Semi-Supervised Multimodal Classification Through Learning from Modal and Strategic Complementarities

Junchi Chen[1], **Richong Zhang**[1,3*], **Junfan Chen**[1,2]

[1] CCSE, School of Computer Science and Engineering, Beihang University, Beijing, China
[2] School of Software, Beihang University, Beijing, China
[3] Zhongguancun Laboratory, Beijing, China
chenjc@buaa.edu.cn, {zhangrc, chenjf}@act.buaa.edu.cn

Abstract

Supervised multimodal classification has been proven to outperform unimodal classification in the image-text domain. However, this task is highly dependent on abundant labeled data. To perform multimodal classification in data-insufficient scenarios, in this study, we explore semi-supervised multimodal classification (SSMC) that only requires a small amount of labeled data and plenty of unlabeled data. Specifically, we first design baseline SSMC models by combining known semi-supervised pseudo-labeling methods with the two most commonly used modal fusion strategies, i.e. feature-level fusion and label-level aggregation. Based on our investigation and empirical study of the baselines, we discover two complementarities that may benefit SSMC if properly exploited: the predictions from different modalities (modal complementarity) and modal fusion strategies for pseudo-labeling (strategic complementarity). Therefore, we propose a Modal and Strategic Complementarity (MSC) framework for SSMC. Concretely, to exploit modal complementarity, we propose to learn reliability weights for the predictions from different modalities and refine the fusion scores. To learn from strategic complementarity, we introduce a dual KL divergence loss to guide the balance of quantity and quality of pseudo-labeled data selection. Extensive empirical studies demonstrate the effectiveness of the proposed framework.

Introduction

Multimodal classification attracts intense interest in the research community. Traditional supervised models utilize a sufficient amount of labeled data and outperform unimodal classifiers (Abavisani et al. 2020; Agarwal et al. 2020; Kiela et al. 2018, 2019; Zou et al. 2023). However, in real-world scenarios, the collection and annotation of large amounts of multimodal data is expensive. A practical solution is semi-supervised learning, which can efficiently use easily accessible unlabeled data. In this paper, we explore a task called semi-supervised multimodal classification (SSMC), which extends semi-supervised learning to multimodal classification scenarios involving image and text modalities.

We first design baseline SSMC models by combining existing semi-supervised approaches and modality fusion methods. The most common semi-supervised learning approach, pseudo-labeling, is chosen as our backbone. Specifically, we leverage the well-known FixMatch (Sohn et al. 2020) and FreeMatch (Wang et al. 2022) as baseline semi-supervised learning methods. For modality fusion, we apply the classical label-level aggregation method (Score Fusion) that fuses the predictions from different modalities and the feature-level fusion method (Feature Concat) that first concatenates the features of varied modalities and then performs classification (Liang et al. 2022; Abavisani et al. 2020).

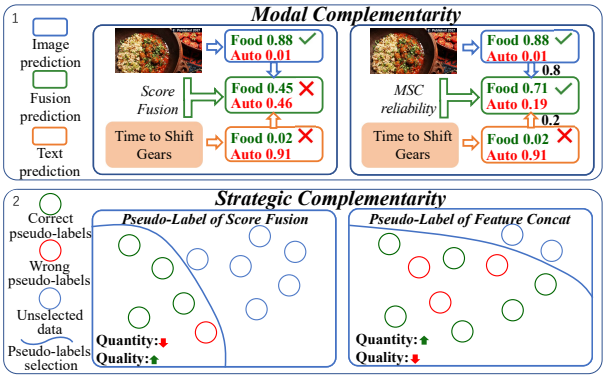

Figure 1: Illustration of the problems in SSMC: (1) The upper shows that existing probability level modality fusion methods encounter modality interference problems and our MSC solves this problem. (2) The lower shows the quality and quantity of pseudo-labels produced by varied strategies.

Based on our investigation of SSMC and baseline models, we discover the following two phenomena. First, both image and text modalities contribute to classification and they form complementarity in SSMC. The two modalities provide uneven information (Zou et al. 2023; Monter-Aldana, Monroy, and Vega 2023; Wang, Tran, and Feiszli 2020) and may cause *modality interference* problem when applying averaged Score Fusion. Under this problem, when a prediction from one of the modalities is incorrect, the fused prediction may also be wrong, even if the prediction from another modality is correct. As shown in Figure 1, this problem often arises when the classification results of different modalities

*Corresponding Author

are inconsistent, and an averaged score easily leads to wrong fusion results. This phenomenon may cause wrong predictions of pseudo-labels and limit the SSMC performance.

Second, pseudo-labels of different strategies display complementarity in SSMC. As shown in Figure 1, when filtering pseudo-labels with threshold, Score Fusion strategy can retain correct pseudo-labels with high accuracy, but the number of selected examples is significantly less than Feature Concat strategy. The situation is reversed for Feature Concat strategy, which retains more pseudo-labeled data but with lower accuracy. Since both the quality and quantity of pseudo-labeled examples are vital in semi-supervised learning, it is necessary to take the complementarity between the two strategies into account when building an SSMC model.

The above analysis motivates us that exploiting complementarities between different modalities and strategies may enhance SSMC performance. Thus, we propose a framework to learn from modal and strategic complementarities (MSC). Specifically, our modal complementarity learning module aims to assign a reliability weight to each modality, as shown in Figure 1. Unlike previous works that treat modality weights as hyper-parameters (Raj and Meel 2021; Chen et al. 2023) or calculate weights based on validation set (Al Obaid et al. 2023), we learn modal reliability weights from different individual instances with calibration from the consistency of predictions. We further leverage a KL divergence to force the less reliable modality to learn from the more reliable modality at the instance level, reducing error accumulation during training. In this way, we successfully solve modality interference problem to improve SSMC.

To learn from the strategic complementarity, we propose a pseudo-labeled data selection approach that utilizes the advantages of both Score Fusion and Feature Concat strategies. Concretely, this approach takes the union of selected data with the two strategies to train the MSC framework. The selected data is treated as guiding signals in a dual KL divergence loss, allowing the weaker strategy branch to learn from the stronger one. This learning process utilizes the strategic complementarity between strategies and makes a trade-off between the quality and quantity of the pseudo-labeled data, thus helping improve SSMC performance.

In summary, our work makes the following contributions: (1) We introduce semi-supervised multimodal classification problem and reveal the complementarities that may help improve SSMC. (2) We propose a modal complementarity learning approach to estimate modal reliability and reduce modality interference problem. (3) We present strategic complementarity learning to balance the quality and quantity of pseudo-labels. (4) We create three benchmarks for SSMC to evaluate our MSC framework. Experimental analyses demonstrate the effectiveness of our method.

Related Works

Semi-supervised Classification A common approach for semi-supervised classification is using regularized information. For example, UDA (Xie et al. 2020) and MixText (Chen, Yang, and Yang 2020) enhance consistency training through data augmentation. Another idea is pseudo-labeling to increase the amount of labeled data during training. For example, a series of works have developed starting with MixMatch (Berthelot et al. 2019; Sohn et al. 2020; Wang et al. 2022), which annotate unlabeled data and choose reliable ones with fixed or variable thresholds. There are approaches leveraging unlabeled data using pseudo-labeling combined with prototype learning (Yang et al. 2023), or constructing dictionaries based on attention (Lee, Ko, and Han 2021).

Multimodal Classification Multimodal classification has been studied in image-text domain (Agarwal et al. 2020; Li et al. 2019; Huang et al. 2020). Mainstream methods include late fusion and early fusion. Late fusion use encoders of different modalities and fuses extracted features. For example, UniS-MMC (Zou et al. 2023) designs splicing and comparison learning mechanisms. Different modalities are embedded with cross-attention mechanism (Abavisani et al. 2020). Early fusion focus on the association between different modal network layers. For example, MI2P (Liang et al. 2022) designs fine-grained plug-and-play modules before encoding features. Previous work has also been done on the significance of different modalities based on prior knowledge of modal strengths (Raj and Meel 2021) or the accuracy of the validation set (Al Obaid et al. 2023).

Preliminaries

Problem Formulation

Semi-supervised multimodal classification (SSMC) aims to classify image-text pairs based on a training set consisting of little labeled data and a large amount of unlabeled data. Formally, let $\mathcal{Z} = \{0, 1\}$ be the modality set, and $\mathcal{K}$ denotes the set of classes. In this work, $\mathcal{Z}$ includes the text and image modality. Let $\mathcal{D}$ be the training set containing a labeled set $\mathcal{X}$ and an unlabeled set $\mathcal{U}$. We denote the labeled set as $\mathcal{X} = \{t_i^l, v_i^l, y_i\}_{i=1}^N$, where t_i^l and v_i^l mean the text and image of the i^{th} data pair and y_i is its corresponding label. The unlabeled set is denoted as $\mathcal{U} = \{t_i^u, v_i^u\}_{i=1}^M$, where t_i^u and v_i^u denote the i^{th} unlabeled image-text pair. For simplicity, we also denote each labeled and unlabeled image-text pair as x_i and u_i. SSMC aims to predict the category of a given data pair with a limited labeled data set $\mathcal{D}$.

Baseline Multimodal Classification Models

Feature Extraction and Unimodal Predictions We follow the setup in previous works (Zou et al. 2023; Liang et al. 2022) that use BERT (Kenton and Toutanova 2019) to encode texts and ViT (Dosovitskiy et al. 2020) to encode images. Specifically, let $f(t)$ and $g(v)$ denote the encoders of text t and image v respectively, we adopt two MLPs $\mathcal{M}_T$ and $\mathcal{M}_V$ with softmax function to acquire probabilities p:

$$p_T(y|t) = \frac{\exp(\mathcal{M}_T(f(t), \phi_T^y))}{\sum_{k\in\mathcal{K}} \exp(\mathcal{M}_T(f(t), \phi_T^k))} \quad (1)$$

$$p_V(y|v) = \frac{\exp(\mathcal{M}_V(g(v), \phi_V^y))}{\sum_{k\in\mathcal{K}} \exp(\mathcal{M}_V(g(v), \phi_V^k))} \quad (2)$$

where ϕ_T^k and ϕ_V^k are parameters corresponding to class k. With the models in Equation (1) and (2), we can respectively make unimodal predictions from the text or image data.

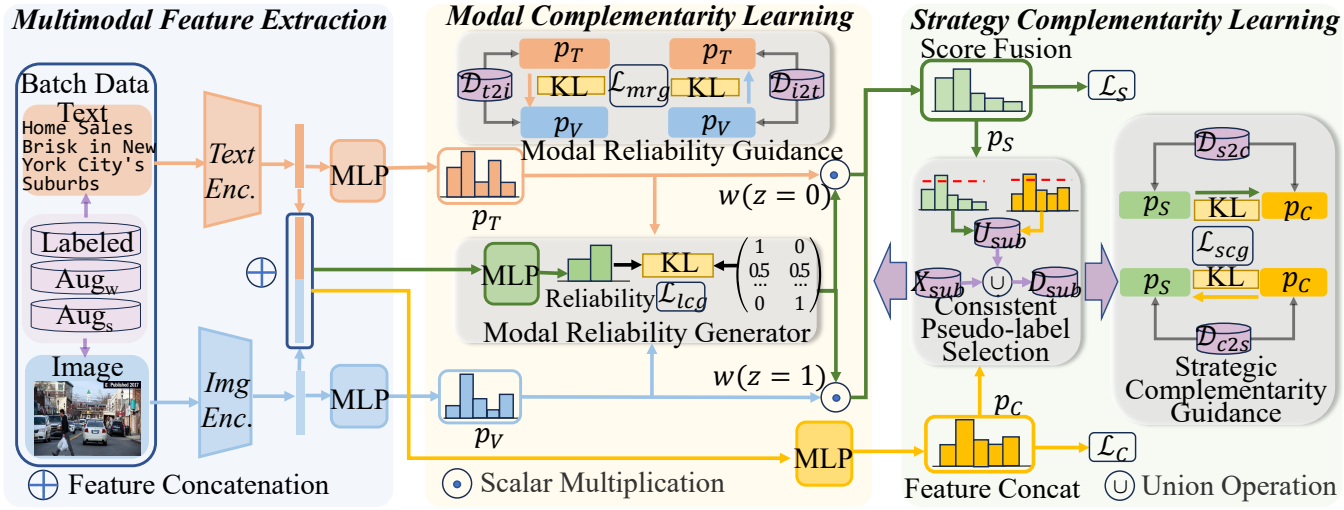

Figure 2: The structure of MSC framework. Modal complementarity learning involves Modal Reliability Generator containing Label Consistency Guidance (LCG) and Modal Reliability Guidance (MRG) module. Strategic complementarity learning includes Consistent Pseudo-label Selection and Strategic Complementarity Guidance (SCG) module.

Score Fusion and Feature Concat Strategies Based on extracted features or unimodal predictions, Score Fusion and Feature Concat can be used to perform multimodal classification. Score Fusion averages unimodal predictions:

$$p_S(y|t,v) = (p_T(y|t) + p_V(y|v))/2 \tag{3}$$

Feature Concat first concatenates extracted features from text and image modalities and then performs softmax classification on the concatenated features. Specifically, an MLP $\mathcal{M}_C$ and softmax function are used to obtain the prediction:

$$p_C(y|t,v) = \frac{\exp(\mathcal{M}_C([f(t),g(v)],\phi_C^y))}{\sum_{k\in\mathcal{K}}\exp(\mathcal{M}_C([f(t),g(v)],\phi_C^k))} \tag{4}$$

where ϕ_C^k are the parameters corresponding to class k. Score Fusion and Feature Concat are the most common modality fusion strategies used to perform multimodal classification, which will be used as baselines to build SSMC models.

Baseline Semi-Supervised Classification Models

We choose the commonly used semi-supervised classification models FixMatch and FreeMatch as our baselines. They both contain losses respectively for labeled data x_i and unlabeled data u_i, where the supervised loss $\mathcal{L}_{sup}$ is

$$\mathcal{L}_{sup} = \frac{1}{B_x}\sum_{i=1}^{B_x}\sum_{k\in K} y_i \log p(k|x_i) \tag{5}$$

where B_x is the batch size of sampled labeled data.

For each unlabeled data u_i, we assign it a pseudo-label and use it as additional training data. We prepare a weakly-augmented version $a(u_i)$ and a strongly-augmented version $\mathcal{A}(u_i)$ for u_i. We utilize the prediction of $a(u_i)$ to select high-quality pseudo-labeled data and calculate cross-entropy loss against $\mathcal{A}(u_i)$. Concretely, we determine the pseudo-label of u_i by $\hat{y}_i = \arg\max_{k\in\mathcal{K}}(p(k|a(u_i)))$ and select high-confidence pseudo-labels to calculate loss $\mathcal{L}_u$:

$$\mathcal{L}_u = \frac{1}{B_u}\sum_{i=1}^{B_u}\mathbf{1}(\max p(a(u_i)) \geq \tau)\sum_{k\in K}\hat{y}_i \log p(k|\mathcal{A}(u_i)) \tag{6}$$

where B_u is the batch size of unlabeled data and τ is a fixed or variable threshold to filter low-quality pseudo-labels.

Pseudo-labeling methods use $\mathcal{L}_{sup} + \lambda\mathcal{L}_u$ or add additional regularization as the loss for optimization. We use $\mathcal{L}_S$ and $\mathcal{L}_C$ to denote the loss when p is specified as p_S and p_C. We create baseline SSMC models by combining baseline multimodal classification models with the known semi-supervised methods.

Motivations

Through analyzing the baselines, we discover two complementarities that may be leveraged to improve SSMC. First, the predictions p_T and p_V from different modalities are complementary. Some predictions are dominant by one of the modalities and existing averaged Score Fusion (as shown in Equation (3)) may encounter *modality interference problem* that causes wrong fusion scores in prediction (as shown in the upper part of Figure 1). To address this problem, we propose to **learn from modal complementarity** that automatically learns reliability weights to weigh the predictions from different modalities and rectify the fusion scores.

Second, our empirical studies demonstrate that the Score Fusion strategy and Feature Concat strategy are complementary when selecting pseudo-labeled data in terms of their quality and quantity. Specifically, as shown in the lower part of Figure 1, score fusion p_S in Equation (3) tends to select less pseudo-labeled data but with high-quality. On the contrary, feature concat p_C in Equation (4) are prone to select more but lower-quality pseudo-labeled data. As the quality

and quantity of pseudo-labeled examples are both important for effective semi-supervised learning, we are motivated to **learn from strategic complementarity** so that we can make a trade-off between the quality and quantity of the selected pseudo-labeled data and allow the strategies to leverage strengths and offset weaknesses.

Method

Our empirical study from baselines reveals that there are modal complementarity and strategic complementarity in SSMC, which may improve performance when they are appropriately utilized. This section will present the design of our MSC framework (the structure is shown in Figure 2) and introduce how we exploit the two complementarities.

Learning from Modal Complementarity

As analyzed in previous sections, to solve *modality interference problem* in SSMC mentioned previously, we design a modal reliability generator to estimate the reliability of predictions from different modalities and use the stronger modality to guide the weaker one.

Modal Reliability Generation and Weighted Fusion Our solution to *modality interference problem* is to assign a weight indicating reliability for each modality and transform averaged fusion to weighted fusion. To acquire modal reliability weights, we propose a Modal Reliability Generator to produce a reliability score for each modality. Specifically, we introduce an MLP $\mathcal{M}_w$ followed by softmax function to obtain reliability distributions $w(z|t,v)$ for different modalities from concatenated feature $[f(t), g(v)]$ as follows:

$$w(z|t,v) = \frac{\exp(\mathcal{M}_w([f(t),g(v)],\phi_w^z))}{\sum_{z'\in\mathcal{Z}}\exp(\mathcal{M}_w([f(t),g(v)],\phi_w^{z'}))} \quad (7)$$

where $\phi_w^{z'}$ is the parameter corresponding to the modality z'. We use $w(z=0|t,v)$ and $w(z=1|t,v)$ to denote the reliability of text modality and image modality. For each image-text pair (t,v), we modify the averaged Score Fusion function in Equation (3) to weighted fusion function:

$$p_S(y|t,v) = w(z=0|t,v)p_T(y|t) + w(z=1|t,v)p_V(y|v) \quad (8)$$

This weighted fusion method can alleviate modality interference problem by assigning a small weight to the prediction of unreliable modality, thus improving SSMC performance. In MSC, we use Equation (8) for Score Fusion branch.

Weight-Bias Reduction via Label Consistency Guidance As previous work mentioned (Monter-Aldana, Monroy, and Vega 2023), strong and weak modalities have varied training convergence rates which result in modal weight-bias towards the stronger one and damage weighted fusion. To tackle this problem, we design a Label Consistency Guidance (LCG) method to calibrate the weight-learning process with limited labels. We use pseudo-labels collected in Strategic Complementarity Learning and labeled data in the same batch to form data set D_{sub} in one batch to train modal reliability generator. Specific rules will be introduced in the Learning from Strategic Complementarity section. We first obtain predicted results with strong-augment of unlabeled input as

$$r_t = \arg\max_{k\in\mathcal{K}} p_T(k|\mathcal{A}(t)), r_v = \arg\max_{k\in\mathcal{K}} p_V(k|\mathcal{A}(v)) \quad (9)$$

Then, we generate supervision signals according to the consistency of pseudo-labels produced from the strong and weak versions of input, which can be formulated as follows:

$$G(z=0|t,v) = \begin{cases} 1 & r_t = \hat{y}, r_v \neq \hat{y} \\ 0 & r_t \neq \hat{y}, r_v = \hat{y} \\ 0.5 & r_t \neq \hat{y}, r_v \neq \hat{y} \\ \frac{p_T(\hat{y}|t)}{p_T(\hat{y}|t)+p_V(\hat{y}|v)} & r_t = r_v = \hat{y} \end{cases} \quad (10)$$

Correspondingly, we have $G(z=1|t,v) = 1 - G(z=0|t,v)$. The intuition of G can be explained as follows:

- When one of the modal predictions r_t or r_v is consistent with pseudo-label $\hat{y}$, we treat the modality with consistent predictions as reliable one and thus assign a probability score 1 for this modality, and vice versa.
- When both modal predictions r_t and r_v are inconsistent with the pseudo-label $\hat{y}$, the predictions from the two modalities are more likely unreliable. Thus we let the contribution of the two modalities be divided equally.
- When both modal predictions r_t and r_v are consistent with the pseudo-label $\hat{y}$, we treat the modal prediction with higher confidence as the more reliable one. Thus, we utilize a softmax operation based on their corresponding maximum prediction to compute the probability score.

KL divergence is used to bring the generated reliability distribution $w(z|t,v)$ close to guidance $G(z|t,v)$. We denote this part of loss $\mathcal{L}_{lcg}$ as follows:

$$\mathcal{L}_{lcg} = \frac{1}{|\mathcal{D}_{sub}|}\sum_{i}^{|\mathcal{D}_{sub}|} KL(G(z|t_i,v_i)||w(z|t_i,v_i)) \quad (11)$$

where $|\mathcal{D}_{sub}|$ is the sum of the subset of the chosen union set in a batch. It is worth noting that we replace $\hat{y}$ with ground-truth labels y and use no-augmented data pairs $(t,v)\in\mathcal{X}$ to get predictions in Equation (9) and (10) for labeled data.

Modal Reliability Guidance Module Because of the small reliability weight, the gradient generated by the unreliable modality may be insufficient to train the corresponding branch. Thus, we allow the unreliable modality to learn from the more reliable one based on the reliability of the instance level and propose a Modal Reliability Guidance (MRG) module.

We adopt the same data $\mathcal{D}_{sub}$ used in LCG. The data subset is divided into two parts. Formally, we introduce $\mathcal{D}_{t2i}\in\mathcal{D}_{sub}$ to denote the data set on which text modality is more reliable and $\mathcal{D}_{i2t}\in\mathcal{D}_{sub}$ to denote the data set on which image modality is more reliable. These two sets can be explained with predicted results of text modality r_t and results of image modality r_v from Equation (9) as follows:

- For each data and label (t,v,y) in $\mathcal{D}_{sub}$, if r_t is consistent with y and r_v is inconsistent with y, or if both are consistent with y but the maximum prediction of p_T is higher than p_V , we add this data into $\mathcal{D}_{t2i}$ which forces p_V to learn from p_T on these data and vice versa for $\mathcal{D}_{i2t}$.

KL divergence is utilized here to complete the training process. We introduce $\mathcal{L}_{mrg}$ to denote the loss of MRG module as follows

$$\mathcal{L}_{mrg} = \frac{1}{|\mathcal{D}_{t2i}|}\sum_{i}^{|\mathcal{D}_{t2i}|} KL(p_T(y|t_i)||p_V(y|v_i))+ \frac{1}{|\mathcal{D}_{i2t}|}\sum_{i}^{|\mathcal{D}_{i2t}|} KL(p_V(y|v_i)||p_T(y|t_i)) \tag{12}$$

where $|\mathcal{D}_{t2i}|$ and $|\mathcal{D}_{i2t}|$ are the sums of two sets. Through the above design, we assign proper reliability weight to modalities and address the modality interference problem in SSMC.

Learning from Strategic Complementarity

Different fusion strategies possess varied performances on SSMC pseudo-label selection, which shows strategic complementarity. To leverage this, we introduce a consistent pseudo-label selection method to construct data sets for guidance information.

Consistent Pseudo-label Selection To make full use of unlabeled data, we propose to aggregate labeled data and unlabeled data with high confidence based on the consistency of the two strategies. For unlabeled data, we retain data on which two strategies predict consistent pseudo-labels and one of the maximum probabilities is higher than threshold η. Concretely, we denote the subset of unlabeled data of one batch $\mathcal{B}$ with $\mathcal{U}_{sub}$ as

$$\mathcal{U}_{sub} = \{(\mathcal{A}(u_i), \hat{y}_s)|(\hat{y}_s = \hat{y}_c) \wedge ((\max p_S(y|a(u_i))) \geq \eta) \vee (\max p_C(y|a(u_i))) \geq \eta) \wedge (u_i \in \mathcal{B})\} \tag{13}$$

where $\hat{y}_s$ and $\hat{y_c}$ are pseudo-labels of weighted Score Fusion and Feature Concat. This selection can be viewed as vote integration and achieves high accuracy. Then we use $\mathcal{X}_{sub}$= $\{(x_i, y_i)|(x_i, y_i) \in \mathcal{B}\}$ to denote all labeled data in the same batch and denote union set as $\mathcal{D}_{sub} = \mathcal{U}_{sub} \cup \mathcal{X}_{sub}$. This set is used in previous sections to form guidance loss and will be utilized in the Strategic Complementarity Guidance Module.

Strategic Complementarity Guidance Module To utilize the strategic complementarity of weighted Score Fusion and Feature Concat introduced in the former sections, we propose a Strategic Complementarity Guidance (SCG) module, promoting two strategies to learn from each other on the instance level and improve SSMC performance.

Specifically, we denote $\mathcal{D}_{s2c} \in \mathcal{D}_{sub}$ as the data on which Feature Concat learn from Score Fusion and denote $\mathcal{D}_{c2s} \in \mathcal{D}_{sub}$ on the contrary. Similar to Equation (9), we use strong-augment to collect $r_s = \arg\max_{k \in \mathcal{K}} p_S(k|\mathcal{A}(t), \mathcal{A}(v))$ and $r_c = \arg\max_{k \in \mathcal{K}} p_C(k|\mathcal{A}(t), \mathcal{A}(v))$ as unlabeled results of two strategies and use no-augment to obtain r_s and r_c for labeled data. The rules of $\mathcal{D}_{s2c}$ and $\mathcal{D}_{c2s}$ are expressed as

- For each data and label (t, v, y) in $\mathcal{D}_{sub}$, if r_s is consistent with y but r_c is inconsistent with y, or if both are consistent with y but the maximum prediction of p_S is higher than p_C , we add this data into $\mathcal{D}_{s2c}$ to push p_C to learn from p_S and vice versa for $\mathcal{D}_{c2s}$.

We use KL divergence to force the strategy with weaker performance to learn from the better one and obtain $\mathcal{L}_{scg}$:

$$\mathcal{L}_{scg} = \frac{1}{|\mathcal{D}_{s2c}|}\sum_{i}^{|\mathcal{D}_{s2c}|} KL(p_S(y|t_i, v_i)||p_C(y|t_i, v_i))+ \frac{1}{|\mathcal{D}_{c2s}|}\sum_{i}^{|\mathcal{D}_{c2s}|} KL(p_C(y|t_i, v_i)||p_S(y|t_i, v_i)) \tag{14}$$

where $|\mathcal{D}_{s2c}|$ and $|\mathcal{D}_{c2s}|$ are the sums of two sets. Through the SCG module, Score Fusion gains more pseudo-labeled data when keeping original quality and Feature Concat obtains higher quality of pseudo-labels when maintaining original quantity. Strategic complementarity makes a trade-off between quantity and quality of pseudo-labels for SSMC. It is worth noting that after applying modal reliability weights, the weighted score fusion still performs similarly to the averaged score fusion in selecting pseudo-labeled data. Further analysis and explanation are provided in the Appendix.

Training Objective

We train the MSC framework using the three loss functions mentioned above combined with the basic pseudo-labeling semi-supervised methods FixMatch and FreeMatch as

$$\mathcal{L} = \mathcal{L}_S + \mathcal{L}_C + \beta_1\mathcal{L}_{lcg} + \beta_2\mathcal{L}_{mrg} + \beta_3\mathcal{L}_{scg} \tag{15}$$

where β_1, β_2, β_3 are weight factors to balance the impact of different losses. In the process of evaluation, we take the average of predicted probabilities p_S and p_C from the MSC framework as final integration results.

Experiments

Datasets and Experiment settings

We select three datasets to construct benchmarks for SSMC: **N24News** (Wang et al. 2021), **UPMC-Food101** (Bossard, Guillaumin, and Van Gool 2014), **CrisisMMD** (Alam, Ofli, and Imran 2018). We reserve 18 classes with more than 1000 data pairs from N24News.We sample data pairs with the same labels on two modalities and only evaluate on task 1 on CrisisMMD. We split the training set into 20, 50, 100 labeled pairs. Dataset statistics are shown in Appendix. We use accuracy and F1 to assess SSMC performance. We use SF and FC as abbreviations of Score Fusion and Feature Concat.

Baselines

We select **BERT** for texts and **ViT** for images as unimodal supervised baselines. For multimodal supervised models, we choose **Score Fusion** and **Feature Concat** as multimodal baselines. To confirm the generalizability of our MSC framework on pseudo-labeling methods, we select two semi-supervised methods **FixMatch** and **FreeMatch**, which are combined with above methods as semi-supervised baselines.

Implementation

We use BERT and ViT to encode texts and images. MSC is combined with FixMatch and FreeMatch. The batch size of labeled data B_x is set to 4 and batch size of unlabeled data

Method	N24News						UPMC-Food101						CrisisMMD					
	20		50		100		20		50		100		20		50		100	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
BERT	74.15	73.62	77.08	76.60	79.82	79.17	80.70	80.89	82.82	82.87	83.66	83.78	66.88	63.91	77.85	73.85	80.92	77.37
ViT	32.22	32.45	37.01	37.46	40.70	40.61	58.35	58.65	62.53	63.00	65.59	65.94	70.94	68.72	82.74	80.54	83.87	81.64
Score Fusion (SF)	73.79	73.03	75.44	74.56	77.75	77.15	67.72	67.35	79.71	79.40	88.81	88.72	72.77	71.44	84.63	81.92	87.70	86.02
Feature Concat (FC)	67.60	67.27	75.81	75.07	79.33	78.63	74.11	74.04	86.30	86.22	87.67	87.63	71.86	70.04	83.22	80.17	85.83	83.90
FixMatch+BERT	76.26	75.42	77.91	**77.68**	80.31	79.89	82.38	82.58	82.92	83.14	82.97	83.37	74.61	69.09	83.13	80.18	83.70	80.64
FixMatch+ViT	33.29	32.06	38.79	38.28	41.81	41.60	60.27	60.04	63.63	63.65	65.07	64.85	74.94	70.86	83.61	80.62	84.05	81.74
FixMatch+SF	70.60	69.04	74.46	73.59	77.28	76.45	67.52	66.42	85.16	84.78	89.13	89.02	77.29	72.08	85.89	83.95	87.85	86.08
FixMatch+FC	74.34	72.84	77.65	76.28	79.95	78.73	86.65	85.57	88.74	88.78	89.71	89.69	77.09	70.34	85.70	83.21	87.67	85.88
FreeMatch+BERT	76.34	**75.87**	78.18	77.21	80.39	79.82	82.39	82.48	83.08	83.50	83.26	83.95	73.46	64.98	83.59	80.89	83.94	81.37
FreeMatch+ViT	33.29	32.16	38.81	37.94	41.69	41.66	60.12	59.89	63.72	63.63	65.07	64.96	75.48	68.97	82.81	79.88	84.24	82.17
FreeMatch+SF	71.18	69.98	75.09	73.76	77.65	76.16	85.40	84.86	88.83	88.62	90.02	89.98	77.29	72.23	86.26	84.24	87.44	85.83
FreeMatch+FC	75.96	74.63	77.65	76.47	80.04	78.92	86.37	86.27	89.34	89.23	89.98	89.98	76.31	68.72	84.02	82.88	87.52	85.72
FixMatch+MSC	76.87	75.17	78.61	77.52	80.94	80.42	91.08	90.88	92.00	91.86	92.24	92.11	**80.31**	**75.40**	**87.07**	84.79	88.00	**86.57**
FreeMatch+MSC	**76.98**	75.18	**78.67**	77.62	**81.05**	**80.48**	**91.12**	**90.88**	**92.21**	**92.09**	**92.41**	**92.27**	79.03	75.04	86.83	**84.92**	**88.09**	86.52

Table 1: The semi-supervised multimodal classification results on N24News, UPMC-Food101 and CrisisMMD.

Dataset	Type	Model	Acc	ΔAcc
UPMC-Food101	SMC	UniS-MMC	94.70	2.29↑
		CMA-CLIP	93.10	0.69↑
		MMBT	92.10	0.31↓
	SSMC	**MSC**	92.41	0
CrsisMMD	SMC	finetuned CLIP	93.15	5.06↑
		DMCC	92.24	4.15↑
		SSE-Cross	89.33	1.24↑
	SSMC	**MSC**	88.09	0

Table 2: Results compared with latest multimodal methods. SMC stands for supervised multimodal classification.

Model	N24News	UPMC-Food101
simple weight	69.77	80.44
only LCG	75.31	83.43
only MRG	71.09	90.21
only SCG	76.00	88.93
w/o LCG	71.18	90.96
w/o MRG	74.23	73.08
w/o SCG	72.26	90.48
MSC	**76.87**	**91.08**

Table 3: Ablation analysis with modules and FixMatch on N24News and UPMC-Food101 with 20 labels per class.

B_u is set to 32. We set $5e-5$ as learning rate. We use RandAugment to gain augmented images and obtain augmented texts with swap and synonym strategy. We employ sentence-bert to calculate similarity between augmented and original texts, choosing text with higher similarity as weak-augment and the other one as strong-augment. We select AdamW as optimizer. η is set to 0.95. The train epoch is set to 20. β_1, β_2, β_3 are set to 1 for N24News and UPMC-Food101 and set to 0.6 for CrisisMMD. Our codes will be released in https://github.com/cjc20000323/SSMC.

Main Results

The SSMC results are shown in Table 1. The upper four supervised methods only use labeled data to train. We observe that the accuracy and F1 of MSC are generally better than supervised and semi-supervised baselines. Especially on UPMC-Food101, our design outperforms the best baselines about $4\%-5\%$. MSC can also outperform the baselines 0.5%-3% on CrisisMMD and N24News. This indicates the effectiveness of MSC to be combined with semi-supervised methods. Another conclusion is that as the number of labeled data gets smaller, the improvement is more distinct, which shows the reliability of MSC in severe label scarcity.

Comparision with Latest Multimodal Methods

Results compared with the latest supervised multimodal classification on the same test set of UPMC-Food101 and CrisisMMD are shown in Table 2. We choose UniS-MMC (Zou et al. 2023), CMA-CLIP (Liu et al. 2021) and MMBT (Kiela et al. 2019) for UPMC-Food101 and finetuned CLIP (Mandal, Khanal, and Caragea 2024) , DMCC (Rezk et al. 2023) and SSE-Cross (Abavisani et al. 2020) for CrisisMMD. In our semi-supervised setting, the sum of labeled data is small (about 16% for UPMC-Food101 and 5% for CrisisMMD). The latest models only outperform MSC 1%-5% on accuracy. These results demonstrate that with far more less labeled data, MSC achieves competitive results compared with supervised models, which confirms the effectiveness of MSC.

Ablation Study

Module Analysis To investigate the contribution of different modules, we evaluate several variants with FixMatch. Table 3 shows the results of variants. *Simple weight* means adding weights to modalities on Score Fusion without other guidance signals, which obtains the worst results. Our MSC

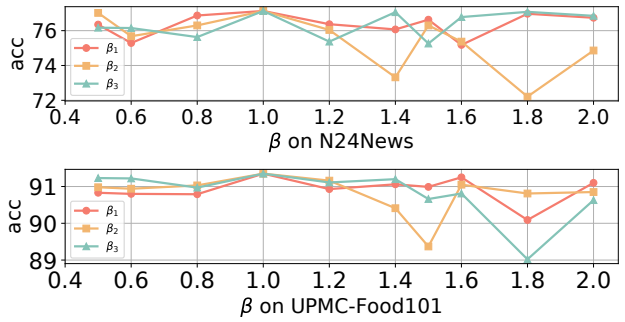

Figure 3: Hyper-parameters of FixMatch+MSC accuracy on N24News and UPMC-Food101 with 20 labels per class

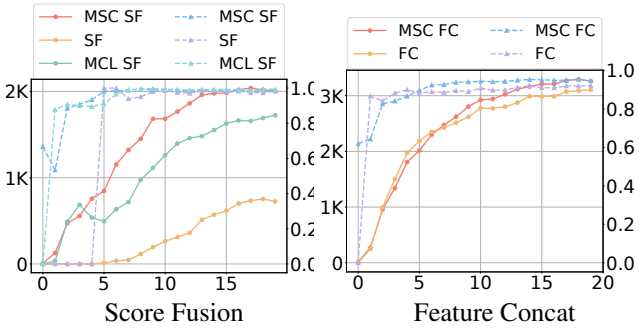

Figure 4: The sum and ratio of correct selected pseudo-labels with FixMatch on UPMC-Food101 with 20 labels per class. Solid lines denote sum curves using the left vertical axis. Dashed lines denote ratio curves using the right vertical axis.

framework behaves not the same on N24News and UPMC-Food101 when removing different modules. LCG and SCG contribute more to N24News and MRG plays a dominant role in UPMC-Food101. The effectiveness of MSC is the contribution of all modules working together.

Hyper-parameter Analysis We make hyper-parameter analysis on the loss weight with FixMatch on UPMC-Food101 and N24News with 20 labeled pairs per class in Figure 3. We display the accuracy when varying values of a hyper-parameter while keeping the other two fixed. The results show that the framework gains different performance with varying hyper-parameter settings. Setting a lower value is likely better than a higher one. For each hyper-parameter, the best configuration is set to 1.

Pseudo-Labels Trade-off Evaluation

We compare the sum and ratio of correct selected pseudo-labels of MSC and baselines with FixMatch. From Figure 4, our method increases the sum of pseudo-labels selected by Score Fusion when maintaining the correct proportion and modal complementarity learning only alleviates not solves trade-off problem. For Feature Concat, MSC increases the ratio of correct pseudo-labels when keeping the sum. This shows that strategic complementarity learning achieves a

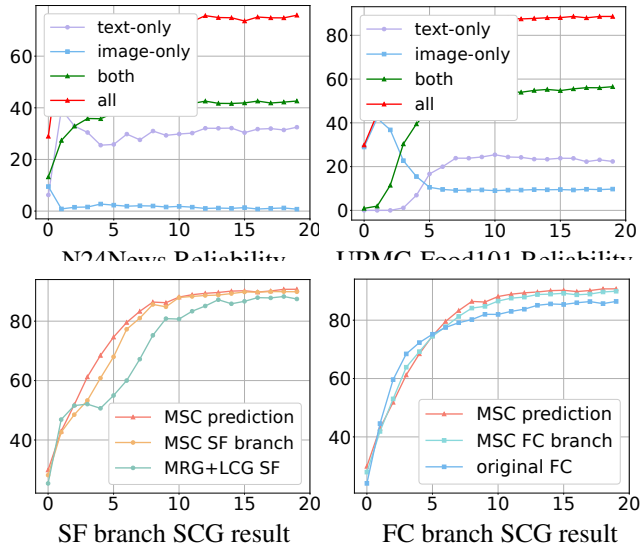

Figure 5: Correct reliability ratio with 20 labels per class (upper two) and classification accuracy under strategic complementarity learning with FixMatch on UPMC-Food101 with 20 labels per class (lower two).

trade-off between quality and quantity of pseudo-labels thus providing SSMC with better pseudo-labels.

Complementarity Analysis

Modal Complementarity From the upper two figures in Figure 5, the ratio that both modalities classify correctly increases as training proceeds, and modal complementarity learning can identify the data correctly classified by only one modality and give higher reliability weight to the corresponding modality branch. This indicates that MSC uses correct modal reliability to solve *modality interference problem* and increase the ratio of correct final predictions.

Strategic Complementarity From the lower in Figure 5, two branches of MSC outperform the corresponding single model separately, and the classification accuracy of the final average prediction is higher than any of the two strategies. This indicates that MSC pushes pseudo-labels of strategies to complement each other during training and includes effective integration with a simple scheme at final prediction.

Conclusion

We propose an MSC framework combined with pseudo-labeling methods to tackle SSMC. Concretely, our design attempts to learn from modal and strategic complementarity. For modal complementarity, MSC learns modal reliability weights to alleviate the modality interference problem when predictions are inconsistent. For strategic complementarity, MSC designs a pseudo-label selection method and balances the quality and quantity of pseudo-labels. Experiment results based on the created benchmarks demonstrate the effectiveness of our framework.

Acknowledgements

This work was supported by the National Science and Technology Major Project under Grant 2022ZD0120202, in part by the National Natural Science Foundation of China (No. U23B2056 and No. 62306026), in part by China Postdoctoral Science Foundation (No. 2023M740184), in part by the Fundamental Research Funds for the Central Universities, and in part by the State Key Laboratory of Complex & Critical Software Environment.

References

Abavisani, M.; Wu, L.; Hu, S.; Tetreault, J.; and Jaimes, A. 2020. Multimodal categorization of crisis events in social media. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14679–14689.

Agarwal, M.; Leekha, M.; Sawhney, R.; and Shah, R. R. 2020. Crisis-dias: Towards multimodal damage analysis-deployment, challenges and assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 346–353.

Al Obaid, A.; Khotanlou, H.; Mansoorizadeh, M.; and Zabihzadeh, D. 2023. Robust semi-supervised fake news recognition by effective augmentations and ensemble of diverse deep learners. *IEEE Access*.

Alam, F.; Ofli, F.; and Imran, M. 2018. Crisismmd: Multimodal twitter datasets from natural disasters. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; and Raffel, C. A. 2019. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32.

Bossard, L.; Guillaumin, M.; and Van Gool, L. 2014. Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, 446–461. Springer.

Chen, H.; Guo, C.; Li, Y.; Zhang, P.; and Jiang, D. 2023. Semi-Supervised Multimodal Emotion Recognition with Class-Balanced Pseudo-labeling. In *Proceedings of the 31st ACM International Conference on Multimedia*, 9556–9560.

Chen, J.; Yang, Z.; and Yang, D. 2020. MixText: Linguistically-Informed Interpolation of Hidden Space for Semi-Supervised Text Classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2147–2157.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Huang, Z.; Zeng, Z.; Liu, B.; Fu, D.; and Fu, J. 2020. Pixelbert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, 4171–4186.

Kiela, D.; Bhooshan, S.; Firooz, H.; Perez, E.; and Testuggine, D. 2019. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*.

Kiela, D.; Grave, E.; Joulin, A.; and Mikolov, T. 2018. Efficient large-scale multi-modal classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Lee, J.-H.; Ko, S.-K.; and Han, Y.-S. 2021. Salnet: Semi-supervised few-shot text classification with attention-based lexicon construction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 13189–13197.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Liang, T.; Lin, G.; Wan, M.; Li, T.; Ma, G.; and Lv, F. 2022. Expanding large pre-trained unimodal models with multimodal information injection for image-text multimodal classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15492–15501.

Liu, H.; Xu, S.; Fu, J.; Liu, Y.; Xie, N.; Wang, C.-C.; Wang, B.; and Sun, Y. 2021. Cma-clip: Cross-modality attention clip for image-text classification. *arXiv preprint arXiv:2112.03562*.

Mandal, B.; Khanal, S.; and Caragea, D. 2024. Contrastive Learning for Multimodal Classification of Crisis related Tweets. In *Proceedings of the ACM on Web Conference 2024*, 4555–4564.

Monter-Aldana, I.; Monroy, A. P. L.; and Vega, F. S. 2023. Dynamic Regularization in UDA for Transformers in Multimodal Classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8700–8711.

Raj, C.; and Meel, P. 2021. ConvNet frameworks for multimodal fake news detection. *Applied Intelligence*, 51(11): 8132–8148.

Rezk, M.; Elmadany, N.; Hamad, R. K.; and Badran, E. F. 2023. Categorizing crises from social media feeds via multimodal channel attention. *IEEE Access*.

Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C. A.; Cubuk, E. D.; Kurakin, A.; and Li, C.-L. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33: 596–608.

Wang, W.; Tran, D.; and Feiszli, M. 2020. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12695–12705.

Wang, Y.; Chen, H.; Heng, Q.; Hou, W.; Fan, Y.; Wu, Z.; Wang, J.; Savvides, M.; Shinozaki, T.; Raj, B.; et al. 2022. Freematch: Self-adaptive thresholding for semi-supervised learning. *arXiv preprint arXiv:2205.07246*.

Wang, Z.; Shan, X.; Zhang, X.; and Yang, J. 2021. N24News: a new dataset for multimodal news classification. *arXiv preprint arXiv:2108.13327*.

Xie, Q.; Dai, Z.; Hovy, E.; Luong, T.; and Le, Q. 2020. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33: 6256–6268.

Yang, W.; Zhang, R.; Chen, J.; Wang, L.; and Kim, J. 2023. Prototype-guided pseudo labeling for semi-supervised text classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 16369–16382.

Zou, H.; Shen, M.; Chen, C.; Hu, Y.; Rajan, D.; and Chng, E. S. 2023. UniS-MMC: Multimodal Classification via Unimodality-supervised Multimodal Contrastive Learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, 659–672.