Balanced Multimodal Learning via On-the-fly Gradient Modulation

Xiaokang Peng[1,†], **Yake Wei**[1,†], **Andong Deng**[2], **Dong Wang**[3], **Di Hu**[1,*]

[1]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing
[1]Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing
[2]Shanghai Jiao Tong University, Shanghai
[3]Shanghai Artificial Intelligence Laboratory, Shanghai
[1]{xiaokangpeng, yakewei, dihu}@ruc.edu.cn, [2]{andongdeng69, dongwang.dw93}@gmail.com

Abstract

Multimodal learning helps to comprehensively understand the world, by integrating different senses. Accordingly, multiple input modalities are expected to boost model performance, but we actually find that they are not fully exploited even when the multimodal model outperforms its uni-modal counterpart. Specifically, in this paper we point out that existing multimodal discriminative models, in which uniform objective is designed for all modalities, could remain under-optimized uni-modal representations, caused by another dominated modality in some scenarios, e.g., sound in blowing wind event, vision in drawing picture event, etc. To alleviate this optimization imbalance, we propose on-the-fly gradient modulation to adaptively control the optimization of each modality, via monitoring the discrepancy of their contribution towards the learning objective. Further, an extra Gaussian noise that changes dynamically is introduced to avoid possible generalization drop caused by gradient modulation. As a result, we achieve considerable improvement over common fusion methods on different multimodal tasks, and this simple strategy can also boost existing multimodal methods, which illustrates its efficacy and versatility. The source code is available at https://github.com/GeWu-Lab/OGM-GE_CVPR2022.

1. Introduction

People perceive the world by collaboratively utilizing multiple senses: eyes to look, ears to listen, and hands to touch. Such a multimodal way can provide more comprehensive information from different aspects. Inspired by the multi-sensory integration ability of humans [11], multimodal data, collected from different sensors, tend to be more considered in machine learning. In recent years, multimodal learning has exhibited a clear advantage in improving the performance of previous uni-modal tasks as well as addressing new challenging problems, such as action recognition [10, 23, 29, 37], audio-visual speech recognition [33], and visual question answering [2, 19, 41].

Multimodal data usually provides more views compared with uni-modal one, accordingly learning with multimodal data should match or outperform the uni-modal case. However, according to the recent study [39], multimodal models that optimize the uniform learning objective for all modalities with joint training strategy can be inferior to the uni-modal model in some situations. Such a phenomenon violates the intention of improving model performance through integrating information from multiple modalities. Previous researchers claimed that various modalities tend to converge at different rates, leading to uncoordinated convergence problem [20, 35, 39]. To cope with this problem, some methods aid the training of multimodal models with the help of additional uni-modal classifiers or pre-trained models [8, 39]. Hence, they inevitably bring extra efforts on training additional neural modules.

However, we further find that even when the multimodal models *outperform* the uni-modal ones, they still cannot fully exploit the potential of multiple modalities. As shown in Figure 1, the joint multimodal model achieves the best event classification performance on VGGSound [7], but the performance of visual and audio modality within it is clearly worse than that in the visual-only and audio-only model, respectively[1]. This interesting observation suggests underoptimized representation in both modalities. We consider the reason could be that, in some multimodal scenarios, the dominated modality [31] with better performance (e.g., sound of wind blowing, vision of playing football, etc.) will suppress the optimization of the other one. Moreover, as il-

---

[†]Equal contribution. [*]Corresponding author.

[1]Here, the visual-only and the audio-only are the models that trained with a single modality. To evaluate the uni-modal encoders of the multimodal model, we observe the performance through fixing the joint-trained uni-modal encoder and finetuning a uni-modal classifier only.
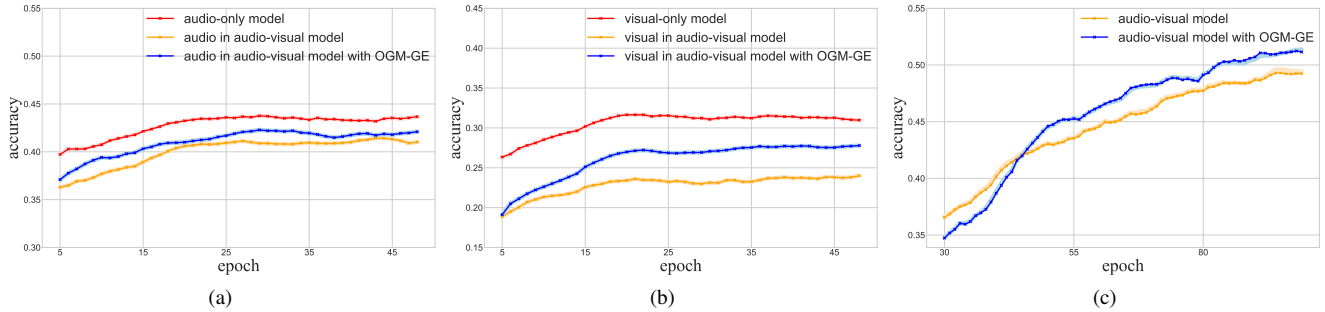
Figure 1. Performance of the uni-modal models, joint-trained multimodal model, and multimodal model with our proposed OGM-GE on the validation set of the VGGSound [7] dataset. (a) Performance of audio modality. (b) Performance of visual modality. (c) Performance of the multimodal model. Best viewed in color.

lustrated in Figure 1(a) and (b), there is another noticeable observation that accuracy drops more remarkably in visual modality compared with audio case, which is consistent with the fact that VGGSound, as a curated sound-oriented dataset, prefers audio modality, even the sound sources are guaranteed to be visible. Generally speaking, such dataset preferences will lead to one modality is usually dominant, resulting in this phenomenon of optimization imbalance.

Aiming to solve the problem above, we first analyze the imbalance phenomenon from the optimization perspective, and find that the modality with better performance contributes to lower joint discriminative loss then dominates the optimization progress via propagating limited gradient over the other modality, thus leading to the under-optimized situation. Then, to ease the situation, we propose to control the optimization process of each modality via the *On-the-fly Gradient Modulation* (OGM) strategy. Specifically, the contribution discrepancy between different modalities to the learning objective is dynamically monitored during training progress, which is then exploited to adaptively modulate the gradients, offering more efforts on the under-optimized modality. However, the modulated gradient may lower the intensity of the stochastic gradient noise, which has been proven to have a positive correlation with generalization ability [21]. Hence, we further introduce extra Gaussian noise that changes dynamically to achieve *Generalization Enhancement* (GE). After applying our method of OGM-GE on the multimodal learning task of VGGSound in Figure 1, we obtain the consistent performance boost for under-optimized uni-modal representation, i.e., the blue curves in Figure 1(a) and (b). More than that, the visual modality gains more improvement. As a result, our method is noticeably superior to the conventional one in the multimodal learning setting, as shown in Figure 1(c). To comprehensively demonstrate the effectiveness of OGM-GE, we test it in various multimodal tasks on different datasets, which brings consistent improvements, working with both vanilla fusion strategies and existing multimodal methods.

To summarize, our contributions are as follows:

- We find the optimization imbalance phenomenon that the performance of the joint multimodal model is limited due to the under-optimized representations, and then analyze it from the optimization perspective.

- The OGM-GE method is proposed to solve the optimization imbalance problem by controlling the optimization process of each modality dynamically as well as enhancing generalization ability.

- The proposed OGM-GE can be plugged in not only vanilla fusion strategies but also existing multimodal frameworks and brings consistent improvement, indicating its promising versatility.

2. Related works

2.1. Multimodal learning

Multimodal learning is a sophisticated learning paradigm in the machine learning community and has been attracting increasing attention because of the growing amount of multimodal data, which naturally contains abundant correlation. There are different research directions according to specific applications. For example, some researchers have explored the correspondence between multimodal data in an unsupervised manner, to learn meaningful representations for downstream tasks [1,4,17,18,25]. Moreover, there is numerous research dedicating to exploiting information of multiple modalities to boost model performance on a certain task compared with uni-modal frameworks, e.g., action recognition [10, 23, 29], audio-visual speech recognition [16, 33], and visual question answering [2, 19]. However, most multimodal methods that utilize joint training strategy could not fully exploit all modalities and produce under-optimized uni-modal representations, making the performance of multimodal models fail to reach where they are expected, even though they sometimes indeed outperform the uni-modal counterparts.

2.2. Imbalanced multimodal learning

The aforementioned defect of previous audio-visual learning methods encourages researchers to explore the reasons behind it. Several studies pointed out that most multimodal learning methods cannot effectively boost the performance even with more information [8,35,39,40], which is caused by the discrepancy between modalities. Wang et al. [39] found that different modalities have different convergence rates, making the jointly trained multimodal model fail to match or outperform its uni-modal counterpart. Winterbottom et al. [40] demonstrated an inherent bias in TVQA dataset towards the textual subtitle modality. Recently, several methods have emerged trying to solve these problems [8, 35, 39]. Wang et al. [39] proposed Gradient-Blending to obtain an optimal blending of modalities based on their over-fitting behaviors. Further, Du et al. [8] trained the multimodal model by distilling knowledge from the well-trained uni-modal models to strengthen the uni-modal encoders. These methods can indeed bring improvement to a degree, but extra efforts are required to introduce additional neural modules, which complicates the training procedure. In this work, from the perspective of optimization, we tackle this problem by adaptively controlling the optimization of each modality without extra modules.

2.3. Stochastic gradient noise

The gradient noise of SGD is considered to have an essential correlation with the generalization ability of the deep models [6, 13, 42, 44, 48]. Such stochastic gradient noise is introduced by random mini-batch sampling, believed to serve as a regularization and help the model to escape from saddle point or local optimum [6,21,42,44]. Zhou et al. [47] further provided theoretical proof that the stochastic gradient algorithms, in conjunction with proper Gaussian noise, are guaranteed to converge to the global optimum in polynomial time with random initialization. In this work, to enhance the generation ability of the multimodal model, we introduce extra Gaussian noise into the gradient and achieve considerable improvement.

3. Method

3.1. Optimization imbalance analysis

We introduce the analysis of the optimization imbalance phenomenon and find that the optimization progress of the multimodal model is dominated by the modality with better performance, leading to another modality being under-optimized. Here we take audio and visual modalities as example. For convenience, we denote training dataset $\mathcal{D} = \{x_i, y_i\}_{i=1,2...N}$. Each x_i consists of two inputs from different modalities as $x_i = (x_i^a, x_i^v)$, where a and v indicate audio and visual modality, respectively. $y_i \in \{1, 2, \cdots, M\}$, where M is the number of cate-

gories. We use two encoders $\varphi^a(\theta^a, \cdot)$ and $\varphi^v(\theta^v, \cdot)$ to extract features, where θ^a and θ^v are the parameters of encoders. Representatively, we take the most widely used vanilla fusion method, concatenation, as the example here[2]. Let $W \in \mathbb{R}^{M \times (d_{\varphi_a} + d_{\varphi_v})}$ and $b \in \mathbb{R}^M$ denote the parameters of the last linear classifier. Then the logits output of the multimodal model is as follows:

$$f(x_i) = W[\varphi^a(\theta^a, x_i^a); \varphi^v(\theta^v, x_i^v)] + b. \quad (1)$$

To observe the optimization process of each modality individually, W can be represented as the combination of two blocks: $[W^a, W^v]$. The Equation 1 can be rewritten as:

$$f(x_i) = W^a \cdot \varphi^a(\theta^a, x_i^a) + W^v \cdot \varphi^v(\theta^v, x_i^v) + b. \quad (2)$$

We denote the logits output for class c as $f(x_i)_c$, and the cross-entropy loss of the discriminative model becomes $L = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{e^{f(x_i)_{y_i}}}{\sum_{k=1}^{M} e^{f(x_i)_k}}$. With the *Gradient Descent* (GD) optimization method, W^a and the parameters of encoder $\varphi^a(\theta^a, \cdot)$ are updated as (similarly for W^v and $\varphi^v(\theta^v, \cdot)$):

$$
\begin{aligned}
W_{t+1}^a &= W_t^a - \eta \nabla_{W^a} L(W_t^a) \\
&= W_t^a - \eta \frac{1}{N} \sum_{i=1}^{N} \frac{\partial L}{\partial f(x_i)} \varphi^a(\theta^a, x_i^a),
\end{aligned}
\quad (3)
$$

$$
\begin{aligned}
\theta_{t+1}^a &= \theta_t^a - \eta \nabla_{\theta^a} L(\theta_t^a) \\
&= \theta_t^a - \eta \frac{1}{N} \sum_{i=1}^{N} \frac{\partial L}{\partial f(x_i)} \frac{\partial (W_t^a \cdot \varphi_t^a(\theta^a, x_i^a))}{\partial \theta_t^a},
\end{aligned}
\quad (4)
$$

where η is the learning rate. According to Equation 3 and 4, we can find that the optimization of W^a and φ^a nearly has no correlation with that of the other modality[3], except the term related to the training loss ($\frac{\partial L}{\partial f(x_i)}$). The uni-modal encoders thus can hardly make adjustment according to the feedback from each other. Then combined with Equation 2, the gradient $\frac{\partial L}{\partial f(x_i)}$ can be rewritten as :

$$\frac{\partial L}{\partial f(x_i)}_c = \frac{e^{(W^a \cdot \varphi_i^a + W^v \cdot \varphi_i^v + b)_c}}{\sum_{k=1}^{M} e^{(W^a \cdot \varphi_i^a + W^v \cdot \varphi_i^v + b)_k}} - 1_{c=y_i}. \quad (5)$$

For convenience, we simplify $\varphi^a(\theta^a, x_i^a)$ and $\varphi^v(\theta^v, x_i^v)$ as φ_i^a and φ_i^v, respectively. Then, we can infer that for sample x_i belonging to class y_i, when one modality, such as the visual modality, has better performance, it contributes more

---

[2]The analysis of another vanilla fusion method, summation, is provided in the *Supp. Materials*. These two simple but effective fusion method are widely used in many existing multimodal method.

[3]Our OGM-GE can also boost methods that contain more complex interaction between modalities, please see details in Section 4.3.
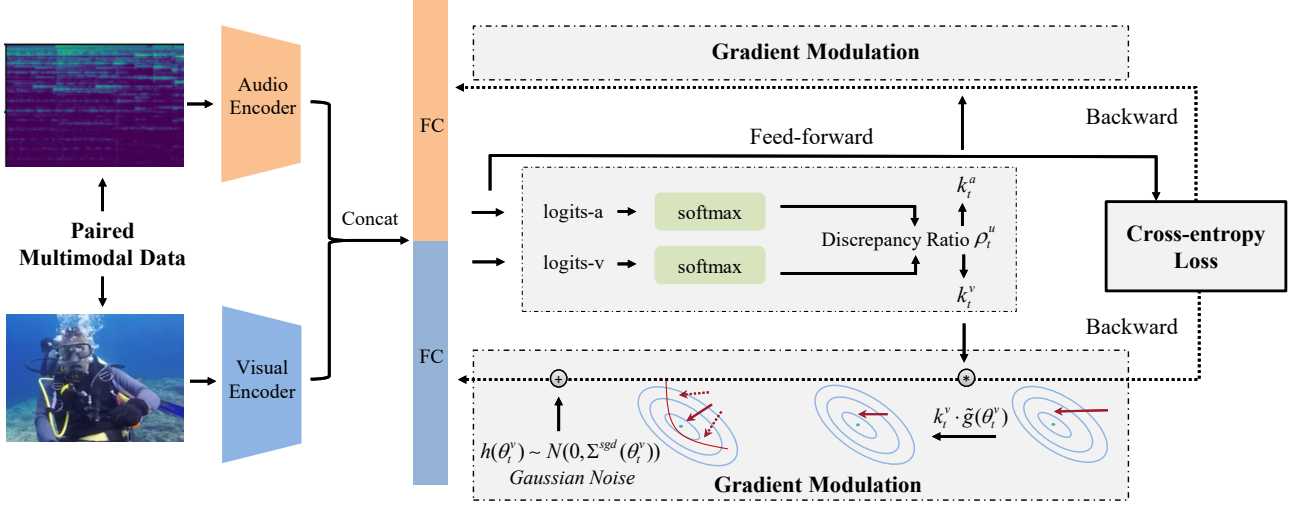
Figure 2. The pipeline of the On-the-fly Gradient Modulation with Generalization Enhancement strategy.

to $\frac{\partial L}{\partial f(x_i)_{y_i}}$ via $W^v \cdot \varphi_i^v$, leading to lower loss globally. Accordingly, the audio modality, which relatively has a lower confidence for the correct category, could obtain limited optimization efforts w.r.t. its modal-specific parameters during the back propagation. This phenomenon indicates that the modality with better performance dominates the optimization progress. As a result, when the training of multimodal model is about to converge, the other modality could still suffer from under-optimized representation and needs further training.

3.2. On-the-fly gradient modulation

As discussed, the optimization progress of multimodal discriminative models is usually dominated by the modality with better performance, leading to the under-optimized representations that limit the model performance. To solve the problem, we aim to amend the optimization process of each modality via OGM strategy, as shown in Figure 2.

In this section, we follow the notation in Section 3.1. The parameters θ^u of the encoder φ^u, where $u \in \{a, v\}$, is updated as follows when using GD method:

$$\theta_{t+1}^u = \theta_t^u - \eta \nabla_{\theta^u} L(\theta_t^u). \tag{6}$$

In practice, we employ the widely used *Stochastic Gradient Descent* (SGD) optimization method and the parameters are updated as:

$$\theta_{t+1}^u = \theta_t^u - \eta \tilde{g}(\theta_t^u), \tag{7}$$

where $\tilde{g}(\theta_t^u) = \frac{1}{m} \sum_{x \in B_t} \nabla_{\theta^u} \ell(x; \theta_t^u)$ is an unbiased estimation of the full gradient $\nabla_{\theta^u} L(\theta_t^u)$. B_t is a random mini-batch which is chosen in the t-th step with size m, and $\nabla_{\theta^u} \ell(x; \theta_t^u)$ is the gradient w.r.t. B_t.

Specific to the optimization imbalance problem discussed in Section 3.1, we propose to adaptively modulate the gradient of each modality via monitoring the discrepancy of their contribution to the learning objective. Here we design the discrepancy ratio ρ_t^u:

$$s_i^a = \sum_{k=1}^{M} 1_{k=y_i} \cdot \text{softmax}(W_t^a \cdot \varphi_t^a(\theta^a, x_i^a) + \frac{b}{2})_k,$$
$$s_i^v = \sum_{k=1}^{M} 1_{k=y_i} \cdot \text{softmax}(W_t^v \cdot \varphi_t^v(\theta^v, x_i^v) + \frac{b}{2})_k, \tag{8}$$

$$\rho_t^v = \frac{\sum_{i \in B_t} s_i^v}{\sum_{i \in B_t} s_i^a}. \tag{9}$$

ρ_t^a is accordingly defined as the reciprocal of ρ_t^v. Here we use $(W_i^u \cdot \varphi_i(\theta^u, x_i^u) + \frac{b}{2})$ as the approximated prediction of modality u to estimate uni-modal performance of the multimodal model[4]. With ρ_t^u to dynamically monitor the contribution discrepancy between audio and visual modalities, we are able to adaptively modulate the gradient via:

$$k_t^u = \begin{cases} 1 - \tanh(\alpha \cdot \rho_t^u) & \rho_t^u > 1 \\ 1 & \text{others,} \end{cases} \tag{10}$$

where α is a hyper-parameter to control the degree of modulation. We integrate the coefficient k_t^u into SGD optimization method, and θ_t^u in iteration t is updated as follows:

$$\theta_{t+1}^u = \theta_t^u - \eta \cdot k_t^u \tilde{g}(\theta_t^u). \tag{11}$$

---

[4]Inspired by Deep Boltzmann Machine [34] that uses the average of the bottom-up weight and the top-down weight to obtain the posterior distribution, we use $\frac{b}{2}$ as the bias term to estimate the uni-modal prediction of the multimodal model.

By means of k_t^u, the optimization of modality with better performance ($\rho_t^u > 1$) is mitigated, while the other modality is not affected and able to get rid of limited optimization efforts, gaining adequate training. Using the SGD optimization method with OGM strategy, the optimization process of each modality is controlled respectively and the imbalance problem can be alleviated.

3.3. Generalization enhancement

In the SGD optimization method, according to the central limit theorem, the gradient $\tilde{g}(\theta_t^u)$ in Equation 7 follows a Gaussian distribution when the batch size m is large enough [27], i.e.,

$$\tilde{g}(\theta_t^u) \sim \mathcal{N}(\nabla_{\theta^u} L(\theta_t^u), \Sigma^{sgd}(\theta_t^u)), \tag{12}$$

$$\Sigma^{sgd}(\theta_t^u) \approx \frac{1}{m}[\frac{1}{N}\sum_{i=1}^{N}\nabla_{\theta^u}\ell(x_i;\theta_t^u)\nabla_{\theta^u}\ell(x_i;\theta_t^u)^{\mathrm{T}} \\ - \nabla_{\theta^u}L(\theta_t^u)\nabla_{\theta^u}L(\theta_t^u)^{\mathrm{T}}]. \tag{13}$$

Then Equation 7 can be rewritten as follows, where ξ_t is the noise term:

$$\theta_{t+1}^u = \theta_t^u - \eta\nabla_{\theta^u}L(\theta_t^u) + \eta\xi_t, \xi_t \sim \mathcal{N}(0, \Sigma^{sgd}(\theta_t^u)). \tag{14}$$

Theorem 1 (SGD generalization ability). *Noise in SGD is closely related to its generalization ability and larger SGD noise often leads to better generalization [21]. The covariance of the SGD noise is proportional to the ratio of learning rate to batch size.*

According to Theorem 1, higher value of the gradient covariance (ξ_t in Equation 14) often brings better generalization ability. When we use the coefficient k_t^u to modulate the gradient, θ^u is updated as:

$$\theta_{t+1}^u = \theta_t^u - \eta\nabla_{\theta^u}L'(\theta_t^u) + \eta\xi_t', \\ \xi_t' \sim \mathcal{N}(0, (k_t^u)^2 \cdot \Sigma^{sgd}(\theta_t^u)), \tag{15}$$

where $\eta\nabla_{\theta^u}L'(\theta_t^u) = k_t^u \cdot \eta\nabla_{\theta^u}L(\theta_t^u)$. Based on Equation 10, $k_t^u \in (0,1]$. Absolutely, when the learning rate and batch-size are fixed, the covariance of ξ_t' is smaller than the original ξ_t. The generalization ability of SGD optimization method could be reduced. Hence, it is desirable to develop a method of controlling the SGD noise, recovering the generalization ability.

To enhance the SGD noise, we introduce a simple but effective *Generalization Enhancement* (GE) method that adds a randomly sampled Gaussian noise $h(\theta_t^u) \sim \mathcal{N}(0, \Sigma^{sgd}(\theta_t^u))$ to the gradient, which owns the same covariance as $\tilde{g}(\theta_t^u)$ and changes dynamically according to the current iteration. Then, combining the OGM method, θ_t^u is updated as:

$$\theta_{t+1}^u = \theta_t^u - \eta(k_t^u\tilde{g}(\theta_t^u) + h(\theta_t^u)) \\ = \theta_t^u - \eta\nabla_{\theta^u}L'(\theta_t^u) + \eta\xi_t' + \eta\epsilon_t, \tag{16}$$

where $\epsilon_t \sim \mathcal{N}(0, \Sigma^{sgd}(\theta_t^u))$. Since ϵ_t and ξ_t are independent, Equation 16 can be rewritten as:

$$\theta_{t+1}^u = \theta_t^u - \eta\nabla_{\theta^u}L'(\theta_t^u) + \eta\xi_t'', \\ \xi_t'' \sim \mathcal{N}(0, ((k_t^u)^2 + 1)\Sigma^{sgd}(\theta_t^u)). \tag{17}$$

Hence, the covariance of the SGD noise is recovered and even enhanced. We provide the overall OGM-GE strategy in Algorithm 1, and the pipeline is shown in Figure 2. Combining with our strategy, the optimization of multimodal model can be more balanced with guaranteed generalization ability for each modality.

Algorithm 1 Multimodal learning with OGM-GE strategy

Input: Training dataset $\mathcal{D} = \{(x_i^a, x_i^v), y_i\}_{i=1,2\ldots N}$, iteration number T, hyper-parameter α, initialized modal-specific parameters θ^u, $u \in \{a, v\}$.

  for $t = 0, \cdots, T-1$ **do**
      Sample a fresh mini-batch B_t from $\mathcal{D}$;
      Feed-forward the batched data B_t to the model;
      Calculate ρ^u using Equation 8 and 9;
      Calculate k_t^u using Equation 10;
      Calculate gradient $\tilde{g}(\theta_t^u)$ using back-propagation;
      Sample $h(\theta_t^u)$ based on covariance of gradient $\tilde{g}(\theta_t^u)$;
      Update using $\theta_{t+1}^u = \theta_t^u - \eta(k_t^u\tilde{g}(\theta_t^u) + h(\theta_t^u))$.
  end for

---

4. Experiments

4.1. Datasets

CREMA-D [5] is an audio-visual dataset for speech emotion recognition, containing 7,442 video clips of 2-3 seconds from 91 actors speaking several short words. This dataset consists of 6 most usual emotions: *angry*, *happy*, *sad*, *neutral*, *discarding*, *disgust* and *fear*. Categorical emotion labels were collected using crowd-sourcing from 2,443 raters. The whole dataset is randomly divided into 6,698-sample training set and validation set according to the ratio of 9/1, as well as a 744-sample testing set.

Kinetics-Sounds (KS) [3] is a dataset containing 31 human action classes selected from Kinetics dataset [22] which contains 400 classes of YouTube videos. All videos are manually annotated for human action using Mechanical Turk and cropped to 10 seconds long around the action. The 31 classes were chosen to be potentially manifested visually and aurally, such as playing various instruments. This dataset contains 19k 10-second video clips (15k training, 1.9k validation, 1.9k test).

VGGSound [7] is a large-scale video dataset that contains 309 classes, covering a wide range of audio events in everyday life. All videos in VGGSound are captured "in the wild" with audio-visual correspondence in the sense that the sound source is visually evident. The duration of each video is 10 seconds, and the division of the dataset is the same as [7]. In our experimental settings, 168,618 videos are used for training and validation, and 13,954 videos are used for testing because some videos are not available now on YouTube.

AVE [36] is an audio-visual video dataset for audio-visual event localization, which covers 28 event classes and consists of 4,143 10-second videos with both auditory and visual tracks as well as frame-level annotations. All videos are collected from YouTube. In experiments, the split of the dataset follows [36].

4.2. Experimental settings

In the experiment, we use the ResNet18-based [14] network as the backbones. Specifically, for the visual encoder, we take multiple frames as input, and put them into the 2D network as [45] does; for the audio encoder, we slightly change the input channel of ResNet18 from 3 to 1 as [7] does, and the rest parts keep unchanged. Videos in AVE, Kinetics-Sounds, and VGGSound lasts 10-second in length and we extract frames with 1 fps. Considering the difference between datasets, 3 frames are uniformly sampled from each 10-second clip as visual inputs. The whole audio data is transformed into a spectrogram of size $257 \times 1,004$ by librosa [28] using a window with length of 512 and overlap of 353. For CREMA-D, we extract 1 frame from each of the clip and process audio data into a spectrogram of size 257×299 with window length of 512 and overlap of 353. We use SGD with 0.9 momentum and 1e-4 weight decay as the optimizer. The learning rate is 1e-3 initially and multiplies 0.1 every 70 epochs, and α is determined from $[0, 1]$ according to the validation set.

4.3. Comparison on the multimodal task

Combination with conventional fusion methods. We first apply OGM-GE in several vanilla fusion methods: baseline, concatenation and summation, and specifically-designed fusion method: FiLM [32], then evaluate the performance on several datasets as shown in Table 1. Baseline is an introduced vanilla fusion method that uses activation function to mark out the related feature component of one modality compared with another. FiLM [32] learns to adaptively influence the output of a neural network by applying an affine transformation to the network's intermediate features. Audio-only and visual-only results are also provided. According to the results, we can find that the performance of each uni-modal model is unbalanced. For example, the audio-only model performance on the VGGSound dataset

Dataset	CREMA-D		VGGSound	
Method	Acc	mAP	Acc	mAP
Audio-only	52.5	54.2	44.3	48.4
Visual-only	41.9	43.0	31.0	34.3
Baseline	50.8	52.6	48.4	51.7
Concatenation	51.7	53.5	49.1	52.5
Summation	51.5	53.5	49.1	52.4
FiLM [32]	50.6	52.1	48.5	51.6
Baseline†	54.4	56.2	50.1	53.5
Concatenation†	61.9	63.9	**50.6**	**53.9**
Summation†	**62.2**	**64.3**	50.4	53.6
FiLM†	55.6	57.4	50.0	52.9

Table 1. Performance on CREMA-D and VGGSound dataset. Combined with OGM-GE, conventional fusion methods consistently gain considerable improvement. † indicates OGM-GE strategy is applied.

greatly outperforms its visual-only counterpart. In addition, the accuracy of audio-only model on the CREMA-D dataset is better than all of the vanilla fusion methods, which indicates that the potential of the multimodal model is indeed suppressed. After combining with OGM-GE, the performance of all the vanilla fusion methods consistently gains considerable improvement on different datasets, indicating the effectiveness and satisfactory flexibility of our method.

Dataset	CREMA-D	KS
Method	Acc	Acc
Concatenation	51.7	59.8
Modality-Drop [9] (audio)	54.4	60.3
Modality-Drop [9] (visual)	53.3	61.3
Grad-Blending [39]	56.8	62.2
OGM	59.0	61.1
OGM-GE	**61.9**	**62.3**

Table 2. Comparison with other modulation strategies on CREMA-D and Kinetics-Sounds dataset. Modality-Drop (audio) method casts the audio input to a certain probability, similarly for Modality-Drop (visual). Compared to concatenation, all other methods make progress, among which our OGM-GE achieves the best performance.

Comparison with other modulation strategies. To demonstrate the advantage of OGM-GE, we make comparisons with two main-stream modulation approaches: Modality-Dropout [43], and Gradient-Blending [39]. To be fair, the same backbone with concatenation is used in all experiments. Results shown in Table 2 prove that all the compared methods achieved better performance than the baseline of concatenation more or less, which indicates that the imbalance phenomenon indeed influences the results but also confirms the effectiveness of these methods. We can also notice that our proposed method show the superior per-

Dataset	KS	VGGSound
Method	Acc	Acc
TSN-AV [38]	58.6	49.0
TSM-AV [26]	60.3	48.8
TBN [24]	60.8	49.4
PSP [46]	59.7	49.2
TSN-AV†	59.1	49.6
TSM-AV†	62.4	49.6
TBN†	**63.1**	**50.4**
PSP†	60.4	49.5

Table 3. Comparison with existing methods on the Kinetics-Sounds and VGGSound. † indicates OGM-GE is applied.

Method	Acc
I-vector [15]	53.6
X-vector [30]	55.6
MWTSM [12]	54.1
I-vector†	55.3
X-vector†	57.1
MWTSM†	**58.0**

Table 4. Comparison with the existing methods on CREMA-D dataset. † indicates OGM-GE is applied.

Audio-visual Event Localization		
w/ or w/o OGM-GE	w/o	w/
AVGA [36]	72.0	**72.8**
PSP [46]	76.2	**76.9**

Table 5. Comparison results on audio-visual event localization between w/ and w/o OGM-GE in two representative frameworks.

formance among all other methods. Firstly, although according to Theorem 1 that OGM might theoretically hurt the generalization ability because it would reduce the intensity of the stochastic gradient noise, it is still observed improvements and comparable to other methods, which surprisingly indicates that its positive effect surpasses its negative impact to some extent. Secondly, when boosted with the introduced GE strategy, our method shows the best performance on all the datasets. It should also be noted that Gradient-Blending [39] requires to train additional uni-modal classifiers for modulation, while our method is free of that hence much more efficient and effective in contrast.

Combination with existing methods. For all the datasets, we combine the proposed OGM-GE with several existing methods to further evaluate its flexibility. For Kinetics-Sounds and VGGSound, TSN [38], TSM [26], TBN [24] and PSP [46] are compared[5]. Considering that TSN and TSM are specifically designed to event classification with visual modality, for fair comparison in multimodal setting, we add a ResNet-18 as the audio encoder and only use 3 RGB frames as the visual input, which we refer to as TSN-AV and TSM-AV. In Table 3 we can find that OGM-GE strengthens these two methods. Moreover, as mentioned before, our method is not limited in the model configurations where the audio and visual encoder have no interaction before fusion process. Concretely, there exist multiple stages of multimodal interaction before the fusion in PSP but OGM-GE can still boost the performance as in other methods, as shown in Table 3. Besides, OGM-GE is also combined with x-vector [30], i-vector [15], and MWTSM [12] on CREMA-D. The results in Table 4 further demonstrate the versatility of OGM-GE.

Application beyond classification. In order to further validate the versatility of OGM-GE in more general scenarios, we employ it in a representative task, audio-visual event localization, which can be viewed as a fine-grained classifi-

cation. The comparison results are shown in Table 5. We verify the effectiveness of our method on commonly used AVE [36] dataset. Specifically, we insert OGM-GE into two representative frameworks AGVA [36] and PSP [46]. The improvement brought by OGM-GE demonstrates that the imbalanced phenomenon is universal for different audio-visual architectures and our method still brings consistent performance gain although there exist interactions between different modalities in these AVEL frameworks, which adds difficulty for accurate estimation of the uni-modal performance. Meanwhile, the slight boosts indicate the cross-modal interactions also alleviate the imbalance. Please refer to *Supp. Materials* for more details of the implementation.

4.4. Ablation study

Imbalance modulation analysis. To further analyze OGM-GE, we monitor the change of discrepancy ratio during training. As shown in Figure 1, both the performance of audio and visual modality under the multimodal models are improved after the modulation. One interesting observation is that our method performs not as well as others in the beginning but outperforms them in the end. That is because we modulate the gradient and mitigate the optimization of the modality with better performance, by which our method can further exploit the information of the other modalities and finally gain improvement. Moreover, it can be noted from Figure 3 that the discrepancy ratio ρ^a obviously decreases after applied OGM-GE, which is a concrete evidence for the effectiveness of our method. However, it should be noted that the two modalities usually do not have equal contribution for the learning objective because of the natural imbalance existed in various curated datasets, such as VGGSound. This can also be indicated by the results in Table 1 and Figure 1.

---

[5]PSP is designed for audio-visual event localization, and here we borrow its backbone to perform the event classification tasks.

Figure 3. Change of discrepancy ratio ρ^a on the VGGSound during training. Our OGM-GE strategy helps the model to have a more balanced optimization.

Dataset	CREMA-D	KS	VGGSound
Method	Acc	Acc	Acc
SGD	51.7	59.8	49.1
SGD†	**61.9**	**63.1**	**50.6**
Adam	49.7	57.4	47.3
Adam†	**54.6**	**58.9**	**48.2**

Table 6. Experiments with SGD and Adam optimizers on CREMA-D, Kinetics-Sounds and VGGSound. † indicates OGM-GE is applied.

Adaptation to other optimizer. To explore the effectiveness of our method in combination with other optimizer, we apply OGM-GE to widely-used Adam optimizer to valid if the OGM-GE strategy works well when optimized by different optimizer. As shown in Table 6, combining with the proposed OGM-GE method, the origin SGD and Adam models gain much performance improvement. In addition, based on the same experimental setting, the SGD optimizer obtains better performance than Adam optimizer. The results show that our method can be well adapted to different optimizers, achieving consistent performance improvement.

Analysis of different noise intensities. As stated in Theorem 1, the intensity of the SGD noise is proportional to the ratio of learning rate to batch-size, and larger noise intensity often leads to better generalization. To empirically validate this theorem, we perform experiments with different batch size and learning rate, using vanilla concatenation fusion method. The introduced GE method, as a noise intensity enhancement strategy, is also solely applied to the vanilla method to explore its effectiveness. As shown in Table 7, smaller batch or larger learning rate, i.e. stronger intensity, brings better performance. These results demonstrate the existence of the potential risk of generalization damage when applying OGM, which further validates the necessity of leveraging GE. After introducing extra Gaussian noise, the performance gains considerable improvement, which demonstrates our GE method plays an important role in facilitating the generalization ability.

Settings	CREMA-D	VGGSound
(b=64, lr=1e-4)	50.4	48.3
(b=64, lr=5e-4)	51.0	48.7
(b=64, lr=1e-3)	51.8	49.1
(b= 64, lr=1e-3)	51.8	49.1
(b=128, lr=1e-3)	50.2	48.8
(b=256, lr=1e-3)	48.6	47.7
(b= 64, lr=1e-3) w/ GE	60.2	50.3

Table 7. Experiments on CREMA-D and VGGSound with different learning rates and batch-size.

5. Discussion

In this paper, we propose a simple but effective multimodal learning strategy called On-the-fly Gradient Modulation with Generalization Enhancement (OGM-GE) to alleviate the optimization imbalance problem, facilitating the exploitation of both modalities. This method achieves consistent performance gain on four representative multimodal datasets under various settings and can generally serve as a flexible plug-in strategy for both vanilla fusion method and specifically-designed fusion method as well as existing multimodal models.

Limitation. However, there is still an unresolved issue that, even equipped with OGM-GE, the uni-modal performance in multimodal model still do not surpass the best uni-modal model. We hypothesize that solely leveraging optimization-oriented method could not thoroughly solve the imbalance problem, thus different approaches, such as more advanced fusion strategy or network architectures, must be investigated. We leave this intriguing challenge to future work. Moreover, the versatility that OGM-GE suggests its tremendous potential to more multimodal scenarios, which may include depth, optical flow, language, etc.

Broader impacts. The proposed method is trained on curated datasets that perhaps contains bias, leading to the model inevitably learned such information. This issue is worth further consideration.

References

[1] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *Advances in Neural Information Processing Systems*, 33, 2020. 2

[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. 1, 2

[3] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 609–617, 2017. 5

[4] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29:892–900, 2016. 2

[5] 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

[6] Pratik Chaudhari and Stefano Soatto. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–10. IEEE, 2018. 3

[7] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020. 1, 2, 6

[8] Chenzhuang Du, Tingle Li, Yichen Liu, Zixin Wen, Tianyu Hua, Yue Wang, and Hang Zhao. Improving multimodal learning with uni-modal teachers. *arXiv preprint arXiv:2106.11059*, 2021. 1, 3

[9] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 6

[10] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10457–10467, 2020. 1, 2

[11] Michael S Gazzaniga, Richard B Ivry, and GR Mangun. Cognitive neuroscience. the biology of the mind,(2014), 2006. 1

[12] Esam Ghaleb, Mirela Popa, and Stylianos Asteriadis. Multimodal and temporal perception of audio-visual cues for emotion recognition. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 552–558. IEEE, 2019. 7

[13] Fengxiang He, Tongliang Liu, and Dacheng Tao. Control batch size and learning rate to generalize well: Theoretical and empirical evidence. *Advances in Neural Information Processing Systems*, 32:1143–1152, 2019. 3

[14] 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*, pages 770–778, 2016. 6

[15] Panikos Heracleous and Akio Yoneyama. A comprehensive study on bilingual and multilingual speech emotion recognition using a two-pass classification scheme. *PloS one*, 14(8):e0220386, 2019. 7

[16] Di Hu, Xuelong Li, et al. Temporal multimodal learning in audiovisual speech recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3574–3582, 2016. 2

[17] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9248–9257, 2019. 2

[18] Di Hu, Chengze Wang, Feiping Nie, and Xuelong Li. Dense multimodal fusion for hierarchically joint representation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3941–3945. IEEE, 2019. 2

[19] Ilija Ilievski and Jiashi Feng. Multimodal learning and reasoning for visual question answering. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 1, 2

[20] Aya Abdelsalam Ismail, Mahmudul Hasan, and Faisal Ishtiaq. Improving multimodal accuracy through modality pretraining and attention. *arXiv preprint arXiv:2011.06102*, 2020. 1

[21] Stanislaw Jastrzebski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*, 2017. 2, 3, 5

[22] 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. 5

[23] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5492–5501, 2019. 1, 2

[24] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 7

[25] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. *arXiv preprint arXiv:1807.00230*, 2018. 2

[26] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093, 2019. 7

[27] Stephan Mandt, Matthew D Hoffman, and David M Blei. Stochastic gradient descent as approximate bayesian inference. *arXiv preprint arXiv:1704.04289*, 2017. 5

[28] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, pages 18–25. Citeseer, 2015. 6

[29] Arsha Nagrani, Chen Sun, David Ross, Rahul Sukthankar, Cordelia Schmid, and Andrew Zisserman. Speech2action: Cross-modal supervision for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10317–10326, 2020. 1, 2

[30] Raghavendra Pappagari, Tianzi Wang, Jesus Villalba, Nanxin Chen, and Najim Dehak. x-vectors meet emotions: A study on dependencies between emotion and speaker recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7169–7173. IEEE, 2020. 7

[31] Kranti Parida, Neeraj Matiyali, Tanaya Guha, and Gaurav Sharma. Coordinated joint multimodal embeddings for generalized audio-visual zero-shot classification and retrieval of videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3251–3260, 2020. 1

[32] 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. 6

[33] Gerasimos Potamianos, Chalapathy Neti, Juergen Luettin, and Iain Matthews. Audio-visual automatic speech recognition: An overview. *Issues in visual and audio-visual speech processing*, 22:23, 2004. 1, 2

[34] Ruslan Salakhutdinov and Geoffrey Hinton. Deep boltzmann machines. In *Artificial intelligence and statistics*, pages 448–455. PMLR, 2009. 4

[35] Ya Sun, Sijie Mai, and Haifeng Hu. Learning to balance the learning rates between various modalities via adaptive tracking factor. *IEEE Signal Processing Letters*, 28:1650–1654, 2021. 1, 3

[36] 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. 6, 7

[37] Dong Wang, Di Hu, Xingjian Li, and Dejing Dou. Temporal relational modeling with self-supervision for action segmentation. *arXiv preprint arXiv:2012.07508*, 2020. 1

[38] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 7

[39] 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, 3, 6, 7

[40] Thomas Winterbottom, Sarah Xiao, Alistair McLean, and Noura Al Moubayed. On modality bias in the tvqa dataset. *arXiv preprint arXiv:2012.10210*, 2020. 3

[41] 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 (NeurIPS)*, 2021. 1

[42] Jingfeng Wu, Wenqing Hu, Haoyi Xiong, Jun Huan, Vladimir Braverman, and Zhanxing Zhu. On the noisy gradient descent that generalizes as sgd. In *International Conference on Machine Learning*, pages 10367–10376. PMLR, 2020. 3

[43] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020. 6

[44] Zeke Xie, Fengxiang He, Shaopeng Fu, Issei Sato, Dacheng Tao, and Masashi Sugiyama. Artificial neural variability for deep learning: On overfitting, noise memorization, and catastrophic forgetting. *Neural computation*, 33(8):2163–2192, 2021. 3

[45] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pages 570–586, 2018. 6

[46] 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. 7

[47] Mo Zhou, Tianyi Liu, Yan Li, Dachao Lin, Enlu Zhou, and Tuo Zhao. Toward understanding the importance of noise in training neural networks. In *International Conference on Machine Learning*, pages 7594–7602. PMLR, 2019. 3

[48] Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from minima and regularization effects. *stat*, 1050:21, 2018. 3