UNIS-MMC: LEARNING UNIMODALITY-SUPERVISED MULTIMODAL CONTRASTIVE REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal learning aims to imitate human beings to acquire complementary information from multiple modalities for final decisions. However, just like a human's final decision can be confused by specific erroneous information from the environment, current multimodal learning methods also suffer from uncertain unimodal prediction when learning multimodal representations. In this work, we propose to contrastively explore reliable representations and increase the agreement among the unimodal representations that alone make potentially correct predictions. Specifically, we first capture task-related representations by directly sharing representations between unimodal and multimodal learning tasks. With the unimodal representations and predictions from the multitask-based framework, we then propose a novel multimodal contrastive learning method to align the representations towards the relatively more reliable modality under the weak supervision of the unimodal predictions. Experimental results on two image-text benchmarks UPMC-Food-101 and N24News, and two medical benchmarks ROSMAP and BRCA, show that our proposed **Uni**modality-**S**upervised **M**ulti**M**odal **C**ontrastive (**UniS-MMC**) learning method outperforms current state-of-the-art multimodal learning methods. The detailed ablation studies further demonstrate the advantage of our proposed method.

1 INTRODUCTION

A prominent point of human intelligence is the ability to handle various information and make better decisions from them. Empowering artificial intelligence with the power of working with multiple modalities is increasingly important with the growing and more accessible data sources, such as images, text, etc (Baltrušaitis et al., 2018). Despite the recent progress in obtaining effective unimodal representations from large pre-trained models (Devlin et al., 2018; Liu et al., 2019; Dosovitskiy et al., 2020), obtaining more trustworthy and complementary multimodal representations remains a challenging fundamental problem for multimodal machine learning.

The mainstream idea for solving multimodal fusion problem is combining unimodal representations directly, including fusing unimodal features (Castellano et al., 2008; Nagrani et al., 2021), fusing unimodal decisions (Ramirez et al., 2011; Tian et al., 2020a), and fusing both (Wu et al., 2022) of them. Those traditional aggregation-based methods conduct multimodal tasks by learning from the final joint representations. To help complement each modality, some methods attempt to check the effectiveness of each participating modalities (Mittal et al., 2020) or capture the reliable parts of the unimodal feature (Han et al., 2022b) to filter potential valid information before fusion.

However, these aggregation-based methods ignore the consistency among different modalities and thus increase the uncertainty of final predictions when unimodal information is prone to decide inconsistently. To solve this issue, the aligned-based fusion methods are further proposed to align the embeddings from different modalities during propagation. Some early works choose to map the unimodal feature to a new space (Wang et al., 2016; Cheng et al., 2017) or use an adaption module (Song et al., 2020) and then apply a regulation loss to minimize the space distance. Although keeping the intra-modal propagation, the introduced alignment losses still lack the ability for inter-modal relationship learning as the weak message exchanging (Wang et al., 2020b). Drawing on the success of contrastive learning in unimodal domain tasks, such as computer vision (Wu et al., 2018; He et al., 2020; Chen et al., 2020; Li et al., 2022) and natural language processing (Gunel et al.,

2020; Qin & Joty, 2022), some researchers are exploring the multimodal contrastive methods for more effectively aligning different modality representations.

Compared to unimodal contrastive learning methods, the multimodal methods do not need to generate enhanced cases (Khosla et al., 2020; Kim et al., 2020) for paired samples. Multi-modality inherently implies paired unimodal features for contrastive learning. Most of the multimodal contrastive methods focus on aligning different modality representations in an unsupervised manner. These multimodal methods directly regard the paired modalities from the same samples as the positive pairs and those modalities from different samples as the negative pairs to encourage the paired modalities to have a similar representation distribution (Tian et al., 2020b; Akbari et al., 2021; Zolfaghari et al., 2021; Liu et al., 2021b; Zhang et al., 2021a; Taleb et al., 2022). Similarly, some methods introduce the supervised contrastive in multimodal area (Zhang et al., 2021b; Pinitas et al., 2022) and treat the sample pairs with the same label in the mini-batch as the positive pairs. Though these alignment-based multimodal contrastive learning methods provide good modality information transfer-ability (Radford et al., 2021; Li et al., 2021b) and competitive performance in some zero-shot learning areas, they still can not provide trustworthy enough multimodal representations for final decisions. The fused features from those inefficient unimodal features after naive alignment are of insufficient assurance for making correct final decisions. The final decisions will be negatively affected by those samples with unimodal representations all making wrong unimodal decisions.

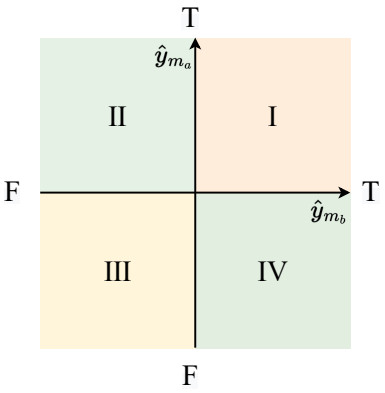

Figure 1: Unimodal predictions $\hat{y}_{m_a}$, $\hat{y}_{m_b}$ of two modality m_a and m_b. Area **I** contains samples with both correct unimodal predictions, area **II** and **III** contain samples with inconsistent unimodal predictions (one correct, one wrong) and area **IV** contains samples with both wrong unimodal predictions. The uncertainty of multimodal decision will increase with inconsistent unimodal information (area **II** and **IV**). On the other hand, multimodal decision will be affected by the samples (area **III**) that make all wrong unimodal predictions after naive alignment.

In this work, we aim to learn trustworthy and complementary multimodal representations for reducing the multimodal decision-making bias when learning with mutually exclusive unimodal representations. We focus on contrastively learning reliable unimodal representations and encouraging the agreement on those unimodal representations with consistent and correct unimodal decisions under the supervision of unimodal predictions. In summary, our contributions are:

- First, we introduce a multi-task-based multimodal learning framework for getting more reliable representations and the respective unimodal predictions. The representations are shared directly between the multimodal learning branch and the unimodal learning branches.
- Next, we propose a novel multimodal contrastive method based on unimodal predictions. We argue that the unimodal information that makes up the multimodal representation should be target-related and consistent. Multimodal contrastive learning keeps exploring more combinations when unimodal representations both give the wrong predictions even if they are paired. It helps different modality features to learn with each other with the respective unimodal supervision.
- Finally, to demonstrate our proposed method, we evaluate our method on four public multimodal classification benchmarks: two image-text datasets UPMC-Food-101 (Wang et al., 2015) and N24News (Wang et al., 2021), and two medical datasets ROSMAP and BRCA Wang et al., 2020a.

2 Related Work

2.1 Contrastive learning

Contrastive learning (Hadsell et al., 2006; Oord et al., 2018) captures distinguishable representations by drawing the positive pairs closer and pushing the negative pairs farther without unitizing the label information. Most self-supervised methods (Chen et al., 2020; Kalantidis et al., 2020; Kim et al., 2020; He et al., 2020; Xiong et al., 2020; Bahri et al., 2021; Van Gansbeke et al., 2021) use the data augmentation methods and treat the corresponding augmented samples as the positive pairs of the

target samples and other samples as the negative pairs. Compared to this, supervised contrastive learning (Gunel et al., 2020; Khosla et al., 2020; Li et al., 2022; Chen et al., 2022; Bai et al., 2022) regards the samples of the same categories as positive pairs and the samples of different categories as negative pairs. Besides, weakly-supervised contrastive methods use auxiliary information as weak labels, such as textual descriptions (Radford et al., 2021), data attributes (Tsai et al., 2022), and document-level co-occurrence information of events (Gao et al., 2022).

In addition to the above single-modality representation learning, contrastive methods for multiple modalities are also widely explored. The common methods (Radford et al., 2021; Jia et al., 2021; Kamath et al., 2021; Li et al., 2021a; Zhang et al., 2022; Taleb et al., 2022) leverage the cross-modal contrastive matching to align two different modalities and learn the inter-modality correspondence. Except the inter-modality contrastive, Visual-Semantic Contrastive (Yuan et al., 2021), XMC-GAN (Zhang et al., 2021a) and CrossPoint (Afham et al., 2022) also introduce the intra-modality contrastive for representation learning. Besides, CrossCLR (Zolfaghari et al., 2021) removes the highly related samples from the negative samples to avoid the bias of false negatives. GMC (Poklukar et al., 2022) builds the contrastive learning process between the modality-specific representations and the global representations of all modalities instead of the cross-modal representations.

2.2 MULTIMODAL LEARNING

Multimodal learning is expected to build models based on multiple modalities and to improve the general performance from the joint representation (Ngiam et al., 2011; Baltrušaitis et al., 2018; Gao et al., 2020). The fusion operation among multiple modalities is one of the key topics in multimodal learning to help the modalities complement each other (Wang, 2021). Multimodal fusion methods are generally categorized into two types: alignment-based fusion and aggregation-based fusion (Baltrušaitis et al., 2018). Alignment-based fusion (Gretton et al., 2012; Song et al., 2020) aligns multimodal features by increasing the modal similarity to extract the modality-invariant features. Aggregation-based methods choose to create the joint multimodal representations by combining the participating unimodal features (early-fusion, Kalfaoglu et al. (2020); Nagrani et al. (2021)), unimodal decisions (late-fusion, Tian et al. (2020a); Huang et al. (2022)) and both (hybrid-fusion, Wu et al. (2022)). In addition to the joint-representation generating methods, some works propose to evaluate the attended modalities and features before fusing, for example, M3ER (Mittal et al., 2020) conducts a modality check and Multimodal Dynamics (Han et al., 2022a) evaluates both the feature-level and modality-level informativeness for trustworthy multimodal fusion.

2.3 MULTI-TASK LEARNING

Compared with single-task learning, multi-task learning aims to optimize several different tasks and share the task-invariant parameters (Caruana, 1997; Ruder, 2017). It is believed to perform better on the original task (Ruder, 2017) when sharing parameters among related tasks. There are two typical parameter-sharing methods in multi-task learning, hard-parameter sharing and soft-parameter sharing. Hard-parameter sharing (Pilault et al., 2020; Bhattacharjee et al., 2022) directly shares the same hidden layer across different tasks and soft-parameter sharing (Misra et al., 2016; Chen et al., 2018) connects the task-specific hidden layers with a special designed mechanism, such as cross-stitch units (Misra et al., 2016) and NDDR layer (Gao et al., 2019). In addition to the wide application in single-modality learning filed, such as (Bao et al., 2022; Fifty et al., 2021) in computer vision and (Sanh et al., 2021) in natural language processing, multi-task learning also provides competitive performance in multimodal learning area (Yu et al., 2021; Abdollahzadeh et al., 2021).

3 METHODOLOGY

In this section, we first introduce the multi-task-based multimodal learning framework to share the learned representations between unimodal predicting tasks and the multimodal predicting task. Then we illustrate how we design the unimodality-supervised multimodal contrastive learning method among modalities to learn the multimodal representations. Finally, we summarize the learning objective for our proposed UniS-MMC.

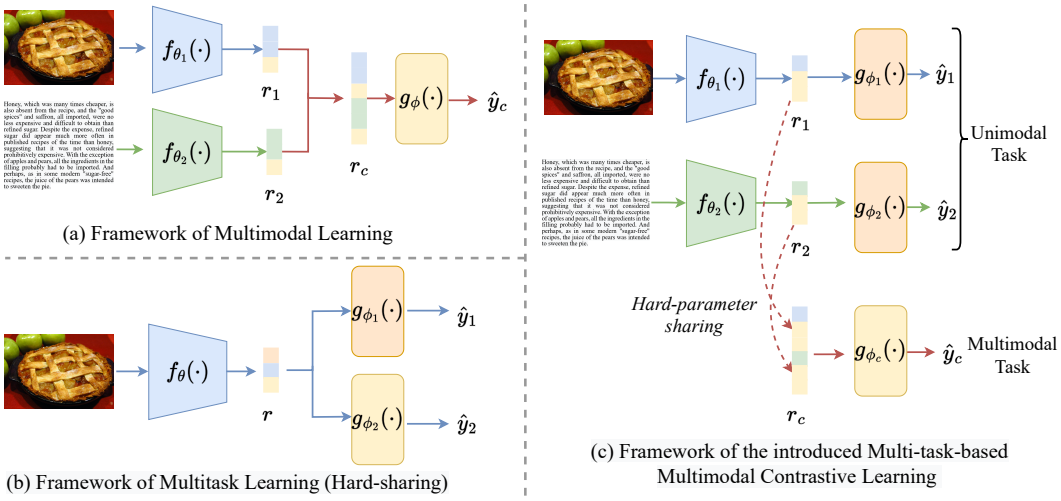

Figure 2: A sketched comparison among (a) multimodal learning, (b) multi-task learning (hard-sharing), and (c) the multitask-based multimodal learning framework.

3.1 Background

Assume we have two modalities m_a and m_b that equally contribute to the multimodal decision. Considering the entropy of final prediction alone, we have the following equation:

$$H(Y) = -p\log(p) - (1-p)\log(1-p), \tag{1}$$

where p is the probability for a correct prediction.

On the one hand, p should be closer to 0 or 1 for reducing the uncertainty of multimodal prediction. This corresponds to the fact that unimodal representations should be consistent to reduce decision bias for opposite unimodal information. On the other hand, the certainty brought by the probability that gives all wrong predictions is meaningless. The goal of multimodal learning is to give a certain and correct prediction through multiple modality data. This consistency among the information from different modalities should appear in those samples whose modalities are all correct for each unimodal prediction. In order to take into account the above two goals for multimodal learning, we propose the following unimodality-supervised multimodal contrastive learning method.

3.2 Multitask based multimodal learning

As shown in Figure 2, we utilize the extracted unimodal representations as inputs to the unimodal classifiers and add the unimodal predicting task to the common aggregation-based multimodal learning method. Following the parameter sharing in the multi-task learning method, the representations are shared directly between unimodal prediction tasks and the multimodal prediction task. Suppose we have data set $\mathcal{D} = \{\{x_m^{(i)}\}_{m=1}^M, y^{(i)}\}_{i=1}^N$ that contains N samples $\mathcal{X} = \{x_m^{(i)} \in \mathbb{R}^{d_m}\}_{m=1}^M$ of M modalities and N corresponding labels $\mathcal{Y} = \{y^{(i)}\}_{i=1}^N$ from K categories.

Unimodal Representation Learning. Given multimodal training data $\{\mathbf{x}_m\}_{m=1}^M$, the raw unimodal data of modality m are firstly processed with respective encoders to obtain the hidden representations. The general encoder network can be written as the mapping function: $f_\theta : \mathcal{X} \to \mathcal{R}$. We denote the learned hidden representation $f_{\theta_m}(\mathbf{x}_m)$ of modality m as $\mathbf{r}_m$. For text and image data in UPMC Food-101 and N24News datasets, we use the pretrained models as feature encoders. Following Multimodal Dynamics Han et al. (2022b), we use one fully-connected layer as the unimodal feature encoder for the medical data in BRCA and ROSMAP.

Unimodal Prediction. Different from the common aggregation-based multimodal learning method, the unimodal learned representations in our framework are also worked as inputs for the unimodal predicting tasks in addition to fusing for multimodal predicting. The classification module can be

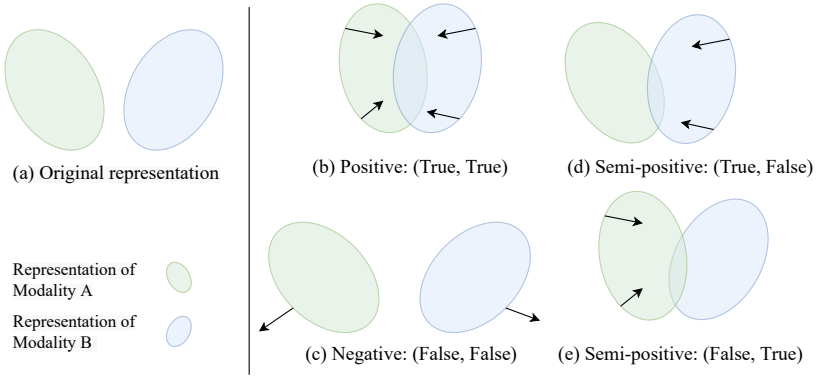

Figure 3: Representation distribution changes for unimodality-supervised multimodal contrastive loss: (a) original modality distribution, (b) positive pairs, (c) negative pairs, (d) semi-positive pairs with True and False predictions, and (e) semi-positive pairs with False and True predictions. The arrows represent the expected alignment directions of our contrastive method.

regarded as a probabilistic model: $g_\phi : \mathcal{R} \to \mathcal{P}$, which maps the hidden representation to a predictive distribution $\mathbf{p}(\mathbf{y} \mid \mathbf{r})$. For a unimodal predicting task, the predictive distribution is only based on the output of the unimodal classifier. The target of the unimodal predicting task is to minimize each unimodal prediction loss:

$$\mathcal{L}_{uni} = -\sum_{m=1}^{M}\sum_{k=1}^{K} y^k \log p_m^k, \tag{2}$$

where y^k is the k-th element category label and $[p_m^1; p_m^2; ...; p_m^K] = \mathbf{p}_m(\mathbf{y} \mid \mathbf{r}_m)$ is the softmax output of unimodal classifiers on modality m.

Multimodal Prediction. When fusing all unimodal representations with concatenation, we get the fused multimodal representations $r_c = r_1 \oplus r_2 \oplus ... \oplus r_m$. Similarly, the multimodal predictive distribution is the output of the multimodal classifier with inputs of the fused multimodal representations. For the multimodal prediction task, the target is to minimize the multimodal prediction loss:

$$\mathcal{L}_{multi} = -\sum_{k=1}^{K} y^k \log p_k^k, \tag{3}$$

where y^k is the k-th element category label and $[p_k^1; p_k^2; ...; p_k^K] = \mathbf{p}_c(\mathbf{y} \mid \mathbf{r}_c)$ is the softmax output of multimodal classifier.

So the goal of multi-task-based multimodal predicting is to minimize the sum of the unimodal predicting loss and the multimodal predicting loss:

$$\mathcal{L}_{mt-mml} = \mathcal{L}_{uni} + \mathcal{L}_{multi}, \tag{4}$$

3.3 Unimodality-supervised multimodal contrastive learning

We aim to reduce the multimodal prediction bias caused by inconsistent unimodal information by learning more certain predictions of each modality. From the multi-task-based multimodal learning framework, we regulate each unimodal representation with the targets. Here we propose a new multimodal contrastive method to enlarge the agreement for those unimodal representations with consistent predictions (shown in Fig 3 (b)). For those samples with both wrong predictions, we encourage they can be more different to explore a larger possibility of correct prediction (shown in Fig 3 (c)). For those samples with mutually exclusive predictions, we encourage them to learn from each other under the supervision of unimodal predictions (shown in Fig 3 (d) and (e)). When considering two specific modalities m_a and m_b of i-th sample, we generate two unimodal hidden representations $r_a^{(i)}$ and $r_b^{(i)}$ from respective unimodal encoders. From the above unimodal predicting step, we also

obtain the unimodal prediction results $\hat{y}_a^{(i)}$ and $\hat{y}_b^{(i)}$. For the designed multimodal contrastive loss, we define the following positive pair, negative pair and semi-positive pair:

Positive Pair. If both the paired unimodal predictions are correct, we define these pairs of unimodal representations are the positive pairs, namely $(r_a^{(i)}, r_b^{(i)}) \in \mathbb{P}$, where $\mathbb{P} = \{\hat{y}_a^{(i)} == y^{(i)}\}_{i=1}^N \bigcap \{\hat{y}_b^{(i)} == y^{(i)}\}_{i=1}^N$ in the mini-batch $\mathbb{B}$. By using the designed contrastive method, the distributions of these positive pairs' representations are encouraged to tend towards each other to get high similarity.

Negative Pair. If both the paired unimodal predictions are wrong, we define these pairs of unimodal representations are the negative pairs, namely $(r_a^{(i)}, r_b^{(i)}) \in \mathbb{N}$, where $\mathbb{N} = \{\hat{y}_a^{(i)} != y^{(i)}\}_{i=1}^N \bigcap \{\hat{y}_b^{(i)} != y^{(i)}\}_{i=1}^N$ in the mini-batch $\mathbb{B}$. By the designed contrastive method, the distributions of these negative pairs' representations are encouraged to be more different to explore more combinations of predictions with updated unimodal representations.

Semi-Positive Pair. If the predictions of the paired unimodal representations are mutually exclusive, one correct and another wrong, we define these pairs of unimodal representations are semi-positive pairs, namely $(r_a^{(i)}, r_b^{(i)}) \in \mathbb{S}$, where $\mathbb{S} = (\{\hat{y}_a^{(i)} == y^{(i)}\}_{i=1}^N \bigcap \{\hat{y}_b^{(i)} != y^{(i)}\}_{i=1}^N) \bigcup (\{\hat{y}_a^{(i)} != y^{(i)}\}_{i=1}^N \bigcap \{\hat{y}_b^{(i)} == y^{(i)}\}_{i=1}^N)$ in the mini-batch $\mathbb{B}$. By the designed contrastive method, the unimodal representations with the correct predictions will be fixed and the unimodal representations with the wrong predictions will be updated to have higher similarity with the correct modality.

With the above definition, we propose the multimodal contrastive loss for two modalities as follows:

$$\mathcal{L}_{b-mmc}(m_a, m_b) = -\log\{\frac{\sum_{i\in\mathbb{P}}(\exp(\cos(r_a^{(i)}, r_b^{(i)})/\tau) + \sum_{i\in\mathbb{S}}(\exp(\cos(r_a^{(i)}, r_b^{(i)})/\tau)}{\sum_{i\in\mathbb{B}}(\exp(\cos(r_a^{(i)}, r_b^{(i)})/\tau)}\}, \tag{5}$$

where $\cos(r_a^{(i)}, r_b^{(i)}) = \frac{r_a^{(i)} \cdot r_b^{(i)}}{\|r_a^{(i)}\| * \|r_b^{(i)}\|}$ is the cosine similarity between paired unimodal representations $r_a^{(i)}$ and $r_b^{(i)}$ for sample i, τ is the temperature coefficient.

The multimodal contrastive loss for M modalities can be computed by:

$$\mathcal{L}_{mmc} = \sum_{i=1}^{M} \sum_{j>i}^{M} \mathcal{L}_{b-mmc}(m_i, m_j), \tag{6}$$

3.4 Learning objective

The overall optimization objective for our proposed UniS-MMC is:

$$\mathcal{L}_{UniS-MMC} = \mathcal{L}_{uni} + \mathcal{L}_{multi} + \lambda \mathcal{L}_{mmc}, \tag{7}$$

where λ is a loss coefficient for balancing the predicting loss and the multimodal contrastive loss.

4 Experiments

4.1 Experimental setup

Datasets and Evaluation Metrics. We evaluate our method on four publicly available multimodal datasets UPMC-Food-101, N24News, BRCA and ROSMAP. UPMC-Food-101 [1] is a multimodal classification dataset that contains textual recipe descriptions and the corresponding images for 101 kinds of food. We get this dataset from their project website and split 5000 samples from the default training set as the validation set. N24News [2] is an news classification dataset with four text types, `Heading`, `Caption`, `Abstract` and `Body`. We choose the first three text sources in this work.

[1] UPMC-Food-101: https://visiir.isir.upmc.fr/

[2] N24News: https://github.com/billywzh717/N24News

Table 1: Multimodal classification performance on **a)** Food101 and **b)** N24News.

a) Model	Fusion		Backbone		Acc
	AGG	ALI	Image	Text	
MMBT	Early	✗	ResNet-152	BERT	$92.1_{\pm 0.1}$
HUSE	Early	✓	Graph-RISE	BERT	92.3
CMA-CLIP	Early	✓	ViT	Transformer	93.1
ME	Early	✗	DenseNet	BERT	94.6
UnSupMMC	Early	✓	ViT	BERT	$94.1_{\pm 0.7}$
SupMMC	Early	✓	ViT	BERT	$94.2_{\pm 0.2}$
UniS-MMC	Early	✓	ViT	BERT	$\mathbf{94.7}_{\pm 0.1}$

b) Model	Fusion		Backbone		Multimodal		
	AGG	ALI	Image	Text	Headline	Caption	Abstract
N24News	Early	✗	ViT	RoBERTa	79.41	77.45	83.33
UnSupMMC	Early	✓	ViT	BERT	$78.6_{\pm 1.2}$	$76.5_{\pm 0.2}$	$81.1_{\pm 0.6}$
SupMMC	Early	✓	ViT	BERT	$78.5_{\pm 1.2}$	$76.9_{\pm 0.4}$	$81.5_{\pm 0.5}$
UniS-MMC	Early	✓	ViT	BERT	$\mathbf{80.2}_{\pm 0.1}$	$\mathbf{77.5}_{\pm 0.3}$	$\mathbf{83.2}_{\pm 0.4}$
UnSupMMC	Early	✓	ViT	RoBERTa	$79.3_{\pm 0.6}$	$77.6_{\pm 0.2}$	$83.9_{\pm 0.4}$
SupMMC	Early	✓	ViT	RoBERTa	$79.4_{\pm 0.3}$	$77.6_{\pm 0.6}$	$84.1_{\pm 0.4}$
UniS-MMC	Early	✓	ViT	RoBERTa	$\mathbf{80.3}_{\pm 0.1}$	$\mathbf{78.1}_{\pm 0.2}$	$\mathbf{84.2}_{\pm 0.1}$

BRCA and ROSMAP are two public multimodal medical datasets, both contain three modalities: mRNA expression, DNA methylation and miRNA expression. BRCA is used for breast invasive carcinoma PAM50 subtype classification with 5 categories and ROSMAP is used for Alzheimer's Disease diagnosis with 2 categories. We get this data online [3] and follow the train-test splitting from previous worksWang et al. (2020a); Han et al. (2022b).

We use classification accuracy (Acc) as evaluation metrics for UPMC-Food-101 and N24News. We report accuracy (Acc), weighted F1 score (WeightedF1) and macro-averaged F1 score (MacroF1) for BRCA and accuracy (Acc), F1 score (F1) and area under the receiver operating characteristic curve (AUC) for ROSMAP. The detailed dataset information can be seen in Appendix A.1.

Implementation. For the image-text dataset UPMC Food-101, we use pretrained BERT Devlin et al. (2018) as a text encoder and pretrained vision transformer (ViT) Dosovitskiy et al. (2020) as an image encoder. For N24News, we utilize two different pretrained language models, BERT and RoBERTa (Liu et al., 2019) as text encoders and also the same vision transformer as an image encoder. All classifiers of these two image-text classification datasets are three fully-connected layers with a ReLU activation function. For two small medical datasets BRCA and ROSMAP, we use one fully-connected layer as a feature encoder for each modality and two fully-connected layers with a ReLU activation function as classifiers.

The default reported results on image-text datasets are obtained with BERT-base (or RoBERTa-base) and ViT-base in this paper. The performance is presented with the average and standard deviation of three runs on Food101 and N24News and five runs on BRCA and ROSMAP. The codes will be available on GitHub. The detailed settings of the hyper-parameter are summarized in Appendix A.2.

4.2 Performance comparison

We first compare our proposed method with the existing state-of-the-art multimodal classification methods, including MMBT (Kiela et al., 2019), ViLT (Kim et al., 2021; Liang et al., 2022), CMA-CLIP (Liu et al., 2021a), ME (Liang et al., 2022) on Food101 and N24News, and MOGONET (Wang et al., 2020a),GMU (Arevalo et al., 2017), TMC (Han et al., 2021), CF (Huang et al., 2021), Dynamics (Han et al., 2022a) on BRCA and ROSMAP. In addition to these methods, we also implement the typical aggregation-based, unsupervised and supervised multimodal contrastive-based methods with the same encoders and classifiers in our method as baseline models.

The final image-text classification performance on Food101 and N24News is presented in Table 1. We have the following findings from the experimental results: (i) the proposed method outperforms all recent state-of-the-art methods on Food101 and produces the best results on every kinds of text sources on N24News with the same encoders; (ii) focusing on the implemented methods, contrastive-based methods with naive alignment outperform many of the recent multimodal methods; (iii) the proposed UniS-MMC has a large improvement compared with both the implemented contrastive-based baseline models and the recent start-of-art multimodal methods.

The multimodal classification results on BRCA and ROSMAP are shown in Table 2. Performance improvement can also be seen in BRCA and ROSMAP. Comparing the implemented unsupervised and supervised contrastive methods with the state-of-the-art methods, the simple concatenation of all unimodal features after alignment with the contrastive strategy can provide competitive results,

[3] BRCA and ROSMAP: https://github.com/txWang/MOGONET

Table 2: Multimodal classification performance on BRCA and ROSMAP

Model	Fusion		BRCA			ROSMAP		
	AGG	ALI	Acc	WeightedF1	MacroF1	Acc	F1	AUC
MOGONET	Early	✗	$80.6_{\pm 0.5}$	$80.0_{\pm 1.5}$	–	$82.5_{\pm 2.2}$	$82.4_{\pm 2.2}$	$87.3_{\pm 2.2}$
GMU	Early	✗	$80.0_{\pm 3.9}$	$79.8_{\pm 5.8}$	$74.6_{\pm 5.8}$	$77.6_{\pm 2.5}$	$78.4_{\pm 1.6}$	$86.9_{\pm 1.6}$
TMC	Late	✗	$84.2_{\pm 0.5}$	$84.4_{\pm 0.9}$	$80.6_{\pm 0.9}$	$82.5_{\pm 0.9}$	$82.3_{\pm 0.6}$	$88.5_{\pm 0.6}$
CF	Early	✗	$81.5_{\pm 0.8}$	$81.5_{\pm 0.9}$	$77.1_{\pm 0.9}$	$78.4_{\pm 1.1}$	$78.8_{\pm 0.5}$	$88.0_{\pm 0.5}$
Dynamics	Early	✗	$87.7_{\pm 0.3}$	$88.0_{\pm 0.5}$	$84.5_{\pm 0.5}$	$84.2_{\pm 1.3}$	$84.6_{\pm 0.7}$	$\mathbf{91.2}_{\pm 0.7}$
UnSupMMC	Early	✓	$87.7_{\pm 0.2}$	$88.0_{\pm 0.2}$	$85.4_{\pm 0.4}$	$82.4_{\pm 1.3}$	$83.0_{\pm 1.1}$	$88.4_{\pm 0.9}$
SupMMC	Early	✓	$87.5_{\pm 0.3}$	$87.9_{\pm 0.3}$	$84.9_{\pm 0.4}$	$83.8_{\pm 0.9}$	$84.7_{\pm 0.9}$	$88.9_{\pm 0.4}$
UniS-MMC	Early	✓	$\mathbf{89.4}_{\pm 0.3}$	$\mathbf{89.7}_{\pm 0.4}$	$\mathbf{86.6}_{\pm 0.4}$	$\mathbf{85.7}_{\pm 0.7}$	$\mathbf{86.3}_{\pm 0.8}$	$91.1_{\pm 0.9}$

Baselines. MOGONET (Wang et al., 2020a),GMU (Arevalo et al., 2017), TMC (Han et al., 2021), CF (Huang et al., 2021), Dynamics (Han et al., 2022a)

even outperform the current best results on some metrics. Besides, our unimodality-supervised multimodal contrastive method outperforms existing methods on most of evaluation metrics.

4.3 Analysis

Classification with Different Combinations of Input Modalities. We first perform an ablation study of classification on N24News with different input modalities. Table 3 provides the classification performance of unimodal learning with image only, text only, traditional multimodal learning with the concatenation of visual and textual features and our proposed UniS-MMC. The text modality is encoded with two different encoders, RoBERTa or BERT. By comparing the models with different language encoders, we find that the feature encoder can significantly affect the multimodal performance, and the RoBERTa-based model usually performs better than the BERT-based model. This is because the multimodal classification task is influenced by each learned unimodal representation. Besides, all the multimodal networks perform better than unimodal networks. It reflects that multiple modalities will help make accurate decisions. Moreover, our proposed UniS-MMC achieves 0.7% to 2.4% improvement over the aggregation-based baseline model with BERT and 0.6% to 1.3% improvement with RoBERTa.

Table 3: Ablation study on N24News with different combinations of input modalities.

Text	Image-only	Bert-based			Roberta-based		
		Text-only	MML	UniSMMC	Text-only	MML	UniSMMC
Headline	$54.1_{\pm 0.2}$	$72.1_{\pm 0.2}$	$78.7_{\pm 1.1}$	$\mathbf{80.2}_{\pm 0.1}$ ↑ 1.5	$71.8_{\pm 0.2}$	$79.0_{\pm 0.4}$	$\mathbf{80.3}_{\pm 0.1}$ ↑ 1.3
Caption		$72.7_{\pm 0.3}$	$76.8_{\pm 0.2}$	$\mathbf{77.5}_{\pm 0.3}$ ↑ 0.7	$72.9_{\pm 0.4}$	$77.5_{\pm 0.2}$	$\mathbf{78.1}_{\pm 0.4}$ ↑ 0.6
Abstract		$78.3_{\pm 0.3}$	$80.8_{\pm 0.2}$	$\mathbf{83.2}_{\pm 0.4}$ ↑ 2.4	$79.7_{\pm 0.2}$	$83.4_{\pm 0.1}$	$\mathbf{84.2}_{\pm 0.1}$ ↑ 0.8

Classification with Different Fusion Strategies. Moreover, we compare the unimodal and multimodal performance with different strategies in Table 4. Frozen-MML and Finetuned-MML both use the concatenation of visual and textual features for multimodal fusion. The difference is that the parameters of pretrained encoders in Frozen-MML are fixed, while in Finetuned-MML are tunable. As the unimodal predicting task is not optimized in both Frozen-MML and Finetuned-MML, we choose to fix the parameter of the respective feature encoder (the raw pretrained encoder in Frozen-MML and the finetuned encoder in Finetuned-MML) and train the unimodal classifier to get the unimodal predicting performance. Unis-MMC is our proposed method and the MT-MML is the method without the proposed multimodal contrastive loss.

Table 4: Unimodal and Multimodal classification on Food101 with Frozen-MML, Finetuned-MML, MT-MML and UniS-MMC.

Method	Unimodal Performance		Multimodal Performance
	Image	Text	
Frozen-MML	67.75	41.67	71.34
Finetuned-MML	68.81	41.77	93.96
MT-MML	72.72	86.72	94.32
UniS-MMC	**72.81**	87.26	94.73

For unimodal performance, both MT-MML and UniS-MMC perform better than Finetuned-MML and Frozen-MML. This is because the multi-task-based learning framework learns unimodal repre-

sentations under the supervision of unimodal predicting tasks. These obtained representations are much more task-related. We also find that the unimodal predicting with image is better than unimodal predicting with text on Finetuned-MML and Frozen-MML. The reason is that image encoder is pretrained on a similar image classification task while text encoder is pretrained on unrelated language modeling tasks. Multimodal performance is continuously improving among the four different models. Finetuned-MML outperforms Frozen-MML means that fine-tuning the unimodal encoders can extract effective representations than directly utilizing the pretrained model as feature encoder. The task-related unimodal representations benefit the multimodal predicting task and MT-MML obtains a further improvement compared with Finetuned-MML. Obviously, UniS-MMC performs best among four methods, which demonstrates the unimodal representations are further optimized based on the multi-task framework with the novel multimodal contrastive alignment method.

Analysis on the Final Multimodal Decision. We summarize unimodal performance on MT-MML and UniS-MMC and present unimodal predictions in Fig 4. The unimodal prediction consistency here is represented by the consistency of the unimodal prediction for each sample. When focusing on the classification details of each modality pair, we find that the proposed UniS-MMC gives a larger proportion of samples with both correct predictions and a smaller proportion of samples with both wrong decisions and opposite unimodal decisions compared with MT-MML.

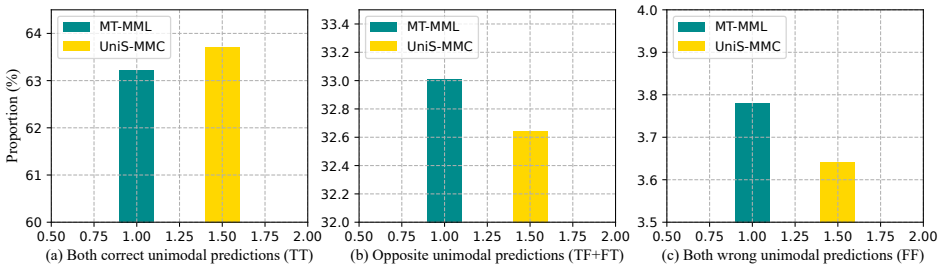

Figure 4: Consistency comparison of unimodal prediction between MT-MML and the UniS-MMC.

Table 5 shows the multimodal predicting results under different unimodal predicting combinations for MT-MML and UniS-MMC. Ideally, all samples should give both correct unimodal predictions and multimodal predictions. As seen from Table 5, our proposed Unis-MMC can give the most correct multimodal classification decisions when both unimodal representations are potentially effective for correct predictions. Also, the promotion of multimodal decisions under positive and both wrong unimodal prediction is smaller. It means that our method is able to reduce the inconsistency and errors of unimodal predictions in multimodal tasks.

Table 5: Multimodal prediction results under different unimodal prediction combinations on MT-MML and UniS-MMC.

Unimodal Prediction	**TT**		**TF+FT**		**FF**		**Multimodal Performance**
Multimodal Prediction	T	F	T	F	T	F	
MT-MML	63.19	0.02	30.54	2.47	0.59	3.19	94.32
UniS-MMC	63.69	0.02	30.43	2.22	0.62	3.02	94.73

5 CONCLUSION

In this work, we propose the Unimodality-Supversied Multimodal Contrastive (UNniS-MMC), a novel method for multimodal fusion to reduce the multimodal decision bias caused by inconsistent unimodal information. Based on the introduced multi-task-based multimodal learning framework, we capture the task-related unimodal representations and evaluate their potential influence on the final decision with the unimodal predictions. Then we contrastively align the unimodal representation towards the relatively reliable modality under the weak supervision of unimodal predictions. This novel contrastive-based alignment method helps to capture more trustworthy multimodal representations. The experiments on four public multimodal classification datasets demonstrate the effectiveness of our proposed method.

REFERENCES

Milad Abdollahzadeh, Touba Malekzadeh, and Ngai-Man Man Cheung. Revisit multimodal meta-learning through the lens of multi-task learning. *Advances in Neural Information Processing Systems*, 34:14632–14644, 2021.

Mohamed Afham, Isuru Dissanayake, Dinithi Dissanayake, Amaya Dharmasiri, Kanchana Thilakarathna, and Ranga Rodrigo. Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9902–9912, 2022.

Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34:24206–24221, 2021.

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017.

Dara Bahri, Heinrich Jiang, Yi Tay, and Donald Metzler. Scarf: Self-supervised contrastive learning using random feature corruption. *arXiv preprint arXiv:2106.15147*, 2021.

Junwen Bai, Shufeng Kong, and Carla P Gomes. Gaussian mixture variational autoencoder with contrastive learning for multi-label classification. In *International Conference on Machine Learning*, pp. 1383–1398. PMLR, 2022.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 423–443, 2018.

Zhipeng Bao, Martial Hebert, and Yu-Xiong Wang. Generative modeling for multi-task visual learning. In *International Conference on Machine Learning*, pp. 1537–1554. PMLR, 2022.

Deblina Bhattacharjee, Tong Zhang, Sabine Süsstrunk, and Mathieu Salzmann. Mult: An end-to-end multitask learning transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12031–12041, June 2022.

Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

Ginevra Castellano, Loic Kessous, and George Caridakis. Emotion recognition through multiple modalities: face, body gesture, speech. In *Affect and emotion in human-computer interaction*, pp. 92–103. Springer, 2008.

Mayee Chen, Daniel Y Fu, Avanika Narayan, Michael Zhang, Zhao Song, Kayvon Fatahalian, and Christopher Ré. Perfectly balanced: Improving transfer and robustness of supervised contrastive learning. In *International Conference on Machine Learning*, pp. 3090–3122. PMLR, 2022.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pp. 794–803. PMLR, 2018.

Yanhua Cheng, Rui Cai, Zhiwei Li, Xin Zhao, and Kaiqi Huang. Locality-sensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3029–3037, 2017.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Chris Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. Efficiently identifying task groupings for multi-task learning. *Advances in Neural Information Processing Systems*, 34:27503–27516, 2021.

Jing Gao, Peng Li, Zhikui Chen, and Jianing Zhang. A survey on deep learning for multimodal data fusion. *Neural Computation*, 32(5):829–864, 2020.

Jun Gao, Wei Wang, Changlong Yu, Huan Zhao, Wilfred Ng, and Ruifeng Xu. Improving event representation via simultaneous weakly supervised contrastive learning and clustering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3036–3049, 2022.

Yuan Gao, Jiayi Ma, Mingbo Zhao, Wei Liu, and Alan L Yuille. Nddr-cnn: Layerwise feature fusing in multi-task cnns by neural discriminative dimensionality reduction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3205–3214, 2019.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. Supervised contrastive learning for pre-trained language model fine-tuning. *arXiv preprint arXiv:2011.01403*, 2020.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pp. 1735–1742. IEEE, 2006.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. *arXiv preprint arXiv:2102.02051*, 2021.

Zongbo Han, Fan Yang, Junzhou Huang, Changqing Zhang, and Jianhua Yao. Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20707–20717, June 2022a.

Zongbo Han, Fan Yang, Junzhou Huang, Changqing Zhang, and Jianhua Yao. Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20707–20717, 2022b.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34:10944–10956, 2021.

Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably). *arXiv preprint arXiv:2203.12221*, 2022.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916. PMLR, 2021.

Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. *Advances in Neural Information Processing Systems*, 33:21798–21809, 2020.

M Kalfaoglu, Sinan Kalkan, and A Aydin Alatan. Late temporal modeling in 3d cnn architectures with bert for action recognition. In *European Conference on Computer Vision*, pp. 731–747. Springer, 2020.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1780–1790, 2021.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*, 2019.

Minseon Kim, Jihoon Tack, and Sung Ju Hwang. Adversarial self-supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:2983–2994, 2020.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pp. 5583–5594. PMLR, 2021.

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

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021a.

Shikun Li, Xiaobo Xia, Shiming Ge, and Tongliang Liu. Selective-supervised contrastive learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 316–325, 2022.

Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pretraining paradigm. *arXiv preprint arXiv:2110.05208*, 2021b.

Tao Liang, Guosheng Lin, Mingyang Wan, Tianrui Li, Guojun Ma, and Fengmao Lv. 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 (CVPR)*, pp. 15492–15501, June 2022.

Huidong Liu, Shaoyuan Xu, Jinmiao Fu, Yang Liu, Ning Xie, Chien-Chih Wang, Bryan Wang, and Yi Sun. Cma-clip: Cross-modality attention clip for image-text classification. *arXiv preprint arXiv:2112.03562*, 2021a.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Yunze Liu, Qingnan Fan, Shanghang Zhang, Hao Dong, Thomas Funkhouser, and Li Yi. Contrastive multimodal fusion with tupleinfonce. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 754–763, 2021b.

Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3994–4003, 2016.

Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 1359–1367, 2020.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34: 14200–14213, 2021.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *ICML*, 2011.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Jonathan Pilault, Amine Elhattami, and Christopher Pal. Conditionally adaptive multi-task learning: Improving transfer learning in nlp using fewer parameters & less data. *arXiv preprint arXiv:2009.09139*, 2020.

Kosmas Pinitas, Konstantinos Makantasis, Antonios Liapis, and Georgios N Yannakakis. Supervised contrastive learning for affect modelling. *arXiv preprint arXiv:2208.12238*, 2022.

Petra Poklukar, Miguel Vasco, Hang Yin, Francisco S Melo, Ana Paiva, and Danica Kragic. Gmc–geometric multimodal contrastive representation learning. *arXiv preprint arXiv:2202.03390*, 2022.

Chengwei Qin and Shafiq Joty. Continual few-shot relation learning via embedding space regularization and data augmentation. *arXiv preprint arXiv:2203.02135*, 2022.

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

Geovany A Ramirez, Tadas Baltrušaitis, and Louis-Philippe Morency. Modeling latent discriminative dynamic of multi-dimensional affective signals. In *International Conference on Affective Computing and Intelligent Interaction*, pp. 396–406. Springer, 2011.

Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*, 2021.

Sijie Song, Jiaying Liu, Yanghao Li, and Zongming Guo. Modality compensation network: Cross-modal adaptation for action recognition. *IEEE Transactions on Image Processing*, 29:3957–3969, 2020.

Aiham Taleb, Matthias Kirchler, Remo Monti, and Christoph Lippert. Contig: Self-supervised multimodal contrastive learning for medical imaging with genetics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20908–20921, 2022.

Junjiao Tian, Wesley Cheung, Nathaniel Glaser, Yen-Cheng Liu, and Zsolt Kira. Uno: Uncertainty-aware noisy-or multimodal fusion for unanticipated input degradation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5716–5723. IEEE, 2020a.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pp. 776–794. Springer, 2020b.

Yao-Hung Hubert Tsai, Tianqin Li, Weixin Liu, Peiyuan Liao, Ruslan Salakhutdinov, and Louis-Philippe Morency. Learning weakly-supervised contrastive representations. *arXiv preprint arXiv:2202.06670*, 2022.

Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc V Gool. Revisiting contrastive methods for unsupervised learning of visual representations. *Advances in Neural Information Processing Systems*, 34:16238–16250, 2021.

Jinghua Wang, Zhenhua Wang, Dacheng Tao, Simon See, and Gang Wang. Learning common and specific features for rgb-d semantic segmentation with deconvolutional networks. In *European Conference on Computer Vision*, pp. 664–679. Springer, 2016.

Tongxin Wang, Wei Shao, Zhi Huang, Haixu Tang, Jie Zhang, Zhengming Ding, and Kun Huang. Moronet: multi-omics integration via graph convolutional networks for biomedical data classification. *bioRxiv*, 2020a.

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

Yang Wang. Survey on deep multi-modal data analytics: Collaboration, rivalry, and fusion. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(1s): 1–25, 2021.

Yikai Wang, Wenbing Huang, Fuchun Sun, Tingyang Xu, Yu Rong, and Junzhou Huang. Deep multimodal fusion by channel exchanging. *Advances in Neural Information Processing Systems*, 33:4835–4845, 2020b.

Zhen Wang, Xu Shan, and Jie Yang. N15news: A new dataset for multimodal news classification. *arXiv preprint arXiv:2108.13327*, 2021.

Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *International Conference on Machine Learning*, pp. 24043–24055. PMLR, 2022.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3733–3742, 2018.

Yuwen Xiong, Mengye Ren, and Raquel Urtasun. Loco: Local contrastive representation learning. *Advances in neural information processing systems*, 33:11142–11153, 2020.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10790–10797, 2021.

Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6995–7004, 2021.

Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 833–842, 2021a.

Miaoran Zhang, Marius Mosbach, David Ifeoluwa Adelani, Michael A Hedderich, and Dietrich Klakow. Mcse: Multimodal contrastive learning of sentence embeddings. *arXiv preprint arXiv:2204.10931*, 2022.

Wenjia Zhang, Lin Gui, and Yulan He. Supervised contrastive learning for multimodal unreliable news detection in covid-19 pandemic. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 3637–3641, 2021b.

Mohammadreza Zolfaghari, Yi Zhu, Peter Gehler, and Thomas Brox. Crossclr: Cross-modal contrastive learning for multi-modal video representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1450–1459, 2021.

A Appendix

A.1 Datasets usage instructions

To make a fair comparison with the previous works, we adopt the following default setting of the split method, as shown in Table 6. Since the UPMC-Food101 dataset does not provide the validation set, we split 5000 samples out of the training set and use them as the validation set.

Table 6: Datasets information and the split results

Dataset	Modalities	#Category	#Train	#Valid	#Test
UPMC-Food-101	2: image, text	101	60085	5000	21683
N24News	2: image, text	24	48988	6123	6124
BRCA	3: mRNA, DNA and miRNA	5	612	-	263
ROSMAP	3: mRNA, DNA and miRNA	2	245	-	106

A.2 EXPERIMENTAL SETTINGS

The model is trained on NVIDIA V100-SXM2-16GB and NVIDIA A100-PCIE-40GB. The corresponding Pytorch version, CUDA version and CUDNN version are 1.8.0, 11.1 and 8005 respectively. We utilize Adam as the optimizer and use ReduceLROnPlateau to update the learning rate. We use Adam Kingma & Ba (2014) as the model optimizer. The temperature coefficient for contrastive learning is set as 0.07 and the loss coefficient in this paper is set as 0.1 to keep loss values in the same order of magnitude. The code is attached and will be available on GitHub. Some key settings of the model implementation are listed as followings:

Table 7: Detailed setting of the hyper-parameter for UPMC-Food-101, BRCA and ROSMAP

Item	UPMC-Food-101	N24News	BRCA	ROSMAP
Batch gradient	128	128	-	-
Batch size	32	32	-	-
Learning rate (m)	2e-5	1e-4	2e-3	2e-3
Dropout (m)	0	0	0.5	0.1
Weight decay	1e-4	1e-4	1e-3	1e-3

A.3 LEARNING WITH A SIGNALE MODALITY

We show the unimodal classification results from different unimodal backbones on text-image datasets in the following Table 8.

Table 8: Unimodal classification performance with different backbones on Food101 and N24News.

Dataset		**Backbone**	**Food101**	**N24News**
Image		ViT	$73.1_{\pm 0.2}$	$54.1_{\pm 0.2}$
Text		BERT	$86.8_{\pm 0.2}$	-
	Heading	BERT	-	$72.1_{\pm 0.2}$
		RoBERTa	-	$71.8_{\pm 0.2}$
	Caption	BERT	-	$72.7_{\pm 0.3}$
		RoBERTa	-	$72.9_{\pm 0.4}$
	Abstract	BERT	-	$78.3_{\pm 0.3}$
		RoBERTa	-	$79.7_{\pm 0.2}$