Complementary-Contradictory Feature Regularization against Multimodal Overfitting

Antonio Tejero-de-Pablos
CyberAgent
Shibuya, Tokyo, Japan
antonio_tejero@cyberagent.co.jp

Abstract

Understanding multimodal learning is essential to design intelligent systems that can effectively combine various data types (visual, audio, etc.). Multimodal learning is not trivial, as adding new modalities does not always result in a significant improvement in performance, i.e., multimodal overfitting. To tackle this, several works proposed regularizing each modality's learning speed and feature distribution. However, in these methods, characterizing quantitatively and qualitatively multimodal overfitting is not intuitive. We hypothesize that, rather than regularizing abstract hyperparameters, regularizing the features learned is a more straightforward methodology against multimodal overfitting. For the given input modalities and task, we constrain "complementary" (useful) and "contradictory" (obstacle) features via a masking operation on the multimodal latent space. In addition, we leverage latent discretization so the size of the complementary-contradictory spaces becomes learnable, allowing the estimation of a modal complementarity measure. Our method successfully improves the performance of datasets with modality overfitting in different tasks, providing insight into "what"and "how much" is learned from each modality. Furthermore, it facilitates transfer learning to new datasets. Our code and a detailed manual are available at https://github.com/CyberAgentAILab/CM-VQVAE.

1. Introduction

Multimodal learning studies how to leverage a variety of data modalities (*e.g.*, color, audio, depth, text, etc.) to solve a given task (*e.g.*, action recognition [34], emotion recognition [15], video grounding [35])[1]. An effective combination of modalities is beneficial for task solving, as different modalities provide unique perceptions on the input.

[1]Our focus is on multimodal *discriminative* models and tasks (e.g., object recognition), not *generative* (e.g., vision-language cross-generation).

Table 1. Multimodal overfitting experiment in emotion recognition (CREMA-D); adding audio to image lowers accuracy.

Image-only	Audio-only	Image+Audio
59.0%	53.49%	54.43%

However, interactions among multimodal data are hard to explain, and thus, designing effective multimodal architectures is challenging. For example, adding new modalities does not directly translate into a performance improvement [19]; this phenomenon is known as multimodal "imbalance" or "overfitting". Table 1 shows an example of how adding the audio modality to the visual modality actually worsens the performance. To solve it, recent works proposed inducing a more harmonic co-learning of modalities by regularizing the learning rate [28] and gradients [19] of one modality over the other. Later it was discovered that adapting the learning pace does not prevent features from one modality interfering with the other, leading to a limited improvement [6]. Instead, they proposed a method to avoid inter-class feature overlaps in the latent space by shrinking each modality's feature distribution. However, such a regularization does not study what the interfering features are nor their removal.

A common technique in discriminative multimodal learning is applying attention weights to the fused multimodal features [5, 9, 34]. This allows emphasizing relevant features and diminishing the unnecessary ones by applying weighting operations (*i.e.*, gated attention) to the multimodal features. We argue that regularizing feature fusion itself can also mitigate modality overfitting and provide more intuitive qualitative and quantitative insights. In this process, we aim to adopt features that support solving the given task in combination with the other modality (*i.e.*, *complementary*), and to ignore features unrelated to the task that hinder the learning of the other modality (*i.e.*, *contradictory*). A similar concept was introduced in [14, 22], in which the multimodal latent space is disentangled into fea-

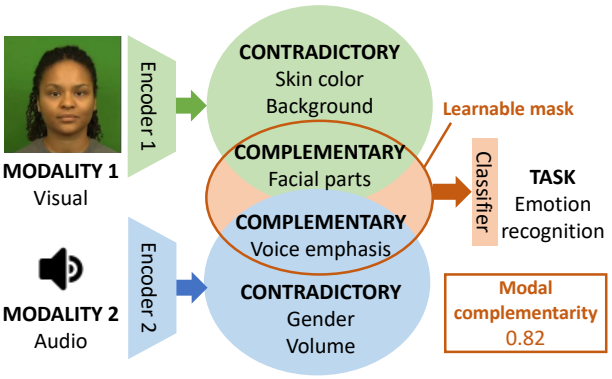

Figure 1. End-to-end pipeline of a generic multimodal discriminative network. To mitigate multimodal overfitting, our novel method learns a mask to separate the features of each modality into *complementary* (supports solving the task) and *contradictory* (hinders solving the task). This helps characterizing multimodal learning quantitatively and qualitatively, and is also transferable to unseen data.

tures that are common among modalities (*shared*) and those that are modality exclusive (*private*). However, they target generative models (*i.e.*, text-to-image and viceversa), and cannot be directly applied to classification tasks. Moreover, their latent space size is fixed as a hyperparameter, having to decide their representation capability manually. This could be solved by discretizing the latent spaces into "feature units" so the amount of complementary and contradictory information becomes measurable. This also allows the definition of a metric to characterize the strong and weak modalities, as well as their "complementarity" for solving the given task.

Figures 1 and 2 depict our proposed methodology and architecture, the Complementary Multimodal VQVAEs (CM-VQVAE). Complementary and contradictory latent spaces are created by learning a mask on the features to solve the given task. They are automatically learned using a combination of modal reconstruction and regularized masking. That is, by penalizing the overuse of features we manage to regularize the latent space to remove features without harming accuracy. Moreover, our separation can facilitate transfer learning on unseen datasets.

To summarize, our contributions are threefold:

- We propose a novel approach against multimodal overfitting via regularization of the multimodal fusion, separating features into complementary and contradictory.
- We implement our methodology via the CM-VQVAE, in which the size of the latent feature spaces is learnable, and propose feature-masking regularization. This allows defining a *modal complementarity* measure.
- We provide thorough evaluation results in various multimodal tasks and datasets that have different degrees of modal complementarity and overfitting, including its application to transfer learning.

2. Related work

2.1. Multimodal overfitting mitigation

In discriminative multimodal learning, multiple modalities are combined with the aim of outperforming unimodal approaches for a given task (*e.g.*, classification, segmentation). In certain combinations of multimodal datasets and tasks, contrary to the intuitions in deep learning, increasing the number of data modalities does not lead to better performance. Recent studies [28] unveiled that the main reason for unsuccessful multimodal learning in discriminative models is the phenomenon of multimodal overfitting (i.e., overfitting of the *strong* modality over the underfitting *weak* modality). Traditional anti-overfitting techniques, such as random feature dropout [32] and reducing the network parameters, are not learnable, and thus, involve a lot of trial and error. Thus, Wang *et al.* [28] proposed regularizing the learning rate for each part of the network corresponding to each modality. In a similar fashion, Peng *et al.* [19] proposed regularizing the backpropagation of gradients to each individual modality during training. This concept of strong-weak modalities was also leveraged to gain robustness against missing/noisy modalities [16].

Motivated by the fact that multimodal DNNs can exploit undesired features [8], Fan *et al.* [6] proved that the performance of learning pace-based regularization approaches is limited by the interference of features learned from each modality, and proposed a feature-based regularization method. Since sparsely distributed feature spaces would interfere among modalities, regularization is applied to densely map unimodal features around their class prototype (*i.e.*, centroid). However, if "interfering" features really exist, we hypothesize that their removal would be a more effective regularization. For example, intuitively, a "prototype" face in a multiracial emotion recognition dataset should not feature any skin color.

The aforementioned methods represent the body of comparison works with our same goal. Unlike previous works, we opt for a more straightforward approach to regularize multimodal features. Our method learns a mask and "encourages" the network to reduce the amount of features employed to solve the task, in particular, those considered unnecessary (*contradictory*) by the network. Unlike dropout, values are not arbitrarily zeroed, but masks are consistent among iterations. The remaining works mentioned in this section were used as a reference to build our method but their goal is different.

2.2. Multimodal feature separation

Feature separation allows understanding data and machine learning processes. When the separated features are interpretable this is called semantic disentanglement [22]. Rather than classification tasks, feature separation has been more extensively studied in the context of generative tasks. Shared/private multimodal latent spaces (SP spaces) were proposed by Shi *et al*. [22] and Lee *et al*. [14] as an analytic method for *generative* multimodal models based on how the brain embeds information across modalities [20, 24]. Specifically, when generating one modality (*e.g*. a black-and-white handwritten digit) from another modality (*e.g*. a colored printed digit), they consider the existence of common features between modalities (shared, *e.g*. the number shape) and exclusive features (private, *e.g*. background color). By separating and then excluding private features from the generative process, not only performance improved, but also semantic disentanglement was possible in some cases. As a result, SP spaces outperformed previous multimodal models in generation tasks (*i.e*., text-to-image and viceversa).

2.2.1 Multimodal feature separation for classification

In multimodal discriminative models, modal-wise features are *fused* at some point of the pipeline (*e.g*. early, late) via element-wise sum/product or concatenation. Unlike cross-modal generation and retrieval, classification tasks do not require finding the mutual information between modalities [1] nor overlapping the features of all modalities in the projected latent space [7]. Rather, techniques such as gated attention weights have been applied when fusing multimodal features in order to emphasize or diminish the amount of information learned from each modality [5,9,34]. Some works in multimodal discriminative models elucidated that the heterogeneity and information contradiction present across modalities hinders the fusion between multimodal features, and considered an attention-based multimodal feature separation approach that allows for an effective fusion [15,33]. However, feature separation is yet to be applied to the problem of multimodal overfitting.

We hypothesize that regularizing the fusion process can allow separating features that can be combined to solve the task (*complementary*) from those that interfere in the task (*contradictory*). As we aim to exclude the contradictory features, instead of directly applying attention, we propose binarizing gating weights into 0 and 1 mask values in order to remove (or keep) those features completely when solving the given multimodal task. For this, we employ a regularized mask that encourages feature separation as much as possible without hindering classification performance. Note that, in biased datasets, contradictory features may contain disentangled semantics such as skin color and gender

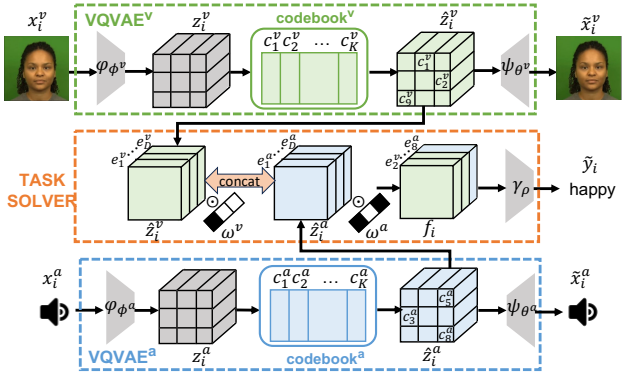

Figure 2. Overview of the proposed method, for the case of two modalities, visual v and audio a. After unimodal features are encoded, they are fused and masked altogether before classification.

(Fig. 1), but in practice, a total feature disentanglement is unfeasible without extra labels for gender, etc. However, our hypothesis is irrespective of whether the separated features have a disentangled semantic meaning or not.

2.3. Transfer learning

As Bengio *et al*. [3] stated, feature separation can be useful in tackling many downstream tasks, and help improve robustness and generalizability of models. This has been proved true for the case of unimodal transfer learning, by separating style and content of images [17]. However, works on multimodal transfer learning rarely leverage feature separation due to its complexity. Instead, they rather rely on adversarial [18] and contrastive [11] losses to solve the gap between source and target data.

Thus, we consider our work as an opportunity to explore the potential of feature separation for transfer learning between multimodal datasets.

3. Methodology

Figure 2 shows the implementation of our proposed method, CM-VQVAE, for the case of two modalities. It contains two main modules. First, VQVAE provides a discretization of all features modality-wise. Then, the task-solver masks part of the features (contradictory) and uses the rest (complementary) for classification. We hypothesize that regularizing this feature-removal process allows optimizing the latent spaces so multimodal overfitting from interfering features is prevented.

3.1. Modality-wise feature extraction

This module is implemented via Vector Quantized-Variational Autoencoders (VQVAEs) [25]. Recently, VQ-VAEs have been utilized to support image-text multimodal generation tasks [21]. Compared to VAEs, discrete vari-

ables have been proved to be more interpretable and space-efficient [4], and they avoid drawbacks from continuous space, such as posterior collapse [12]. In particular, training a multimodal VAE latent space is challenging as the number and disparity of the modalities increases [31].

Let our training dataset (N samples) be $\{x_i, y_i\}_{i=1,2...N}$, where a sample i consists of M modalities $x_i = \{x_i^m\}_{m=1,2...M}$ and a label $y_i \in \{1, 2...J\}$ indicating one of J classes. Each x_i^m is input to and reconstructed by a VQVAE^m module. VQVAE^m learns an encoder φ with parameters ϕ^m, a decoder ψ with parameters θ^m, and a codebook $C^m = \{c_k^m\}_{k=1,2...K}$ of size K, where $c_k^m \in \mathbb{R}^D$. Modality features are extracted by $z_i^m = \varphi(x_i^m|\phi^m) \in \mathbb{R}^{D\times H\times W}$, where D is the number of feature maps, and $H \times W$ is their size. Then, VQVAE discretizes the original features into the quantized code space by replacing them with their nearest-neighbor in the learned codebook C^m, $z_i^m \rightarrow \hat{z}_i^m = c_q^m$, where $q = \arg\max_k ||z_i^m - c_k^m||$. Finally, the reconstruction is output by the decoder $\tilde{x}_i^m = \psi(c_i^m|\theta^m)$.

The original modality is reconstructed so the encoder does not discard the *contradictory* features, since keeping all features is necessary for an effective separation [14, 22, 27, 31]. Also, although the decoder is not used during inference, it allows visualizing the feature space (Figs. 3 and 4).

3.2. Multimodal feature separation

The task-solver fuses the encoded features of the modalities $\hat{z}_i^m$ ($m = 1...M$) and masks them before solving the given multimodal task. As previous works [9, 19, 28, 30] we choose concatenation as our preliminary fusion method. Formally, $\hat{z}_i = [\hat{z}_i^1; ...; \hat{z}_i^M]$. Masking is performed after concatenation, so features are separated considering all multimodal information. In the multimodal feature fusion step, masking allows entirely removing part of the features (contradictory) versus those that are used in the task (complementary). This is stricter than attention, which emphasizes or diminishes the effect of individual values. For separation, we consider the channel dimensions of $\hat{z}$ our units for masking, as feature maps in convolutional networks are considered to contain independent semantic information [30] (*e.g.*, colors, shapes).

Formally, we learn a mask $\Omega \in \mathbb{R}^{M\cdot D}$ and apply it to the encoding $\hat{z}_i$. During masking, Ω is binarized (ω), so each weight Ω_d take the value 0 if they are lower than a threshold t, or 1 if they are higher than t. Formally, $f_i = \hat{z}_i \odot \omega$, where each value of ω masks an entire channel of $\hat{z}_i$. In the resulting $f_i \in \mathbb{R}^{(M\cdot D)\times H\times W}$ remain the multimodal complementary features, as those considered contradictory by the network were zeroed (Fig. 2 omits them for simplicity). Finally, the multimodal classifier γ with parameters ρ models a joint distribution over the separated features, from which to obtain the label predictions $\tilde{y}_i = \gamma(f_i|\rho)$.

3.3. Training

Training the proposed network end-to-end requires optimizing several objectives:

$$\mathcal{L} = \mathcal{L}_{\text{recons}} + \mathcal{L}_{\text{code}} + \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{compl}} \tag{1}$$

Reconstruction loss. $\mathcal{L}_{\text{recons}}$ is the mean square error between the input and the reconstruction:

$$\mathcal{L}_{\text{recons}} = \sum_{m=1}^{M} ||x_i^m - \hat{x}_i^m||_2^2. \tag{2}$$

Codebook loss. $\mathcal{L}_{\text{code}}$ is the error between the features and the learned codes:

$$\mathcal{L}_{\text{code}} = \sum_{m=1}^{M} ||z_i^m - \hat{z}_i^m||_2^2. \tag{3}$$

Task loss. $\mathcal{L}_{\text{task}}$ is the learning objective corresponding to the given multimodal task, *e.g.* for classification, the cross entropy loss of a batch B:

$$\mathcal{L}_{\text{task}} = -\sum_{i=1}^{B} y_i \log(\text{Softmax}(\tilde{y}_i)). \tag{4}$$

Regularization term. This term is key in our method for feature-separation of interfering features to prevent multimodal overfitting. $\mathcal{L}_{\text{compl}}$ penalizes high mask values Ω_d in order to encourage the masking of features.

$$\mathcal{L}_{\text{compl}} = \lambda \sum_{d=1}^{M\cdot D} \Omega_d \tag{5}$$

Instead of assuming that the network will automatically ignore all contradictory features in the multimodal latent space, we explicitly promote masking via this term. This allows for better separation (see Sec. 4.4), and the encouraged feature pruning improves generalization [2]. This regularization is based on the l_1-norm, since the l_0-norm (using ω^m instead of Ω^m) is hard to optimize in practice, as observed in [9]. The hyperparameter λ (empirically $\lambda = 0.0001$) adjusts the balance with $\mathcal{L}_{\text{task}}$.

In summary, optimizing this loss equals to *reducing the number of features as much as possible while aiming for the highest task accuracy*. In our experiments, we utilize SGD optimizer and learning rate 0.001 (see Appendix 3 for the rest of the hyperparameter values and architectural details).

3.4. Complementarity ζ

Previous works [14, 22] set the number of private/shared features as a hyperparameter based on the complexity of the generated modality (*e.g.*, ImageNet $\gg$ MNIST). However, having to manually hardcode the size of the separated

spaces for a classification task is a big limitation, as usually it is unsure "how much information of each modality is necessary to solve the given task". As learning ω is an optimization of the size of our feature spaces, we can measure *how much* information from one modality is being learned over the other by tracking which values in ω masked which modality m: ω^m. Then, the ratio (%) of feature units of modality m (*i.e.*, $\hat{z}_i^m$) contained in the *complementary* space is:

$$\begin{aligned} \text{compl}^m &= 100 \cdot \frac{1}{D} \sum_{d=1}^{D} \omega_d^m, \\ \omega_d^m &= \begin{cases} 1, & \text{if } \Omega_d^m > t \\ 0, & \text{otherwise} \end{cases} \end{aligned} \tag{6}$$

Thus, the parameter compl^m measures how much information modality m contributes to the given task. Similarly, the *contradictory* space size contr^m measures the information from that modality that is not used to solve the task, and can be calculated by for the case $\Omega_d^m \leq t$.

In our hypothesis, if one modality is redundant, the other will contribute most of the features (*i.e.*, $\text{compl}^1 \gg \text{compl}^2$). Likewise, if both are equally important the complementary feature space will be more balanced (*i.e.*, $\text{compl}^1 \approx \text{compl}^2$). Therefore, we can define an estimation of complementarity ζ between modalities as:

$$\zeta^{1,2} = \frac{min(\text{compl}^1, \text{compl}^2)}{max(\text{compl}^1, \text{compl}^2)} \tag{7}$$

Complementarity ζ is a ratio in the range $[0, 1]$. Note that the size of our learned separated spaces meets the following:

$$\sum_{m=1}^{M} \text{compl}^m + \text{contr}^m = 100(\%) \tag{8}$$

4. Experiments

First, we evaluate the efficacy of our method against multimodal overfitting in a variety of tasks and modalities, and explore their learning dynamics (*i.e.*, separation, complementarity). Based on the public implementation of VQ-VAE[2], we use the ResNet-based [10] network as the backbone of our modules, which is also used in the comparison methods [6,19,28]. Please see Appendix 5 for a comparison with other architectures. Mask values are initialized to 0.1 and $t = 0.05$, so learning starts with all features available in the task solver (details on the selection of t in Appendix 6). In addition, the VQVAE codebooks contain a set of 512 VQVAE codes of size $D = 64$.

We used the following datasets for evaluation (see Appendix 2 for data preprocessing and other details). While the related work focuses on audio-visual datasets, we opted for a more varied combination of modalities to study different types of modal complementarity.

CREMA-D [29] uses color video frames and audio modalities for emotion recognition. We chose this dataset since it is the benchmark in all previous methods for modality overfitting mitigation.

PennAction [36] is a dataset for pose recognition via color and pose modalities. We purposely chose PennAction because, unlike CREMA-D, modalities are very redundant (the task can be almost exclusively solved via pose).

NYUv2 [23] is a dataset for semantic segmentation of indoor scenes via color and depth modalities. We purposely chose NYUv2 to evaluate our method in a task different than recognition and with no multimodal overfitting.

4.1. Quantitative evaluation

Table 2 summarizes the task accuracy obtained by our method, and the ratio of complementary features for each modality. Our method prevents multimodal overfitting (Multi $<$ max(unimodal) $<$ Ours), where Multi means using all features (*i.e.*, vanilla concatenation). Our method also outperforms the baselines, *feature dropout* and *gated attention*. First, the dropout [32] configuration masks a set of feature maps chosen randomly in each iteration with equal probability. The results show that our learnable masking is more effective than randomly removing multimodal features. Then, we compare attention weights instead of masking, that is, we use Ω instead of ω (*i.e.*, no binarization). This methodology has been used in many previous works [9, 34], but similar to [19], *gated attention* did not provide the best performance against multimodal overfitting. Finally, our method proves effective when masking is properly regularized (Ours† $<$ Ours).

Regarding the amount of features learned, in general, the feature space gets compressed. For CREMA-D, the ratio of image vs. audio that make up the latent spaces is balanced, which results in a high modal complementarity ζ. This means both image and audio are considered equally necessary for emotion recognition by our method. In contrast, for PennAction, the pose information is mostly used to solve the task, while image information is largely ignored, resulting in a low ζ, as expected. For NYUv2, the network uses slightly more depth features than color. As color and depth have spatial consistency, part of the features are considered redundant.

Our experiments allow observing an interesting phenomenon: the modalities with the lower performance (*i.e.*, audio in CREMA-D, and color in PennAction and NYUv2) have a bigger ratio of complementary features in the unimodal setting. However, when combined in the multimodal setting, their complementary ratio is lower. Thus, including a *stronger* modality allows to make up for contradic-

[2] https://github.com/zalandoresearch/pytorch-vq-vae

Table 2. Task accuracy and modal complementarity of the proposed method. The upper rows are the unimodal and vanilla multimodal (all features concatenated) cases, the mid rows are the baselines, and the lower row is the proposed method († indicates no regularization).

CREMAD	**Acc.(%)**	**compl^c**	**compl^a**	ζ
Color	59.0	10.16	0.0	0.0
Audio	53.49	0.0	21.09	0.0
Multi	53.89	50.0	50.0	1.0
Drop.	53.36	25.0	25.0	1.0
Gated	57.66	—	—	—
Ours†	54.43	41.41	25.78	0.62
Ours	**65.32**	**13.28**	**13.28**	**1.0**

PennAc	**Acc.(%)**	**compl^c**	**compl^p**	ζ
Color	31.18	34.37	0.0	0.0
Pose	94.02	0.0	18.75	0.0
Multi	85.95	50.0	50.0	1.0
Drop.	90.1	25.0	25.0	1.0
Gated	91.86	—	—	—
Ours†	88.04	27.34	44.53	0.61
Ours	**95.22**	**5.47**	**14.06**	**0.39**

NYUv2	**mIoU**	**compl^c**	**compl^d**	ζ
Color	21.16	19.53	0.0	0.0
Depth	33.1	0.0	16.4	0.0
Multi	34.32	50.0	50.0	1.0
Drop.	37.28	25.0	25.0	1.0
Gated	34.41	—	—	—
Ours†	36.78	31.25	45.31	0.68
Ours	**38.66**	**8.59**	**10.94**	**0.78**

tory features in the *weakest* modality. This can be also observed since the amount of complementary features in the regularized space is lower than the sum of its respective unimodal counterparts (*e.g.*, Ours[compl^c+compl^a] $<$ Color[compl^c]+Audio[compl^a]). Another interesting phenomenon is that, by applying the regularization term (*i.e.*, $\mathcal{L}_{\text{compl}}$) in the loss, the amount of complementary features (*i.e.*, those used to solve the task) is greatly reduced, resulting in a boost in performance. That is, without regularization the network converges to suboptimal solutions that employ more features. The reason why feature dropout has a constant size is that it is set so each modality has the same probability (0.5) of being masked ($25.0 = 100\% * 0.5/2$ modalities).

4.2. Qualitative evaluation

Figures 3 and 4 visualize the multimodal separation in our method, for all the features $\psi(\hat{z}_i^m|\theta) = \tilde{x}_i^m$, the complementary features $\psi(\hat{z}_i^m \odot \omega^m|\theta)$, and the contradictory features $\psi(\hat{z}_i^m \odot \overline{\omega^m}|\theta)$. The $\hat{x}_i^m$ are successfully reconstructed from the the original data, which indicates that the VQ-VAE codebooks learned faithful representations. The reconstructed complementary and contradictory spaces are a visualization of *what* features the network employs to solve the given task. Note that, while in some cases they have a semantic meaning (*e.g.*, gender), this is not always the case neither is a requirement for the applicability of our method. In CREMA-D (Fig. 3), the faces (a) reconstructed from the complementary features are missing skin color variations. This means that the network was able to automatically learn that those features are not helpful to solve the task. The audio modality (b) also seems to represent more basic voice signals in the complementary space rather than the contradictory, which contains a lot of detail. In PennAction (Fig. 4), complementary features (a) display very basic bi-color shapes, while most visual information is left in the contradictory space. In the pose modality (b), complementary features better reflect the input pose, while the contradictory space seems to contain pose information useful to reconstruct the input but irrelevant to the action class. The reconstructed poses from Penn Action show similar results to [6], in which the complementary features seem to illustrate the *prototypical* figure of an action, while the contradictory features appear as the user-specific noise.

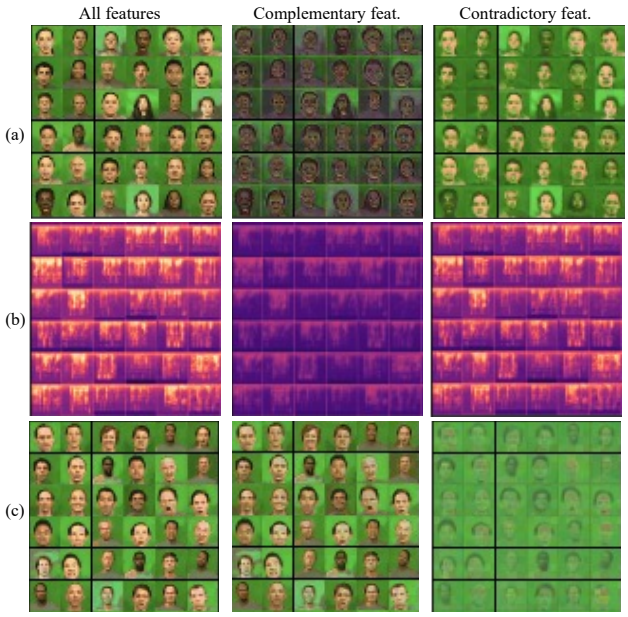

Figure 3. All features, complementary features and contradictory features reconstructed with the learned decoder for modalities (a) Color, (b) Audio and (c) Color (w/o regularization) in CREMA-D.

We must keep in mind that the decoder parameters ψ_θ already model some prior knowledge of the modalities, and thus, they can add information to the reconstructions (*e.g.*, basic shapes, etc.), but rarely remove. Appendix 4 and 1 contain the segmentation and separation results for NYUv2, and a basic example on the Digits dataset for further comprehension of the conceptual differences between our method and the related work [6, 14].

4.3. Ablation study

Table 3 shows a comparison with the related work in multimodal overfitting on the common setting (*i.e.*, ResNet backbone and audio-visual dataset CREMA-D). Our method provides the highest performance boost compared to the vanilla multimodal case. These results also indicate that feature-based regularization approaches are a promising novel methodology against multimodal overfitting.

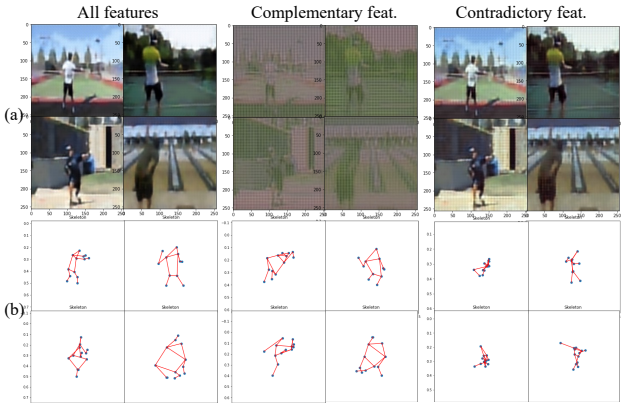

Figure 4. All features, complementary features and contradictory features reconstructed with the learned decoder for modalities (a) Color and (b) Pose in PennAction.

Table 3. Comparison with the previous works in multimodal overfitting on the common setting.

Method	Acc. (%)	Improv.	Reg. type
Multi (vanilla)	53.89	—	—
Grad-Blend [28]	56.8	Δ2.91	Learn.
OGM-GE [19]	57.7	Δ3.81	Learn.
PMR [6]	61.1	Δ7.21	Feat.
Ours	**65.32**	**Δ11.43**	Feat.

Table 4. Ablation study on the main components of our method.

Config.	$C^m + \psi_\theta^m$	Ω	Acc. (%)	ζ
(i)	×	×	53.89	—
(ii)	✓	×	51.88	—
(iii)	×	✓	57.12	0.45
Ours	✓	✓	**65.32**	**1.0**

Table 4 examines the effectiveness of the main components of CM-VQVAE. The lack of a mask implies the impossibility of calculating complementarity ζ. We show the results on CREMA-D, as it is more complex than PennAction but less challenging than NYUv2. Config. (i) indicates not using reconstruction in our pipeline (*e.g.*, omitting the codebook and decoder) as well as not learning the mask of each modality. The lack of a mask implies using all features, complementary and contradictory to solve the given multimodal task. However, since the reconstruction of the original modality is not necessary, irrelevant features are likely not encoded, and thus, their interference is reduced. Config. (ii) includes reconstruction but omits masking, resulting in the worst accuracy. The reason is that contradictory features are kept but not masked, so they hinder task accuracy. Config. (iii) omits reconstruction but keeps the mask. Finally, our proposed configuration displays the highest accuracy.

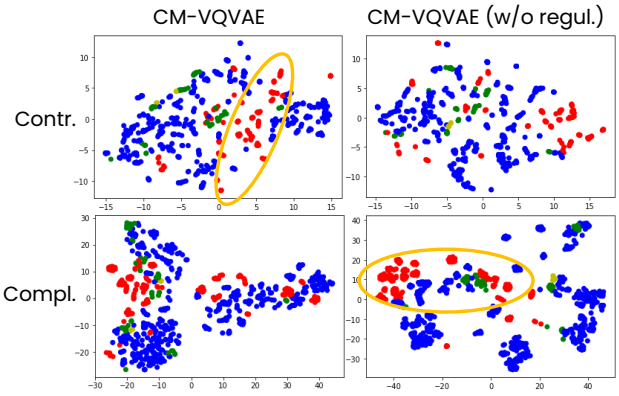

Figure 5. TSNE [26] plot of the learned complementary/contradictory visual features in the CREMA-D dataset. Each color represents a different ethnicity: African American (red), Asian (green), Caucasian (blue) and Unknown (green).

Table 5. Linear separability of complementary and contradictory features for ethnicity (E) and gender (G) in the color and audio modalities respectively of CREMA-D.

Acc. (%)	E (compl^c)	E (contr^c)	G (compl^a)	G (contr^a)
Ours†	0.79	0.60	0.51	0.43
Ours	0.53	**0.82**	0.42	**0.69**

4.4. Application to transfer learning

We take the example of the visual modality in the emotion recognition task to study the effect of encouraging feature separation via regularization. Fig. 3 (c) shows the reconstruction of complementary and contradictory features when no regularization term is applied. Compared with (a), we can observe that skin color features seem not fully removed, and thus, they are still used for classification. Fig. 5 provides an alternative visualization by plotting the complementary and contradictory features of the visual modality in a 2D space. Particularly, the red samples (labeled in the dataset as African American) seem to be more clearly separable in the complementary space (*i.e.*, used to solve the task) when regularization is not used. For further clarification, we trained a linear probe on the complementary/contradictory features for ethnicity (E) and gender (G) labels in the color and audio modalities respectively. Table 5 shows that those features are stronger in the contradictory space when applying regularization.

Relying on such features would harm the generalizability of the trained model to, *e.g.*, be finetuned with new data. We evaluate the transferability of our multimodal learning on the RML emotion database [29]. RML shares 5 of the 6 emotion labels with CREMA-D, which makes it a good target for transfer learning. Tab. 6 shows the accuracies by training our method with RML from scratch and by finetuning our method pretrained on CREMA-D, with and without

Table 6. Transferability of our multimodal learning from CREMA-D to RML emotion datasets.

Method	**Accuracy** (%)	ζ
No transfer (RML only)	74.35	0.75
OGM-GE [19]	66.67	–
Ours w/o regul.	71.21	0.59
Ours	**77.27**	**0.94**

regularization (the accuracy on the unshared labels, *i.e. neutral* and *surprise*, is not calculated). As we hypothesized, using regularization successfully transfers the features from CREMA-D, outperforming the base performance of RML. Additionally, we finetuned the pretrained model in [19] provided by the authors[3] for comparison.

5. Discussion and conclusions

General performance. Our results indicate that our method is capable of learning which features are useful to solve the task (complementary) and which ones are not (contradictory), so that the performance is boosted by mitigating interference between multimodal features. In addition, our methodology allows visualizing and quantifying the learned features of each modality, which provides more insight regarding the phenomenon of overfitting than previous works. As shown in the ablation study, including modality reconstruction during feature masking is required for better accuracy, which adds computational cost during training (model parameters and loss functions). However, the decoder is not necessary during inference, when the pipeline behaves as a regular discriminative model.

Regularization term. As shown in the experiments, removing regularization affects our both quantitatively and qualitatively. Currently, the hyperparameters λ and t are calculated empirically, but we will study optimal ways of choosing this parameter in the future. It may seem that encouraging masking could eventually zero out all features, causing gradients to vanish and stopping learning. However, that is not the case. Zero values in ω still have a weight value in Ω, which means gradients still have an effect even in masked features, with a chance to be unmasked again.

Feature separation. In our method, the network disentangled features such as skin color in CREMA-D and colored lines in Digits (see Appendix 1) based on their significance for the task to solve, without the need of any ad-hoc discriminative loss (*e.g.*, adversarial loss [13]). This separation was also transferable to a similar task without access to the downstream data during pretraining. We are aware that the presence of complementary/contradictory features with semantic meaning may not occur in all problems (*e.g.*, there is no equivalent to CREMA-D's skin color in NYUv2). Nevertheless, our method still provided a boost regardless. **Note:** *The purpose of our method is to approach multimodal overfitting via feature regularization, which indirectly results in feature compression, and sometimes in bias-reduction and feature disentanglement. However, we are not proposing novel feature compression, disentanglement nor bias mitigation methods.*

Complementarity. Results in CREMA-D and NYUv2 indicate that a more balanced latent space between modalities (higher ζ in Tab. 2) leads to a better performance. These results are in line with learning pace-based multimodal overfitting methods [19, 28]. However, for datasets where one modality is negligible (color in PennAction or SVHN in Digits), best results involve lower complementarity. Appendix 7 contains additional graphs on how complementarity varies during training, and references to other definitions of multimodal complementarity found in the related work. In the future, we will study the relationship between modal complementarity and how to select appropriate multimodal learning techniques that lead to better performance.

Extension to $M > 2$. All the processes described in Sec. 3 are valid for any number of modalities. Thus, adding new modalities (*e.g.* text) is theoretically plausible, although it would imply more discrepancies among modalities. A possible extension of our method is discussed in Appendix 8. We leave these intriguing challenges to future work.

5.1. Conclusions

We propose a novel method for regularizing discriminative models against multimodal overfitting, based on learnable masking for feature separation. We extrapolated the concept of shared/private latent spaces from multimodal generative models to multimodal discriminative models: the complementary/contradictory spaces. We encouraged the network to learn the essential features for the given task via regularization, and observed that (i) including a stronger modality allows to make up for contradictory features in the weakest modality, and (ii) the amount of complementary features is greatly reduced, converging to a more optimal solution. This results on a boost in performance, which is more noticeable in the presence of complementary/contradictory features with semantic meaning (*e.g.*, CREMA-D > PennAction > NYUv2). Moreover, transfer learning is also benefited from this regularization. Future work involves exploring the usefulness of this method for multimodal learning regardless of multimodal overfitting.

6. Acknowledgements

I would like to thank Mayu Otani for her advice and support in this research.

[3] https://zenodo.org/record/6778788

References

[1] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. MultiMAE: Multi-modal multi-task masked autoencoders. In *Proceedings of the European Conference on Computer Vision*, pages 348–367, 2022. 3

[2] Brian Bartoldson, Ari Morcos, Adrian Barbu, and Gordon Erlebacher. The generalization-stability tradeoff in neural network pruning. *Advances in Neural Information Processing Systems*, 33:20852–20864, 2020. 4

[3] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. 3

[4] Ruben Cartuyvels, Graham Spinks, and Marie-Francine Moens. Discrete and continuous representations and processing in deep learning: looking forward. *AI Open*, 2:143–159, 2021. 4

[5] Xiao Dong, Xunlin Zhan, Yangxin Wu, Yunchao Wei, Michael C Kampffmeyer, Xiaoyong Wei, Minlong Lu, Yaowei Wang, and Xiaodan Liang. M5Product: Self-harmonized contrastive learning for e-commercial multi-modal pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21252–21262, 2022. 1, 3

[6] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. PMR: Prototypical modal rebalance for multimodal learning. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2, 5, 6, 7

[7] Yuan Gong, Andrew Rouditchenko, Alexander H Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James R Glass. Contrastive audio-visual masked autoencoder. In *Proceedings of the International Conference on Learning Representations*, 2023. 3

[8] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2017. 2

[9] 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*, pages 20707–20717, 2022. 1, 3, 4, 5

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 5

[11] Donghyun Kim, Yi-Hsuan Tsai, Bingbing Zhuang, Xiang Yu, Stan Sclaroff, Kate Saenko, and Manmohan Chandraker. Learning cross-modal contrastive features for video domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13618–13627, 2021. 3

[12] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *Proceedings International Conference on Learning Representations*, 2014. 4

[13] Gihyun Kwon and Jong Chul Ye. Diagonal attention and style-based GAN for content-style disentanglement in image generation and translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13980–13989, 2021. 8

[14] Mihee Lee and Vladimir Pavlovic. Private-shared disentangled multimodal VAE for learning of latent representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1692–1700, 2021. 1, 3, 4, 6

[15] Yong Li, Yuanzhi Wang, and Zhen Cui. Decoupled multimodal distilling for emotion recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2023. 1, 3

[16] Huan Ma, Qingyang Zhang, Changqing Zhang, Bingzhe Wu, Huazhu Fu, Joey Tianyi Zhou, and Qinghua Hu. Calibrating multimodal learning. In *Proceedings of the International Conference on Machine Learning*, 2023. 2

[17] Yu Mitsuzumi, Go Irie, Daiki Ikami, and Takashi Shibata. Generalized domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1084–1093, 2021. 3

[18] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 122–132, 2020. 3

[19] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8238–8247, 2022. 1, 2, 4, 5, 7, 8

[20] Rodrigo Quian Quiroga, Alexander Kraskov, Christof Koch, and Itzhak Fried. Explicit encoding of multimodal percepts by single neurons in the human brain. *Current Biology*, 19(15):1308–1313, 2009. 3

[21] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proceedings International Conference on Machine Learning*, pages 8821–8831, 2021. 3

[22] Yuge Shi, N. Siddharth, Brooks Paige, and Philip H.S. Torr. Variational mixture-of-experts autoencoders for multi-modal deep generative models. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 3, 4

[23] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *European Conference on Computer Vision*, pages 746–760, 2012. 5

[24] Barry E Stein, Terrence R Stanford, and Benjamin A Rowland. The neural basis of multisensory integration in the midbrain: its organization and maturation. *Hearing research*, 258(1-2):4–15, 2009. 3

[25] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3

[26] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. 7

[27] Ramakrishna Vedantam, Ian Fischer, Jonathan Huang, and Kevin Murphy. Generative models of visually grounded imagination. In *Proceedings International Conference on Learning Representations*, 2018. 4

[28] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12695–12705, 2020. 1, 2, 4, 5, 7, 8

[29] Yongjin Wang and Ling Guan. Recognizing human emotional state from audiovisual signals. *IEEE Transactions on Multimedia*, 10(5):936–946, 2008. 5, 7

[30] 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, 2020. 4

[31] Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. *Advances in Neural Information Processing Systems*, 31, 2018. 4

[32] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020. 2, 5

[33] Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. Disentangled representation learning for multimodal emotion recognition. In *Proceedings of the ACM International Conference on Multimedia*, 2022. 3

[34] Lijin Yang, Yifei Huang, Yusuke Sugano, and Yoichi Sato. Interact before align: Leveraging cross-modal knowledge for domain adaptive action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14722–14732, 2022. 1, 3, 5

[35] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10287–10296, 2020. 1

[36] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2248–2255, 2013. 5