DOMFN: A Divergence-Orientated Multi-Modal Fusion Network for Resume Assessment

Yang Yang[†]
Nanjing University of Science and Technology,
MIIT Key Lab. of Pattern Analysis and Machine Intelligence, NUAA
State Key Lab. for Novel Software Technology, NJU

Jingshuai Zhang[†]
Baidu Talent Intelligence Center, Baidu Inc

Fan Gao
Tokyo Institute of Technology

Xiaoru Gao
Rutgers University

Hengshu Zhu[*]
Baidu Talent Intelligence Center, Baidu Inc

ABSTRACT

In talent management, resume assessment aims to analyze the quality of a job seeker's resume, which can assist recruiters to discover suitable candidates and benefit job seekers improving resume quality in return. Recent machine learning based methods on large-scale public resume datasets have provided the opportunity for automatic assessment for reducing manual costs. However, most existing approaches are still content-dominated and ignore other valuable information. Inspired by practical resume evaluations that consider both the content and layout, we construct the multi-modalities from resumes but face a new challenge that sometimes the performance of multi-modal fusion is even worse than the best uni-modality. In this paper, we experimentally find that this phenomenon is due to the cross-modal divergence. Therefore, we need to consider *when is it appropriate to perform multi-modal fusion?* To address this problem, we design an instance-aware fusion method, i.e., Divergence-Orientated Multi-Modal Fusion Network (DOMFN), which can adaptively fuse the uni-modal predictions and multi-modal prediction based on cross-modal divergence. Specifically, DOMFN computes a functional penalty score to measure the divergence of cross-modal predictions. Then, the learned divergence can be used to decide whether to conduct multi-modal fusion and be adopted into an amended loss for reliable training. Consequently, DOMFN rejects multi-modal prediction when the cross-modal divergence is too large, avoiding the overall performance degradation, so as to achieve better performance than uni-modalities. In experiments, qualitative comparison with baselines on real-world dataset demonstrates the superiority and explainability of the proposed DOMFN, e.g., we find a meaningful phenomenon that multi-modal fusion has positive effects for assessing resumes from UI Designer and Enterprise Service positions, whereas affects the assessment of Technology and Product Operation positions.

CCS CONCEPTS

• **Computing methodologies → Supervised learning**.

KEYWORDS

Multi-Modal Learning, Instance-Aware Fusion, Resume Assessment

ACM Reference Format:
Yang Yang[†], Jingshuai Zhang[†], Fan Gao, Xiaoru Gao, and Hengshu Zhu. 2022. DOMFN: A Divergence-Orientated Multi-Modal Fusion Network for Resume Assessment. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisbon, Portugal.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3548203

1 INTRODUCTION

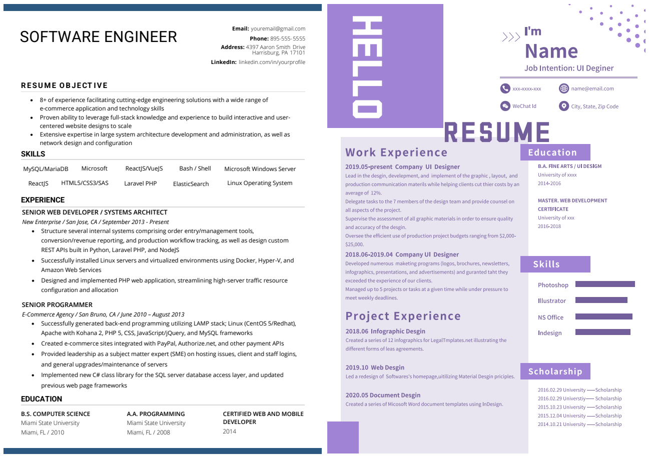
With the development of Internet, online recruitment, where employers and job seekers can interact conveniently and efficiently[8], is gaining popularity in recent talent management. In professional recruiting platforms, recruiters always pre-screen potential hires based on job seekers' resumes. Therefore, a high-quality resume can help job seekers stand out, which leads the emergence of resume assessment service as the demands increase. In detail, third-party organizations conduct assessment of the job seeker's resume to assist improving quality. For example, Ladders website [1] provides a convenient platform for recruiters and job seekers, and can polish user-uploaded resumes. However, with the rapid expansion of talent market, e.g., LinkedIn [2] receives 77 job applications every second and 49 million user access each week, it is time-consuming and burdensome to evaluate high volume of resumes. Obviously, effective techniques are required for intelligent resume assessment, which aims to alleviate the burden of manual cost and offer faster feedback to job seekers.

Considering resume content is often presented in text form, we naturally refer to the natural language processing (NLP) techniques to automatically assess the resumes. Early NLP approaches mainly

[*] is the corresponding author. [†] Both authors contributed equally.

MM '22, October 10–14, 2022, Lisbon, Portugal.
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548203

[1] https://www.theladders.com/about-us
[2] https://news.linkedin.com



SOFTWARE ENGINEER

Email: youremail@gmail.com
Phone: 895-555-5555
Address: 4397 Aaron Smith Drive
Harrisburg, PA 17101
LinkedIn: linkedin.com/in/yourprofile

RESUME OBJECTIVE

- 8+ of experience facilitating cutting-edge engineering solutions with a wide range of e-commerce application and technology skills
- Proven ability to leverage full-stack knowledge and experience to build interactive and user-centered website designs to scale
- Extensive expertise in large system architecture development and administration, as well as network design and configuration

SKILLS

MySQL/MariaDB	Microsoft	ReactJS/VueJS	Bash / Shell	Microsoft Windows Server
ReactJS	HTML5/CSS3/SAS	Laravel PHP	ElasticSearch	Linux Operating System

EXPERIENCE

SENIOR WEB DEVELOPER / SYSTEMS ARCHITECT
New Enterprise / San Jose, CA / September 2013 - Present

- Structure several internal systems comprising order entry/management tools, conversion/revenue reporting, and production workflow tracking, as well as design custom REST APIs built in Python, Laravel PHP, and NodeJS
- Successfully installed Linux servers and virtualized environments using Docker, Hyper-V, and Amazon Web Services
- Designed and implemented PHP web application, streamlining high-server traffic resource configuration and allocation

SENIOR PROGRAMMER
E-Commerce Agency / San Bruno, CA / June 2010 – August 2013

- Successfully generated back-end programming utilizing LAMP stack; Linux (CentOS 5/Redhat), Apache with Kohana 2, PHP 5, CSS, JavaScript/jQuery, and MySQL frameworks
- Created e-commerce sites integrated with PayPal, Authorize.net, and other payment APIs
- Provided leadership as a subject matter expert (SME) on hosting issues, client and staff logins, and general upgrades/maintenance of servers
- Implemented new C# class library for the SQL server database access layer, and updated previous web page frameworks

EDUCATION

B.S. COMPUTER SCIENCE Miami State University Miami, FL / 2010	A.A. PROGRAMMING Miami State University Miami, FL / 2008	CERTIFIED WEB AND MOBILE DEVELOPER 2014

(*a*) Software Engineer

HELLO

>>> I'm
Name
Job Intention: UI Deginer

xxx-xxxx-xxx
name@email.com
WeChat Id
City, State, Zip Code

RESUME

Work Experience

2019.05-present Company UI Designer
Lead in the desgin, develepment, and implement of the graphic , layout, and production communication materils while helping clients cut thier costs by an average of 12%.
Delegate tasks to the 7 members of the design team and provide counsel on all aspects of the project.
Supervise the assessment of all graphic materials in order to ensure quality and accuracy of the desgin.
Oversee the efficient use of production project budgets ranging from $2,000-$25,000.

2018.06-2019.04 Company UI Designer
Developed numerous maketing programs (logos, brochures, newsletters, infographics, presentations, and advertisements) and guranted taht they exceeded the experience of our clients.
Managed up to 5 projects or tasks at a given time while under pressure to meet weekly deadlines.

Project Experience

2018.06 Infographic Desgin
Created a series of 12 infographics for LegalTmplates.net illustrating the different forms of leas agreements.

2019.10 Web Desgin
Led a redesign of Softwares's homepage,uitilizing Material Desgin priciples.

2020.05 Document Desgin
Created a series of Micosoft Word document templates using InDesign.

Education

B.A. FINE ARTS / UI DESIGM
University of xxxx
2014-2016

MASTER. WEB DEVELOPMENT CERTIFICATE
University of xxx
2016-2018

Skills

Photoshop
Illustrator
NS Office
Indesign

Scholarship

2016.02.29 University ——Scholarship
2016.02.29 Universtiy—— Scholarship
2015.10.23 University ——Scholarship
2015.12.04 University ——Scholarship
2014.10.21 University ——Scholarship

(*b*) User Interface (UI) Designer

Figure 1: (Best view in color.) Resume examples of different positions. In addition to the content, the layout of resume can also reflect the skills of job seekers, especially for positions that require related techniques like UI Designer.

focus on learning to represent words into vectors in low continuous vector space, and then achieve global representation by fusing the word embeddings for analysis [24]. With the development of deep learning, deep natural language models (e.g., LSTM, BERT) [5, 6] are designed to directly model global representation, and have demonstrated excellent performance. Along this line, several attempts [19, 23, 26, 38] are proposed to employ state-of-the-art deep networks to analyze the resumes as a classification task (i.e., binary classification considering whether it is qualified or not). Note that almost all existing approaches are content-dominated, which develop models with content information only. However, in practical manual assessment, experts will evaluate the resumes from different perspectives, including the content and layout [1]. Using Figure 1 as an example, software engineer presents capability with direct and monochromatic layout while UI Designer applies colorful and elaborate layout that reflects design abilities. Therefore, although texutal information is critical for the evaluation of resume quality, we should not ignore the contribution of visual information in the assessment.

To integrate information from multiple modalities, multi-modal fusion is one of the highly researched aspects in multi-modal machine learning [40–43]. Most existing fusion approaches have been developed using model-agnostic methods [2], which includes early (i.e., feature-based) fusion, late (i.e., decision-based) fusion and hybrid fusion. In detail, early fusion integrates features with designed operator after they are extracted, and late fusion performs integration after each modality has made a decision. At present, each modality also uses the corresponding deep networks, e.g., the Transformer [32] for text modality, ResNet [13] for image modality, to replace the original linear models as backbones. However, several recent studies have reported unsatisfactory performance of multi-modal DNNs in various tasks [10, 14, 34], that the performance of multi-modal fusion is inferior to single modality. In fact, this phenomenon also exists when using traditional multi-modal fusion for resume assessment, i.e., the best uni-modal model often outperforms the joint model (Table 1 in Experiments). Upon inspection, the problem appears to be modal divergence, i.e., cross-modal information of partially instances is mismatch, and even has negative effects on fusion. These findings compel us to ask, *when is it appropriate to perform multi-modal fusion?* In other words, the gap between text and image predictions stimulates us to process instance-aware multi-modal fusion.

Therefore, in this paper, we put forward the *differentiated multi-modal fusion* for resume assessment, and propose the Divergence-Orientated Multi-Modal Fusion Network (DOMFN), with the purpose of modeling instance-aware multi-modal fusion based on their divergence. In detail, DOMFN builds independent neural networks for textual and visual modalities, and computes a functional penalty score to measure the cross-modal divergence, which can decide whether to integrate multi-modal prediction or only fuse uni-modal predictions for final prediction. Meanwhile, DOMFN applies a novel amended loss for reliable training based on the learned divergence. In results, we evaluate DOMFN with a real-world dataset collecting from a recruitment website, the experimental results prove that our model achieves promising performance compared with both uni-modal and multi-modal networks.

2 RELATED WORK

Resume Assessment. Recently, the newly available big data in recruitment field allow researchers to conduct resume analysis through more quantitative ways [12, 26, 28], which largely depend on resume content using NLP techniques. Considering that the traditional word2vec-based [21] models always failed to capture the semantics with entire context, RNN [7] and LSTM [15] were applied to utilize the sequential information of a sentence/document [4, 16, 21]. Furthermore, the Pre-trained Language Models (PLMs) such as BERT [6] and GPT [27] achieved the state-of-art performance in global representation learning. For example, BERT extracted the relational features of words in sentence concurrently to reflect the global semantics [6]. Inspired by the deep NLP models, [31] used a hierarchical model to extract entities from the resume for prediction. [44] presented a text mining-based approach to extract resume semantic data and devised a set of visualizations to represent the semantic information. [23] addressed that the consistency of different parts in the resume would affect resume quality. However, the practical assessments usually consider multiple perspectives (e.g., content and layout design), rather than single content information.

Multi-Modal Fusion. Traditional multi-modal fusion methods can be categorized based on the stage in which fusion occurs, namely early, late and hybrid fusion [45]. Considering the types of modal interaction, early fusion includes simple operation-based and bilinear-based fusion. Simple operation-based fusion employs the concatenation/add/weighted methods to integrate different modal features [9, 17, 37, 45]. Though simple operation-based fusion is direct and has few parameters, it fails to explore complicated correlations between modalities. Therefore, bilinear pooling fusion is constructed to calculate the outer product over input modalities to promote the higher-dimensional interaction [45]. For example, [22] proposed to decomposes the weights into low-rank factors to compute tensor-based fusion. Moreover, hybrid fusion [2, 18] combines outputs from early fusion and individual uni-modal predictors. In contrast, late fusion refers to integrate multi-modal predictions. In addition to the simple add/mean/majority voting operator, recent approaches always adopt the attention mechanism. For example, [39] introduced an adaptive weighting method for multi-modal fusion. [25] built novel transformer based architecture for modal fusion at multiple layers using bottlenecks. However, [10, 14, 34] observed that the best uni-modal network often outperforms the multi-modal network. Therefore, [34] proposed to estimate the uni-modal generalization and overfitting speeds in order to calibrate the learning through loss re-weighing. [10] promoted the reliability and robustness by integrating evidence that explained the prediction of each modality. Nevertheless, these methods still do not give up on multi-modal fusion, and ignore an important issue that excessive divergence between modalities has side effects on the prediction of fused multi-modal representation.

3 PROPOSED METHOD

In this section, we will introduce our DOMFN for resume assessment. As shown in the Figure 2, the overall framework are composed of two parts: *Multi-Modal Feature Extraction* and *Instance-Aware Multi-Modal Fusion.* In detail, Multi-Modal Feature Extraction firstly processes the statistical and contextual information as textual features, and regards the layout design of the resume as visual features. Then, Instance-Aware Multi-Modal Fusion considers the divergence of cross-modal predictions and aims to distinguish which examples are suitable for multi-modal fusion, where we introduce two independent modal predictors and a trusted multi-modal fusion predictor for ensemble to acquire the final prediction.

3.1 Multi-Modal Feature Extraction

For each resume, it involves textual and visual modality. We exploit different neural networks to capture the features from each modality, which includes Textual Feature Extraction and Visual Feature Extraction.

Textual Features. Specifically, the textual modality consists of two kinds of information, namely statistical information (e.g., school, degree, major, etc) and context (i.e., work and project experiences). Statistical data is discontinuous and unordered words that express information from different perspectives. We summarize values of each description. For example, we divide the degree data in *Junior high school and below*, *High school*, *Associate*, *Bachelor*, *Master*, *Ph.D.*, *MBA*, *EMBA*, *MPA*. To relate and combine all statistical data, we map them into an one-hot vector which encodes categorical features as a numeric array based on the unique values. In detail, we have 16 (i.e, degree, major, school, working years, intended position, etc) categories of statistical data and finally get a 58-dimension vector as the statistical representation. In result, given the content T, we can extract the contained statistical feature as:

$$\mathbf{x}_a = One-hot(T) \tag{1}$$

where $\mathbf{x}_a \in \mathbb{R}^{d_a}$, $d_a = 58$ denotes the dimension.

In resume, the context information indicates sentences describing one's work and project experiences. Normally, these sentences are relatively independent and non-interfering, so that we don't concatenate them wholly to acquire global features but input single sentence into the extractor. Therefore, as shown in the Figure 2, we adopt the pre-trained BERT [6] as encoder. For each raw sentence s_i in content T, we add special symbol $[CLS]$ in the front, and $[SEP]$ in the end of sentence as [6] to guide the model generating contextualized representation. We use the $[CLS]$ embedding from the output of last layer to represent each sentence because it integrates the semantics of each word:

$$\mathbf{z}_i = BERT(s_i) \tag{2}$$

where $\mathbf{z}_i \in \mathbb{R}^{d_s}$ denotes the global feature, $d_s = 768$ represents the dimension. Assuming we have N experiences in the resume, we can get N representations after encoding. Consequently, we combine the statistical and textual features as:

$$\begin{aligned} \mathbf{x}_t &= \Phi_z(\mathbf{x}_z \oplus \mathbf{x}_a) \\ \mathbf{x}_z &= \mathbf{z}_1 \oplus \mathbf{z}_2 \oplus \cdots \oplus \mathbf{z}_N \end{aligned} \tag{3}$$

where the $\oplus$ denotes concatenation operator, and Φ_z is the learnable mapping layer, i.e., the fully connected network, which maps the concatenated features to d_s dimension. $\mathbf{x}_t \in \mathbb{R}^{d_s}$ can be regarded the extracted text modal representation.

Visual Feature. The visual information in the resume refers to the layout in each page and the attached design portfolio, both of which are in the format of images. Basically, these images of each page are not consecutive and contain individual information. In this case, we leverage the pre-trained ResNet [13] to encode each image. We take the output of penultimate layer as visual feature vector. Therefore, as shown in the Figure 2, Given $I = \{I_1, I_2, \cdots, I_M\}$, where M denotes the maximum number of pages in the resume, we collect a stacked visual feature matrix. Similarly, we incorporate a concatenation operator with a learnable mapping layer to get the final visual vector:

$$\begin{aligned} \mathbf{z}_m &= ResNet(I_m) \\ \mathbf{x}_v &= \Phi_v(\mathbf{z}_1 \oplus \mathbf{z}_2 \oplus \cdots \oplus \mathbf{z}_M) \end{aligned} \tag{4}$$

where $\mathbf{z}_m \in \mathbb{R}^{d_v}$ denotes the hidden representation, $\mathbb{R}^{d_v} = 2048$. Φ_v is the learnable mapping layer, i.e., the fully connected network, which maps the concatenated features to d_v dimension.

3.2 Instance-Aware Multi-Modal Fusion

Overview. This section introduces the process of modal fusion in our work. Specifically, we first build two independent fully connected neural networks for textual and visual modal prediction, aiming to perform prediction from single perspective. In order to distinguish whether different modalities observe the same phenomenon, we calculate a functional penalty score to quantify the divergence of cross-modal prediction results. Based on the learned divergence, we define a instance-aware fusion mechanism to connect the divergence and strategic modal fusion, i.e., if the functional penalty score is higher than the threshold, which means textual and visual modality don't share the common ground, we choose to reject the fusion. Otherwise, we believe the fused modality will promote the robustness. Lastly, the independent modal predictions and multi-modal prediction are merged for the final prediction.

Uni-modal and Multi-modal Prediction. We first incorporate individual predictors for each modality. As we have extracted

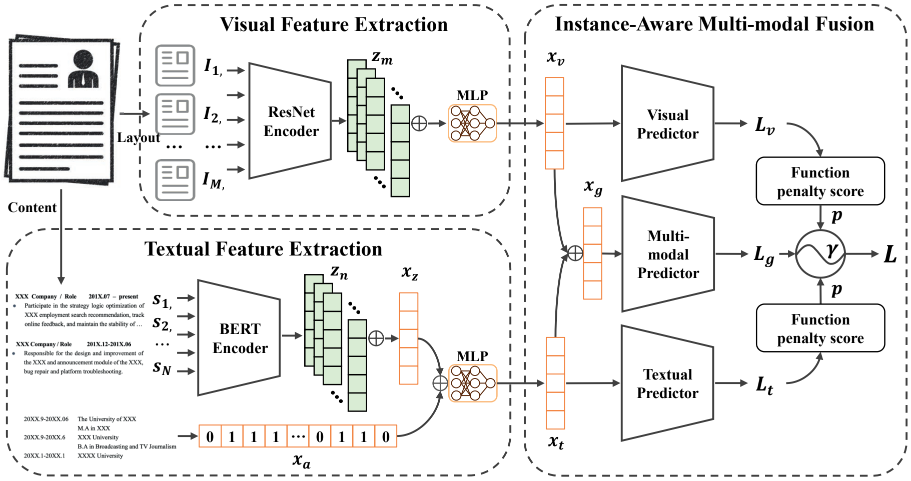

Visual Feature Extraction
Instance-Aware Multi-modal Fusion
Layout
Content
$I_{1,}$
$I_{2,}$
$I_{M,}$
ResNet Encoder
z_m
MLP
x_v
Visual Predictor
L_v
Function penalty score
x_g
Multi-modal Predictor
L_g
p
γ
L
Function penalty score
Textual Feature Extraction
XXX Company / Role 201X.07 – present
Participate in the strategy logic optimization of XXX employment search recommendation, track online feedback, and maintain the stability of ...
XXX Company / Role 201X.12-201X.06
Responsible for the design and improvement of the XXX and announcement module of the XXX, bug repair and platform troubleshooting.
$s_{1,}$
$s_{2,}$
s_N
BERT Encoder
z_n
x_z
MLP
20XX.9-20XX.06 The University of XXX M.A in XXX
20XX.9-20XX.6 XXX University B.A in Broadcasting and TV Journalism
20XX.1-20XX.1 XXXX University
0 1 1 1 ⋯ 0 1 1 0
x_a
x_t
Textual Predictor
L_t

Figure 2: Illustration of the proposed Divergence-Orientated Multi-Modal Fusion Network in Resume Assessment. Firstly, multi-modal feature extraction includes the text and visual feature extractions. Then, the instance-aware multi-modal fusion can adaptively fuse the uni-modal predictions and multi-modal prediction based on the learned cross-modal divergence.

textual and visual features using the well-explored pre-trained models, we use the fully connected networks to classify the text and image independently. The textual output y_t and visual output y_v are formulated as following:

$$y_t = f_t(\mathbf{x}_t) \quad y_v = f_v(\mathbf{x}_v) \tag{5}$$

where f_t, f_v denote the predictors with fully connected networks of text and image. Furthermore, in the process of multi-modal fusion, we choose to utilize early fusion technique (i.e, concatenation) to interact different modal features, and the predictor is also built with the fully neural network to predict the results. Accordingly, the fused modality and the predictor can be formulated as:

$$\mathbf{x}_g = \mathbf{x}_t \oplus \mathbf{x}_v \quad y_g = f_g(\mathbf{x}_g) \tag{6}$$

where $\mathbf{x}_g \in \mathbb{R}^{d_s+d_v}$ denotes the multi-modal fused features. f_g denotes the predictor with fully connected network for $\mathbf{x}_g$.

Overall Loss. In this part, we answer the question that when is it appropriate to perform multi-modal fusion. Obviously, different modalities from unrelated neural networks have divergence. Besides, textual and visual modalities focusing on complementary characteristics have different contributions to the final output, so the common practices that train the multi-modal models simultaneously without considering the errors from divergence are unfair. Therefore, we employ a novel amended loss function for uni-modal predictors.

In detail, the commonly adopted loss for textual and visual predictors is always defined as binary cross entropy:

$$L_o = -[y \log y_o + (1-y)\log(1-y_o)], \quad o \in \{t, v\} \tag{7}$$

where L_o represents the uni-modal loss of o-th modality, and y is the ground-truth label, $y = 1$ represents qualified and otherwise is unqualified. Meanwhile, the average value of uni-modal predictions is: $\bar{y} = \frac{1}{2}\sum_{o\in\{t,v\}} y_o$, where y_o denotes the uni-modal prediction.

Traditional approaches train multiple uni-modal predictors independently, which can be regarded as a ensemble process by considering the uni-modal predictor as independent classifier. However, this may not be optimal as demonstrated in [3] that the quadratic error of the ensemble estimator is guaranteed to be less than or equal to the average quadratic error of the component estimators, which advocates learning a set of both accurate and decorrelated models. Therefore, the ensemble error consists of both the individual error and the interactions (i.e., divergence) within ensemble, and we need to regularize the model correlation in ensemble to learn a divided and conquered approach. Based on this idea, in order to encourage cooperation of different modalities to generate reliable estimation, we are inspired by [20, 46] to employ negative correlation learning with a correlation penalty term into the error function of each individual network, so that all the networks can be trained simultaneously and interactively. In result, we reformulate the loss function of uni-modal $o \in \{t, v\}$-th predictor in Equation 7, which is defined as amended loss:

$$\begin{aligned} \widetilde{L}_o &= L_o + \eta p_o \\ p_o &= (y_o - \bar{y}) \sum_{k=\{t,v\}/o} (y_k - \bar{y}), \end{aligned} \tag{8}$$

where p_o is defined as functional penalty score, which represents the divergence between modalities in the decision level. Clearly, the loss function of each predictor is composed of the errors of individual modality and the divergence. The η is to adjust the influence of divergence on the loss function. On the other hand, take the average value of uni-modal predictions $\bar{y}$ into Equation 8, the functional penalty scores of each modality are the same:

$$p = -(\frac{y_t - y_v}{2})^2 \tag{9}$$

Therefore, Equation 8 explains that the model learns to negatively correlate the individual errors and the divergence. On the other hand, to support the intensive instance-aware multi-modal fusion, we set a fixed threshold γ to control the process of fusion.

Specifically, we claim that the appropriate case to merge modality is when the value of p is lower than γ, which means the textual and visual modality have the same observation, and the cross-modal divergence supports the fusion. Therefore, in this case we use the sum of all loss functions to train the overall model. Conversely, if the value of p is higher than γ, we believe that the multi-modal fusion is unnecessary because the information from different modalities is mutually exclusive, influencing the learning of multi-modal predictors. Then we only use the losses of individual modalities to train the encoder. In summary, this process can be formalized as:

$$L = \begin{cases} \widetilde{L}_t + \widetilde{L}_v + L_g, & p \leq \gamma \\ \widetilde{L}_t + \widetilde{L}_v, & p > \gamma \end{cases} \tag{10}$$

where L_g has the same loss function as Equation 7.

In addition to the training process, the threshold also plays an alternative role in the prediction selection with the comparison between p and γ. In test phase, the final output is undoubtedly the fused results of uni-modal and multi-modal predictions when p is lower than γ. In contrast, when the divergence of textual and visual modality rejects modal fusion, we employ the confidence of predictions from uni-modal classifier to select the final output. We hold the opinion that prediction is reliable when it has a high confidence. Therefore, the output layer is as follows:

$$\hat{y} = \begin{cases} \max\{y_t, y_v, y_g\}, & p \leq \gamma \\ \max\{y_t, y_v\}, & p > \gamma \end{cases} \tag{11}$$

where y_g indicates the output of multi-modal predictor. The detail process of training is shown in Algorithm 1.

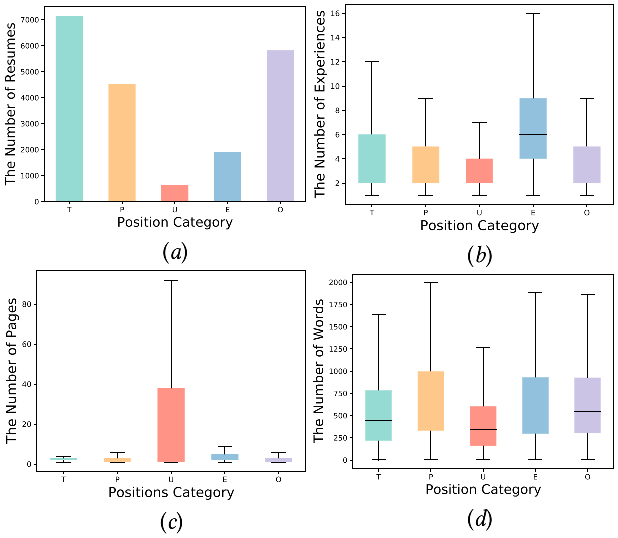
Algorithm 1 Training

```
Input: The resume set $\mathcal{D}$, threshold $\gamma$, the training epochs $E$.
Output: The updated encoders $f_t, f_v, f_g$
  for $e = 1, 2, \cdots, E$ do
    Randomly sample mini-batch from $\mathcal{D}$.
    Calculating the $y_t$ and $y_v$ according to Equation 5
    Calculating the $p$ according to Equation 9
    if $p < \gamma$ then
      $\widetilde{L}_t + \widetilde{L}_v + L_g \xrightarrow{update} f_t, f_v, f_g, \Phi_z, \Phi_v$
    else
      $\widetilde{L}_t \xrightarrow{update} f_t, \Phi_z$
      $\widetilde{L}_v \xrightarrow{update} f_v, \Phi_v$
    end if
  end for
```

4 EXPERIMENTS

4.1 Dataset Description

In this section, we will introduce the real-world resume dataset we used, which is in-firm data of talents from a high-tech company in China (note that all sensitive information in the dataset has been removed or anonymized). In order to remove the bias caused by name, age, and gender, all the personal information is not under consideration. In detail, we have 5 categories of resumes from different positions, including Technology, Product Operation, UI Designer, Enterprise Service, and Others, respectively represented by $\mathcal{T}, \mathcal{P}, \mathcal{U}, \mathcal{E}$, and $\mathcal{O}$. Note that the job contents of UI Designer and Enterprise Service include product design and promotion, which require higher photoshop and exhibition skills. Each resumes contains: candidates’ school, degree, working years, project & work experiences, layout design, etc. We label the resume that attracted recruiter’s attention and received interview invitation as the positive examples and otherwise negative. In order to enrich the textual information of the resume, we employed the resumes that have at least one project or work experience description. In result, we sampled 10,000 positive and 10,000 negative examples as full dataset for experiments. The dataset is split into 80% for training, 10% for validation, and remaining as testing data.



(a) (b) (c) (d)

Figure 3: (Best view in color.) Basic statistics. (a):The number of resumes applying for different positions. (b): The number distribution of experiences in a resume with regard to different positions. (c) The number distribution of pages in a resume with regard to different positions. (d) The number distribution of words describing a experience.

Furthermore, we present the basic statistics of the dataset in Figure 3. As shown in the Figure 3(a), different kinds of resumes have imbalance distribution in which technical resumes have the largest number. Figure 3(b) has displayed the number distribution of experiences based on different position categories. Besides, Compared with Figure 3(c) and (d), we can find that the UI design-oriented resumes tend to offer abundant design portfolio rather than write long text. In experiments, we will deeply analyze the effect of instance-aware multi-modal fusion on different positions.

4.2 Experimental Setup

4.2.1 Feature Processing. **Attribute Data.** As the collected resume is PDF format, we produce the text with PaddleOCR [33]. For each resume, we summarize the attributes in 5 aspects: school, degree, major, working years, positions. Then we grouped each attribute into different levels and map them into one-hot vectors. We finally get a 58-dimension vector to represent attribute features. **Contextual Data.** As for contextual information (project & work experiences), we consider the top 15 experiences because few resumes contains more than 15 slides of experiences according to Figure 3(b). Similarly, we set the maximum number of words describing a single experience as 300 and deleted the words beyond limitation according to Figure 3(d). We utilize the pre-trained 12-layer BERT to extract the 758-dimension contextual embedding for each experience. **Image Data.** We use the tool of pdf-to-img in Python to transfer every page of the resume into a picture. Analysing the number distribution of pages in Figure 3(c), we only study the top 10 images of every resume sample. To acquire the visual modal representation, we apply 50-layer Resnet to extract visual feature.

4.2.2 Implementations. In the experiments, textual and visual predictors f_t, f_v are both 4-layer fully connected networks while the predictor f_g for multi-modal fusion is a 2-layer fully connected network. The mapping functions Φ_v and Φ_t are two independent 2-layer fully connected networks. To speedup the training, we first pre-train the textual and visual encoder with the collecting training dataset by 15 epochs, the learning rate is set as 10^{-2} for both predictors. Then we fine-tune all the predictors with 75 epochs. We optimize with Adam Optimizer using learning rates 10^{-3} and a decay weight 1×10^{-2}. We performed the best hyper-parameters ($\eta = 0.5, \gamma = 0.8$) for training and testing. The model is trained on a single 12G Nvidia K-40 GPU. We will publish the code and dataset after the paper is accepted.

4.2.3 Evaluation Metrics. Identifying whether a resume is qualified to pass the screening is a binary classification problem. AUC is a significant and widely used metric that reflects model performance within different boundary values between classes[38]. Besides, we also exploite the standard evaluation scores of Accuracy, Recall, Precision, and Macro-F1 to evaluate the effectiveness.

4.3 Baseline Models

To validate the effectiveness of proposed DOMFN, we compare it with existing state-of-the-art uni-modal approaches and multi-modal fusion methods. Note that we also include the state-of-the-art reliable multi-modal fusion methods for comparison. In detail, the uni-modal approaches include LSTM [15] and BERT [6] for text modality, and VGG [30] and ResNet [13] for image modality. Moreover, we compare our model with multi-modal fusion methods, i.e., LMF [22], MIMN [36], OGR [34], MIMM [9], TMC [11], LF [29] and CMML [39]. Note that the OGR, TMC, and CMML are reliable multi-modal fusion methods considering modal imbalance. To study the superiority of our method in calculating the divergence of different modalities, we also evaluate the results of DP-Ensemble (i.e., an ensemble method) [35].

4.4 Experimental Results

Performance of Resume Assessment. We first conduct comprehensive experiments to evaluate performance of baseline models and our proposed DOMFN on resume assessment. The results are recorded in Table 1. We find that: 1) BERT performs significantly better than visual modality, which reveals that text performs major impact on the resume assessment task. On the other hand, the models developed by the image modality, i.e., VGG and ResNet are also weak predictors, which perform better than random prediction, so they can also provide auxiliary information. 2) Several multi-modal fusion methods, e.g., MIMN, and LF perform worse than the BERT on partial criteria, especially the AUC and Macro-F1. This

Table 1: The Overall Performance of Uni-modal and Multi-modal Methods. The best results are highlighted in bold.

Methods	AUC	Accuracy	Precision	Recall	Macro-F1
LSTM	0.837	0.756	0.759	0.756	0.756
BERT	0.842	0.745	0.746	0.845	0.744
VGG	0.656	0.621	0.626	0.621	0.618
ResNet	0.666	0.630	0.634	0.631	0.628
LMF	0.850	0.773	0.782	0.774	0.771
MIMN	0.830	0.739	0.740	0.739	0.739
MIMM	0.840	0.773	0.786	0.773	0.770
LF	0.825	0.744	0.748	0.745	0.743
OGR	0.850	0.770	0.796	0.771	0.765
TMC	0.853	0.766	0.785	0.767	0.763
CMML	0.849	0.776	0.786	0.777	0.774
DP-Ensemble	0.840	0.780	0.787	0.780	0.778
DOMFN	**0.863**	**0.785**	**0.796**	**0.786**	**0.783**

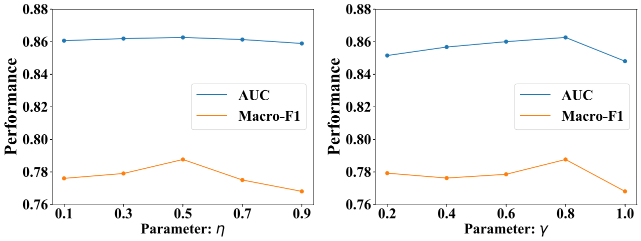
Table 2: Performance of comparison methods on different position categories. The best results are highlighted in bold.

Metrics	Methods	$\mathcal{T}$	$\mathcal{P}$	$\mathcal{U}$	$\mathcal{E}$	$\mathcal{O}$
AUC	BERT	0.721	0.851	0.882	0.606	0.882
	ResNet	0.539	0.68	0.792	0.655	0.639
	TMC	0.754	0.867	0.898	0.647	0.864
	OGR	0.743	0.855	0.881	0.610	0.882
	CMML	0.745	0.826	0.785	0.701	0.857
	DOMFN	**0.760**	**0.879**	**0.903**	**0.703**	**0.882**
Accuracy	BERT	0.75	0.739	0.631	0.572	0.809
	ResNet	0.635	0.631	0.736	0.611	0.619
	TMC	0.773	0.774	0.807	0.594	0.801
	OGR	**0.785**	0.787	0.842	0.561	0.795
	CMML	0.758	0.739	0.701	0.672	0.806
	DOMFN	0.779	**0.796**	**0.859**	**0.677**	**0.809**
Macro-F1	BERT	0.676	0.738	0.607	0.524	0.704
	ResNet	0.550	0.631	0.736	0.609	0.574
	TMC	0.638	0.770	0.806	0.594	0.751
	OGR	0.671	0.784	0.840	0.547	0.755
	CMML	0.658	0.739	0.698	0.672	0.749
	DOMFN	**0.677**	**0.794**	**0.857**	**0.677**	**0.755**

Table 3: Ablation Study. The best results are highlighted.

Model	AUC	Accuracy	Precision	Recall	Macro-F1
DOMFN	**0.863**	**0.790**	**0.802**	**0.793**	**0.788**
w/o $\widetilde{L}_t$	0.852	0.772	0.773	0.772	0.772
w/o $\widetilde{L}_v$	0.852	0.774	0.776	0.774	0.774
w/o $\widetilde{L}_t\widetilde{L}_v$	0.844	0.753	0.754	0.752	0.751
w/o instance	0.860	0.782	0.800	0.782	0.778
w/o all	0.825	0.744	0. 748	0.744	0.743

phenomenon indicates that directly complex multi-modal fusion techniques may fail to promote the results without considering the cross-modal gap. 3) OGR, TMC and CMML can improve the prediction performance and perform better than other multi-modal fusion approaches on most criteria, which reveals the effectiveness of robust fusion by considering cross-modal imbalance. 4) DOMFN outperforms all the comparison baselines under the metrics of AUC, Accuracy, Recall, Precision and Macro-F1. This phenomenon indicates that multi-modal fusion is not always useful, so we need to carefully consider when it is appropriate to conduct the fusion.



(a) Influence of η (b) Influence of γ

Figure 4: Prediction performance vs. hyper-parameters.

Performance on Different Positions. In order to identify the improvements of DOMFN in different positions, we further conduct more experiments. The results are shown in Table 2, note that considering the effectiveness, we only exhibit the results of best uni-modal and multi-modal fusion approaches. We find that: 1) In positions of Technology, Product Operation, Enterprise Service and Others, the existing state-of-the-art uni-modal and multi-modal fusion approaches perform worse than DOMFN on most criteria. The reason is that textual information contributes more than visual information to the prediction with a large margin. Therefore, we don't always need multi-modal fusion for these positions with high probability. 2) In UI Designer and Enterprise Service positions, existing multi-modal fusion approaches have a promoting effect, for the reason that visual information plays an important role in assessing. 3) DOMFN performs consistently better for various positions on most criteria, except Accuracy on Technology position. For example, DOMFN obtains remarkable improvements in Product Operation, UI Designer and Enterprise Service resume assessment. Note that the textual and multi-modal results in Technology are almost the same, so we conclude that DOMFN tends to reject multi-modal fusion and choose to use textual information in this position. Therefore, DOMFN can adaptively learn the distinction of different resume positions and adjust the fusion strategies accordingly by considering the cross-modal divergence.

Furthermore, the average functional penalty scores of different positions are 0.825/0.753/0.367/0.564/0.858 for $\mathcal{T}$, $\mathcal{P}$, $\mathcal{U}$, $\mathcal{E}$, and $\mathcal{O}$, the results validate that the multi-modal fusion is important for positions (e.g., $\mathcal{U}$ and $\mathcal{E}$ positions) with small functional penalty scores, whereas having little impact on positions with large functional penalty scores. This is in line with the realistic evaluation because UI Designer (i.e. $\mathcal{U}$) and Enterprise Service (i.e. $\mathcal{E}$) positions do require a comprehensive evaluation of resume content and layout, while other positions mainly consider content.

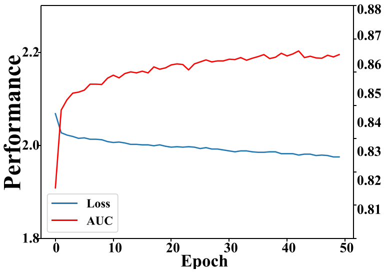


Figure 5: Objective function value vs. number of iterations.

4.5 Ablation Study

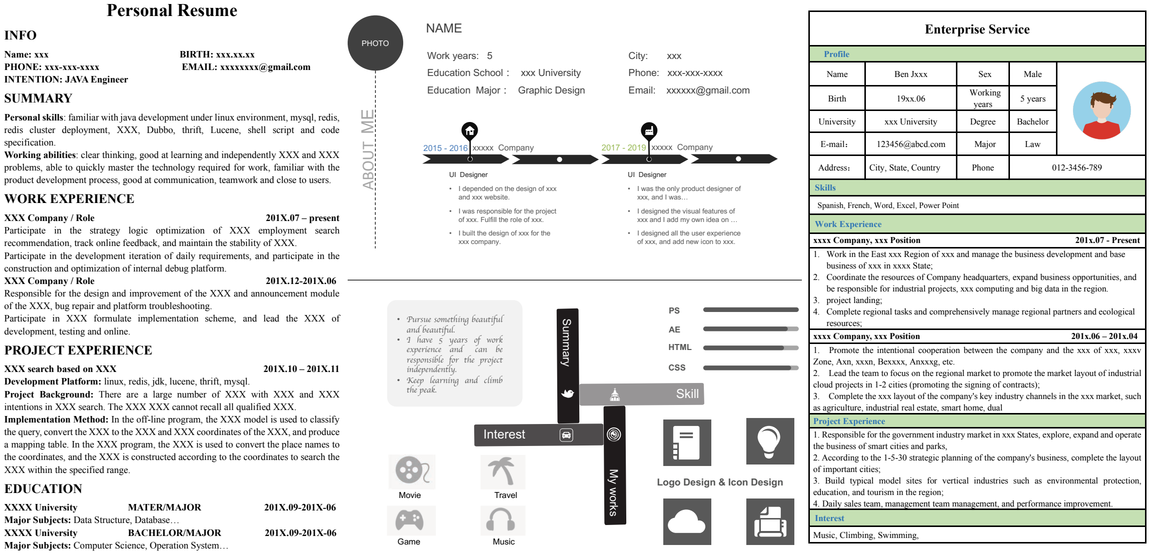
We conduct more ablation studies to verify the effectiveness of proposed DOMFN method: 1) w/o $\widetilde{L}_t$, we replace amended loss with the original loss function as Equation 7 for textual modality. 2) w/o $\widetilde{L}_v$, we replace amended loss with the original loss function as Equation 7 for visual modality. 3) w/o $\widetilde{L}_t\widetilde{L}_v$, we replace amended loss with the original loss function as Equation 7 for two modalities. 4) w/o instance, we remove the pre-defined threshold γ in instance-aware multi-modal fusion. 5) w/o all, we remove the amended loss and functional penalty score. Table 3 records the results, we find that: 1) DOMFN performs better than the w/o $\widetilde{L}_t$, w/o $\widetilde{L}_v$, and w/o $\widetilde{L}_t, \widetilde{L}_v$ on various criteria, and w/o $\widetilde{L}_t$, w/o $\widetilde{L}_v$ perform better than w/o $\widetilde{L}_t, \widetilde{L}_v$ on various criteria. This phenomenon validates the effectiveness of amended loss for prediction. 2) DOMFN performs better than w/o instance, which verifies that the instance-aware multi-modal fusion can effectively eliminate the interference of examples with large cross-modal divergence. DOMFN performs better than w/o all, which verifies the importance of functional penalty score on multi-modal fusion.

4.6 Parameter Sensitivity

The main parameters include the η in Equation 8 and γ in Equation 10. To explore the sensitivity of these parameters, we vary the parameter η in $\{0.1, 0.3, 0.5, 0.7, 0.9\}$, and γ in $\{0.2, 0.4, 0.6, 0.8, 1\}$ and record the results in Figure 4. The illustrations in figure verifies that DOMFN achieves the best performance when the functional penalty score p is 0.8, which reveals that the multi-modal fusion is not recommended when the divergence between two modalities is too large. Meanwhile, the DOMFN achieves best performance when η is 0.5, which indicates that the penalty score is important for improving performance, but too large penalty score will affect the optimization of classification loss.

4.7 Convergence Analysis

To investigate the convergence of DOMFN empirically. We record the objective function value and the classification AUC in each iteration (i.e., each epoch). The Figure 5 records results of DOMFN. It clearly reveals that the objective function value decreases as the iterations increase, and the classification performance becomes stable after several iterations. Moreover, these additional experiment results indicate that our DOMFN can converge fast, i.e., DOMFN converges after 20 epochs.



Personal Resume

INFO

Name: xxx **BIRTH:** xxx.xx.xx
PHONE: xxx-xxx-xxxx **EMAIL:** xxxxxxxx@gmail.com
INTENTION: JAVA Engineer

SUMMARY

Personal skills: familiar with java development under linux environment, mysql, redis, redis cluster deployment, XXX, Dubbo, thrift, Lucene, shell script and code specification.
Working abilities: clear thinking, good at learning and independently XXX and XXX problems, able to quickly master the technology required for work, familiar with the product development process, good at communication, teamwork and close to users.

WORK EXPERIENCE

XXX Company / Role **201X.07 – present**
Participate in the strategy logic optimization of XXX employment search recommendation, track online feedback, and maintain the stability of XXX.
Participate in the development iteration of daily requirements, and participate in the construction and optimization of internal debug platform.
XXX Company / Role **201X.12-201X.06**
Responsible for the design and improvement of the XXX and announcement module of the XXX, bug repair and platform troubleshooting.
Participate in XXX formulate implementation scheme, and lead the XXX of development, testing and online.

PROJECT EXPERIENCE

XXX search based on XXX **201X.10 – 201X.11**
Development Platform: linux, redis, jdk, lucene, thrift, mysql.
Project Background: There are a large number of XXX with XXX and XXX intentions in XXX search. The XXX XXX cannot recall all qualified XXX.
Implementation Method: In the off-line program, the XXX model is used to classify the query, convert the XXX to the XXX and XXX coordinates of the XXX, and produce a mapping table. In the XXX program, the XXX is used to convert the place names to the coordinates, and the XXX is constructed according to the coordinates to search the XXX within the specified range.

EDUCATION

XXXX University **MATER/MAJOR** **201X.09-201X-06**
Major Subjects: Data Structure, Database…
XXXX University **BACHELOR/MAJOR** **201X.09-201X-06**
Major Subjects: Computer Science, Operation System…

$y = 1, \hat{y} = 0.90$
$y_{w/o} = 0.19, p = 0.88$
(a) resume of position $\mathcal{T}$

PHOTO

ABOUT ME

NAME

Work years: 5 City: xxx
Education School : xxx University Phone: xxx-xxx-xxxx
Education Major : Graphic Design Email: xxxxxx@gmail.com

2015 - 2016 xxxxx Company

UI Designer

- I depended on the design of xxx and xxx website.
- I was responsible for the project of xxx. Fulfill the role of xxx.
- I built the design of xxx for the xxx company.

2017 - 2019 xxxxx Company

UI Designer

- I was the only product designer of xxx, and I was…
- I designed the visual features of xxx and I add my own idea on …
- I designed all the user experience of xxx, and add new icon to xxx.

Summary

- *Pursue something beautiful and beautiful.*
- *I have 5 years of work experience and can be responsible for the project independently.*
- *Keep learning and climb the peak.*

Skill

PS
AE
HTML
CSS

Interest

Movie
Travel
Game
Music

My works

Logo Design & Icon Design

$y = 1, \hat{y} = 0.96$
$y_B = 0.53, y_R = 0.86, p = 0.12$
(b) resume of position $\mathcal{U}$

Enterprise Service

Profile

Name	Ben Jxxx	Sex	Male	
Birth	19xx.06	Working years	5 years	
University	xxx University	Degree	Bachelor	
E-mail:	123456@abcd.com	Major	Law	
Address:	City, State, Country	Phone	012-3456-789	

Skills

Spanish, French, Word, Excel, Power Point

Work Experience

xxxx Company, xxx Position **201x.07 - Present**

1. Work in the East xxx Region of xxx and manage the business development and base business of xxx in xxxx State;
2. Coordinate the resources of Company headquarters, expand business opportunities, and be responsible for industrial projects, xxx computing and big data in the region.
3. project landing;
4. Complete regional tasks and comprehensively manage regional partners and ecological resources;

xxxx Company, xxx Position **201x.06 – 201x.04**

1. Promote the intentional cooperation between the company and the xxx of xxx, xxxv Zone, Axn, xxxn, Bexxxx, Anxxxg, etc.
2. Lead the team to focus on the regional market to promote the market layout of industrial cloud projects in 1-2 cities (promoting the signing of contracts);
3. Complete the xxx layout of the company's key industry channels in the xxx market, such as agriculture, industrial real estate, smart home, dual

Project Experience

1. Responsible for the government industry market in xxx States, explore, expand and operate the business of smart cities and parks,
2. According to the 1-5-30 strategic planning of the company's business, complete the layout of important cities;
3. Build typical model sites for vertical industries such as environmental protection, education, and tourism in the region;
4. Daily sales team, management team management, and performance improvement.

Interest

Music, Climbing, Swimming,

$y = 0, \hat{y} = 0.03$
$y_B = 0.38, y_R = 0.15, p = 0.06$
(c) resume of position $\mathcal{E}$

Figure 6: (Best view in color.) Three resume examples. To protect the privacy of job-seekers, we delete all personal information and mask some sensitive words.

4.8 Case Study

To explore the contribution of multi-modal fusion and instance-aware selection, we present several visualization cases for illustration. In Figure 6, (a) exhibits the resume example of Technology position to exhibit the importance of instance-aware selection, (b) and (c) present the superior of multi-modal fusion on two related positions especially appreciating layout, i.e., UI Designer and Enterprise Service positions. Specifically, for Figure 6 (a), the ground-truth $y = 1$, the prediction score of DOMFN, and w/o instance are 0.9 and 0.19 respectively, and the functional penalty score $p = 0.88$. The results reveal that w/o instance (i.e., direct fusing uni-modal predictions and multi-modal prediction without rejection) achieves error prediction, for the reason that the cross-modal divergence is large, i.e., $p > \gamma$, so that the multi-modal fusion provides negative effect for prediction. In contrast, DOMFN can predict correctly by discarding the multi-modal fusion with pre-defined γ.

Moreover, for Figure 6 (b), the ground-truth $y = 1$, the prediction score of DOMFN, BERT, Resnet are 0.96, 0.53, and 0.86 respectively, and the penalty function score $p = 0.12$. For Figure 6 (c), the ground-truth $y = 0$, the prediction score of DOMFN, BERT, and Resnet are 0.03, 0.38, and 0.15 respectively, and the penalty function score $p = 0.06$. From these two cases, we find that DOMFN can achieve more confident prediction than single textual modality, as the visual modality can act as auxiliary information for prediction. This phenomena validates that multi-modal information can actually promote the prediction. Meanwhile, after the expert assessment, we also confirm that layout of Enterprise Service resume in Figure 6 (c) is too monotonous and has no new conception, whereas layout of UI Designer resume in Figure 6 (b) highlights its professionalism in design. In result, we can conclude that it is necessary to conduct instance-aware multi-modal fusion for acquiring more reliable predictions in resume assessment task.

5 CONCLUSION

In this study, we focused on the intelligent resume assessment, and improved the traditional content-dominated classification into multi-modal classification by adding the layout as visual modality. Different form existing multi-modal fusion approaches that always incorporate all modalities, we believe that the cross-modal divergence will affect the multi-modal fusion, and proper multi-modal fusion is needed. Therefore, we designed a novel Divergence-Orientated Multi-Modal Fusion Network (DOMFN) with instance-aware fusion. In detail, DOMFN computed a functional penalty score to measure the divergence between textual and visual modal predictions, and applied a novel training process with an amended loss for reliable multi-modal fusion based on the learned divergence, which can reject multi-modal prediction when it is dangerous. Qualitative comparison with baselines on real-world datasets demonstrated the superiority and explainability of DOMFN.

ACKNOWLEDGMENT

This research was supported by NSFC (62006118, 61906092, 61836013), Natural Science Foundation of Jiangsu Province of China under Grant (BK20200460, BK20190441). Jiangsu Shuangchuang (Mass Innovation and Entrepreneurship) Talent Program. Young Elite Scientists Sponsorship Program by CAST, and CAAI-Huawei MindSpore Open Fund (CAAIXSJLJJ-2021-014B).

REFERENCES

[1] Jan Ketil Arnulf, Lisa Tegner, and Øyunn Larssen. 2010. Impression making by résumé layout: Its impact on the probability of being shortlisted. *European Journal of Work and Organizational Psychology* 19, 2 (2010), 221–230.

[2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2019), 423–443.

[3] Gavin Brown, Jeremy L. Wyatt, and Peter Tiňo. 2005. Managing Diversity in Regression Ensembles. *J. Mach. Learn. Res.* 6 (2005), 1621–1650.

[4] Kyunghyun Cho, B van Merrienboer, Caglar Gulcehre, F Bougares, H Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Doha, Qatar, 1724–1734.

[5] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2019. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *Proceedings of the International Conference on Learning Representations*. Addis Ababa, Ethiopia.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*. Minneapolis, 4171–4186.

[7] Jeffrey L. Elman. 1990. Finding Structure in Time. *Cognitive Science* 14, 2 (1990), 179–211.

[8] Ben Greiner. 2004. An online recruitment system for economic experiments. (2004).

[9] Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Punta Cana, Dominican, 9180–9192.

[10] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. 2021. Trusted Multi-View Classification. In *Proceedings of the International Conference on Learning Representations*. Virtual Event.

[11] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. 2021. Trusted Multi-View Classification. In *Proceedings of the International Conference on Learning Representations*. Austria.

[12] Christopher G Harris. 2017. Finding the best job applicants for a job posting: A comparison of human resources search strategies. In *2017 IEEE International Conference on Data Mining Workshops*. IEEE, New Orleans, 189–194.

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

[14] Jack Hessel and Lillian Lee. 2020. Does my multimodal model learn cross-modal interactions? It’s harder to tell than you might think!. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Virtual Event, 861–877.

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

[16] Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991* (2015).

[17] Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L Iuzzolino, and Kazuhito Koishida. 2020. MMTM: Multimodal transfer module for CNN fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle, 13289–13299.

[18] Zhen-zhong Lan, Lei Bao, Shoou-I Yu, Wei Liu, and Alexander G Hauptmann. 2014. Multimedia classification and event detection using double fusion. *Multimedia tools and applications* 71, 1 (2014), 333–347.

[19] Hao Lin, Hengshu Zhu, Yuan Zuo, Chen Zhu, Junjie Wu, and Hui Xiong. 2017. Collaborative Company Profiling: Insights from an Employee’s Perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*. San Francisco, California, 1417–1423.

[20] Yong Liu and Xin Yao. 1999. Ensemble learning via negative correlation. *Neural networks* 12, 10 (1999), 1399–1404.

[21] Zhiyuan Liu, Yankai Lin, and Maosong Sun. 2020. *Representation learning for natural language processing*. Springer Nature.

[22] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient Low-rank Multimodal Fusion With Modality-Specific Factors. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia, 2247–2256.

[23] Yong Luo, Huaizheng Zhang, Yongjie Wang, Yonggang Wen, and Xinwen Zhang. 2018. ResumeNet: A learning-based framework for automatic resume quality assessment. In *Proceedings of the IEEE International Conference on Data Mining*. Singapore, 307–316.

[24] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* 26 (2013).

[25] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. 2021. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems* 34 (2021).

[26] Chuan Qin, Hengshu Zhu, Tong Xu, Chen Zhu, Chao Ma, Enhong Chen, and Hui Xiong. 2020. An Enhanced Neural Network Approach to Person-Job Fit in Talent Recruitment. *ACM Trans. Inf. Syst.* 38, 2 (2020), 15:1–15:33.

[27] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. (2018).

[28] Dazhong Shen, Hengshu Zhu, Chen Zhu, Tong Xu, Chao Ma, and Hui Xiong. 2018. A Joint Learning Approach to Intelligent Job Interview Assessment. In *Proceedings of the International Joint Conference on Artificial Intelligence*. Stockholm, Sweden, 3542–3548.

[29] Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*. San Diego, 160–170.

[30] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. , 14 pages.

[31] Amit Singh, Catherine Rose, Karthik Visweswariah, Vijil Chenthamarakshan, and Nandakishore Kambhatla. 2010. PROSPECT: a system for screening candidates for recruitment. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. Toronto, Ontario, Canada, 659–668.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017), 5998–6008.

[33] Pengfei Wang, Chengquan Zhang, Fei Qi, Shanshan Liu, Xiaoqiang Zhang, Pengyuan Lyu, Junyu Han, Jingtuo Liu, Errui Ding, and Guangming Shi. 2021. PGNet: Real-time Arbitrarily-Shaped Text Spotting with Point Gathering Network. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Virtual Event, 2782–2790.

[34] Weiyao Wang, Du Tran, and Matt Feiszli. 2020. What makes training multimodal classification networks hard?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle, 12695–12705.

[35] Yanzhao Wu, Ling Liu, Zhongwei Xie, Ka-Ho Chow, and Wenqi Wei. 2021. Boosting Ensemble Accuracy by Revisiting Ensemble Diversity Metrics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Virtual Event, 16469–16477.

[36] Nan Xu, Wenji Mao, and Guandan Chen. 2019. Multi-interactive memory network for aspect based multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 371–378.

[37] Zhen Xu, David R So, and Andrew M Dai. 2021. MUFASA: Multimodal Fusion Architecture Search for Electronic Health Records. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. Virtual Event, 10532–10540.

[38] Rui Yan, Ran Le, Yang Song, Tao Zhang, Xiangliang Zhang, and Dongyan Zhao. 2019. Interview choice reveals your preference on the market: to improve job-resume matching through profiling memories. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Anchorage, 914–922.

[39] Yang Yang, Ke-Tao Wang, De-Chuan Zhan, Hui Xiong, and Yuan Jiang. 2019. Comprehensive Semi-Supervised Multi-Modal Learning.. In *Proceedings of the International Joint Conference on Artificial Intelligence*. Macao, China, 4092–4098.

[40] Yang Yang, Yi-Feng Wu, De-Chuan Zhan, Zhi-Bin Liu, and Yuan Jiang. 2018. Complex Object Classification: A Multi-Modal Multi-Instance Multi-Label Deep Network with Optimal Transport. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. London, UK, 2594–2603.

[41] Yang Yang, De-Chuan Zhan, Ying Fan, and Yuan Jiang. 2017. Instance Specific Discriminative Modal Pursuit: A Serialized Approach. In *Proceedings of The 9th Asian Conference on Machine Learning*. Seoul, Korea, 65–80.

[42] Yang Yang, De-Chuan Zhan, Ying Fan, Yuan Jiang, and Zhi-Hua Zhou. 2017. Deep Learning for Fixed Model Reuse. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. San Francisco, California, 2831–2837.

[43] Yang Yang, De-Chuan Zhan, Yi-Feng Wu, Zhi-Bin Liu, Hui Xiong, and Yuan Jiang. 2021. Semi-Supervised Multi-Modal Clustering and Classification with Incomplete Modalities. *IEEE Trans. Knowl. Data Eng.* 33, 2 (2021), 682–695.

[44] Chen Zhang and Hao Wang. 2018. Resumevis: A visual analytics system to discover semantic information in semi-structured resume data. *ACM Transactions on Intelligent Systems and Technology* 10, 1 (2018), 1–25.

[45] Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. 2020. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing* 14, 3 (2020), 478–493.

[46] Le Zhang, Zenglin Shi, Ming-Ming Cheng, Yun Liu, Jia-Wang Bian, Joey Tianyi Zhou, Guoyan Zheng, and Zeng Zeng. 2019. Nonlinear regression via deep negative correlation learning. *IEEE transactions on pattern analysis and machine intelligence* 43, 3 (2019), 982–998.