Improving Multimodal Fusion with Multi-stage Feature Selection

Anonymous ACL submission

Abstract

Multimodal fusion benefits from multimodal information, involving integrating multiple modal features into one representation. However, multimodal fusion in multi-modal network suffers from two problems. First, multi-modal networks are often prone to overfitting due to their increased modal capacity. Second, different modalities contribute differently to performance improvement. In order to address the two issues, we propose a two-stage multimodal feature selection method, named MFS, to improve multimodal fusion by selecting those features with the max mutual information with intra-modal and inter-modal features, and those features that have causal relationships with labels and features to be involved in modal fusion. On one hand, MFS effectively reduces the representation dimension by filtering out redundant information from each modality, alleviating the overfitting problem due to over capacity modal representation. On the other hand, MFS selects modal- and task-specific representation dimension for each modality of each downstream task, which dynamically selects a specific number of features as modal representation, thereby better adapting to downstream tasks. We conduct extensive experiments on 3 multimodal tasks, 5 datasets, across vision, language, and audio modalities. Our experimental results demonstrate the generality and effectiveness of two-stage multimodal feature selection of our MFS[1] model.

1 Introduction

Multimodal data, including verbal, acoustic, and visual information, is widely used to enhance machine learning systems like dialog systems. Multimodal fusion plays a crucial role by integrating information from various modalities for decision making.

With additional information, multi-modal networks are expected to match or outperform their uni-modal counterparts (Huang et al., 2021). However, multi-modal networks are often observed to underperform uni-modal networks (AlAmri et al., 2019; Thomason et al., 2019; Wang et al., 2020). To explore this deficiency, (Wang et al., 2020) prove that multimodal network with increased capacity causes overfitting, which further hurts the performance. (Liang et al., 2022) argue that the gradients of different modalities should be further adjusted during training. (Du et al., 2023) also notice the phenomenon of modality laziness, which causes insufficient modal fusion. In summary, multimodal fusion generally suffers from two problems. First, it tends to overfit due to the increased capacity of modal representation. The combination of different modalities significantly increases the complexity of the model, making the training process complicated and unstable. Second, it is difficult to select the most informative modality or features for the decision making as different modalities contribute differently to the performance at different rates (Wang et al., 2020). Additionally, the difference across different downstream tasks impacts the dimension or feature of each modality required for solving the specific task. For example, a vision-dominated task, such as image captioning, would require more information from visual modality while textual information might contribute little. As illustrated in Figure 1, the situation will be dif-

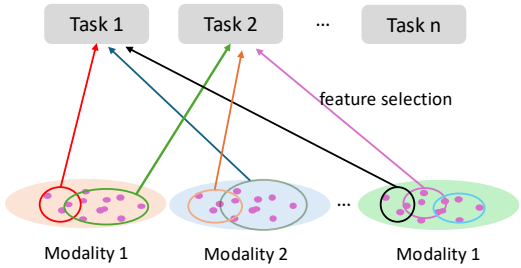


Figure 1: Illustration of how different features of each modality contribute to the specific task differently.

[1]Code will be released after acceptance

ferent if the task at hand differs.

To align with downstream tasks, multimodal fusion should dynamically adjust each modality's contribution while reducing the modal representation's dimension. Modal fusion is the process of integrating information from multiple data sources or sensory modalities into one representation, and aims to maximally leverages the strengths of each modality. In this work, we regard the modal fusion process as an optimization problem, where we set mutual information maximization among hidden units as optimization objectives. Two types of modal features are regarded as informative features: 1) features with the maximum mutual information with intra-modal and inter-modal features (denoted as **F1 features**), and 2) features that have causal relationships with labels (denoted as **F2 features**). In regards, we propose a two-stage feature selection method, named MFS, to improve multimodal fusion by selecting F1 and F2 features from the modal representation. Our MFS is designed to consist of two stages: 1) F1 feature selection with mutual information, and 2) F2 feature selection with Bayesian Network (BN). On one hand, MFS reduces the representation dimension of each modality, alleviating the overfitting problem caused by over-capacity modal representation. On the other hand, MFS dynamically selects modal- and task-specific representation dimensions for each modality of each downstream task, enhancing the flexibility to task adaptation.

Specifically, first stage takes the mutual information maximization as the optimization objective to select features those maximize the intra- and inter-modal mutual information, thereby filtering out redundant features in representation. In contrast with first stage, the second stage leverage bayesian network as basic to model the causal relationships among features from multiple modalities, where features and labels are viewed as nodes and the edges is constructed based on the conditional dependencies between nodes. For example, the child node's probability distribution is conditionally dependent on the states of parent nodes. Through this stage, we can further filters out the features that are without causal relationship with the nodes in BN.

To sum up, our contributions are two-folds:

1. We propose a two-stage multimodal feature selection method (MFS) for improving multimodal fusion, which selects i) features with the maximum mutual information with intra-modal and inter-modal features, and ii) features that have the causal relationships with labels and other features, revealed by Bayesian network.

2. We view the first stage of feature selection as an optimization problem, with maximizing mutual information between features under specific constrains. In the second stage, we propose to use Bayesian network to learn causal structure of features, thereby filtering out redundant features.

2 Related Work

Featrue Selection Feature selection is a technique that effectively reduces the dimensionality of the feature space by eliminating irrelevant and redundant features without significantly affecting the quality of decision-making of the trained model. (Theng and Bhoyar, 2024) group the works of feature selection into three categories. The first is filter-based feature selection which extracts feature subsets using statistical metrics, such as correlation, consistency, information, or distance, without involving any learning algorithms. For example, (Shimamura and Hirata, 2019) filter out the features with high inconsistency rate, where high consistency rate represent the high relevance of the feature to the class target variable. The second is wrapper-based feature selection. Wrapper-based feature selection combines the search strategies and machine learning algorithms (classifier as an evaluation metric), which adopts the black-box approach where the classifier's output is used as feedback to find the appropriate subset of features. For example, (Shimamura and Hirata, 2019) treat the combination of features as a candidate solution and generating new combinations through crossover and mutation to select feature. The third is embedded-based feature selection. embedded-based feature selection integrates the feature selection process as one inseparable learning process into the model estimation process. For example, (Wang et al., 2022) propose a method to assess the ridge regression for feature selection, aiming to address the issue that linear discriminant analysis and generalized discriminant analysis limit the number of relevant features to be generated.

Model Fusion Multimodal fusion involves integrating information from different modalities. The fusion methods can be categorized into three types:

early fusion, late fusion, and hybrid fusion. Early fusion aggregates the raw features from different modalities into a joint representation before processing by deep learning models. (Pérez-Rúa et al., 2019) proposes MFAS that focuses on early fusion to learn how to combine features from different modalities effectively. (Fukui et al., 2016) perform feature-level fusion by compactly representing the joint feature space using bilinear pooling techniques. (Le et al., 2023) manifest the application of feature-level fusion in emotion recognition. Late fusion combines the decisions from different classifiers into one final decision. (Grover et al., 2023) define a late fusion multimodal deep learning model and proved that late fusion technique outperforms the single architecture-based prediction. (Ding et al., 2022) propose a multi-level late fusion learning framework with residual connections for sarcasm detection that surpasses other fusion models. Hybrid fusion leverages both early fusion and late fusion, integrating the strengths of both approaches. (Choi and Lee, 2019) propose the EmbraceNet to combining both early and late features in modal fusion. (Tan and Bansal, 2019) propose a LXMERT framework to learn vision-and-language connections, which leverages Transformer architecture to handle early and late feature fusion simultaneously.

3 Methodology

3.1 Task Formalization

Given a dataset $D = \{(s_1, y_1), \cdots, (s_N, y_N)\}$, each sample contains M modal representations $s_i = \{I_1, I_j, \ldots, I_M\}$, where $I_j = \{e_j^1, e_j^2, \cdots, e_j^{d_j}\}$ and $I_j \in \mathbf{R}^{1\times d_j}$ represents the modal representation, which is extracted based on modal encoder (e.g., T5 (Raffel et al., 2020) for text, CLIP (Radford et al., 2021a) for image and AST (Gong et al., 2021) for audio). N denotes the number of data samples, d_j denotes the dimension of modal representation I_j. Before feature selection, we extract modal representation by modal encoder in the offline and then feed the multimodal representations into MFS to select F1 and F2 features. $F1 = \{F1_{I_1}, \cdots, F1_{I_M}\}, F1 \in \mathbf{R}^{n_{select}}$ denotes the selected features after first stage of MFS, n_{select} denotes the number of selected features, $F2 = \{F2_{I_1}, \cdots, F2_{I_M}\}$ denotes the selected features after MFS.

3.2 Overall Architecture

The framework of multimodal feature selection (MFS) is shown in Fig 2. MFS receives multiple modal representations and outputs the selected feature used for modal fusion. Furthermore, we take the fusion representation based on selected feature to train downstream tasks. Specifically, MFS consists of two stages: 1) F1 feature selection using mutual information and 2) F2 feature selection using a Bayesian Network. In the first stage, MFS selects F1 features based on the maximum mutual information with both intra-modal and inter-modal features. In the second stage, MFS filters out the features without causal relationships with other features using bayesian network, and the remaining features (i.e.,F2 features) is used for modal fusion and is input into the model for downstream task learning.

3.3 Feature Selection with Mutual Information

Mutual information is used for measuring the relevance among variables, making it an ideal metric for feature selection. Given two features (e.g., X and Y) of the modal representation, their joint probability density function $p(x, y)$ and the marginal probability density function $p(x)$ and $p(y)$. The mutual information between neurons X and Y is given below:

$$I(X;Y) = \int_Y \int_X p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) dx\, dy \tag{1}$$

Considering that the feature correlates within a single modality and between different modalities, we propose to select the features based on intra-modal and inter-modal mutual information maximization, respectively. Intra-modal with mutual information maximization selects features having maximum mutual information with other features within the same modal representation, and inter-modal with mutual information maximization selects features that have max mutual information with features from other modal representation within a same sample. For clarity, we define a binary decision variable $X_{m,j} \in \{0, 1\}$ to indicate whether the feature is selected in representation m, where $X_{m,j} = 1$ denotes the feature with index j is selected, otherwise it is not selected. Take modal representation as an example, we define $X_{I_i,j} \in \{0, 1\}, j \in \{1, \cdots, d_i\}$ as the binary decision variable to record which features within

the modal representation I_i are selected.

Intra-modality Mutual Information Maximization Given a modal representation $I_i = \{e_i^1, e_i^2, \cdots, e_i^{d_i}\}$, $e_i^j \in I_i$ denotes a feature in the representation. Ideally, we can calculate the mutual information for all feature combination, i.e., $e_i^j, j \in [1, d_i]$ with every other feature in feature vector I_i. Yet, considering that pair-wise time complexity, we calculate the mutual information between each feature and the principal components of modal representation. We adopt principle component analysis (PCA) (Greenacre et al., 2022) to extract K principle components for each modal representation. For modality representation I_i, we calculate the mutual information between feature $e_i^j, j \in [1, d_i]$ and K principle components $PC_i = \{pc_i^1, pc_i^2, \ldots, pc_i^K\}$ as feature selection metric, and store the mutual information between features and PC_i with matrix MI_i. The matrix $MI_i \in \mathbf{R}^{d_i \times K}$ is defined as:

$$\begin{aligned} &MI_i(j,k) = I(e_i^j; pc_i^k), \\ &\forall j \in \{1, \ldots, d_i\}, \forall k \in \{1, \ldots, K\} \end{aligned} \tag{2}$$

We formalize the intra-modality optimization as an integer linear programming problem, which is a mathematical optimization problem in which some of the variables are required to be integers (Graver, 1975). Its optimization objective is to maximize the intra-modality mutual information, i.e., $\mathscr{O}_{intra}$, as follows:

$$\begin{aligned} Max \quad & \mathscr{O}_{intra} = \sum_{i=1}^{M} \sum_{j=1}^{d_i} \sum_{k=1}^{K} MI_i(j,k) X_{I_i,j} \\ s.t. \quad & X_{I_i,j} \in \{0,1\}, \\ & \forall i \in \{1, \ldots, M\}, \forall j \in \{1, \ldots, d_i\} \end{aligned} \tag{3}$$

We adopt Gurobi Optimizer [2] to solve for the solution to Equation (3).

Inter-modality Mutual Information Maximization In contrast to the feature selection based on intra-modality mutual information maximization, inter-modality mutual information maximization focuses on the inter-feature relevance across modalities. In this part, we select features in the modal representation that maximize mutual information across modalities to reduce feature dimension for each modal representation. The inter-modality mutual information is calculated based on the selected features from intra-modality optimization. Given two modal representations I_i and I_j $(i \neq j)$, their mutual information matrix MI_{ij} is defined as:

$$MI_{ij}(\hat{i}, k) = I(e_i^{\hat{i}}; pc_j^k), \tag{4}$$

$$\forall \hat{i} \in \{1, \ldots, d_i\}, \forall k \in \{1, \ldots, K\} \tag{5}$$

where $pc_j^k \in PC_j$ and PC_j contains K principle components of modal representation I_j. $\hat{i}$ denotes feature index of modal representation I_i. Similarly, the inter-modality feature selection can be viewed as an integer linear programming problem and its objective function of inter-modality model is formalized as:

$$\begin{aligned} Max \quad & \mathscr{O}_{inter} = \sum_{I_i, I_j \in s_l, i \neq j} \sum_{\hat{i}=1}^{d_i} \sum_{k=1}^{K} MI_{ij}(\hat{i}, k) X_{I_i,\hat{i}} \\ s.t. \quad & X_{I_i,\hat{i}} \in \{0,1\}, \\ & \forall i, j \in \{1, \ldots, M\}, \forall \hat{i} \in \{1, \ldots, d_i\} \end{aligned} \tag{6}$$

where $s_l \in D$. Similarly, we adopt Gurobi Optimizer to solve for the solution to Equation (6). Feature selection based on intra-modality and inter-modality mutual information maximization tends to select features at identical positions across different modalities, which limits the flexibility and diversity. In order to select various features, we introduce the regularization terms $\mathscr{L}_1$ and $\mathscr{L}_2$ for intra-modality and inter-modality optimization respectively:

$$\begin{aligned} \mathscr{O}_{intra} &= \sum_{i=1}^{M} \sum_{m=1}^{d_i} \sum_{n=1}^{K} MI_i(j,k) X_{I_i,j} - \mathscr{L}_1 \\ \mathscr{O}_{inter} &= \sum_{I_i, I_j \in s_l, i \neq j} \sum_{\hat{i}=1}^{d_i} \sum_{k=1}^{K} MI_{ij}(\hat{i}, k) X_{I_i,\hat{i}} - \mathscr{L}_2 \end{aligned} \tag{7}$$

where $\mathscr{L}_1 = \sum_i \sum_{j=1}^{d_i} X_{I_i,j}$, $\mathscr{L}_2 = \sum_i \sum_{j=1}^{d_i} X_{I_i,\hat{i}}$. To this, we obtain F1 feature $F1 = \{F_{I_1}^1, \cdots, F_{I_M}^1\}$ for each sample, where $F_{I_i}^1 \in \mathbf{R}^{\hat{d}_i}$ denotes the selected feature from modal representation I_i, and $\hat{d}_i$ denotes the number of selected features in the first stage.

3.4 Feature Selection with Bayesian Network

First stage feature selection based on mutual information focuses on the correlation among features, ignoring the cross-modal interactions in fusion process. To address this problem, we introduce

[2] https://pypi.org/project/gurobipy/

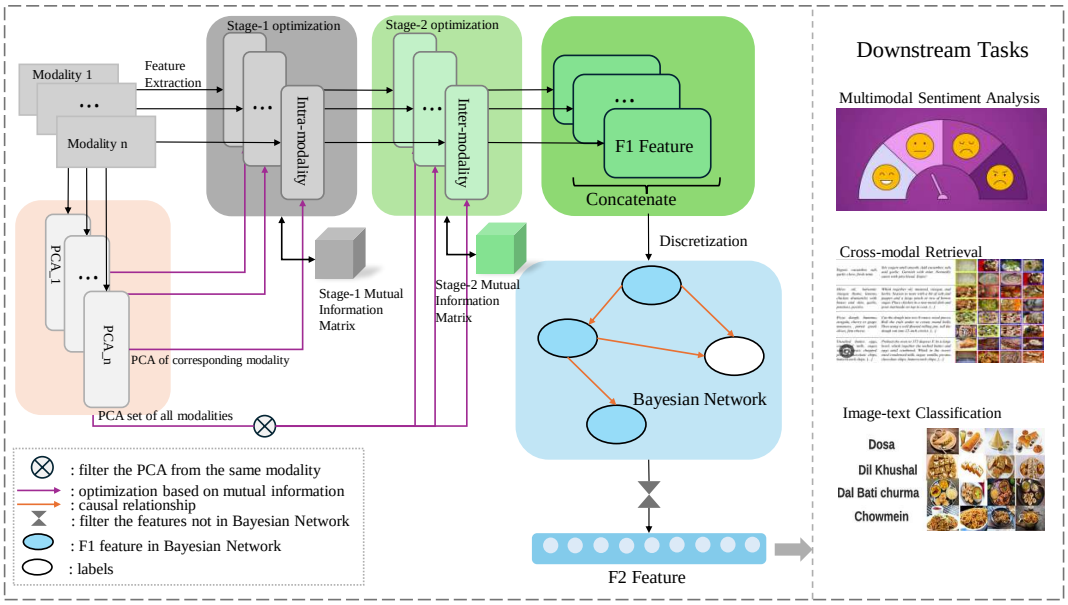


Figure 2: The overview of multimodal feature selection.

# Task	# Dataset	# Total Instances	# Number of Annotations	# Modality
Multimodal Sentiment Analysis	CH-SIMS	1825	3	A+V+T
	CH-SIMS v2	4403	3	A+V+T
	MOSI	2199	2	A+V+T
Cross-modal Retrieval	Wikipedia	2866	10	V+T
Image-text Classification	UMPC Food 101	90686	101	V+T

Table 1: The details of datasets in different multimodal tasks. A, V, T denote Audio, Vision and Text modality.

Bayesian Network (BN) as to discover the causal relationships between features, selecting features that have causal relationships with labels and other features. BN is a probabilistic graphical model, which can be used for modeling direct causal dependence between variables (Kitson et al., 2023).

We concatenate the selected features in $F1$, i.e., $[F^1_{I_1}, \cdots, F^1_{I_M}]$, where $[,]$ is the concatenation operation. Assuming $F_{I_i} \in \mathbf{R}^{1\times\hat{d}_i}$, then we have $\sum_i(\hat{d}_i)$ features after concatenation for one sample, where $\hat{d}_i$ denotes the dimension of I_i after first-stage feature selection. We take these features as nodes in BN, modeling the causal relation among features. Additionally, we also consider sample labels $\{y_1, \cdots, y_N\}$ as the nodes, aiming to capture the feature-to-label dependencies. Due to the different distributions of continuous feature and discrete label, we discretize the features in $\hat{X}$ using KDE [3] before BN structure learning.

After discretization, we adopt hill climbing algorithm [4] to learn network structure from data, where hill climbing algorithm takes BIC score as the score function to measures how much a given variable is "influenced" by a given list of potential parents. BIC score is generally used as a score metric for Bayesian structure scoring for Bayesian Networks with Dirichlet priors. For clarity, the algorithm of learning BN from modality feature with Hill Climing is shown in Technical Appendix. With BN, we can capture the interactions between features and their influence on the labels $y_1, \cdots, y_N$ which aids in discarding useless information. Thus, we can capture the causal relationships of feature to feature and feature to label, further select those informative features.

After two-stage feature selection, F2 features $F2 = \{F^2_{I_1}, \cdots, F^2_{I_M}\}$ are selected from multimodal representation. In the training phase, we concatenate the features in F2 as the fused modal representation and feed it into a Multi-layer Perceptron classifier for training downstream tasks. Dur-

[3] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.gaussian_kde.html

[4] https://pgmpy.org/structure_estimator/hill.html

ing the inference phase, as labels are not available, we fuse the features in modal representation based on the indices of features that are selected in the training phase as the modal fusion representation to infer.

4 Experiments

4.1 Tasks and Datasets

The proposed multimodal feature selection is model-agnostic and highly flexible, making it applicable to multimodal learning task. We conduct extensive experiments across **multimodal sentiment analysis** on CH-SIMS (Yu et al., 2020), CH-SIMS v2 (Liu et al., 2022), MOSI datasets (Zadeh et al., 2016), **image-text retrieval** on Wikipedia dataset (Rasiwasia et al., 2010), and **image-text classification** on UMPC Food 101 dataset (Wang et al., 2015). The details of the tasks and datasets are shown in Table 1.

Multimodal Sentiment Analysis Multimodal sentiment analysis (MSA) predicts sentiment polarity using audio, visual, and text modalities. We evaluate MFS performance on three datasets: **CH-SIMS**, a Chinese MSA dataset; **CH-SIMS v2**, its extended version; and **MOSI**, a widely used MSA benchmark.

Cross-modal Retrieval Cross-modal retrieval (CR) is the process of finding the relevant items in one modality based on the query in another modality. We use the **Wikipedia-based** dataset (Rasiwasia et al., 2010) to test the performance of MFS in cross-modal information retrieval task.

Image-text Classification Image-text classification task (ITC) involves using both visual and textual information to classify the given information into 101 categories. We evaluate the performance of MFS on **UMPC Food 101** dataset (Wang et al., 2015).

4.2 Evaluation Metrics

Following previous work, we adopt accuracy (Acc) and weighted F1 score to evaluate the performance of MFS on CH-SIMS, CH-SIMS v2, UPMC Food 101 datasets, and we adopt binary classification accuracy (Acc-2,i.e., positive and negative sentiment polarity) and F1 as the metric on MOSI dataset (Han et al., 2021; Yang et al., 2023; Gallo et al., 2020). For cross-modal retrieval task, we use the mean average precision (MAP) and Precision@10 (the average precision of the first 10 retrieved items) as the evaluation metric.

4.3 Implementation Details

For the CH-SIMS and CH-SIMS v2 datasets, we utilize Wav2Vec (Schneider et al., 2019), CLIP (Radford et al., 2021b), and BERT (Devlin et al., 2018) to extract modal representation from raw audio, vision, and text data, respectively. For the MOSI dataset, we employ AST (Gong et al., 2021), ViT (Radford et al., 2021a), and T5 (Raffel et al., 2020) for extracting features from audio, visual, and text data, respectively. For the Wikipedia dataset, we extract visual and text modalities using CLIP and BERT, respectively. For UMPC Food 101 dataset, we utilize CLIP and BERT to extract visual and text data, respectively. For more implementation detail, please refer to Technical Appendix. Additionally, we also offer bayesian network visualization on multiple datasets in Technical Appendix.

4.4 Baselines

Since MFS can be used for improving multimodal fusion based on existing methods, we then compare it with intuitive multimodal fusion methods like Concatenation, Add, Element-wise Maximum, Linear Fusion and MLP. The details of baselines lare as follows:

- Concatenation: simply concatenate feature vectors from different modalities into one feature vector.
- Add: add feature vectors from different modalities into one feature vector.
- Element-wise Maximum: combine different modal representations into one vector by selecting the element-wise maximum feature among modal representations.
- Linear Fusion: apply linear projection after concatenating different modal representations.
- MLP: apply two-layer perceptron to fuse features after concatenating different modal representations.

4.5 Main Results

We compare the proposed MFS with existing fusion methods under the same experimental settings, including single-modality results as shown in Table 2.

#Model	#Multimodal Sentiment Analysis						#Cross-modal Retrieval		#Image-text Classification	
	CH-SIMS		CH-SIMS v2		MOSI		Wikipedia		UMPC Food 101	
	Acc	Weighted F1	Acc	Weighted F1	Acc-2	F1	MAP	Precison@10	Acc	F1
Single-Modal(Audio)	48.58	44.31	45.36	45.07	51.44	52.72	-	-	-	-
Single-Modal(Visual)	65.86	65.97	64.89	64.24	51.75	54.06	-	-	71.37	71.35
Single-Modal(Text)	64.33	63.20	44.39	37.38	70.21	72.38	-	-	69.92	69.89
Concatenation	69.15	67.26	64.41	62.70	74.64	76.16	62.84	56.29	82.43	80.49
Add	67.61	66.13	64.60	62.43	73.58	74.67	61.60	55.05	81.67	81.66
Element-wise Maximum	68.16	65.16	64.12	61.87	71.09	74.89	61.99	56.96	80.11	80.09
Linear-Fusion	66.52	64.58	65.18	62.68	73.72	74.89	61.60	54.14	77.10	77.05
MLP	66.30	65.41	65.18	62.69	74.16	75.88	61.71	55.06	84.82	84.85
MFS*	**72.43**	**69.06**	**66.34**	**64.08**	**80.32**	**82.01**	**63.35**	**57.82**	**85.86**	**85.84**

Table 2: Comparison of baseline models and MFS on various datasets and tasks.

#Ablation Stage	#Multimodal Sentiment Analysis						#Cross-modal Retrieval		#Image-text Classification	
	CH-SIMS		CH-SIMS v2		MOSI		Wikipedia		UMPC Food 101	
	Acc	Weighted F1	Acc	Weighted F1	Acc-2	F1	MAP	Precison@10	Acc	Weighted F1
Stage-1 Ablation	68.71	67.23	64.60	62.50	75.64	77.81	57.33	49.28	82.69	80.71
Stage-2 Ablation	68.27	67.25	65.67	63.54	77.17	78.70	62.46	56.42	83.91	83.89
MFS*	**72.43**	**69.06**	**66.34**	**64.08**	**80.32**	**82.01**	**63.35**	**57.82**	**85.86**	**85.84**

Table 3: Ablation studies on the two stages in MFS.

Performance on Multimodal Sentiment Analysis Task MFS achieves significant improvements across three datasets on MSA task. MFS improves the metric of Acc at least 3.28%, 1.16% and 5.68% on CH-SIMS, CH-SIMS and MOSI datasets, respectively. Regarding Weighted F1 (F1) metric, MFS gains at least 1.8% on CH-SIMS, 1.38% on CH-SIMS v2 and 5.85% on MOSI improvements.

Performance on Cross-modal Retrieval Task As shown in the results of cross-modal retrieval task, MFS obtains the highest MAP of 63.35% and precision@10 of 57.82%, achieving on average 1.4% and 2.3% improvements in MAP and precision@10 respectively.

Performance on Image-text Classification Task MFS obtains the best performance with 85.86% in accuracy and 85.84% in weighted F1 score with an average of 4.6% and 5% improvements in accuracy and weighted F1 score respectively compared with baselines.

In summary, MFS outperforms the previous modal fusion methods spanning MSA, CR and ITC tasks. The results illustrate the superiority of MFS and demonstrate the effectiveness of performing feature selection before modal fusion.

4.6 Ablation Study

We further conduct an ablation study to validate the effectiveness of each stage in the proposed feature selection method. Table 3 reports the results. First, we remove the first stage of MFS, i.e., F1 feature selection with mutual information, and directly select F2 features using a bayesian network (BN) on the original features. For the sentiment analysis and image-text classification tasks, accuracy and weighted F1 scores decreases by an average of 3% and 2.8%, respectively. For the cross-modal retrieval task, mean average precision (MAP) drops by 6.02%, and Precision@10 drops by 8.54%. These findings suggest that F1 feature selection with mutual information is crucial for selecting informative features. Second, we eliminate the second stage of MFS. We use the F1 features to fuse modal representations and feed them into the model for downstream tasks, resulting in performance drop across all three downstream tasks. For instance, in the multimodal sentiment analysis and image-text classification tasks, performance drops by at least 1.5% on average across all metrics. These results indicate that F2 feature selection with BN can further effectively filter out unrelated features based on the F1 features. Overall, these results demonstrate the necessity of the two-stage feature selection process.

4.7 Visualization of Feature Importance

To verify the importance of the selected features, we visualize the feature importance with original features and selected features using a Random For-

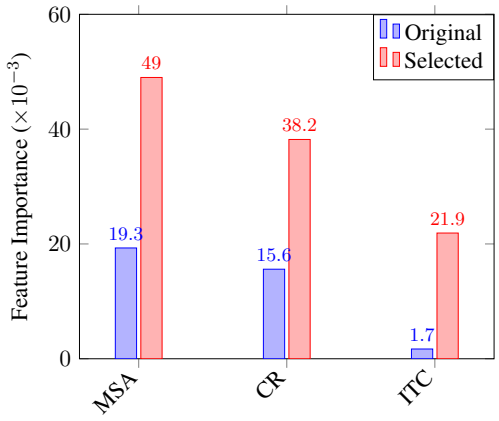
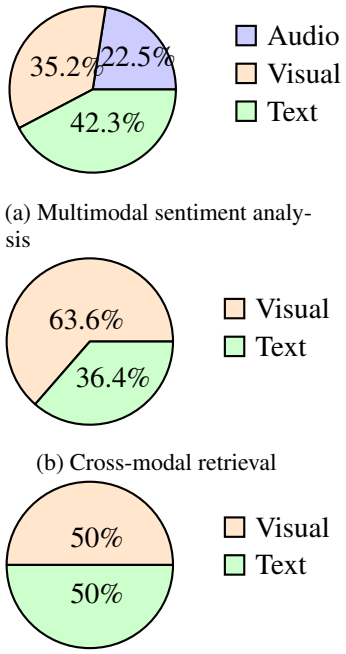


Figure 3: Feature importance for various tasks. MSA represents multimodal sentiment analysis, CR represents cross-modal retrieval and ITC represents image-text classification.



(a) Multimodal sentiment analysis

(b) Cross-modal retrieval

(c) Image-text classification

Figure 4: Distribution of features selected from different modalities for each task.

est Classifier (RFC)[5], as shown in Figure 3. Random Forest is a widely used non-linear method for assessing feature importance (Saarela and Jauhiainen, 2021). Specifically, we input the features and corresponding labels to train the RFC model, which then measures the importance of each feature in achieving correct classifications. It can be observed that the classifier values the features after our two-stage feature selection more than the original features across all modalities and tasks. This demonstrates that MFS still retains features that have significant impacts on model performance after discarding redundant features of modal representation.

4.8 Distribution of Selected Features

We further visualize the distribution of features after two-stage selection to analyze how different modalities contribute differently across downstream tasks. The results are shown in Figure 4. In the multimodal sentiment analysis task, 42.3% of the F2 features come from the text modality, 35.2% from the visual modality, and only 22.5% from the audio modality. For the cross-modal retrieval task, F2 features are primarily selected from the visual modality. Unlike these tasks, which are dominated by a specific modality, both visual and text modalities play significant roles in the image-text classification task. These findings indicate that the difference among downstream tasks indeed influences the dimensionality or features required from each modality for multimodal fusion. Furthermore, they demonstrate that the proposed MFS can dynamically select features based on the specific requirements of downstream tasks by shedding light on causal relationship and conditional dependence among modalities with Bayesian Network, making the modality fusion process more flexible and better suited to the task at hand.

[5]https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

5 Conclusion

This paper proposes a novel two-stage multimodal feature selection (MFS) method to improve multimodal fusion. The proposed MFS effectively selects informative features (F1 and F2 features) to reduce the dimensionality of modal representations. It further enables dynamic feature selection, allowing automatic adjustments of features based on different downstream tasks. We conduct extensive experiments across three multimodal tasks and five datasets, spanning vision, language, and audio modalities, and achieve substantial improvements on all datasets. This demonstrates the effectiveness and generalizability of the proposed two-stage multimodal feature selection. We believe this work establishes a new experimental framework that offers a fresh perspective to the multimodal fusion research community.

6 Limitations

The correct functioning of the MFS model requires tuning multiple parameters. For each downstream task, we design and optimize these parameters based on extensive experimental results. Although this process is currently manual, it can be automated using a self-adapting search algorithm to streamline the parameter selection process and improve performance. Automating this step would allow for more efficient exploration of parameter spaces, potentially enhancing the model's adaptability and generalization across different tasks without the need for manual intervention. This opens up possibilities for further improving the scalability and robustness of the MFS model in diverse applications.

References

Huda AlAmri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K. Marks, Chiori Hori, Peter Anderson, Stefan Lee, and Devi Parikh. 2019. Audio visual scene-aware dialog. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 7558–7567. Computer Vision Foundation / IEEE.

Jun-Ho Choi and Jong-Seok Lee. 2019. Embracenet: A robust deep learning architecture for multimodal classification. *Information Fusion*, 51:259–270.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ning Ding, Sheng-wei Tian, and Long Yu. 2022. A multimodal fusion method for sarcasm detection based on late fusion. *Multimedia Tools and Applications*, 81(6):8597–8616.

Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, and Hang Zhao. 2023. On uni-modal feature learning in supervised multi-modal learning. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 8632–8656. PMLR.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*.

Ignazio Gallo, Gianmarco Ria, Nicola Landro, and Riccardo La Grassa. 2020. Image and text fusion for upmc food-101 using bert and cnns. In *2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pages 1–6. IEEE.

Yuan Gong, Yu-An Chung, and James R. Glass. 2021. AST: audio spectrogram transformer. In *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, Brno, Czechia, August 30 - September 3, 2021*, pages 571–575. ISCA.

Jack E Graver. 1975. On the foundations of linear and integer linear programming i. *Mathematical Programming*, 9(1):207–226.

Michael Greenacre, Patrick JF Groenen, Trevor Hastie, Alfonso Iodice d'Enza, Angelos Markos, and Elena Tuzhilina. 2022. Principal component analysis. *Nature Reviews Methods Primers*, 2(1):100.

Nitin Grover, Aviral Chharia, Rahul Upadhyay, and Luca Longo. 2023. Schizo-net: A novel schizophrenia diagnosis framework using late fusion multimodal deep learning on electroencephalogram-based brain connectivity indices. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:464–473.

Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 9180–9192. Association for Computational Linguistics.

Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. 2021. What makes multi-modal learning better than single (provably). In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 10944–10956.

Neville Kenneth Kitson, Anthony C Constantinou, Zhigao Guo, Yang Liu, and Kiattikun Chobtham. 2023. A survey of bayesian network structure learning. *Artificial Intelligence Review*, 56(8):8721–8814.

Hoai-Duy Le, Guee-Sang Lee, Soo-Hyung Kim, Seungwon Kim, and Hyung-Jeong Yang. 2023. Multi-label multimodal emotion recognition with transformer-based fusion and emotion-level representation learning. *IEEE Access*, 11:14742–14751.

Paul Pu Liang, Yiwei Lyu, Gunjan Chhablani, Nihal Jain, Zihao Deng, Xingbo Wang, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2022. Multiviz: An analysis benchmark for visualizing and understanding multimodal models. *CoRR*, abs/2207.00056.

Yihe Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao. 2022. Make acoustic

and visual cues matter: Ch-sims v2. 0 dataset and av-mixup consistent module. In *Proceedings of the 2022 international conference on multimodal interaction*, pages 247–258.

Juan-Manuel Pérez-Rúa, Valentin Vielzeuf, Stéphane Pateux, Moez Baccouche, and Frédéric Jurie. 2019. Mfas: Multimodal fusion architecture search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6966–6975.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021a. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021b. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. 2010. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 251–260.

Mirka Saarela and Susanne Jauhiainen. 2021. Comparison of feature importance measures as explanations for classification models. *SN Applied Sciences*, 3(2):272.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Sho Shimamura and Kouichi Hirata. 2019. The reselection of adjacent sets by consistency-based feature selection algorithm. In *Proceedings of the 2nd international conference on information science and systems*, pages 210–214.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Dipti Theng and Kishor K. Bhoyar. 2024. Feature selection techniques for machine learning: a survey of more than two decades of research. *Knowl. Inf. Syst.*, 66(3):1575–1637.

Jesse Thomason, Daniel Gordon, and Yonatan Bisk. 2019. Shifting the baseline: Single modality performance on visual navigation & QA. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1977–1983. Association for Computational Linguistics.

Chun-Chun Wang, Chi-Chi Zhu, and Xing Chen. 2022. Ensemble of kernel ridge regression-based small molecule–mirna association prediction in human disease. *Briefings in Bioinformatics*, 23(1):bbab431.

Weiyao Wang, Du Tran, and Matt Feiszli. 2020. What makes training multi-modal classification networks hard? In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 12692–12702. Computer Vision Foundation / IEEE.

Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frederic Precioso. 2015. Recipe recognition with large multimodal food dataset. In *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE.

Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. 2023. Confede: Contrastive feature decomposition for multimodal sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7617–7630.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3718–3727.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.

Technical Appendix

.1 MFS Algorithm

MFS contains two-stage feature selection and each stage is responsible for specific-type feature selection. The full process of MFS is shown as follows:

Algorithm 1 Algorithm of MFS

Require: Data set D, Number of features to select for each modal in feature reduction $\hat{d}_i$, Number of PCA components K, Number of bins for discretization n_{bins}

Ensure: Multimodal Feature $F2 = \{F^2_{I_1}, \cdots, F^2_{I_M}\}$

Stage 1: F1 feature selection
for each modality $I_i \in \{I_1, I_2, \ldots, I_M\}$ **do**
  Compute PCA for modality I_i: $PC_i \leftarrow PCA(I_i)$
  Calculate mutual information: $\mathbf{MI}_i(j,k) = I(e^j_i, pc^k_i)$
  Solve optimization objective with Gurobi Optimizer: Equation (3)
end for
for each pair of modalities (I_i, I_j) where $i \neq j$ **do**
  Compute PCA for modality I_i
  Calculate mutual information: $\mathbf{MI}_{ij}(\hat{i},k) = I(e^{\hat{i}}_i, pc^k_j)$
  Solve optimization objective with Gurobi Optimizer: Equation (6) to obtain F1 feature
end for
Stage 2: F2 Feature Selection
Concatenate the selected features in $F1$
Discretization of labels
Adopt hill climbing algorithm to learn network structure from data
Filter out the features without causality and output F2 feature
return $F2 = \{F^2_{I_1}, \cdots, F^2_{I_M}\}$

.2 Discretization

The label vector L contains discrete values while features in $\hat{X}$ are continuous, thus we need to discretize the features to learn the BN structure from D. When it comes to discretization for BN, one intuitive method is K-means clustering, other common method includes Minimum Description Length (MDL) algorithm. But we propose another method of using Kernel Density Estimation (KDE) to discretize the continuous feature values. KDE doesn't assume any underlying distribution of the data, unlike MDL, which relies on predefined models. Also KDE provides a smoother estimation of the probability density function than K-means, which helps in capturing the underlying structure of the data more precisely.

Given the jth feature x^j_i of the ith modality, $x^j_i = \{c_1, \ldots, c_s\}$, perform KDE to estimate the probability density function (PDF). The KDE is given :

$$\hat{f}(c) = \frac{1}{sh} \sum_{i=1}^{s} K\left(\frac{c - c_i}{h}\right) \tag{8}$$

where K is the kernel function, h is the band width and s is the number of data points. Then we define a range of points $T = \{t_1, \ldots, t_m\}$ that spans the minimum and maximum values of x^j_i:

$$t_j = x_{\min} + j\left(\frac{x_{\max} - x_{\min}}{m-1}\right), \quad j = 0, 1, \ldots, m-1 \tag{9}$$

where m is a large number of points. We then compute the estimated density values $\hat{f}(t_j)$ for each point t_j in T and construct the cumulative distribution function (CDF) from the estimated density values:

$$\hat{F}(t_j) = \sum_{k=1}^{j} \hat{f}(t_k)\Delta t \tag{10}$$

where Δt is the spacing between points t. We define the bin edges $B = \{b_0, b_1, \ldots, b_{n_{\text{bins}}}\}$ based on the quantiles of the CDF:

$$b_k = \hat{F}^{-1}\left(\frac{k}{n_{\text{bins}}}\right), \quad k = 0, 1, \ldots, n_{\text{bins}} \tag{11}$$

where n_{bins} is the desired number of bins. Assign each value c_i in x^j_i to a bin based on the bin edges:

$$d_i = \sum_{k=1}^{n_{\text{bins}}} \mathbf{1}_{[b_{k-1}, b_k)}(c_i) \tag{12}$$

where $\mathbf{1}_{[b_{k-1}, b_k)}(c_i)$ is the indicator function that returns 1 if c_i is within the interval $[b_{k-1}, b_k)$, and 0 otherwise.

.3 Implementation Details

For clarity, we give implementation details of MFS on multimodal sentiment analysis, cross-modal retrieval and image-text classification tasks. The hyperparameters on five datasets are reported in Table 4.

Algorithm 2 Learning BN from Modality Features with Hill Climbing

Require: Data set D, Score function BIC, Maximum iterations max_iter
Ensure: Bayesian Network Structure BN

```
Initialize:
    BN ← Empty Network
    current_score ← BIC(BN, D)
    iteration ← 0
while convergence not reached and iteration < max_iter do
    iteration ← iteration + 1
    best_neighbor ← None
    best_score ← current_score
    Generate Neighbors:
    for each possible single edge addition, deletion, or reversal in BN do
        neighbor ← Apply the operation to BN
        neighbor_score ← BIC(neighbor, D)

        if neighbor_score > best_score then
            best_neighbor ← neighbor
            best_score ← neighbor_score
        end if
    end for
    if best_score > current_score then
        BN ← best_neighbor
        current_score ← best_score
    else
        break    // Local maximum found
    end if
end while
return BN
```

Multimodal Sentiment Analysis For the CH-SIMS and CH-SIMS v2 datasets, we utilize Wav2Vec, CLIP, and BERT to extract feature vectors from raw audio, vision, and text data, respectively. After feature extraction, we project them into a lower-dimensional space of 64 dimensions $d_i = 64$ using a linear activation function for modal representation I_i. For CH-SIMS, we extract 15-dimension features (i.e., $K = 15$) as PCA module outputs for each modality and the dimension of F1 for each modality is set to 30, i.e., $\hat{d}_i = 30, \forall i \in \{t, a, v\}$. In discretization phase, the sample rate is set $n_{bins} = 15$. For CH-SIM v2, we set $d_i = 64$,$K = 16$,$\hat{d}_i = 25$, and $n_{bins} = 15$ in MFS. For the MOSI dataset, we employ AST, ViT, and T5 models for extracting features from audio, visual, and text data, respectively. We set $d_i = 64$,$K = 20$,$\hat{d}_i = 30$,and $n_{bins} = 15$ in MFS.

.3.1 Cross-modal Retrieval

For the Wikipedia dataset, we extract features from both images and text using CLIP and BERT, respectively. After obtaining these features, we project them into a 64-dimensional feature space using a linear activation function. The hyperparameters are set to $\hat{d}_i = 45$, $K = 16$, and $n_{bins} = 10$.

.3.2 Image-text Classification

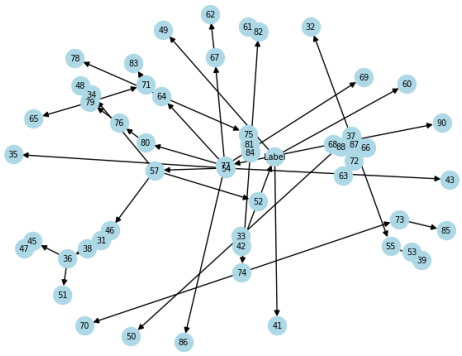
For UMPC Food 101 dataset, we utilize CLIP and BERT to extract features from image and text data, respectively. These extracted features are then projected into an 80-dimensional feature space using a linear activation function. The hyperparameters are set to $\hat{d}_i = 32$, $K = 17$, and $n_{bins} = 15$.

.4 Visualization of Bayesian Network

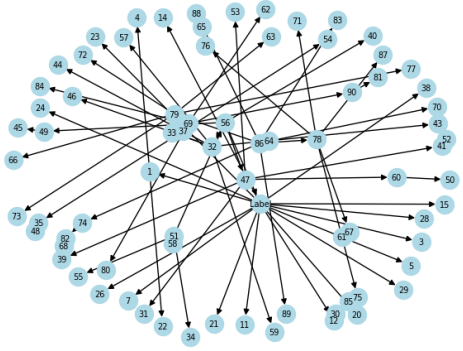
We visualize the bayesian network of CH-SIMS, CH-SIMS v2, Wikipedia and UMPC Food 101 datasets, and number of nodes mean the feature index. The visualizations are given in Figure 5.

Table 4: Hyperparameters for Each Dataset.

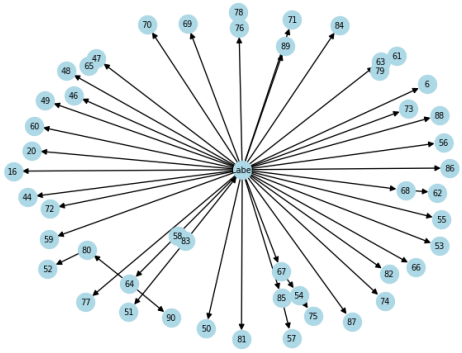
	#Multimodal Sentiment Analysis			#Cross-modal Retrieval	#Image-text Classification
	CH-SIMS	**CH-SIMS v2**	**MOSI**	**Wikipedia**	**UMPC Food 101**
$d_i, \forall i \in t, a, v$	64	64	64	64	64
K	15	17	20	16	17
$\hat{d}_i, \forall i \in t, a, v$	30	30	30	45	32
n_{bins}	15	15	15	10	15



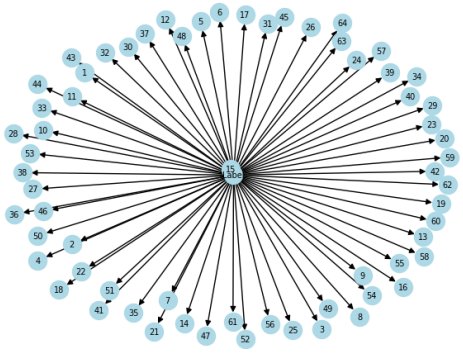
(a) Bayesian Network Visualization on CH-SIMS.



(b) Bayesian Network Visualization on CH-SIMS v2.



(c) Bayesian Network Visualization on Wikipedia.



(d) Bayesian Network Visualization on UMPC Food 101.

Figure 5: Bayesian Network visualization on different datasets.