Article

Not peer-reviewed version

Multimodal Information Fusion with Neural Gating

Olivia Smith , Ava Martinez , Noah Brown *

Posted Date: 24 September 2024

doi: 10.20944/preprints202409.1917.v1

Keywords: multimodal learning; neural networks; genre classification; information fusion; gated mechanisms

Article

Multimodal Information Fusion with Neural Gating

Olivia Smith, Ava Martinez and Noah Brown *

University of Central Oklahoma

* Correspondence: nbrown@uco.edu

Abstract: In this study, we introduce an innovative framework for multimodal learning that leverages enhanced fusion gate units within gated neural network architectures. The proposed Fusion Gate Unit (FGU) serves as a pivotal component in neural network designs, aiming to derive a comprehensive intermediate representation by amalgamating data from diverse modalities. The FGU is adept at determining the extent to which each modality influences the unit's activation through the utilization of multiplicative gating mechanisms. We conducted evaluations on a multilabel genre classification task for movies, utilizing both plot summaries and poster images as input modalities. The results demonstrate that the FGU significantly elevates the macro F-score compared to single-modality approaches and surpasses existing fusion techniques, including mixture of experts models. Additionally, we present the MM-IMDb dataset alongside this publication, which, to our knowledge, represents the most extensive publicly accessible multimodal dataset for movie genre prediction to date. This dataset is expected to facilitate further research and development in the field of multimodal information processing.

Keywords: multimodal learning; neural networks; genre classification; information fusion; gated mechanisms

1. Introduction

Representation learning techniques have garnered substantial attention from both researchers and industry practitioners due to their remarkable success in addressing intricate challenges across various domains, including computer vision, speech recognition, and natural language processing [7,39,66]. Despite the natural occurrence of multimodal data in real-world scenarios, where information is often presented through multiple channels such as images, text, and audio, the majority of existing efforts have predominantly focused on unimodal data. Multimodality encapsulates the concept that a single real-world entity can be depicted through different perspectives or data types. For instance, collaborative platforms like Wikipedia provide comprehensive descriptions of notable individuals by integrating textual narratives, visual images, and occasionally audio recordings. Similarly, users on social media platforms often comment on events such as concerts or sporting events using concise phrases accompanied by multimedia attachments, including images, videos, and audio clips. In the medical field, patient records are typically represented by a heterogeneous collection of images, sounds, textual notes, and various physiological signals.

The burgeoning availability of multimodal datasets from diverse sources has propelled the advancement of automated analytical techniques designed to harness the rich potential of such data. These techniques aim to uncover patterns and structures that reveal complex interrelationships among different data modalities [15,18,67]. In recent years, the research community focused on representation learning has increasingly prioritized multimodal tasks. Notable applications include visual question answering [14] and image captioning [32,60,62,74], where innovative methods for integrating various representation learning architectures have been developed to effectively combine information from multiple modalities.

Most contemporary models in multimodal learning emphasize either mapping information from one modality to another or addressing auxiliary tasks to construct a unified representation that encapsulates the information from all involved modalities [4–6]. In contrast, our work proposes a novel module specifically designed to integrate multiple information sources in a manner that is directly

 doi:10.20944/preprints202409.1917.v

optimized with respect to the primary objective function of the task at hand. The cornerstone of our proposed module is the concept of gating mechanisms, which selectively determine the contribution of each input modality to the final representation. By employing multiplicative gates, the Fusion Gate Unit (FGU) assigns varying degrees of importance to different features concurrently [1,2], thereby generating a rich and dynamic multimodal representation. This approach obviates the need for manual tuning of feature weights, as the FGU autonomously learns the optimal gating patterns directly from the training data.

One of the key advantages of our gated model is its versatility; the FGU can be seamlessly integrated into various neural network architectures tailored to different tasks. Furthermore, it can be optimized in an end-to-end fashion alongside other components of the network using standard gradient-based optimization techniques, ensuring cohesive and efficient learning across the entire model.

To illustrate the practical applicability of our approach, we investigate the task of movie genre identification based on two distinct modalities: plot summaries and poster images. Genre classification is a fundamental task with wide-ranging applications, including document categorization [33,82], recommendation systems [43], and information retrieval systems. [83,84] Previous predictive models, such as *MaxoutMLP_w2v* and *VGG_transfer* (detailed in Section 3), which are grounded in representation learning, demonstrate that even human evaluators might struggle to accurately classify genres without access to both plot and visual information.

The central hypothesis of this research is that integrating Fusion Gate Units into the neural network architecture, as opposed to employing manually engineered multimodal fusion strategies, will enable the model to learn input-dependent gating patterns. These patterns will effectively determine the relative contribution of each modality to the activation of hidden units, thereby enhancing the model's ability to generate accurate and meaningful predictions.

The remainder of this paper is structured as follows: Section 2 provides a comprehensive review of related literature and discusses the context of previous research in multimodal learning and fusion techniques. Section 3 delineates the methodologies employed as baselines and elaborates on our proposed representation-learning-based model incorporating Fusion Gate Units. Section 4 outlines the experimental setup, including the specifications of the MM-IMDb dataset. Section 5 presents and analyzes the results obtained from the movie genre classification experiments. Finally, Section 6 summarizes the key findings, discusses their implications, and proposes avenues for future research.

2. Related Work

2.1. Multimodal Fusion Research

Extensive reviews [15,18,20,40,86,97] have consolidated various methodologies addressing the challenges of multimodal analysis. A prevailing theme across these studies is the demonstrated advantage of multimodal approaches over their unimodal counterparts in performing automatic analysis tasks. Typically, a multimodal analysis framework processes two or more distinct modalities—such as video, audio, images, and text—that collectively describe a specific concept or entity. In recent developments, there has been a growing consensus on leveraging representation learning models to effectively capture and represent information from these diverse data sources [39,88]. Despite advancements in feature extraction, the optimal strategy for integrating these extracted features remains an active area of research.

The primary objective of multimodal fusion is to construct a unified representation that facilitates more efficient and accurate automatic analysis tasks, such as building classifiers or predictors. A rudimentary yet commonly employed technique involves concatenating features from different modalities to form a single, comprehensive feature vector [35,51,57]. While this approach is straightforward and easy to implement, it often overlooks the intrinsic correlations and interdependencies between the various modalities, potentially limiting the model's performance.

 doi:10.20944/preprints202409.1917.v

To address the limitations of simple concatenation, more sophisticated fusion strategies have been proposed, including the use of Restricted Boltzmann Machines (RBMs) and autoencoders. For instance, Ngiam et al. [48], Fei et al. [94] demonstrated the efficacy of concatenating higher-level representations by training two separate RBMs to independently reconstruct audio and video representations. Furthermore, they developed a model capable of reconstructing both modalities from just one, effectively capturing the interplay between them. Notably, their approach was able to emulate the McGurk effect, a perceptual phenomenon illustrating the interaction between auditory and visual stimuli in speech perception. Building on this, Srivastava & Salakhutdinov [56], Xu et al. [75] extended the framework by incorporating Deep Boltzmann Machines, modifying both the feature learning and reconstruction phases. They argued that such a strategy could exploit vast amounts of unlabeled data, thereby enhancing performance in tasks like retrieval and annotation.

Similar methodologies have emerged that utilize neural network architectures for multimodal fusion [13,21,34,37,41,45,58,61,102,103]. These approaches typically involve separate input layers for each modality, which are then integrated into a final supervised layer, such as a softmax regression classifier. This modular design allows for flexibility in handling various types of input data while maintaining a unified prediction mechanism.

An alternative fusion paradigm focuses on designing objective or loss functions tailored to the specific target task [11,22,38,45,54,55,64]. These strategies often operate under the assumption that a shared latent space exists, where different modalities can represent the same semantic concepts through appropriate transformations of the raw data. Semantic embedding representations are crafted such that semantically similar concepts are mapped to proximate locations in this latent space [49]. For example, Socher et al. [54] proposed a multimodal approach for zero-shot classification by training a word-based neural network to represent textual information and employing unsupervised feature learning models for image representation. Their fusion technique involved learning a linear mapping to project images into the semantic word space, supplemented by a Bayesian framework to discern whether an image belonged to a seen or unseen class. Similarly, Frome et al. [22] utilized a Convolutional Neural Network (CNN) trained on the ImageNet dataset for image representation and a word-based neural language model [47,105] for text. They achieved fusion by retraining the CNN with text representations as targets, significantly improving scalability from 2 to 20,000 unknown classes in zero-shot learning tasks. Norouzi et al. [49] further refined this approach by constructing a convex combination using classifier-estimated probabilities and semantic embedding vectors for unseen labels, achieving state-of-the-art results. These task-specific fusion models, however, tend to be tightly coupled with their respective tasks, necessitating adaptations when applied to different domains or objectives.

Our proposed model, the Gated Multimodal Unit (FGU), shares conceptual similarities with the Mixture of Experts (MoE) framework [31]. While MoE is predominantly utilized for decision fusion—combining multiple predictors to tackle supervised learning problems [63]—the FGU is designed as an integral component within the representation learning paradigm. Unlike traditional MoE models, the FGU operates independently of the final task, such as classification, regression, or unsupervised learning, provided that the associated cost function remains differentiable. This design allows the FGU to seamlessly integrate into various neural network architectures, enhancing their ability to learn rich, multimodal representations without being confined to specific application domains.

2.2. Movie Genre Classification Application

In the realm of movie genre classification, a diverse array of methodologies has been explored, leveraging various modalities to characterize each film. These modalities include textual features, image-based features, and multimedia elements such as audio and video. One of the pioneering studies in this area was conducted by Huang et al. [27], who classified movie previews into three genres by extracting handcrafted video features and training a decision tree classifier. Their evaluation

 doi:10.20944/preprints202409.1917.v

was based on a relatively small dataset of 44 films, highlighting the nascent stage of research in this domain.

Focusing solely on textual data, Shah et al. [52] tackled single-label genre classification using clustering algorithms applied to movie scripts from a dataset comprising 260 movies. This approach underscored the potential of textual analysis in genre prediction but was limited by its single-label framework. Subsequently, Pais et al. [50] extended this work by incorporating both visual and textual features to classify movies as either drama or non-drama, utilizing a dataset of 107 samples. Their bimodal approach demonstrated improved classification performance over unimodal methods, albeit within a binary genre classification context.

Expanding the scope to include multiple modalities, Hong & Hwang [25] investigated the use of Probabilistic Latent Semantic Analysis (PLSA) models to integrate audio, image, and text data for predicting the genre of movie previews. Their study focused on single-label classification across four genres using a dataset of 140 movies sourced from IMDb, highlighting the challenges of multimodal integration in genre prediction tasks.

More recently, Fu et al. [23] employed a combination of handcrafted visual features for poster analysis and bag-of-words representations for synopses. They trained separate Support Vector Machines (SVMs) for each modality and subsequently combined their predictions. Their approach was evaluated on a substantially larger dataset of 2,400 movies, each annotated with a single genre out of four possible categories. This study underscored the scalability of multimodal fusion techniques but remained constrained by its single-label classification framework.

The aforementioned studies predominantly addressed genre classification within a single-label setup. However, the multilabel scenario is arguably more reflective of real-world conditions, as most movies encompass multiple genres (e.g., "The Matrix" (2000) is categorized as both Sci-fi and Action). In this context, Anand [12] explored the efficacy of utilizing keywords and user-generated tags for multilabel genre classification, employing the MovieLens 1M dataset, which includes 1,700 movies. Additionally, Ivasic-Kos et al. [29,30] conducted multilabel classification using handcrafted features derived from movie posters, working with a dataset of 1,500 samples across six genres. Further extending this line of research, Makita & Lenskiy [43,44] leveraged the movie ratings matrix and genre correlation matrix to predict genres within a smaller subset of the MovieLens dataset, encompassing 18 movie genres.

A common limitation across these studies is the reliance on publicly available MovieLens datasets, which lack a standardized experimental setup, making systematic comparisons between different methodologies challenging. Moreover, the scale of these datasets is relatively modest, with none exceeding 10,000 samples. Addressing these limitations, our work introduces a new dataset derived from the MovieLens 20M dataset, significantly expanding the number of samples and genres. This dataset not only includes genre annotations, poster images, and plot summaries but also encompasses over 50 additional characteristics sourced from the IMDb website. By releasing both the dataset and the accompanying source code, we aim to facilitate the inclusion of more movies and genres in future research, thereby providing a robust benchmark for evaluating and comparing multimodal genre classification models.

Our FGU-based approach stands out in this landscape by offering a flexible and scalable solution for multilabel genre classification. Unlike previous models that often require task-specific adaptations, the FGU can dynamically learn to weigh the contributions of different modalities, enhancing classification performance across a diverse set of genres and dataset scales. This adaptability is particularly beneficial in multilabel scenarios, where the interplay between various genres and their corresponding modalities can be complex and multifaceted.

3. Methods

This study introduces a sophisticated neural network-based framework tailored for the multilabel classification of multimodal datasets. Central to this framework is the innovative Gated Multimodal

 doi:10.20944/preprints202409.1917.v

Unit (FGU), a novel hidden unit designed to autonomously determine the influence of each modality on the unit's activation through the use of gating mechanisms. Detailed elaboration of the FGU architecture is provided in Subsection 3.1.

Recognizing that statistical properties often vary significantly across different modalities [56], it is essential to adopt distinct representation strategies that align with the inherent characteristics of each data type. This research explores a variety of techniques for representing textual and visual data. For textual information, we evaluated multiple approaches, including word2vec models, n-gram models, and Recurrent Neural Network (RNN) models, as discussed comprehensively in Subsection 3.2. Conversely, for visual data, we assessed two different Convolutional Neural Network (CNN) architectures, detailed in Subsection 3.3, to effectively capture and represent visual features.

3.1. Gated Multimodal Unit for Multimodal Fusion

Multimodal learning is intrinsically linked to the concept of data fusion, which seeks to integrate multiple information sources into a cohesive representation that encapsulates more comprehensive information than any individual source could provide [18]. Data fusion can be categorized broadly into two primary strategies: feature fusion and decision fusion. Feature fusion, often referred to as early fusion, involves selecting and combining subsets of features from different modalities or creating new feature combinations that better represent the underlying information necessary to address a specific problem. This approach aims to create a unified feature space that captures the complementary strengths of each modality. On the other hand, decision fusion, or late fusion, entails combining the outputs or decisions from separate systems or classifiers, typically through methods such as averaging, voting, or more sophisticated Bayesian frameworks, to reach a consensus decision.

In this research, we propose the Gated Multimodal Unit (FGU), a novel model that amalgamates principles from both feature and decision fusion methodologies through the use of gated neural networks. The FGU draws inspiration from the gating mechanisms found in recurrent architectures like Gated Recurrent Units (GRU) and Long Short-Term Memory (LSTM) networks, which are renowned for their ability to control information flow within neural networks. The primary function of the FGU is to serve as an internal component within a larger neural network architecture, facilitating the creation of intermediate representations by effectively combining data from multiple modalities.

Each modality is represented by a feature vector, denoted as x_i, where i indexes the different modalities. These feature vectors are processed through neurons equipped with hyperbolic tangent (tanh) activation functions, which encode modality-specific internal representations. For every input modality x_i, there exists an associated gate neuron, represented by sigmoid (σ) activation functions. These gate neurons play a crucial role in regulating the contribution of each modality's features to the overall output of the FGU. Specifically, when a new data sample is introduced to the network, each gate neuron evaluates the collective feature vectors from all modalities to determine the extent to which its corresponding modality should influence the internal encoding of that particular sample.

The mathematical formulation governing the FGU is as follows:

$$
\begin{aligned}
h_v &= \tanh\left(W_v \cdot x_v\right) \\
h_t &= \tanh\left(W_t \cdot x_t\right) \\
z &= \sigma\left(W_z \cdot [x_v, x_t]\right) \\
h &= z \odot h_v + (1 - z) \odot h_t \\
\Theta &= \{W_v, W_t, W_z\}
\end{aligned}
$$

Here, h_v and h_t represent the internal representations derived from the visual and textual modalities, respectively. The gate z is computed by applying a sigmoid function to a linear combination of the concatenated input modalities, effectively determining the weighting between h_v and h_t. The symbol $\odot$ denotes element-wise multiplication, ensuring that each feature is appropriately scaled based on the

 doi:10.20944/preprints202409.1917.v

gate's output. The parameter set Θ comprises the weight matrices W_v, W_t, and W_z, which are learnable parameters optimized during the training process.

One of the significant advantages of the FGU is its differentiable nature, which allows seamless integration with other neural network components. This compatibility ensures that the entire network, including the FGU, can be trained end-to-end using standard gradient-based optimization algorithms, such as stochastic gradient descent (SGD). By enabling the network to learn optimal gating patterns directly from the training data, the FGU obviates the need for manual tuning of feature weights, thereby enhancing the model's adaptability and performance across diverse multimodal tasks.

3.2. Text Representation

Effective text representation is pivotal for classification tasks leveraging machine learning techniques. Traditional text representation methods, such as n-gram models and bag-of-words approaches, rely on counting the frequency of word occurrences or sequences of characters. While these methods are straightforward and computationally efficient, they often fail to capture the nuanced relationships between words and their contextual usage within the text. This limitation hampers the model's ability to understand semantic and syntactic nuances essential for accurate classification.

To address these shortcomings, more advanced representation techniques have been developed. Notably, Bengio et al. [16] introduced a neural network-based language model (NNLM) capable of learning distributed representations of words that encapsulate contextual information. Building upon this foundation, the word2vec model [46] emerged as a simplified yet powerful unsupervised learning algorithm that generates vector representations for words based on their contextual surroundings. The word2vec model leverages large corpora of unlabeled text to learn these embeddings, capturing both semantic and syntactic relationships through vector arithmetic operations.

In this study, we evaluated three distinct text representation strategies:

n-gram Inspired by the methodology proposed by Kanaris & Stamatatos [33], we employed the n-gram approach for text representation. Despite its simplicity, the n-gram model serves as a robust baseline, effectively capturing local word dependencies and sequences within the text data.

Word2Vec Word2Vec is an unsupervised learning framework that generates dense vector representations for words by analyzing their contextual co-occurrences [46]. These vector embeddings are capable of capturing complex semantic and syntactic relationships, enabling the model to perform operations such as word analogies through vector arithmetic. In our approach, each movie is represented by the average of the word vectors corresponding to the words in its plot outline. This averaging process leverages the additive compositionality property of word2vec, where the combined representation retains meaningful semantic information. By averaging rather than summing the vectors, we mitigate the risk of excessively large input values, thereby maintaining numerical stability during subsequent neural network processing.

Recurrent Neural Network For a more context-aware representation, we explored the use of Recurrent Neural Networks (RNNs) to model the sequential nature of textual data. Specifically, we investigated two variants:

- *RNN_w2v*: This variant employs transfer learning by utilizing pre-trained word2vec vectors as input embeddings. The RNN processes the sequence of word vectors, capturing temporal dependencies and contextual information to generate a comprehensive representation of the plot outline.
- *RNN_end2end*: In contrast, this variant learns word embeddings from scratch in an end-to-end manner, allowing the RNN to optimize the embeddings jointly with the classification task. This approach enables the model to tailor the embeddings specifically to the genre classification objective, potentially enhancing performance by capturing task-relevant features.

 doi:10.20944/preprints202409.1917.v

Each of these text representation methods offers unique advantages, and their comparative effectiveness is evaluated within our experimental framework to determine the most suitable approach for multimodal genre classification.

3.3. Visual Representation

In the domain of computer vision, Convolutional Neural Networks (CNNs) have established themselves as the cornerstone for visual representation learning. CNNs excel at automatically extracting hierarchical feature representations from raw image data, making them indispensable for a wide array of vision-related tasks. A key characteristic of CNNs is their ability to leverage large-scale datasets to learn transferable features that generalize well across different domains, a property that is extensively utilized in transfer learning paradigms.

For visual representation in this study, we explored two primary strategies:

VGG Transfer This approach leverages the VGG Network [53], a deep CNN architecture renowned for its performance on the ImageNet dataset. By utilizing the pre-trained VGG model as a feature extractor, we extract the activations from the last hidden layer as the visual representation for each movie poster. This transfer learning strategy capitalizes on the rich feature representations learned from extensive image data, enabling effective utilization of visual information without the need for training a network from scratch.

End-to-End CNN In contrast to the transfer learning approach, the end-to-end CNN strategy involves training a custom CNN architecture from the ground up, tailored specifically to our dataset and classification task. Our architecture comprises five convolutional layers designed to progressively extract higher-level features from the input images, followed by a Multi-Layer Perceptron (MLP) that serves as the classifier. This end-to-end training allows the CNN to learn feature representations that are highly specialized and optimized for the task of genre classification based on poster imagery.

By evaluating both transfer learning and end-to-end training methodologies, we aim to ascertain the most effective approach for capturing and representing visual features pertinent to movie genre classification.

3.4. Classification Model

Following the extraction and representation of multimodal data, the next critical step involves mapping these feature vectors to their corresponding genre labels. To achieve this, we explored two distinct classification methodologies: Logistic Regression and a more complex Neural Network architecture.

Logistic Regression As a baseline classification approach, we employed Logistic Regression, a well-established statistical method for binary and multiclass classification tasks. Logistic Regression models the probability of each genre label given the input feature vectors, making it a straightforward yet effective choice for multilabel classification scenarios.

Neural Network Architecture To harness the expressive power of deep learning, we implemented a Multilayer Perceptron (MLP) with two fully connected layers, incorporating the Maxout activation function. The Maxout activation function, defined as:

$$h_i(\mathbf{s}) = \max_{j \in [1,k]} z_{i,j} \tag{1}$$

where $\mathbf{s} \in \mathbb{R}^n$ is the input vector, $z_{i,j} = \mathbf{s}^T \mathbf{W}_{\ldots ij} + \mathbf{b}_{ij}$ represents the output of the j-th linear transformation for the i-th hidden unit, and $\mathbf{W} \in \mathbb{R}^{d \times m \times k}$ and $\mathbf{b} \in \mathbb{R}^{m \times k}$ are the learnable parameters, offers several advantages. Maxout networks have been demonstrated to act as universal approximators with as few as two hidden units, providing the capability to model

 doi:10.20944/preprints202409.1917.v

complex, non-linear functions [24]. Additionally, the Maxout activation function mitigates the issue of unit saturation, allowing for more stable and efficient training dynamics. By incorporating Maxout into the MLP, we aim to enhance the model's capacity to capture intricate patterns and relationships within the multimodal feature space, thereby improving classification performance.

Both classification models were rigorously evaluated to determine their efficacy in mapping the rich, multimodal feature representations to accurate genre predictions. The comparative analysis between Logistic Regression and the Maxout-based MLP provides insights into the trade-offs between simplicity and expressiveness in the context of multilabel genre classification.

4. Experiments

4.1. Dataset

In this study, we introduce the Multimodal IMDb (**MM-IMDb**)[1] dataset, which will be made accessible to the public. The MM-IMDb dataset was meticulously constructed by leveraging IMDb identifiers sourced from the Movielens 20M dataset[2], encompassing ratings for approximately 27,000 films. Utilizing the IMDbPY library[3], we systematically excluded any movies that lacked corresponding poster images, ensuring that each entry in the final dataset is complete with essential visual and textual information. Consequently, the MM-IMDb dataset comprises 25,959 movies, each annotated with detailed plot summaries, poster images, genre classifications, and an extensive collection of 50 additional metadata fields. These metadata include attributes such as the release year, language, writer, director, aspect ratio, and more, providing a rich context for each film.

It is important to note that films can be classified under multiple genres simultaneously. The co-occurrence matrix of genre tags illustrates the frequency with which different genres appear together, highlighting the interconnected nature of film classifications. The distribution of movie poster dimensions and the lengths of plot summaries are also characterized, with poster sizes varying and plot lengths averaging 92.5 words. The longest plot summary in the dataset contains 1,431 words, and on average, each movie is associated with 2.48 genres. The primary task defined in this work is the prediction of movie genres based on plot summaries and poster images. However, the rich metadata available opens avenues for additional tasks, such as predicting movie ratings and enabling content-based retrieval systems.

4.2. Experimental Configuration

The MM-IMDb dataset was divided into three distinct subsets to facilitate model training and evaluation. Specifically, the training set comprises 15,552 samples, the development set includes 2,608 samples, and the test set consists of 7,799 samples. This distribution ensures that the model is trained on a substantial portion of the data while retaining adequate samples for validation and testing. The distribution of genre tags across these subsets is detailed in Table 1. To maintain a balanced representation of genres, the dataset was stratified such that the training, development, and test sets contain 60%, 10%, and 30% of samples for each genre, respectively.

---

1 http://lisi1.unal.edu.co/mmimdb/
2 http://grouplens.org/datasets/movielens/
3 http://imdbpy.sourceforge.net/

 doi:10.20944/preprints202409.1917.v

Table 1. Distribution of Genres Across Training, Development, and Test Subsets

Genre	Train	Dev	Test	Genre	Train	Dev	Test
Drama	8424	1401	4142	Family	978	172	518
Comedy	5108	873	2611	Biography	788	144	411
Romance	3226	548	1590	War	806	128	401
Thriller	3113	512	1567	History	680	118	345
Crime	2293	382	1163	Music	634	100	311
Action	2155	351	1044	Animation	586	105	306
Adventure	1611	278	821	Musical	503	85	253
Horror	1603	275	825	Western	423	72	210
Documentary	1234	219	629	Sport	379	64	191
Mystery	1231	209	617	Short	281	48	142
Sci-Fi	1212	193	586	Film-Noir	202	34	102
Fantasy	1162	186	585				

Evaluation Metrics

Multilabel classification introduces complexities beyond those found in traditional multiclass classification, particularly in the realm of performance evaluation. Various metrics can yield significantly different insights into model performance [42]. In this context, we report four distinct averages of the f-score (f_1):

- **Sample-based Average (f_1^{sample})**: This metric calculates the f-score for each individual sample and subsequently averages these scores across all samples.
- **Micro-average (f_1^{micro})**: This approach aggregates the contributions of all classes to compute the f-score globally by considering all true positives, false positives, and false negatives.
- **Macro-average (f_1^{macro})**: Here, the f-score is computed independently for each genre and then averaged, treating all genres equally regardless of their prevalence.
- **Weighted Macro-average ($f_1^{weighted}$)**: Similar to the macro-average, but each genre's f-score is weighted by the number of true instances it has, providing a balance between rare and common genres.

The precise mathematical formulations for these metrics are as follows [42]:

$$f_1^{sample} = \frac{1}{N}\sum_{i=1}^{N}\frac{2\times|\hat{y}_i\cap y_i|}{|\hat{y}_i|+|y_i|} \quad f_1^{macro} = \frac{1}{Q}\sum_{j=1}^{Q}\frac{2\times \mathrm{p}_j\times \mathrm{r}_j}{\mathrm{p}_j+\mathrm{r}_j} \quad f_1^{weighted} = \frac{1}{Q^2}\sum_{j=1}^{Q}Q_j\frac{2\times \mathrm{p}_j\times \mathrm{r}_j}{\mathrm{p}_j+\mathrm{r}_j}$$

$$\mathrm{p}^{micro} = \frac{\sum_{j=1}^{Q} tp_j}{\sum_{j=1}^{Q} tp_j+\sum_{j=1}^{Q} fp_j} \quad \mathrm{r}^{micro} = \frac{\sum_{j=1}^{Q} tp_j}{\sum_{j=1}^{Q} tp_j+\sum_{j=1}^{Q} fn_j} \quad f_1^{micro} = \frac{2\times \mathrm{p}^{micro}\times \mathrm{r}^{micro}}{\mathrm{p}^{micro}+\mathrm{r}^{micro}}$$

where:

- N denotes the total number of samples.
- Q represents the total number of genres.
- Q_j is the count of true instances for the j-th genre.
- p_j and r_j are the precision and recall for the j-th genre, respectively.
- $\hat{y}_i$ and y_i are the predicted and true binary label vectors for the i-th sample.
- tp_j, fp_j, and fn_j correspond to the true positives, false positives, and false negatives for the j-th genre.

 doi:10.20944/preprints202409.1917.v

Textual Feature Representation

For the textual data, we utilized the pre-trained Google Word2vec embeddings[4] to represent words in a dense vector space. By intersecting the Word2vec vocabulary with the words present in the MM-IMDb plot summaries, we established a final vocabulary comprising 41,612 unique terms. Prior to embedding, all text was converted to lowercase, and no additional preprocessing steps were applied, preserving the original linguistic structures. To assess the impact of network depth on model performance, we also evaluated a simplified architecture featuring a single fully connected layer.

To benchmark the effectiveness of our textual representation, we conducted evaluations on two publicly available datasets: the *7genre* dataset, which contains 1,400 web pages classified into 7 distinct genres, and the *ki-04* dataset, encompassing 1,239 samples across 8 genres. Our model was compared against state-of-the-art results presented by Kanaris & Stamatatos [33], who employed character n-grams combined with structured HTML tag information for genre prediction in web pages.

Visual Feature Representation

For visual data, our initial approach involved utilizing the VGG network as a feature extractor, referred to as *VGG_Transfer*. This method leverages the deep features learned by the VGG architecture, which are then used for downstream tasks without further modification. The second approach involves processing raw images directly through a Convolutional Neural Network (CNN). Given the variability in image sizes, all poster images were standardized by scaling and cropping them to a resolution of 160×256 pixels while maintaining the original aspect ratio. The CNN architecture comprises five convolutional layers with filter sizes of 5×5, 3×3, 3×3, 3×3, and 3×3, respectively, each followed by a pooling layer with a 2×2 window. Each convolutional layer contains 16 hidden units. The convolutional layers are subsequently connected to the *MaxoutMLP* classifier, which aggregates the extracted features for genre prediction.

Multimodal Feature Integration

To effectively combine textual and visual modalities, we explored four distinct fusion strategies as baselines:

Average Probability (Late Fusion) This strategy involves averaging the probabilities output by the best-performing model for each modality and then applying a threshold to determine the final genre predictions.

Concatenation Building on findings from prior research [35,51,57], we concatenated the feature representations from both modalities and fed the combined vector into the *MaxoutMLP* architecture for classification.

Linear Sum Inspired by the approach of Vinyals et al. [60], this method applies a linear transformation to each modality's representation to align their dimensionalities before summing them. The resultant vector is then processed by the *MaxoutMLP* classifier.

Mixture of Experts (MoE) We adapted the MoE model [31] for multilabel classification by exploring two gating mechanisms: *tied* gating, where a single gate influences all logistic outputs, and *untied* gating, where each logistic output has its own dedicated gate. Both logistic regression and *MaxoutMLP* were evaluated as expert models within this framework.

Neural Network Training Procedures

All neural network models were trained using the Batch Normalization technique [28], which normalizes the inputs of each layer across the mini-batch, ensuring that each hidden unit maintains a zero mean and unit variance. This approach facilitates faster training and more stable convergence. We

[4] https://code.google.com/archive/p/word2vec/

 doi:10.20944/preprints202409.1917.v

employed Stochastic Gradient Descent (SGD) with the ADAM optimization algorithm [36] to update the network weights. To mitigate overfitting, dropout regularization and max-norm constraints were applied.

The hyperparameters explored during training included hidden layer sizes ($\{64, 128, 256, 512\}$), learning rates ($[10^{-3}, 10^{-1}]$), dropout rates ($[0.3, 0.7]$), max-norm values ($[5, 20]$), and weight initialization ranges ($[10^{-3}, 10^{-1}]$). We trained 25 models with randomly initialized hyperparameters, selecting the best-performing model based on validation set performance. This random search strategy is advantageous over grid search, particularly for training deep models, as it more efficiently explores the hyperparameter space [17]. All implementations were carried out using the Blocks framework [59][5].

During training, Batch Normalization significantly accelerated the training process and improved convergence rates, reducing the model's sensitivity to hyperparameter settings such as learning rates and initialization ranges. Additionally, the incorporation of dropout and max-norm regularization techniques contributed to enhanced generalization performance on the test set.

5. Experimental Results

5.1. Performance on Synthetic Data

To assess the capability of our model to discern the modality that provides the most informative features for classification, we devised a synthetic task based on a generative model. This model defines a binary target variable C and two input feature vectors x_v and x_t, each residing in $\mathbb{R}^2$. A latent binary variable M determines which modality—visual or textual—contains the relevant information for classifying the sample.

The generative process is defined as follows:

$$
\begin{aligned}
C &\sim \text{Bernoulli}(p_C) & x_v &= My_v + (1-M)\hat{y}_v \\
M &\sim \text{Bernoulli}(p_M) & y_t &\sim \mathcal{N}(\gamma_t^C) \\
y_v &\sim \mathcal{N}(\gamma_v^C) & \hat{y}_t &\sim \mathcal{N}(\hat{\gamma}_t) \\
\hat{y}_v &\sim \mathcal{N}(\hat{\gamma}_v) & x_t &= M\hat{y}_t + (1-M)y_t
\end{aligned}
$$

In this setup:

- C is the binary class label.
- M decides which modality holds the class-informative features.
- y_v and y_t are class-dependent features drawn from Gaussian distributions centered at γ_v^C and γ_t^C, respectively.
- $\hat{y}_v$ and $\hat{y}_t$ are noise features drawn from Gaussian distributions centered at $\hat{\gamma}_v$ and $\hat{\gamma}_t$, respectively.
- The input features x_v and x_t are composites of either informative or noise features based on the value of M.

We trained a single FGU model with a sigmoid activation function applied to the hidden state h, optimizing it using binary cross-entropy loss. For each experiment, we generated 200 samples per class, and conducted $1,000$ independent trials with varying random seeds. The FGU model outperformed a logistic regression classifier in 370 instances and matched its performance in the remaining trials. This demonstrates the FGU's ability to effectively learn the latent variable M that dictates the relevance of each modality for classification tasks.

[5] https://github.com/johnarevalo/FGU-mmimdb

 doi:10.20944/preprints202409.1917.v

Moreover, we investigated the relationship between the gate activations z and the latent variable M. The analysis revealed a perfect correlation of 1, indicating that the FGU successfully inferred the modality relevance solely from the input features x_v and x_t.

To visualize how the gate activations z correspond to different regions of the feature space, we projected z back onto the input feature dimensions in a synthetic experiment where $x_v, x_t \in \mathbb{R}$. The left plot illustrates the regions where $z > 0.5$ (indicating a preference for the visual modality) and $z \leq 0.5$ (indicating a preference for the textual modality). The right plot shows the model's predictions, confirming that the gating mechanism effectively isolates noise and maintains decision boundaries consistent with modality relevance.

5.2. Genre Classification Performance

Prior to integrating our textual representation into the multimodal framework, we validated its efficacy on separate classification tasks using two public datasets. The *MaxoutMLP_w2v* model achieved state-of-the-art performance on the *ki-04* dataset and improved the f-score on the *7Genre* dataset from 0.841 to 0.854 compared to the baseline presented by Kanaris & Stamatatos [33]. It is noteworthy that the baseline method utilized additional HTML structural information from web pages, whereas our representation relies solely on textual data, highlighting its robustness and effectiveness.

Table 2 presents a comprehensive overview of the classification performance across different modalities and representations. For the textual modality, the highest performance was achieved using the *MaxoutMLP_w2v* representation, indicating the effectiveness of Word2vec embeddings combined with a multilayer perceptron classifier. The performance trends were consistent across all evaluation metrics. In contrast, models trained from scratch, such as the RNN-based architectures, exhibited inferior performance, likely due to insufficient data to capture meaningful word relationships. This observation underscores the importance of leveraging pre-trained embeddings for tasks with limited labeled data.

Table 2. Summary of Genre Classification Performance on the MM-IMDb Dataset

Modality	**Representation**	**F-Score**			
		weighted	samples	micro	macro
Multimodal	FGU	**0.617**	**0.630**	**0.630**	**0.541**
	Linear_sum	0.600	0.607	0.607	0.530
	Concatenate	0.597	0.605	0.606	0.521
	AVG_probs	0.604	0.616	0.615	0.491
	MoE_MaxoutMLP	0.592	0.593	0.601	0.516
	MoE_MaxoutMLP (tied)	0.579	0.579	0.587	0.489
	MoE_Logistic	0.541	0.557	0.565	0.456
	MoE_Logistic (tied)	0.483	0.507	0.518	0.358
Text	MaxoutMLP_w2v	0.588	0.592	0.595	0.488
	RNN_transfer	0.570	0.580	0.580	0.480
	MaxoutMLP_w2v_1_hidden	0.540	0.540	0.550	0.440
	Logistic_w2v	0.530	0.540	0.550	0.420
	MaxoutMLP_3grams	0.510	0.510	0.520	0.420
	Logistic_3grams	0.510	0.520	0.530	0.400
	RNN_end2end	0.490	0.490	0.490	0.370
Visual	VGG_Transfer	0.410	0.429	0.437	0.284
	CNN_end2end	0.370	0.350	0.340	0.210

Regarding the visual modality, models utilizing pre-trained networks like *VGG_Transfer* outperformed those trained end-to-end, suggesting that the dataset size is insufficient for learning complex visual features from scratch. This aligns with the general understanding that deep visual models require large amounts of data to generalize effectively.

 doi:10.20944/preprints202409.1917.v

5.3. Detailed Genre-wise Analysis

To gain deeper insights into model performance, we conducted a genre-wise evaluation, as detailed in Table 3. The table compares the macro f-scores for single-modality approaches (textual and visual) against the multimodal FGU approach across various genres.

Table 3. Macro F-Score per Genre for Single-Modality and Multimodal FGU Approaches

Genre	Textual	Visual	FGU	Genre	Textual	Visual	FGU
Drama	0.74	0.67	**0.77**	Fantasy	0.42	0.25	**0.46**
Comedy	0.65	0.59	**0.68**	Family	0.50	0.46	**0.58**
Romance	**0.53**	0.33	0.51	Biography	**0.40**	0.02	0.25
Thriller	0.57	0.39	**0.62**	War	0.57	0.19	**0.64**
Crime	**0.61**	0.25	0.59	History	**0.35**	0.06	0.29
Action	0.58	0.37	**0.60**	Animation	0.43	0.61	**0.68**
Adventure	**0.51**	0.32	**0.51**	Musical	0.14	0.18	**0.28**
Horror	0.65	0.41	**0.69**	Western	0.52	0.37	**0.65**
Documentary	0.67	0.18	**0.76**	Sport	0.64	0.11	**0.70**
Mystery	0.38	0.11	**0.39**	Short	0.20	0.24	**0.27**
Sci-Fi	0.63	0.30	**0.66**	Film-Noir	0.02	0.11	**0.37**
Music	**0.51**	0.01	0.48				

The FGU-based multimodal approach consistently outperformed single-modality models across the majority of genres, demonstrating the efficacy of integrating textual and visual information. Notably, in genres such as *Animation* and *Family*, the visual modality alone sometimes surpassed the textual representation, underscoring the significance of visual features in these categories. Conversely, genres like *Biography* and *War* showed substantial improvements when both modalities were combined, highlighting the complementary nature of textual and visual data.

5.4. Analysis of Modality Influence

To further understand the contribution of each modality to the FGU model's predictions, we analyzed the gate activations (z) across test samples. Specifically, we examined the proportion of samples for which the model predominantly relied on the textual modality ($z \leq 0.5$) versus the visual modality ($z > 0.5$) when assigning genre labels, which illustrates the percentage distribution of gate activations corresponding to each genre.

As anticipated, the textual modality generally plays a more significant role in genre prediction. However, certain genres like *Animation* and *Family* exhibit a higher reliance on visual features, aligning with their inherent visual characteristics. This observation is consistent with the genre-wise performance results, where the visual modality demonstrated superior performance in these categories. The gating mechanism effectively identifies and leverages the most informative modality for each genre, enhancing overall classification performance.

5.5. Qualitative Analysis of Predictions

To provide a qualitative perspective on the FGU model's performance, we examined specific test examples where the model exhibited notable improvements in genre prediction. Table 4 showcases instances where the FGU successfully utilized the most accurate modality, thereby eliminating false positives and enhancing prediction accuracy. These examples highlight the model's ability to integrate and prioritize information from both textual and visual sources effectively.

 doi:10.20944/preprints202409.1917.v

Table 4. Illustrative Examples of Genre Predictions on the Test Set. Red genres denote false positives, while blue genres indicate true positives.

The World According to Sesame Street	
Ground Truth	Documentary
Textual	Documentary, History
Visual	Comedy, Adventure, Family, Animation
FGU	Documentary
Babar: The Movie	
Ground Truth	Adventure, Fantasy, Family, Animation, Musical
Textual	Adventure, Documentary, War, Music
Visual	Comedy, Adventure, Family, Animation
FGU	Adventure, Family, Animation
Letters from Iwo Jima	
Ground Truth	Drama, War, History
Textual	Drama, Action, War, History
Visual	Thriller, Action, Adventure, Sci-Fi
FGU	Drama, War, History
The Last Elvis	
Ground Truth	Drama
Textual	Comedy, Documentary, Family, Biography, Music
Visual	Drama, Romance
FGU	Drama

These examples demonstrate how the FGU model effectively integrates textual and visual information to produce accurate genre classifications. In cases where the textual modality alone may introduce false positives, the incorporation of visual data allows the model to refine its predictions, ensuring that only the most relevant genres are assigned. This qualitative analysis underscores the practical benefits of multimodal integration in complex classification tasks.

6. Conclusions and Future Directions

In this study, we introduced an innovative approach for learning fusion transformations from multiple data modalities. Drawing inspiration from the mechanisms employed by recurrent neural networks to regulate information flow, our proposed model leverages multiplicative gating mechanisms. The Gated Multimodal Unit (FGU) is designed to accept two or more distinct input sources and autonomously learn the extent to which each modality influences the activation of the unit. Through a series of synthetic experiments, the FGU demonstrated its capability to uncover and utilize hidden latent variables effectively. Furthermore, when applied to real-world scenarios, the FGU consistently outperformed models that relied solely on single modalities, highlighting the advantages of multimodal integration.

A notable attribute of the FGU is its differentiable nature, which allows it to be seamlessly integrated into a wide array of neural network architectures. This compatibility ensures that the FGU can be trained using standard gradient-based optimization techniques, facilitating its adoption in diverse applications. The flexibility and robustness of the FGU make it a valuable component for enhancing the performance of complex neural networks dealing with heterogeneous data sources.

In addition to the methodological advancements, this work also contributes to the research community by releasing a comprehensive new dataset. This dataset encompasses approximately

 doi:10.20944/preprints202409.1917.v

27,000 movie plots, accompanying images, and a wealth of metadata, making it the largest of its kind for the task of movie genre classification based on multimodal information. To our knowledge, this is the first publicly available dataset that amalgamates such a substantial volume of textual and visual data for this specific application, providing a valuable resource for future research and development in the field.

Looking ahead, our future work aims to explore deeper and more sophisticated architectures of FGU layers. By stacking multiple FGU layers, we anticipate capturing more intricate interactions between different modalities, thereby further enhancing the model's ability to learn complex fusion transformations. Additionally, we plan to integrate attention mechanisms into the FGU framework. Attention mechanisms have shown great promise in allowing models to focus on the most relevant parts of the input data, and their incorporation could enable the FGU to dynamically prioritize information from different modalities based on the context of each input instance.

Another promising direction for future research involves delving deeper into the interpretability of the features learned by the FGU. Understanding how the FGU weighs and combines different modalities can provide valuable insights into the decision-making processes of multimodal models. This transparency is crucial for applications where explainability is essential, such as in recommendation systems or automated content moderation.

Moreover, we intend to extend the application of the FGU beyond movie genre classification to other domains that inherently involve multimodal data. Potential areas of application include healthcare, where combining medical images with patient records can lead to more accurate diagnoses, and autonomous driving, where integrating visual data from cameras with sensor information can enhance vehicle perception systems.

In summary, the FGU presents a robust and flexible framework for multimodal fusion, demonstrating superior performance over single-modality models in both synthetic and real-world settings. The release of the MM-IMDb dataset further empowers researchers to explore and expand upon our work. By pursuing the outlined future directions, we aim to advance the capabilities of multimodal learning systems, fostering the development of more intelligent and context-aware applications.

References

1. Anson Bastos, Abhishek Nadgeri, Kuldeep Singh, Isaiah Onando Mulang, Saeedeh Shekarpour, Johannes Hoffart, and Manohar Kaul. 2021. RECON: Relation Extraction using Knowledge Graph Context in a Graph Neural Network. In *Proceedings of the Web Conference 2021*. 1673–1685.
2. Philipp Christmann, Rishiraj Saha Roy, Abdalghani Abujabal, Jyotsna Singh, and Gerhard Weikum. 2019. Look before You Hop: Conversational Question Answering over Knowledge Graphs Using Judicious Context Expansion. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management CIKM*. 729–738.
3. 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: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 4171–4186.
4. Endri Kacupaj, Kuldeep Singh, Maria Maleshkova, and Jens Lehmann. 2022. An Answer Verbalization Dataset for Conversational Question Answerings over Knowledge Graphs. *arXiv preprint arXiv:2208.06734* (2022).
5. Magdalena Kaiser, Rishiraj Saha Roy, and Gerhard Weikum. 2021. Reinforcement Learning from Reformulations In Conversational Question Answering over Knowledge Graphs. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 459–469.
6. Yunshi Lan, Gaole He, Jinhao Jiang, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. A Survey on Complex Knowledge Base Question Answering: Methods, Challenges and Solutions. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conferences on Artificial Intelligence Organization, 4483–4491. Survey Track.

7. Yunshi Lan and Jing Jiang. 2021. Modeling transitions of focal entities for conversational knowledge base question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.
8. Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7871–7880.
9. Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
10. Pierre Marion, Paweł Krzysztof Nowak, and Francesco Piccinno. 2021. Structured Context and High-Coverage Grammar for Conversational Question Answering over Knowledge Graphs. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2021).
11. Zeynep Akata, Honglak Lee, and Bernt Schiele. Zero-Shot Learning with Structured Embeddings. *CoRR*, abs/1409.8, 2014. URL http://arxiv.org/abs/1409.8403.
12. Deepa Anand. Evaluating folksonomy information sources for genre prediction. In *Advance Computing Conference (IACC), 2014 IEEE International*, pp. 887–892, feb 2014. doi:10.1109/IAdCC.2014.6779440.
13. Galen Andrew, Raman Arora, Jeff A Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *ICML (3)*, pp. 1247–1255, 2013.
14. Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *International Conference on Computer Vision (ICCV)*, 2015.
15. Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16(6):345–379, April 2010. ISSN 0942-4962. doi:10.1007/s00530-010-0182-0.
16. Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137–1155, 2003.
17. James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305, 2012.
18. Chidansh Bhatt and Mohan Kankanhalli. Multimedia data mining: state of the art and challenges. *Multimedia Tools and Applications*, 51(1):35–76, 2011. ISSN 1380-7501. doi:10.1007/s11042-010-0645-5.
19. Adam Coates and Andrew Y Ng. The importance of encoding versus training with sparse coding and vector quantization. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 921–928, 2011.
20. Li Deng. A tutorial survey of architectures, algorithms, and applications for deep learning. *APSIPA Transactions on Signal and Information Processing*, 3, 2014. ISSN 2048-7703. doi:10.1017/atsip.2013.9. URL http://journals.cambridge.org/article{_}S2048770313000097.
21. Fangxiang Feng, Ruifan Li, and Xiaojie Wang. Constructing hierarchical image-tags bimodal representations for word tags alternative choice. *arXiv preprint arXiv:1307.1275*, 2013.
22. Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc\textquotesingle Aurelio Ranzato, and Tomas Mikolov. DeViSE: A Deep Visual-Semantic Embedding Model. In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger (eds.), *Advances in Neural Information Processing Systems 26*, pp. 2121–2129. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/5204-devise-a-deep-visual-semantic-embedding-model.pdf.
23. Zhikang Fu, Bing Li, Jun Li, and Shuhua Wei. Fast Film Genres Classification Combining Poster and Synopsis. In Xiaofei He, Xinbo Gao, Yanning Zhang, Zhi-Hua Zhou, Zhi-Yong Liu, Baochuan Fu, Fuyuan Hu, and Zhancheng Zhang (eds.), *Lecture Notes in Computer Science*, volume 9242 of *Lecture Notes in Computer Science*, pp. 72–81. Springer International Publishing, Cham, 2015. doi:10.1007/978-3-319-23989-7_8. URL http://link.springer.com/10.1007/978-3-319-23862-3.
24. Ian Goodfellow, David Warde-farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. Maxout networks. In Sanjoy Dasgupta and David Mcallester (eds.), *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pp. 1319–1327. JMLR Workshop and Conference Proceedings, May 2013.
25. Hao-Zhi Hong and Jen-Ing G Hwang. Multimodal PLSA for Movie Genre Classification. In Friedhelm Schwenker, Fabio Roli, and Josef Kittler (eds.), *Multiple Classifier Systems: 12th International Workshop, MCS 2015, G{ü}nzburg, Germany, June 29 - July 1, 2015, Proceedings*, pp. 159–167. Springer International Publishing,

Cham, 2015. ISBN 978-3-319-20248-8. doi:10.1007/978-3-319-20248-8_14. URL http://dx.doi.org/10.1007/978-3-319-20248-8{_}14.

26. Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pp. 873–882. Association for Computational Linguistics, 2012.
27. Hui-Yu Huang, Weir-Sheng Shih, and Wen-Hsing Hsu. A Film Classifier Based on Low-level Visual Features. In *2007 IEEE 9th Workshop on Multimedia Signal Processing*, volume 3, pp. 465–468. IEEE, 2007. ISBN 978-1-4244-1273-0. doi:10.1109/MMSP.2007.4412917. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4412917.
28. Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of The 32nd International Conference on Machine Learning*, pp. 448–456, 2015.
29. Marina Ivasic-Kos, Miran Pobar, and Luka Mikec. Movie posters classification into genres based on low-level features. In *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, volume i, pp. 1198–1203. IEEE, may 2014. ISBN 978-953-233-077-9. doi:10.1109/MIPRO.2014.6859750. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6859750.
30. Marina Ivasic-Kos, Miran Pobar, and Ivo Ipsic. Automatic Movie Posters Classification into Genres. In Madevska Ana Bogdanova and Dejan Gjorgjevikj (eds.), *ICT Innovations 2014: World of Data*, pp. 319–328. Springer International Publishing, Cham, 2015. ISBN 978-3-319-09879-1. doi:10.1007/978-3-319-09879-1_32. URL http://dx.doi.org/10.1007/978-3-319-09879-1{_}32.
31. Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.
32. Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. *arXiv preprint arXiv:1511.07571*, 2015.
33. Ioannis Kanaris and Efstathios Stamatatos. Learning to recognize webpage genres. *Information Processing and Management*, 45(5):499–512, 2009. ISSN 03064573. doi:10.1016/j.ipm.2009.05.003. URL http://dx.doi.org/10.1016/j.ipm.2009.05.003.
34. Yoonseop Kang, Saehoon Kim, and Seungjin Choi. Deep learning to hash with multiple representations. In *2012 IEEE 12th International Conference on Data Mining*, pp. 930–935. IEEE, 2012.
35. Douwe Kiela and Léon Bottou. Learning Image Embeddings using Convolutional Neural Networks for Improved Multi-Modal Semantics. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-14)*, 2014.
36. Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
37. Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Multimodal neural language models. In *ICML*, volume 14, pp. 595–603, 2014.
38. Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. *arXiv preprint arXiv:1411.2539*, 2014.
39. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. doi:10.1038/nature14539. URL http://dx.doi.org/10.1038/nature14539.
40. Dong Yu Li Deng. *Deep Learning: Methods and Applications*. NOW Publishers, May 2014. URL https://www.microsoft.com/en-us/research/publication/deep-learning-methods-and-applications/.
41. Xinyan Lu, Fei Wu, Xi Li, Yin Zhang, Weiming Lu, Donghui Wang, and Yueting Zhuang. Learning multimodal neural network with ranking examples. In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 985–988. ACM, 2014.
42. Gjorgji Madjarov, Dragi Kocev, Dejan Gjorgjevikj, and Sašo Džeroski. An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition*, 45(9):3084–3104, 2012. ISSN 0031-3203. doi:http://dx.doi.org/10.1016/j.patcog.2012.03.004. URL http://www.sciencedirect.com/science/article/pii/S0031320312001203.
43. Eric Makita and Artem Lenskiy. A movie genre prediction based on Multivariate Bernoulli model and genre correlations. (May), mar 2016. URL http://arxiv.org/abs/1604.08608.

 doi:10.20944/preprints202409.1917.v

44. Eric Makita and Artem Lenskiy. A multinomial probabilistic model for movie genre predictions. 2016. URL http://arxiv.org/abs/1603.07849.
45. Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*, 2014.
46. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
47. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pp. 3111–3119, 2013.
48. J Ngiam, A Khosla, and M Kim. Multimodal Deep Learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 689—-696, 2011. URL http://ai.stanford.edu/{~}ang/papers/icml11-MultimodalDeepLearning.pdf.
49. Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeff Dean. Zero-Shot Learning by Convex Combination of Semantic Embeddings. *CoRR*, abs/1312.5, dec 2014. URL http://arxiv.org/abs/1312.5650.
50. Gregory Pais, Patrick Lambert, Daniel Beauchene, Francoise Deloule, and Bogdan Ionescu. Animated movie genre detection using symbolic fusion of text and image descriptors. In *2012 10th International Workshop on Content-Based Multimedia Indexing (CBMI)*, number 1, pp. 1–6. IEEE, jun 2012. ISBN 978-1-4673-2369-7. doi:10.1109/CBMI.2012.6269813. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6269813.
51. Deli Pei, Huaping Liu, Yulong Liu, and Fuchun Sun. Unsupervised multimodal feature learning for semantic image segmentation. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6. IEEE, aug 2013. ISBN 978-1-4673-6129-3. doi:10.1109/IJCNN.2013.6706748. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6706748.
52. Dharak Shah, Saheb Motiani, and Vishrut Patel. Movie Classification Using k-Means and Hierarchical Clustering. Technical report, 2013.
53. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
54. Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-Shot Learning Through Cross-Modal Transfer. In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger (eds.), *Advances in Neural Information Processing Systems 26*, pp. 935–943. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/5027-zero-shot-learning-through-cross-modal-transfer.pdf.
55. Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. Grounded Compositional Semantics for Finding and Describing Images with Sentences. *Transactions of the Association for Computational Linguistics (TACL)*, 2(April):207–218, 2014. URL http://nlp.stanford.edu/{~}socherr/SocherLeManningNg{_}nipsDeepWorkshop2013.pdf.
56. Nitish Srivastava and Ruslan Salakhutdinov. Multimodal Learning with Deep Boltzmann Machines. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 25*, pp. 2222–2230. Curran Associates, Inc., 2012. URL http://papers.nips.cc/paper/4683-multimodal-learning-with-deep-boltzmann-machines.pdf.
57. Heung Il Suk and Dinggang Shen. Deep learning-based feature representation for AD/MCI classification. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8150 LNCS, pp. 583–590, 2013. ISBN 9783642407628. doi:10.1007/978-3-642-40763-5_72.
58. Jian Tu, Zuxuan Wu, Qi Dai, Yu-Gang Jiang, and Xiangyang Xue. Challenge Huawei challenge: Fusing multimodal features with deep neural networks for Mobile Video Annotation. In *Multimedia and Expo Workshops (ICMEW), 2014 IEEE International Conference on*, pp. 1–6, 2014. doi:10.1109/ICMEW.2014.6890609.
59. Bart Van Merriënboer, Dzmitry Bahdanau, Vincent Dumoulin, Dmitriy Serdyuk, David Warde-Farley, Jan Chorowski, and Yoshua Bengio. Blocks and fuel: Frameworks for deep learning. *arXiv preprint arXiv:1506.00619*, 2015.
60. Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3156–3164, 2015.

61. Pengcheng Wu, Steven C.H. Hoi, Hao Xia, Peilin Zhao, Dayong Wang, and Chunyan Miao. Online multimodal deep similarity learning with application to image retrieval. In *Proceedings of the 21st ACM international conference on Multimedia - MM '13*, MM '13, pp. 153–162, New York, New York, USA, 2013. ACM Press. ISBN 9781450324045. doi:10.1145/2502081.2502112. URL http://doi.acm.org/10.1145/2502081.2502112{%}5Cnhttp://dl.acm.org/citation.cfm?doid=2502081.2502112.
62. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2(3):5, 2015.
63. Seniha Esen Yuksel, Joseph N Wilson, and Paul D Gader. Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems*, 23(8):1177–1193, 2012.
64. Yin Zheng, YJ Zhang, and Hugo Larochelle. Topic Modeling of Multimodal Data: an Autoregressive Approach. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. ISBN 2011000211. URL http://www.dmi.usherb.ca/{~}larocheh/publications/ZhengY2014.pdf.
65. Matthew J Smith. Getting value from artificial intelligence in agriculture. *Animal Production Science*, 2018.
66. Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, Tat-Seng Chua, and Shuicheng Yan. Enhancing video-language representations with structural spatio-temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
67. Hao Fei, Yafeng Ren, and Donghong Ji. Retrofitting structure-aware transformer language model for end tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2151–2161, 2020.
68. Shengqiong Wu, Hao Fei, Fei Li, Meishan Zhang, Yijiang Liu, Chong Teng, and Donghong Ji. Mastering the explicit opinion-role interaction: Syntax-aided neural transition system for unified opinion role labeling. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, pages 11513–11521, 2022.
69. Wenxuan Shi, Fei Li, Jingye Li, Hao Fei, and Donghong Ji. Effective token graph modeling using a novel labeling strategy for structured sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4232–4241, 2022.
70. Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. Latent emotion memory for multi-label emotion classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7692–7699, 2020.
71. Fengqi Wang, Fei Li, Hao Fei, Jingye Li, Shengqiong Wu, Fangfang Su, Wenxuan Shi, Donghong Ji, and Bo Cai. Entity-centered cross-document relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9871–9881, 2022.
72. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
73. Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*, 2018.
74. Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. *arXiv preprint arXiv:2305.11719*, 2023.
75. Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. Faithful logical reasoning via symbolic chain-of-thought. *arXiv preprint arXiv:2405.18357*, 2024.
76. Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. SearchQA: A new Q&A dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*, 2017.
77. Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, Fei Li, Libo Qin, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Lasuie: Unifying information extraction with latent adaptive structure-aware generative language model. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2022*, pages 15460–15475, 2022.
78. Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27, 2011.
79. Hao Fei, Yafeng Ren, Yue Zhang, Donghong Ji, and Xiaohui Liang. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in Bioinformatics*, 22(3), 2021.
80. Shengqiong Wu, Hao Fei, Wei Ji, and Tat-Seng Chua. Cross2StrA: Unpaired cross-lingual image captioning with cross-lingual cross-modal structure-pivoted alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2593–2608, 2023.

81. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
82. Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. Reasoning implicit sentiment with chain-of-thought prompting. *arXiv preprint arXiv:2305.11255*, 2023.
83. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi:10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.
84. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
85. Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
86. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Proceedings of the International Conference on Machine Learning*, 2024.
87. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
88. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, and Tat-Seng Chua. Dysen-vdm: Empowering dynamics-aware text-to-video diffusion with llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7641–7653, 2024.
89. Mihir Momaya, Anjnya Khanna, Jessica Sadavarte, and Manoj Sankhe. Krushi–the farmer chatbot. In *2021 International Conference on Communication information and Computing Technology (ICCICT)*, pages 1–6. IEEE, 2021.
90. Hao Fei, Fei Li, Chenliang Li, Shengqiong Wu, Jingye Li, and Donghong Ji. Inheriting the wisdom of predecessors: A multiplex cascade framework for unified aspect-based sentiment analysis. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 4096–4103, 2022.
91. Shengqiong Wu, Hao Fei, Yafeng Ren, Donghong Ji, and Jingye Li. Learn from syntax: Improving pair-wise aspect and opinion terms extraction with rich syntactic knowledge. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 3957–3963, 2021.
92. Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Chong Teng, Tat-Seng Chua, Donghong Ji, and Fei Li. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5923–5934, 2023.
93. Hao Fei, Qian Liu, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Scene graph as pivoting: Inference-time image-free unsupervised multimodal machine translation with visual scene hallucination. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5980–5994, 2023.
94. Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. 2024.
95. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.
96. Abbott Chen and Chai Liu. Intelligent commerce facilitates education technology: The platform and chatbot for the taiwan agriculture service. *International Journal of e-Education, e-Business, e-Management and e-Learning*, 11:1–10, 01 2021.
97. Shengqiong Wu, Hao Fei, Xiangtai Li, Jiayi Ji, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Towards semantic equivalence of tokenization in multimodal llm. *arXiv preprint arXiv:2406.05127*, 2024.
98. Jingye Li, Kang Xu, Fei Li, Hao Fei, Yafeng Ren, and Donghong Ji. MRN: A locally and globally mention-based reasoning network for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1359–1370, 2021.
99. Hao Fei, Shengqiong Wu, Yafeng Ren, and Meishan Zhang. Matching structure for dual learning. In *Proceedings of the International Conference on Machine Learning, ICML*, pages 6373–6391, 2022.

100. Hu Cao, Jingye Li, Fangfang Su, Fei Li, Hao Fei, Shengqiong Wu, Bobo Li, Liang Zhao, and Donghong Ji. OneEE: A one-stage framework for fast overlapping and nested event extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1953–1964, 2022.
101. Isakwisa Gaddy Tende, Kentaro Aburada, Hisaaki Yamaba, Tetsuro Katayama, and Naonobu Okazaki. Proposal for a crop protection information system for rural farmers in tanzania. *Agronomy*, 11(12):2411, 2021.
102. Hao Fei, Yafeng Ren, and Donghong Ji. Boundaries and edges rethinking: An end-to-end neural model for overlapping entity relation extraction. *Information Processing & Management*, 57(6):102311, 2020.
103. Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10965–10973, 2022.
104. Mohit Jain, Pratyush Kumar, Ishita Bhansali, Q Vera Liao, Khai Truong, and Shwetak Patel. Farmchat: a conversational agent to answer farmer queries. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(4):1–22, 2018.
105. Shengqiong Wu, Hao Fei, Hanwang Zhang, and Tat-Seng Chua. Imagine that! abstract-to-intricate text-to-image synthesis with scene graph hallucination diffusion. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 79240–79259, 2023.
106. Hao Fei, Tat-Seng Chua, Chenliang Li, Donghong Ji, Meishan Zhang, and Yafeng Ren. On the robustness of aspect-based sentiment analysis: Rethinking model, data, and training. *ACM Transactions on Information Systems*, 41(2):50:1–50:32, 2023.
107. Yu Zhao, Hao Fei, Yixin Cao, Bobo Li, Meishan Zhang, Jianguo Wei, Min Zhang, and Tat-Seng Chua. Constructing holistic spatio-temporal scene graph for video semantic role labeling. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5281–5291, 2023.
108. Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14734–14751, 2023.
109. Hao Fei, Yafeng Ren, Yue Zhang, and Donghong Ji. Nonautoregressive encoder-decoder neural framework for end-to-end aspect-based sentiment triplet extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):5544–5556, 2023.
110. Yu Zhao, Hao Fei, Wei Ji, Jianguo Wei, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Generating visual spatial description via holistic 3D scene understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7960–7977, 2023.