Boosting Multi-modal Model Performance with Adaptive Gradient Modulation

Hong Li[1,7,*] Xingyu Li[2,*] Pengbo Hu[3] Yinuo Lei[1,7] Chunxiao Li[6] Yi Zhou[3,4,5,†]
[1]School of Information Science and Technology, ShanghaiTech University
[2]Shanghai Center for Brain Science and Brain-Inspired Technology, Shanghai, China
[3]SIST, University of Science and Technology of China, Hefei, China
[4]NEL-BITA, University of Science and Technology of China, Hefei, China
[5]Key Laboratory of Brain-inspired Intelligent Perception and Cognition
(University of Science and Technology of China), Ministry of Education
[6]School of Management, University of Science and Technology of China, Hefei, China
[7]Shanghai Innovation Center for Processor Technologies
{lihong1, leiyn2022}@shanghaitech.edu.cn xyli1905@bsbii.cn
pbhu@mail.ustc.edu.cn {chunxiao.li, yi_zhou}@ustc.edu.cn

Abstract

While the field of multi-modal learning keeps growing fast, the deficiency of the standard joint training paradigm has become clear through recent studies. They attribute the sub-optimal performance of the jointly trained model to the modality competition phenomenon. Existing works attempt to improve the jointly trained model by modulating the training process. Despite their effectiveness, those methods can only apply to late fusion models. More importantly, the mechanism of the modality competition remains unexplored. In this paper, we first propose an adaptive gradient modulation method that can boost the performance of multi-modal models with various fusion strategies. Extensive experiments show that our method surpasses all existing modulation methods. Furthermore, to have a quantitative understanding of the modality competition and the mechanism behind the effectiveness of our modulation method, we introduce a novel metric to measure the competition strength. This metric is built on the mono-modal concept, a function that is designed to represent the competition-less state of a modality. Through systematic investigation, our results confirm the intuition that the modulation encourages the model to rely on the more informative modality. In addition, we find that the jointly trained model typically has a preferred modality on which the competition is weaker than other modalities. However, this preferred modality need not dominate others. Our code will be available at `https://github.com/lihong2303/AGM_ICCV2023`.

1. Introduction

Recent years have seen tremendous progress in deep multi-modal learning. Despite these advances, integrating information from multiple modalities remains challenging. Many efforts have been made to design sophisticated fusion methods for better performance. However, adding additional modalities only slightly improves accuracy in some multi-modal tasks. For example, trained on the CMU-MOSEI [5] dataset, the accuracy of the text-based single-modal model is only about 1% point lower than that of the multi-modal model based on both text and audio modalities. Similar phenomena have also been observed across a wide variety of multi-modal datasets [25, 4].

Such an inefficiency in exploiting and integrating information from multiple modalities presents a great challenge to the multi-modal learning field. It is commonly believed that this inefficiency is a consequence of the existence of the dominant modality, which prevents the model from fully exploiting the other relatively weak modalities [18, 14]. Recent studies [1, 15, 10] theoretically investigate the training process of late fusion models and explain the production of the dominant modality with the concept of modality competition. In addition to the theoretical studies, there is a group of empirical works that attempts to develop methods to modulate the training of a multi-modal model and balance the learning of different modalities and, thus, achieve better performance. To our best knowledge, existing modulation methods are confined to late fusion models which greatly limits their application. More importantly, little effort has been paid to the study of the mechanism behind the effectiveness of those modulation methods.

[*] These authors contributed equally.
[†] Corresponding author: Yi Zhou.

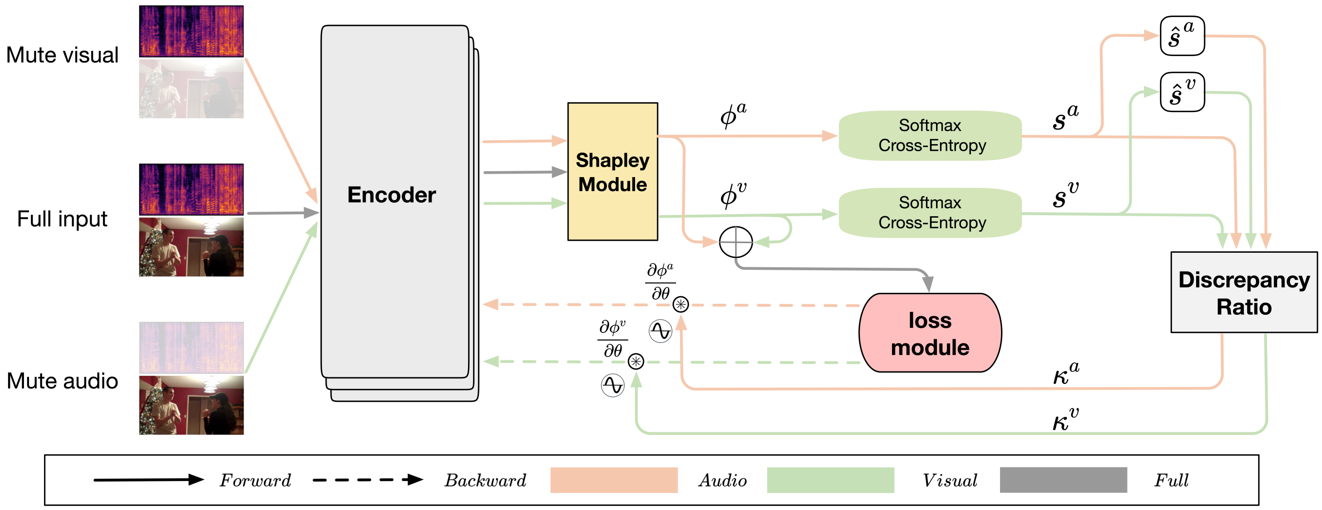

Figure 1. Schematic diagram of the adaptive gradient modulation (AGM) method. Firstly, based on the full input and corresponding muted inputs, the Shapley module produces mono-modal outputs ϕ^m, which disentangle the responses of the multi-modal model to individual modalities. Next, ϕ^m are used to compute the mono-modal cross-entropy s^m that reflect the amount of information in modality m. At last, s^m and their running average $\hat{s}^m$ are fed to the Discrepancy Ratio module to compute the modulation coefficients κ^m for each modality, which in turn modulate the strength of corresponding gradient signals during back-propagation.

It is natural to ask *Can we design a modulation method that applies to more complex fusion strategies?* and *Is it possible to understand the working mechanism of modulation in terms of modality competition?* To this end, we propose an adaptive gradient modulation method, which utilizes a Shapley value-based attribution technique, that can in principle apply to any fusion strategy. Our approach achieves better performance compared with the current modulation methods. Moreover, we introduce the mono-modal concept to represent the competition-less state of a modality in a multi-modal model and build a metric on top of it to directly measure the competition strength of a modality in this multi-modal model. This novel metric lay the base for us to quantitatively study the behavior of modality competition and the working mechanism of our adaptive gradient modulation method.

Our main contributions are three-fold:

1. We propose an adaptive gradient modulation method that can boost the performance of multi-modal models with various fusion strategies and justify its effectiveness through extensive experiments.

2. We introduce the mono-modal concept to capture the competition-less state of a modality and build a novel metric to measure the modality competition strength.

3. We systematically analyze the behavior of modality competition and study the mechanism of how our modulation method works.

2. Related work

2.1. Multi-modal learning

Multi-modal learning is a fast-growing research area. It addresses the needs of effectively processing multi-sensory data in real-world tasks and has applications in various fields, such as multi-modal sentiment classification [31, 4], audio-visual localization [23] and visual question answering [2, 16, 28]. According to the fusion strategy, one distinguishes three types [3], i.e., the late fusion, the early fusion, and the hybrid fusion, when the fusion happens at the output stage, at the input stage, and in a complex manner, respectively. From another perspective, existing models can be divided into two categories, either jointly training different modalities in an end-to-end fashion or exploiting pretrained models and building a multi-stage pipeline.

In this paper, we focus on the multi-modal joint training models for the multi-modal classification task, and we will compare models with different fusion strategies.

2.2. Modality-specific modulation

Recent studies [26, 15] reveal the deficiency of the multimodal joint training paradigm that information on the input modalities is often under-exploited. To address this deficiency, existing works commonly propose to intervene in the training process. Geng *et al.* [8] propose to obtain noise-free multi-view representations with the help of uncertainty in Dynamic Uncertainty-Aware Networks. Wang *et al.* [27] devise the Gradient-blending technique which addresses the overfitting in a multi-modal model by optimally blending modalities. Wu *et al.* [29] propose to balance the speed of learning from different modalities based on their conditional utilization rates. Fujimori *et al.* [6] emphasize the heterogeneity of different network branches in joint training and propose to avoid overfitting through modality-specific early stopping. Yao and Mihalcea [30] advocate using modality-specific learning rates for different branches in a multi-modal model to fully explore the capacity of the corresponding network architecture. More recently, Peng *et al.* [20] proposes to adjust the gradients of individual modalities based on their output magnitudes. The assumption is that in an ideal multi-modal model, the outputs of individual modalities should be balanced, i.e., having similar magnitudes. Consequently, the gradient of the modality with larger outputs will be modulated on-the-fly towards a lower magnitude during each training iteration.

Despite the effectiveness of the above-mentioned methods, they are all confined to late fusion models, limiting their practical use. More importantly, the mechanism of why those methods work to improve the multi-modal model remains unexplored.

2.3. Mono-modal behavior

One way to investigate the mechanism underlying a multi-modal model is to quantify how much modalities affect each other in the model. In a recent theoretical analysis, Huang *et al.* [15] term this interaction among modalities as the modality competition.

Due to the complexity and non-linearity of neural network models, it is infeasible to isolate a part of the computations that account for the competition. Existing works instead attempt to measure the mono-modal behavior inside a multi-modal model, which can partly reflect the interactions among modalities. Hessel and Lee [13] design the empirical multimodally-additive function projection (EMAP) that implicitly reflects the mono-modal behavior by averaging out all other modalities. Yao *et al.* [30] employ the layer conductance [22] to evaluate the importance of individual modalities in late fusion models. Gat *et al.* [7] propose the perceptual scores to measure the mono-modal importance directly. The key idea of their method is the input permutation, which removes the influence of modalities other than the targeting one. What is most related to the goal of measuring the modality competition is the recently proposed SHAPE scores [14]. The authors devise a way to compute the mono-modal marginal contribution and the cross-modal cooperation strength based on the Shapley values.

It is worth noting that all the above-mentioned methods are self-oriented in the sense that they only utilized the multi-modal model, where competition already presents. The lack of information about how each modality behaves without competition prevents those models from faithfully reflecting the modality competition strength.

3. Method

3.1. Adaptive gradient modulation

Drawing inspiration from the Shapley value-based attribution method [14] and the On-the-fly gradient modulation generalization enhancement (OGM-GE) algorithm [20], we propose an adaptive gradient modulation (AGM) method that modulates the level of participation of individual modalities. Figure 1 presents the illustration of the proposed AGM. Our approach is in line with the OGM-GE algorithm in the sense that both attempt to balance the mono-modal responses in a multi-modal model.

Nonetheless, our approach differs from the OGM-GE in the following three important aspects: 1) We adopt a Shapley value-related method to compute the mono-modal responses. In this way, our approach applies to complex fusion strategies rather than being limited to the late fusion case. 2) We extend the method to calculate the discrepancy ratios so that our approach can deal with situations with more than two modalities. 3) In our approach, the discrepancy ratios are modulated towards their running average rather than 1, reflecting the distinctions among different modalities.

3.1.1 Isolating the mono-modal responses

The core component of our approach is the algorithm to isolate the mono-modal responses, which enables us to further compute the mono-modal cross entropy and the mono-modal accuracy.

Let $\phi(x), x = (x^{m_1}, \dots, x^{m_k})$ be a multi-modal model on the data with k modalities and $\mathcal{M} := \{m_i\}_{i\in[k]}$ be the set of all modalities. Same as in [14] we use zero-padding 0^m to represent the absence of features of modality m. When S is a subset of $\mathcal{M}$, $\phi(S)$ denotes that if $m \in S$, the component x^m is substituted with 0^m. Then the mono-modal response for m is defined as

$$\phi^m(x) = \sum_{S\subseteq\mathcal{M}/\{m\};S\neq\emptyset} \frac{|S|!(k-|S|-1)!}{k!} V_m(S;\phi), \tag{1}$$

where $V_m(S;\phi) = \phi(S \cup \{m\}) - \phi(S)$. Note that we exclude the empty subset from the above summation. In this way, we ensure the relation

$$\phi(x) = \sum_m \phi^m(x). \tag{2}$$

As an example, for the two-modality case eq. (1) is simplified to

$$\begin{aligned}\phi^{m_1}(x) = \frac{1}{2}\left[\phi(\{m_1, m_2\}) - \phi(\{0^{m_1}, m_2\})\right.\\ \left.+\phi(\{m_1, 0^{m_2}\})\right].\end{aligned} \tag{3}$$

The mono-modal cross entropy and mono-modal accuracy are then defined subsequently,

$$s^m = \mathbb{E}_{x\sim\mathcal{D}}\left[-\log\left(\text{Softmax}(\phi^m(x))_y\right)\right], \tag{4}$$

and

$$Acc_m = \mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{1}_{y=y_p(x)}\right], \tag{5}$$

where y is the ground-truth class of x and y_p the model prediction, $y_p(x) = \arg\max_{y'\in[K]} \phi^m_{y'}(x)$.

3.1.2 Modulating the training process

We modulate the level of participation of individual modalities through adjusting the intensity of the back-propagation signal of each modality,

$$\theta_{t+1} = \theta_t - \eta\frac{\partial\mathcal{L}}{\partial\phi}\cdot\sum_m \kappa^m_t \frac{\partial\phi^m}{\partial\theta}\bigg|_t, \tag{6}$$

where t refers to a specific iteration of training, θ denotes the trainable network parameters, η is the learning rate and $\mathcal{L}$ is the loss function.

Coefficient κ^m_t controls the magnitude of the update signal for modality m at iteration t. Intuitively, if a modality is too strong (weak) we want to suppress (amplify) its update signal. The strength of a modality is measured by the averaged differences relative to the other modalities

$$r^m_t = \exp\left(\frac{1}{K-1}\sum_{m'\in[K];m'\neq m}\left(s^m_t - s^{m'}_t\right)\right). \tag{7}$$

We choose to compare different modalities based on their mono-modal cross-entropy, since s^m_t reflects the amount of information attributed to modality m within the full model outputs. Then κ^m_t is defined as follows

$$\kappa^m_t = \exp\left(-\alpha * (r^m_t - \tau^m_t)\right), \tag{8}$$

where $\alpha > 0$ is a hyper-parameter that controls the degree of modulation and τ^m_t is the reference for modulation. Consequently, when a modality is too strong ($r^m_t > \tau^m_t$), we lower its update signal ($\kappa^m_t < 1$).

In the current implementation, we choose τ^m_t to be

$$\tau^m_t = \exp\left(\frac{1}{K-1}\sum_{m'\in[K];m'\neq m}\left(\hat{s}^m(t) - \hat{s}^{m'}(t)\right)\right), \tag{9}$$

where $\hat{s}^m(t)$ denotes the running average of mono-modal cross-entropy at iteration t,

$$\hat{s}^m(t) = \hat{s}^m(t-1)\cdot\frac{t-1}{t} + \frac{s^m_t}{t}. \tag{10}$$

The above steps are summarized in Algorithm 1 below.

Algorithm 1 Adaptive Gradient Modulation

```
Training dataset $\mathcal{D} = \{(x^{m_1}, x^{m_2}, .., x^{m_k}), y_i\}$, iteration number $T$, logits output of a modality $o^m_t$, model logits output $o_t$, softmax output of a modality $p^m_t$, batch size $N$, mono-modal information $s^m_t$, batch information discrepancy $r^m_t$, running average information discrepancy $\tau^m_t$, modulation coefficient $\kappa^m_t$, $m \in \{m_1, m_2, ..., m_k\}$.
$\hat{s}^m = 0$.
for t=$1, 2, \ldots, T$ do
    $o^{m_1}_t, o^{m_2}_t, ..., o^{m_k}_t, o_t = \text{net}(x^{m_1}, x^{m_2}, ..., x^{m_k})$
    $p^m_t = \text{Softmax}(o^m_t)$
    $s^m_t = \frac{1}{N}\sum_{i=1}^N log^{p^m_t[i][y[i]]}$
    $\overline{s}_t = \frac{s^{m_1}_t + s^{m_2}_t +, ..., + s^{m_k}_t}{k}$, $\overline{\hat{s}}_t = \frac{\hat{s}^{m_1}_t + \hat{s}^{m_2}_t +, ..., + \hat{s}^{m_k}_t}{k}$
    $r^m_t = e^{((s^m_t - \overline{s}_t)\cdot\frac{k}{k-1})}$, $\tau^m_t = e^{((\hat{s}^m - \overline{\hat{s}}_t)\cdot\frac{k}{k-1})}$
    $\kappa^m_t = e^{(-\alpha*(r^m_t - \tau^m_t))}$
    $\hat{s}^m = \frac{\hat{s}^m\cdot t}{t+1} + \frac{s^m_t}{t+1}$
    Update using $\theta_{t+1} = \theta_t - \eta\frac{\partial\mathcal{L}}{\partial\phi}\cdot\sum_m \kappa^m_t\frac{\partial\phi^m}{\partial\theta}\bigg|_t$
end for
```

3.2. Mono-modal competition strength

The empirical study [26] demonstrates that multi-modal joint training can lead to suboptimal performance that is even worse than the mono-modal model. Recently, Huang *et al.* [15] theoretically study this phenomenon in a simplified setting and attribute it to the modality competition mechanism that the representation learning of a modality is generally affected by the presence of other modalities. The authors further suggest that modality competition potentially explains the effectiveness of the adaptive learning methods [26, 20], which are designed to improve the performance of joint training.

However, the above-mentioned studies are all confined to late fusion cases. It remains unexplored whether the modality competition mechanism can generalize to other fusion strategies and how it alters the representation learning in realistic multi-modal models. This leads to an urgent need for methods that directly measure competition strength.

To quantify modality competition, one must specify the competition-less state for each modality. Previous attribution methods [13, 30, 7, 14] only utilize the responses of the underlying multi-modal model where the competition already took place and, hence, is in principle incapable of reflecting modality competition. To address this challenge, we introduce the mono-modal concept, which defines how the corresponding modality in a given multi-modal model will behave in the absence of modality competition. Then the competition strength is estimated based on the deviation of the multi-modal model outputs with respect to this mono-modal concept.

3.2.1 Mono-modal concept

Let $x = (x^{m_1}, x^{m_2})$ denote a multi-modal input feature, where x^{m_1} and x^{m_2} refer to the mono-modal components. We focus on two modalities case below and the extension to more modalities is straightforward.

The processing of x^{m_1} by a multi-modal model is determined by the complementary component x^{m_2}, the network architecture ϕ[1] , the training settings $\mathcal{T}$[2] and the dataset $\mathcal{D}$. We call this quadruple $\mathcal{E}_{m_1} := (x^{m_2}, \phi, \mathcal{T}, \mathcal{D})$ as the environment of mono-modal input x^{m_1}. Roughly speaking, in the competition-less state we want to remove the effects of x^{m_2} while retaining the "normal" processing of x^{m_1}. This can be formally denoted as $\mathcal{E}_{m_1}/m_2$.

With the above notations, we abstract the competition-less state for m_1 as a function $\mathcal{C}^{m_1}(x^{m_1}; \mathcal{E}_{m_1}/m_2)$ that maps the inputs to vectors in $\mathbb{R}^K$, where K is the number of classes. Intuitively, $\mathcal{C}^{m_1}$ captures the responses, of a given multi-modal model, to the mono-modal inputs without modality competition. Following the terminology in [19], $\mathcal{C}^{m_1}$ is referred as the *mono-modal concept* of modality m_1. In the following, we elaborate the construction of $\mathcal{C}^m, m \in \{m_1, m_2\}$ under different situations.

Late fusion case. In late fusion the multi-modal model can be written as $\phi(x) = \phi^{m_1}(x^{m_1}) + \phi^{m_2}(x^{m_2})$. It is natural to set $\mathcal{E}_{m_1}/m_2 = (\mathbf{0}^{m_2}, \phi^{m_1}, \mathcal{T}_{m_1}, \mathcal{D}_{m_1})$. $\mathbf{0}^{m_2}$ denotes the null input of modality m_2, which is realized, in the current case, by simply discarding the branch ϕ^{m_2}. $\mathcal{T}_{m_1}$ refers to the same training set for the m_1 branch as it was during the training of the multi-modal model ϕ. At last, $\mathcal{D}_{m_1}$ denotes the set of mono-model feature components $\{x_i^{m_1}\}_{i \in [N]}$, where N is the number of data samples and $[N] := \{1, \ldots, N\}$. In practice, we need to *train* ϕ^{m_1} on $\mathcal{D}_{m_1}$ with settings $\mathcal{T}_{m_1}$, and $\mathcal{C}^{m_1}$ is nothing but the resulting network function.

Early and hybrid fusion cases. In these situations, the model can only be written as $\phi(x^{m_1}, x^{m_2})$. There is no apparent way to separate the processing of x^{m_1} and x^{m_2} at the architecture level. In order to mute the influence from m_2, we substitute x^{m_2} with a zero vector of the same dimension. Since the zero vector bears no information about the task, it won't introduce modality competition. Therefore, one can formally write $\mathcal{E}_{m_1}/m_2 = (\mathbf{0}^{m_2}, \phi, \mathcal{T}, \mathcal{D}_{m_1})$, indicating that the architecture and training settings are the same as for the multi-modal model. This time $\mathbf{0}^{m_2}$ refers to the zero input of m_2 feature components [3]. Practically, to construct $\mathcal{C}^{m_1}$, we need to *train* ϕ on $\mathcal{D}' := \mathcal{D}_{m_1} \otimes \{\mathbf{0}^{m_2}\}$ with $\mathcal{T}$. Samples in $\mathcal{D}'$ are of form $(x^{m_1}, \mathbf{0}^{m_2})$.

3.2.2 Competition strength

With the mono-modal concepts as a reference, we are ready to quantify the deviation of the multi-modal model responses from those competition-less states. A linear probing method [19] is employed to estimate this deviation. Specifically, let z be the latent feature before the last classifier layer in the multi-modal model, we train a linear predictor from z to the targeting mono-modal concept $\mathcal{C}^m$,

$$f^m(z) = \mathbf{W}z + \mathbf{b}, \tag{11}$$

whose parameters $\mathbf{W}$ and $\mathbf{b}$ are determined by minimizing the empirical mean square error of the predictions,

$$\begin{aligned} \mathbf{W}^{m,*}, \mathbf{b}^{m,*} =& \arg\min_{\mathbf{w},b} \frac{1}{N} \sum_{i \in [N]} \|f^m(z_i) - \mathcal{C}^m(x_i^m)\|_2^2 \\ &+ \lambda \left(\|\mathbf{W}\|_2 + \|\mathbf{b}\|_2\right), \end{aligned} \tag{12}$$

where $\|\cdot\|_p$ denotes the L_p norm, i refers to the index of data samples and λ is the regularization strength. The L_2 regularization term is introduced to avoid overfitting.

The quality of the above linear fitting reflects how much the multi-modal features deviate from their competition-less states. Thus we define the competition strength as

$$d^m = \frac{\sum_i \left(\mathcal{C}^m(x_i^m) - f^m(z_i)\right)^2}{\sum_i (\mathcal{C}^m\left(x_i^m\right) - \overline{\mathcal{C}^m})^2}, \tag{13}$$

where $\overline{\mathcal{C}^m}$ is the mean mono-modal concept value over data samples. d^m measures the quality of the linear predictions with respect to the naive baseline, i.e., simply predicting the mean value. Its value ranges from 0 to 1, indicating the weakest and strongest competition levels respectively.

In practice, we reserve two hold-out datasets for computing the competition strength. One of them is used to train the linear predictor and the other to calculate d^m.

4. Experiments and discussion

4.1. Experimental settings

In this paper, we systematically apply our adaptive gradient modulation approach to situations that cover different fusion strategies, different modality combinations, and different network architectures. For the late fusion case, our approach is compared with existing modulation methods. Moreover, we also include the mono-modal accuracy and the modality competition strength for all the situations.

[1] we abuse the symbol ϕ a little so that it may refer to both the network architecture and the corresponding network function.

[2] $\mathcal{T}$ includes the initialization, the loss function, hyper-parameters, and specific techniques, e.g., the learning rate scheduler, used in training.

[3] We also try to use the random inputs for $\mathbf{0}^m$. Our results suggest that there is no big difference between these two implementations. Please refer to the supplementary material for the detailed sanity check of the definition of the mono-modal concept.

We carry out experiments [4] on five popular multi-modal datasets. The AV-MNIST [25] is collected for a multi-media classification task that involves disturbed images and audio features. The CREMA-D [4] is an audio-visual dataset for speech emotion recognition which consists of six emotional labels. The UR-Funny [11] is created for humor detection, involving words (text), gestures (vision), and prosodic cues (acoustic) modalities. The AVE [23] is devised for an audio-visual event localization classification task, including 28 event classes. The CMU-MOSEI [31] is collected for sentence-level emotion recognition and sentiment analysis, including audio, visual, and text modalities. Here we only use text and audio modalities.

The experiments can be grouped into two classes. The first one concerns the performance of our approach and the behavior of modality competition in the late and early fusion strategies across different multi-modal datasets. We adopt a unified design of the multi-modal models in this class. The fusion module in the early fusion case is all built with the MAXOUT [9] network. In addition, for each dataset, the network models for both fusion strategies use the same encoder architecture. Specifically, for the AV-MNIST, the CREMA-D, and the Kinetics-Sound datasets, ResNet18 [12] is used as an encoder for both the audio and visual modalities. For the UR-Funny dataset, we use Transformer [24] for the encoder for all three modalities.

In the second class, we carry out experiments with current SOTA models and show that our approach can also enhance more complex models in a realistic application. For the AVE dataset, the PSP [32] network is used, which features elaborately designed methods that align the audio and visual representations during fusion. For the CMU-MOSEI dataset, we adopt the Transformer-based joint-encoding (TBJE) [5] as the model. TBJE jointly encodes input modalities through the modular co-attention and the glimpse layer.

Our code is implemented in Pytorch 1.2, and experiments are run on a single NVIDIA 3090 GPU. For the detailed experimental settings and hyper-parameters, please refer to the supplementary material.

4.2. The effectiveness of AGM

In this subsection, we focus on the Acc column in all the tables and demonstrate the universal effectiveness of our AGM method in improving the model performance.

Tables 1 to 3 summarize the results on the AV-MNIST, the CREMA-D, and the UR-Funny dataset, respectively. In the late fusion cases, our approach is compared with the Modality-Specific Early Stopping (MSES) and Modality-Specific Learning Rate (MSLR) methods. For situations with only two modalities, we also include the results of the Gradient Blending (G-Blending), Characterizing and Overcoming the Greedy Nature of Learning (Greedy), and On-the-fly Gradient Modulation Generalization Enhancement (OGM-GE) method.

	AV-MNIST	Acc	Acc_a	Acc_v	d^a	d^v
Late fusion	$\mathcal{C}^a$	-	39.61	-	-	-
	$\mathcal{C}^v$	-	-	65.14	-	-
	Joint-Train	69.77	16.05	55.83	0.7838	0.1408
	G-Blending [27]	70.32	14.36	56.59	0.7963	0.1359
	Greedy [29]	70.65	18.80	63.46	0.7358	0.1340
	MSES [6]	70.68	27.50	63.34	0.7538	0.1372
	MSLR [30]	70.62	22.72	62.92	0.7300	0.1437
	OGM-GE [20]	71.08	24.53	55.85	0.7445	0.1617
	AGM	**72.14**	38.90	63.65	0.6787	0.1197
Early fusion	$\mathcal{C}^a$	-	41.60	-	-	-
	$\mathcal{C}^v$	-	-	65.46	-	-
	Joint-Train	71.15	24.28	60.14	0.7668	0.1825
	AGM	**72.26**	47.79	68.48	0.7146	0.1796

Table 1. The accuracy (Acc, Acc_a, Acc_v) and the competition strength (d^a, d^v) on the AV-MNIST dataset for multi-modal models using different fusion strategies. In late fusion, comparison with several modality-specific intervention methods: Modality-Specific Early Stop (MSES), Modality-Specific Learning Rate(MSLR), and On-the-fly Gradient Modulation Generalization Enhancement (OGM-GE). The results of Joint-Train are included as baselines. $\mathcal{C}_a$ and $\mathcal{C}_v$ indicate the performance of audio and visual modality concepts, respectively. The best results are shown in **bold**.

It is evident that our approach constantly improves the performance w.r.t. the Joint-Train case and achieves the best accuracy in all situations. In the late fusion case, while all modulation methods generally boost the performance compared to the Joint-Train baseline, our approach exceeds the second-best one for a gap of at least 1.06%. It is notable that the improvement in the early fusion case by our approach is comparable with the ones in late fusion cases. We note the significant increase in accuracy on CREMA-D, where, after modulating, the results of our approach are 17.34% and 19.58% higher than the ones of Joint-Train in late and early fusion, respectively. There is also a gap of 10.34% between our approach and OGM-GE. Such supersizing effectiveness may be attributed to the fact that the most informative modality in CREMA-D, i.e., the visual modality, is considerably under-exploited in the Joint-Train. In fact, the mono-modal accuracy of the visual modality is only 22.72%, which is much lower than its potential performance of the mono-modal concept, i.e., 75.93%. We observe that the improvement from MSES and MSLR is often very lim-

[4] To better demonstrate the universal effectiveness of AGM, we further carry out experiments on the Kinetics-Sound [17] using both the late fusion and the FiLM [21] fusion strategies. These results are included in the supplementary material due to the space limit.

UR-Funny		Acc	Acc_a	Acc_v	Acc_t	d^a	d^v	d^t
Late fusion	$\mathcal{C}^a$	-	59.23	-	-	-	-	-
	$\mathcal{C}^v$	-	-	53.16	-	-	-	-
	$\mathcal{C}^t$	-	-	-	63.46	-	-	-
	Joint-Train	64.50	50.31	51.53	49.78	0.5558	0.1058	0.4513
	MSES [6]	64.23	50.31	49.69	57.87	0.5605	0.1028	0.4592
	MSLR [30]	64.74	50.31	48.62	49.69	0.5257	0.0975	0.4316
	AGM	**65.97**	54.87	49.36	62.22	0.5234	0.0725	0.5147
Early fusion	$\mathcal{C}^a$	-	58.25	-	-	-	-	-
	$\mathcal{C}^v$	-	-	53.29	-	-	-	-
	$\mathcal{C}^t$	-	-	-	61.07	-	-	-
	Joint-Train	65.15	54.87	50.86	54.14	0.7217	0.2672	0.2906
	AGM	**66.07**	64.87	55.20	63.36	0.6962	0.2697	0.3200

Table 2. The same as Table 1, but for UR-Funny dataset. The involved modalities are audio, visual, and text.

CREMA-D		Acc	Acc_a	Acc_v	d^a	d^v
Late fusion	$\mathcal{C}^a$	-	62.63	-	-	-
	$\mathcal{C}^v$	-	-	75.93	-	-
	Joint-Train	61.14	57.10	22.72	0.4593	0.7555
	G-Blending [27]	62.03	19.58	16.89	0.4706	0.8005
	Greedy [29]	63.08	43.05	16.89	0.4598	0.7661
	MSES [6]	60.99	54.86	22.57	0.4607	0.7546
	MSLR [30]	64.42	54.86	26.31	0.4614	0.7150
	OGM-GE [20]	68.16	55.16	36.32	0.5448	0.6929
	AGM	**78.48**	48.58	57.85	0.6624	0.5067
Early fusion	$\mathcal{C}^a$	-	61.29	-	-	-
	$\mathcal{C}^v$	-	-	75.78	-	-
	Joint-Train	61.88	42.60	16.89	0.5345	0.9905
	AGM	**81.46**	76.53	80.42	0.8753	0.6496

Table 3. The same as Table 1, but for CREMA-D dataset.

ited. Actually, on CREMA-D the accuracy of MSES in the late fusion case is worse than the one of Joint-Train. This could be the consequence that MSES only controls the time to stop training and, thus, can only provide limited guidance to the weights update.

We next show that our approach can also boost the performance of existing SOTA models. Those models normally equip with elaborately designed fusion modules to ensure higher prediction accuracy. Table 4 shows the results on the AVE dataset and CMU-MOSEI dataset, on which the improvements are 1.09% and 0.85%, respectively. It is worth noting that all other modulation methods can not apply to such complex situations, as there are no separable branches in the network models for different modalities.

AGM adjusts the modulation coefficients based on the running average of the mono-modal cross entropy which serves as a reference of idea relative strengths of individual

AVE	Acc	Acc_a	Acc_v	d^a	d^v
$\mathcal{C}^a$	-	65.00	-	-	-
$\mathcal{C}^v$	-	-	64.69	-	-
PSP [32]	76.02	52.58	50.18	0.6223	0.6232
AGM	**77.11**	72.34	70.68	0.6198	0.6337

CMU-MOSEI	Acc	Acc_t	Acc_a	d^t	d^a
$\mathcal{C}^t$	-	80.92	-	-	-
$\mathcal{C}^a$	-	-	74.46	-	-
TBJE [5]	80.91	73.59	73.08	0.5794	0.9450
AGM	**81.76**	79.41	73.08	0.5774	0.9540

Table 4. The accuracy and competition strength on the AVE and the MOSEI dataset for the general joint-training network with elaborating fusion structures network. Audio and visual are involved in the AVE dataset and audio and text in MOSEI. PSP stands for general joint training network for the AVE dataset and TBJE for the CMU-MOSEI dataset. $\mathcal{C}_a$, $\mathcal{C}_v$ and $\mathcal{C}_a$ indicate the performance of audio, visual, and text modality, respectively. The best results are shown in **bold**.

modalities. Additional experiments demonstrate that this reference is better than the brutal force requirement of equal contribution from all modalities. Further, we consider an in-depth comparison between AGM and the OGM-GE as their performance outstands in our experiments. Specifically, we investigate whether the Generalization Enhancement (GE) technique can hence AGM and, in turn, whether a running average reference can boost the performance of OGM-GE. We find that neither provides an improvement. The details of the above-mentioned results can be found in the supplementary material.

Combining all the above results, we conclude that our modulation approach can help boost the model performance regardless of the fusion strategy, the number and types of involved modalities, and the network architecture.

4.3. Modality competition

The competition strength metric provides us a base to analyze the states of individual modalities in a joint-trained model and understand the mechanism of how the modulation methods work.

In the following, we first compare the changes in competition strength before and after modulating and investigate what is brought to the multi-modal model by our adaptive gradient modulation. This follows a discussion of the modality competition behavior.

4.3.1 Gradient modulation & modality competition

Our primary concern is how the modulation affects the model performance in terms of changing the competition state. The modality competition directly measures the deviation from the competition-less state and provides more accurate information about the competition state compared to the mono-modal accuracy, which mainly reflects the information in a single modality. Generally, we distinguish two different types of change in competition strength.

In the first type, modality competition is mitigated by modulation. The results on AV-MNIST (Table 1) exemplify this situation. For both fusion strategies, the competition strengths of audio (d^a) and visual (d^v) modalities decrease, and their mono-modal accuracy (Acc_a and Acc_v) increases as well as the multi-modal performance. This suggests that suppressing the competition, allows the model to better utilize inputs from different modalities. Figure 2 illustrates the change in performance and competition strength along with training. For the joint training baseline (left panel in Figure 2), d^a increases while d^v decreases in the initial training stage up to the 9-th epoch. Hence, the model initially learns information from the visual modality. Indeed, Acc_a is almost the random guess while Acc_v is close to the full multi-modal accuracy. In later epochs, d^a starts to decrease and its mono-modal accuracy increases accordingly. On the other hand, the increase of d^v is accompanied by the decrease of Acc_v. When adaptive gradient modulation is applied (right panel in Figure 2), the competition strength of both modalities decreases along training and converges to lower values than their counterpart in the joint training case. At the same time, their mono-modal accuracy keeps increasing. We find that the model starts to learn the audio modality at a relatively earlier epoch and Acc_a is boosted considerably.

In the second type, the competition of some modalities could be strengthened. Results in Tables 2 to 4 belong to this type. For CREMA-D, d^v decreases while d^a increases. This allows the model to better exploit the visual modality [5], which is more informative [6]. Similar behaviors are observed on the AVE and CMU-MOSEI datasets. In both cases, the modulation leads to a decrease in competition strength of the more informative modality, i.e., the audio modality of AVE and the text modality of CMU-MOSEI. The results for UR-Funny differ from previous cases. It mainly reflects a balance in information usage between the audio and text modalities. Interestingly, we note that even though the text modality possesses better information, its d^t increases after modulation. We suspect this could be due to a high-order effect when multiple modalities are present. In other words, combining the text and the visual modalities could be more informative than combining the audio and visual modalities.

In summary, the results quantitatively demonstrate the behavior behind the effectiveness of our modulation method. In most cases, the picture is clear that while the raw model possesses a certain bias towards some modalities, the modulation pushes the model to rely on the more informative modalities [7].

4.3.2 Behavior of modality competition

In the following, we proceed to investigate the modality competition in the joint training situation. We systematically study the competition's behavior from various perspectives that cover the model's preference towards individual modalities, the relation to the fusion strategy, and the relation to the input data.

Existence of preferred modality. Our results reveal that modality competition is commonly present in multi-modal models. In fact, there is at least one modality with non-trivial competition strength in all situations. However, we emphasize that it is not necessary for a multi-modal model to have a dominant modality. The results on AVE (Table 4) provide such an example. The balance of the two modalities, in this case, could be attributed to the elaborately designed fusion method in the PSP model. In addition, we recognize a trend in all the experiments that the modality with the lowest competition strength always has a higher mono-modal accuracy. This suggests that there exists the model-preferred modality, which the raw multi-modal model tends to explore. This preference will be broken by the modulation which encourages more efficient usage of modality information.

Relation to fusion strategy. The modality competition strengths are similar in the late and early fusion cases. For

[5] We remark that, in this case, the modality collapse in joint training on CREMA-D can be attributed to the modality competition.

[6] The accuracy of the visual mono-modal concept is higher than the one of the audio modality.

[7] Note that better use of informative modalities does not necessarily lead to low competition strengths of these modalities.

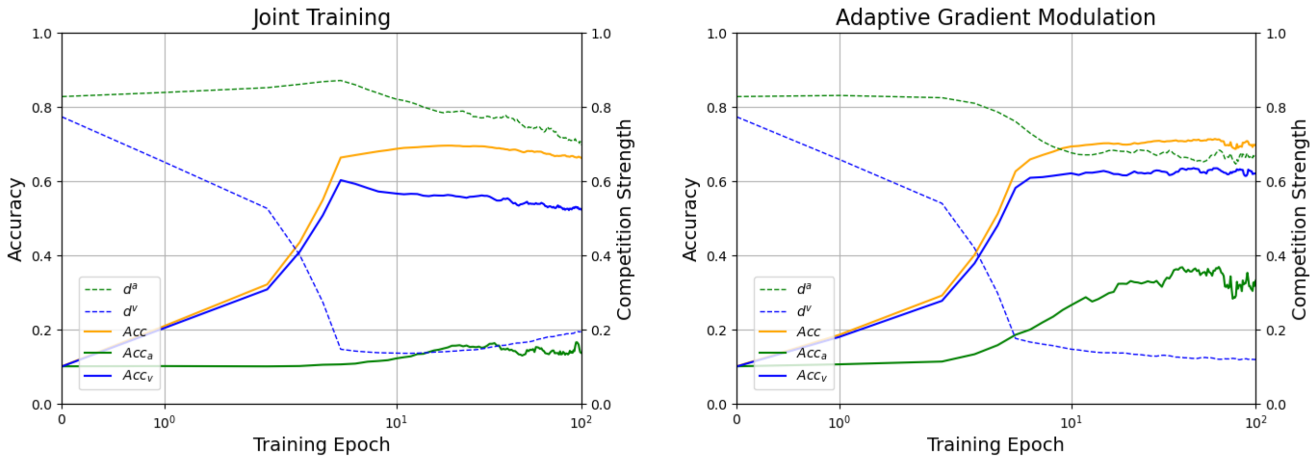

Figure 2. Accuracy (Acc, Acc_a, Acc_v) and competition strength (d^a, d^v) of joint-training multimodal model and multimodal model with AGM using addition fusion method on the validation set of the AV-MNIST dataset. The left is the joint-training multimodal model and the right is the multimodal model with our proposed AGM.

example, in Table 2 for the UR-Funny dataset, audio modality is always with the strongest competition, the text modality the second, and the visual modality the weakest. Other results show similar behavior. As this tendency is independent of the fusion strategy, our results suggest that the strength of competition may depend more on the task and the input data.

Relation to modality information. It is intuitive to expect that the modality with higher information for the task will have lower competition strength, i.e., being better exploited by the model. However, it is not always the case. While the above intuition applies to the results on AV-MNIST and CREMA-D datasets, the visual modality in CREMA-D is under-explored in the joint training case even though it is more informative. Moreover, for the UR-Funny dataset, the visual modality, which contains less information, has a very low competition strength in the joint training case. In conclusion, current results do not support any correlation between the modality information and the competition strength.

Relation to modality type. To study whether the modality type affects the competition states, we compare the results of CREMA-D and AV-MNIST. Both datasets are composed of visual and audio modalities, and the visual modality is more informative. In addition, our experiments on these two datasets share the same network architecture. Nonetheless, the competition state of the visual modality in CREMA-D is opposite to the one in AV-MNIST. Therefore, the strength of modality competition tends to be unrelated to the modality type.

5. Conclusion

In this paper, we propose an adaptive gradient modulation method to boost the performance of jointly trained multi-modal models. With the Shapley value-based approach to estimate the mono-modal responses, our modulation method can apply to models of all possible fusion strategies. The experiments show that our method beats the existing modulation methods and can improve the model performance across different fusion strategies, different modality combinations and different network architectures. In addition, we devise a novel metric based on the mono-modal concept to directly measure the competition strength in a multi-modal model. With this metric, we systematically analyze the behavior of modality competition in joint training and investigate the mechanism underlying the effectiveness of our modulation method. Our results reveal more complex patterns of modality competition than those proposed by previous studies.

We hope this work can promote the community's understanding of the modality competition and the modulation methods and inspire better designs of multi-modal models.

Acknowledgments

This work was supported by the National Science and Techno- logy Innovation 2030 Project of China (Nos. 2021ZD0202600) and the National Natural Science Foundation of China (NSFC) (Nos. U22B2063).

References

[1] Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816*, 2020.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[3] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy.

[4] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014.

[5] Jean-Benoit Delbrouck, Noé Tits, Mathilde Brousmiche, and Stéphane Dupont. A transformer-based joint-encoding for emotion recognition and sentiment analysis. *arXiv preprint arXiv:2006.15955*, 2020.

[6] Naotsuna Fujimori, Rei Endo, Yoshihiko Kawai, and Takahiro Mochizuki. Modality-Specific Learning Rate Control for Multimodal Classification. In Shivakumara Palaiahnakote, Gabriella Sanniti di Baja, Liang Wang, and Wei Qi Yan, editors, *Pattern Recognition*, Lecture Notes in Computer Science, pages 412–422, Cham, 2020. Springer International Publishing.

[7] Itai Gat, Idan Schwartz, and Alexander Schwing. Perceptual Score: What Data Modalities Does Your Model Perceive?, Oct. 2021.

[8] Yu Geng, Zongbo Han, Changqing Zhang, and Qinghua Hu. Uncertainty-aware multi-view representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7545–7553, 2021.

[9] Ian Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. Maxout networks. In *International conference on machine learning*, pages 1319–1327. PMLR, 2013.

[10] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification with dynamic evidential fusion. *IEEE transactions on pattern analysis and machine intelligence*, 45(2):2551–2566, 2022.

[11] Md Kamrul Hasan, Wasifur Rahman, Amir Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, et al. Ur-funny: A multimodal language dataset for understanding humor. *arXiv preprint arXiv:1904.06618*, 2019.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[13] Jack Hessel and Lillian Lee. Does my multimodal model learn cross-modal interactions? It's harder to tell than you might think!, Oct. 2020.

[14] Pengbo Hu, Xingyu Li, and Yi Zhou. SHAPE: An Unified Approach to Evaluate the Contribution and Cooperation of Individual Modalities. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 3064–3070, Vienna, Austria, July 2022. International Joint Conferences on Artificial Intelligence Organization.

[15] 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). In *Proceedings of the 39th International Conference on Machine Learning*, pages 9226–9259. PMLR, June 2022.

[16] Ilija Ilievski and Jiashi Feng. Multimodal learning and reasoning for visual question answering. *Advances in neural information processing systems*, 30, 2017.

[17] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[18] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18177–18186, 2022.

[19] Thomas McGrath, Andrei Kapishnikov, Nenad Tomašev, Adam Pearce, Demis Hassabis, Been Kim, Ulrich Paquet, and Vladimir Kramnik. Acquisition of Chess Knowledge in AlphaZero, Aug. 2022.

[20] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced Multimodal Learning via On-the-fly Gradient Modulation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8228–8237, New Orleans, LA, USA, June 2022. IEEE.

[21] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[22] Avanti Shrikumar, Jocelin Su, and Anshul Kundaje. Computationally Efficient Measures of Internal Neuron Importance, July 2018.

[23] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 247–263, 2018.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[25] Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. Centralnet: a multilayer approach for multimodal fusion. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[26] Weiyao Wang, Du Tran, and Matt Feiszli. What Makes Training Multi-Modal Classification Networks Hard? In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12692–12702, Seattle, WA, USA, June 2020. IEEE.

[27] 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.

[28] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. In *Thirty-fifth Conference on Neural*

Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021.

[29] 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*, pages 24043–24055. PMLR, 2022.

[30] Yiqun Yao and Rada Mihalcea. Modality-specific Learning Rates for Effective Multimodal Additive Late-fusion. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1824–1834, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[31] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, 2018.

[32] Jinxing Zhou, Liang Zheng, Yiran Zhong, Shijie Hao, and Meng Wang. Positive sample propagation along the audio-visual event line. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8436–8444, 2021.