followed by CNN, DT, MLP, and SVMs. However, our model does not have a substantial advantage over the CNN on the small dataset.

Table 5: Comparison among different models on a large dataset.

Model	Precision	Recall	F1	Accuracy
MLP	0.86	0.80	0.83	0.80
DT	0.88	0.84	0.86	0.85
SVMs	0.79	0.56	0.65	0.79
CNN	0.90	0.90	0.90	0.90
GFN	**0.92**	**0.92**	**0.92**	**0.92**



(a) Mean ROC curve

(b) PR curve

Figure 12: Learning curves of different models on a large dataset.

Our model has a more obvious advantage when validated on a large dataset, as shown in Table 5. Three main measurement metrics further increase. Figure 12(a) and, especially, Figure 12(b) show a substantial performance improvement for our framework. The proposed framework has the highest true positive rate and lowest false positive rate, which indicates that the proposed approach ensures the highest probability of identifying faculty homepages and the lowest probability of falsely recognizing non-faculty pages as faculty homepages.

5 Conclusions and Future Work

The recognition of faculty homepages is typically a binary classification problem with multimodal features, namely, textual contents, HTML layout, and images, that interact in a complicated fashion. In this study, an interactive mechanism is introduced to capture the intrinsic relations among these multimodal features. The Kronecker product is applied to preserve the intrinsic relationships among different feature modes and to optimize them iteratively. To overcome the noise and model complexity caused by this mechanism, a gate mechanism is proposed. Faculty webpage recognition is also a class-imbalanced problem, in which the total sample number of the minority class is far less than the total number of the majority class. In this study, multimodal generative adversarial nets (MGANs) is introduced to rebalance the dataset. MGANs generate fake features in terms of each feature mode via iterative adversarial training. This procedure can make the fusion distribution of fake features approach the distribution of real features while preserving the independence of each feature set, as well as the interactions among features.

The proposed multimodal generative and fusion framework is generalizable to many other multimodal learning problems with class-imbalanced data and interactive feature modes. A good example is the prediction of media-aware stock movements, in which the market information space consists of several modes, including transaction data, news articles, and investors' mood in bear markets [37]. However, the effectiveness in related fields remains to be explored in the near future.

References

[1] Ugo Fiore, Alfredo De Santis, Francesca Perla, Paolo Zanetti, and Francesco Palmieri. Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. *Information Sciences*, 2017.

[2] Xiaoguang Qi and Brian D Davison. Web page classification: Features and algorithms. *ACM computing surveys (CSUR)*, 41(2):12, 2009.

[3] Min-Yen Kan and Hoang Oanh Nguyen Thi. Fast webpage classification using url features. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 325–326. ACM, 2005.

[4] M Indra Devi and K Selvakuberan. Fast web page categorization without the web page. In *International Conference on Semantic Web and Digital Libraries (ICSD-2007). ARD Prasad & Devika P. Madalli*. Citeseer, 2007.

[5] Eda Baykan, Monika Henzinger, Ludmila Marian, and Ingmar Weber. Purely url-based topic classification. In *Proceedings of the 18th international conference on World wide web*, pages 1109–1110. ACM, 2009.

[6] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 256–263. ACM, 2000.

[7] Floriana Esposito, Donato Malerba, Luigi Di Pace, and Pietro Leo. A machine learning approach to web mining. In *Congress of the Italian Association for Artificial Intelligence*, pages 190–201. Springer, 1999.

[8] Eric J Glover, Kostas Tsioutsiouliklis, Steve Lawrence, David M Pennock, and Gary W Flake. Using web structure for classifying and describing web pages. In *Proceedings of the 11th international conference on World Wide Web*, pages 562–569. ACM, 2002.

[9] Pável Calado, Marco Cristo, Edleno Moura, Nivio Ziviani, Berthier Ribeiro-Neto, and Marcos André Gonçalves. Combining link-based and content-based methods for web document classification. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 394–401. ACM, 2003.

[10] In-Ho Kang and GilChang Kim. Query type classification for web document retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 64–71. ACM, 2003.

[11] Xiaoguang Qi and Brian D Davison. Knowing a web page by the company it keeps. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 228–237. ACM, 2006.

[12] Thorsten Joachims, Nello Cristianini, and John Shawe-Taylor. Composite kernels for hypertext categorisation. In *ICML*, volume 1, pages 250–257, 2001.

[13] Rung-Ching Chen and Chung-Hsun Hsieh. Web page classification based on a support vector machine using a weighted vote schema. *Expert Systems with Applications*, 31(2):427–435, 2006.

[14] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge & Data Engineering*, (9):1263–1284, 2008.

[15] Miroslav Kubat, Robert C Holte, and Stan Matwin. Machine learning for the detection of oil spills in satellite radar images. *Machine learning*, 30(2-3):195–215, 1998.

[16] Ronald Pearson, Gregory Goney, and James Shwaber. Imbalanced clustering for microarray time-series. In *Proceedings of the ICML*, volume 3, 2003.

[17] Haibo He and Xiaoping Shen. A ranked subspace learning method for gene expression data classification. In *IC-AI*, pages 358–364, 2007.

[18] Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449, 2002.

[19] Francisco Louzada, Paulo H Ferreira-Silva, and Carlos AR Diniz. On the impact of disproportional samples in credit scoring models: An application to a brazilian bank data. *Expert Systems with Applications*, 39(9):8071–8078, 2012.

[20] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[21] Andrea Dal Pozzolo, Olivier Caelen, Reid A Johnson, and Gianluca Bontempi. Calibrating probability with undersampling for unbalanced classification. In *Computational Intelligence, 2015 IEEE Symposium Series on*, pages 159–166. IEEE, 2015.

[22] Georgios Douzas and Fernando Bacao. Effective data generation for imbalanced learning using conditional generative adversarial networks. *Expert Systems with applications*, 91:464–471, 2018.

[23] Jie Sun, Jie Lang, Hamido Fujita, and Hui Li. Imbalanced enterprise credit evaluation with dte-sbd: Decision tree ensemble based on smote and bagging with differentiated sampling rates. *Information Sciences*, 425:76–91, 2018.

[24] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[25] Hoo-Chang Shin, Neil A Tenenholtz, Jameson K Rogers, Christopher G Schwarz, Matthew L Senjem, Jeffrey L Gunter, Katherine P Andriole, and Mark Michalski. Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 1–11. Springer, 2018.

[26] Wen-tau Yih, Xiaodong He, and Christopher Meek. Semantic parsing for single-relation question answering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 643–648, 2014.

[27] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[28] Chunrong Ai and Edward C Norton. Interaction terms in logit and probit models. *Economics letters*, 80(1):123–129, 2003.

[29] Asoke Basu. *Elementary statistical theory in sociology*, volume 12. Brill Archive, 1976.

[30] Manohar Narhar Vartak et al. On an application of kronecker product of matrices to statistical designs. *The Annals of Mathematical Statistics*, 26(3):420–438, 1955.

[31] Steffen Rendle. Factorization machines. In *IEEE International Conference on Data Mining*, 2011.

[32] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[33] Charles E Metz. Basic principles of roc analysis. In *Seminars in nuclear medicine*, volume 8, pages 283–298. Elsevier, 1978.

[34] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[35] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[36] Kevin Roth, Aurelien Lucchi, Sebastian Nowozin, and Thomas Hofmann. Stabilizing training of generative adversarial networks through regularization. In *Advances in Neural Information Processing Systems*, pages 2018–2028, 2017.

[37] Qing Li, Yuanzhu Chen, Li Ling Jiang, Ping Li, and Hsinchun Chen. A tensor-based information framework for predicting the stock market. *ACM Transactions on Information Systems (TOIS)*, 34(2):11, 2016.