Balanced Multimodal Learning via On-the-fly Gradient Modulation

Presenter:

Connor Weeks

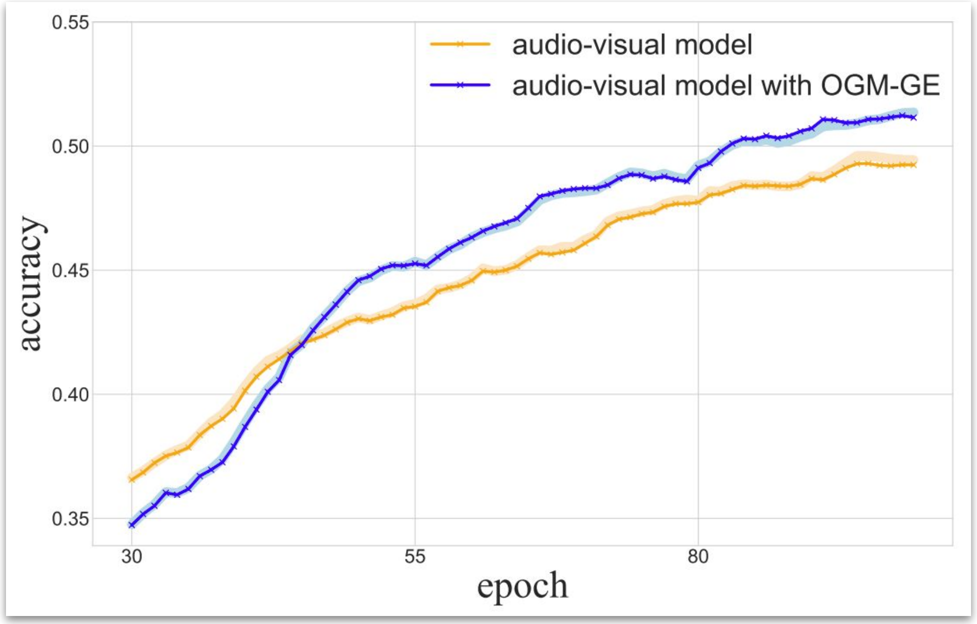
Overview

Problem Statement

Existing multimodal models could have under-optimized representations due to another dominant modality.

Contribution

Proposes a novel training procedure which measures the discrepancy between modalities to balance training and improve performance.



Related Works | Methodology | Experimental Setup | Results | Discussion

Related Work

Multimodal Learning

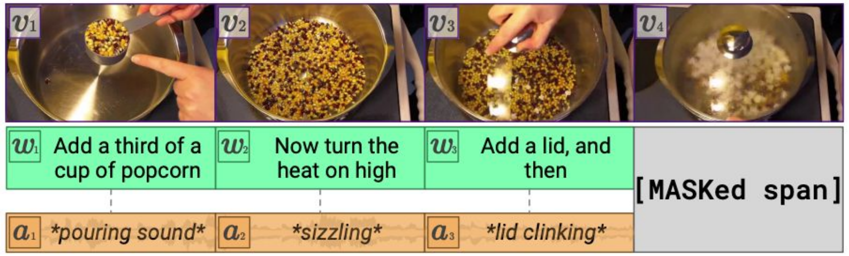
New modalities can boost performance in

- *Action recognition*
- *Audio-visual speech recognition*
- *Visual question answering*

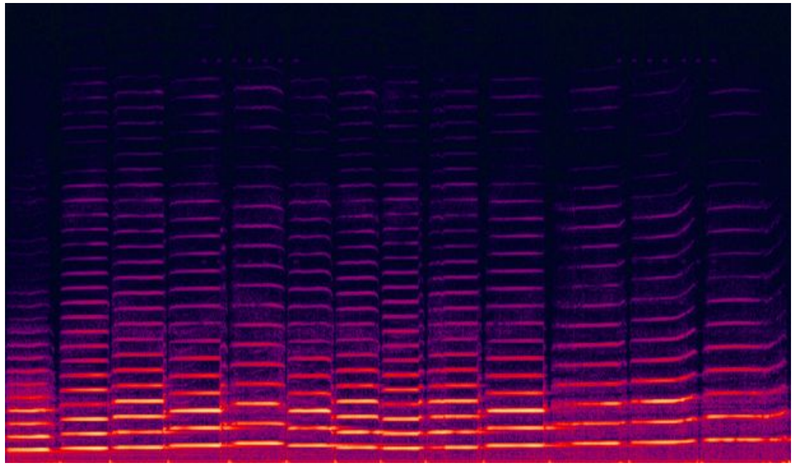
New modalities can have:

- *Different convergence rates*
- *Information discrepancies*

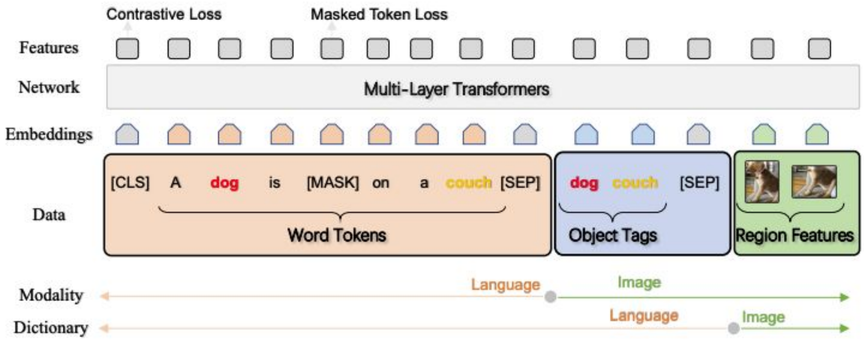
These can bias training toward one modality

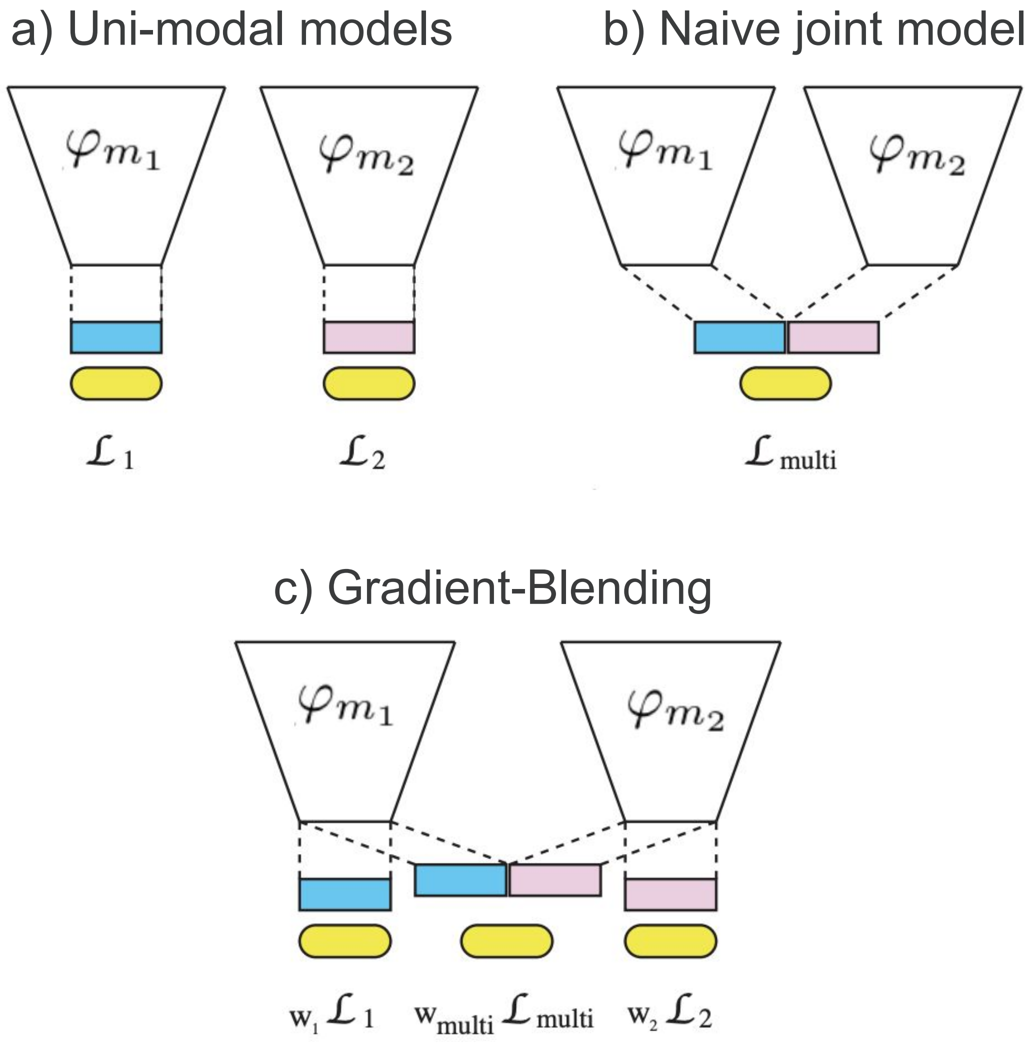


What color are her eyes?
What is the mustache made of?



video cluster #27, purity: 0.36





Gradient-Blending

- is an approach for improving multimodal balancing .
- *"computes an optimal blending of modalities based on their overfitting behaviors."*
- *targets overfitting rather than penalizing a dominant modality.*

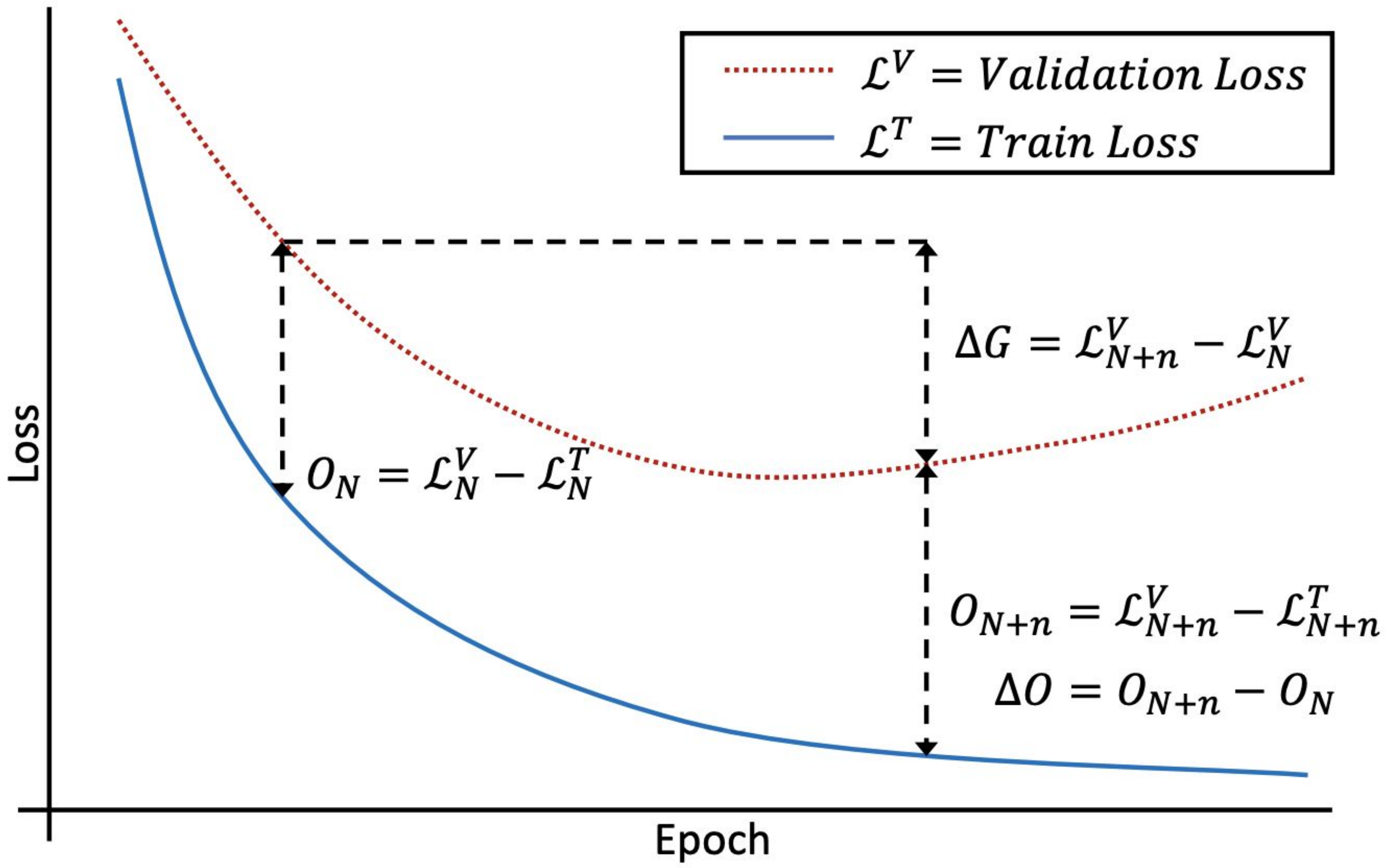
Gradient-Blending

Definition of Overfitting-to-Generalization Ratio (OGR)

$$OGR \equiv \left| \frac{\Delta O_{N,n}}{\Delta G_{N,n}} \right| = \left| \frac{O_{N+n} - O_N}{\mathcal{L}^*_N - \mathcal{L}^*_{N+n}} \right|$$

Gradient-Blending: computes an optimal blending of multiple gradients to minimize OGR^2

Uses small checkpoints to allow each gradient step to be more easily calculated.



Gradient-Blending

Equation for optimal gradient blend

$$w^* = \arg\min_w \mathbb{E}\left[\left(\frac{\langle \nabla\mathcal{L}^{\mathcal{T}} - \nabla\mathcal{L}^*, \sum_k w_k v_k\rangle}{\langle \nabla\mathcal{L}^*, \sum_k w_k v_k\rangle}\right)^2\right]$$

per-modality weights

$$w_k^* = \frac{1}{Z}\frac{\langle \nabla\mathcal{L}^*, v_k\rangle}{\sigma_k^2}$$

final loss calculation

$$\mathcal{L}_{blend} = \sum_{i=1}^{k+1} w_i \mathcal{L}_i$$

Online vs Offline Blending

Algorithm 2: Offline Gradient-Blending

input: φ^0, Initialized model
N, # of epochs
Result: Trained multi-head model φ^N
Compute per-modality weights
$\{w_i\}_{i=1}^k = GB\_Estimate(\varphi^0, N)$;
Train φ^0 with $\{w_i\}_{i=1}^k$ for N epochs to get φ^N ;

Gradient-Blending

Settings
- Audio represented with log-Mel
- RGB encoder is ResNet3D-based
- Modalities fused with two-layer FC-layer
- SGD optimizer

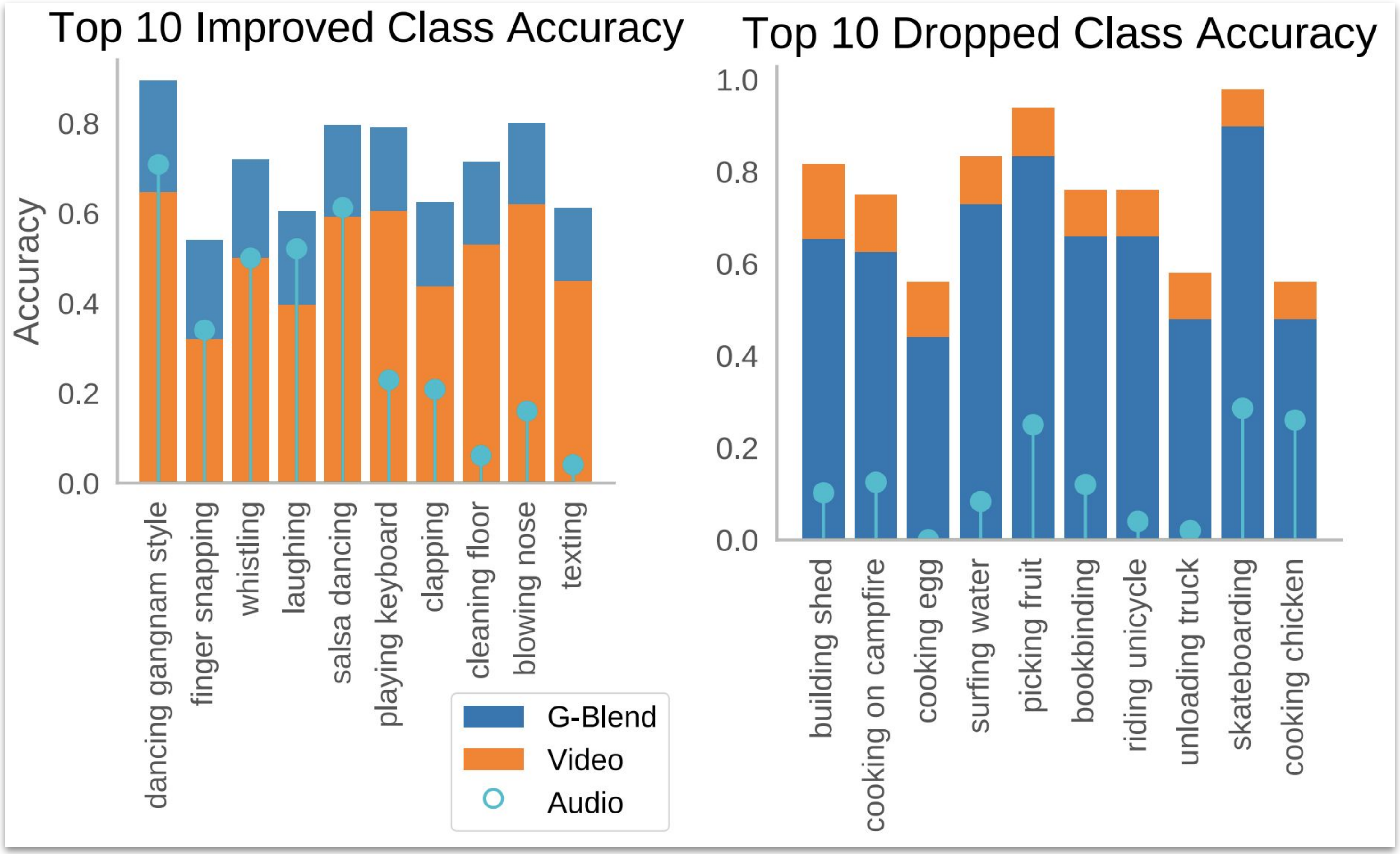
Results on Kinetics dataset

Method	Clip	V@1	V@5
Naive Training	61.8	71.7	89.6
RGB Only	63.5	72.6	90.1
Offline G-Blend	65.9	74.7	91.5
Online G-Blend	**66.9**	**75.8**	**91.9**

Results across 3 modalities: RGB image, Audio, Optical Flow.

Modal	RGB + A			RGB + OF			OF + A			RGB + OF + A		
Weights	[RGB,A,Join]=[0.630,0.014,0.356]			[RGB,OF,Join]=[0.309,0.495,0.196]			[OF,A,Join]=[0.827,0.011,0.162]			[RGB,OF,A,Join]=[0.33,0.53,0.01,0.13]		
Metric	Clip	V@1	V@5	Clip	V@1	V@5	Clip	V@1	V@5	Clip	V@1	V@5
Uni	63.5	72.6	90.1	63.5	72.6	90.1	49.2	62.1	82.6	63.5	72.6	90.1
Naive	61.8	71.4	89.3	62.2	71.3	89.6	46.2	58.3	79.9	61.0	70.0	88.7
G-Blend	**65.9**	**74.7**	**91.5**	**64.3**	**73.1**	**90.8**	**54.4**	**66.3**	**86.0**	**66.1**	**74.9**	**91.8**

Gradient-Blending



- Some accuracy drops compared to single-modality predictions
- Achieves SotA results on:
  - Kinetics
  - Sports1M
  - AudioSet
- Monitors overfitting separately for each modality

Other Related Methods

Improving Multimodal Learning with Uni-modal Teachers.

Proposes the Uni-Modal Teacher (**UMT**) method to combine uni-modal knowledge. Separate networks for each modality, then are used as teachers to distill a multimodal model.

Learning to Balance the Learning Rates Between Various Modalities via Adaptive Tracking Factor

*Defines an adaptive tracking factor (**ATF**) to adjust the learning rate of each modality. Proposes other methods to update the ATF, avoiding unimodal overfitting or underfitting.*

Methodology (OGM)

Overview of Methodology

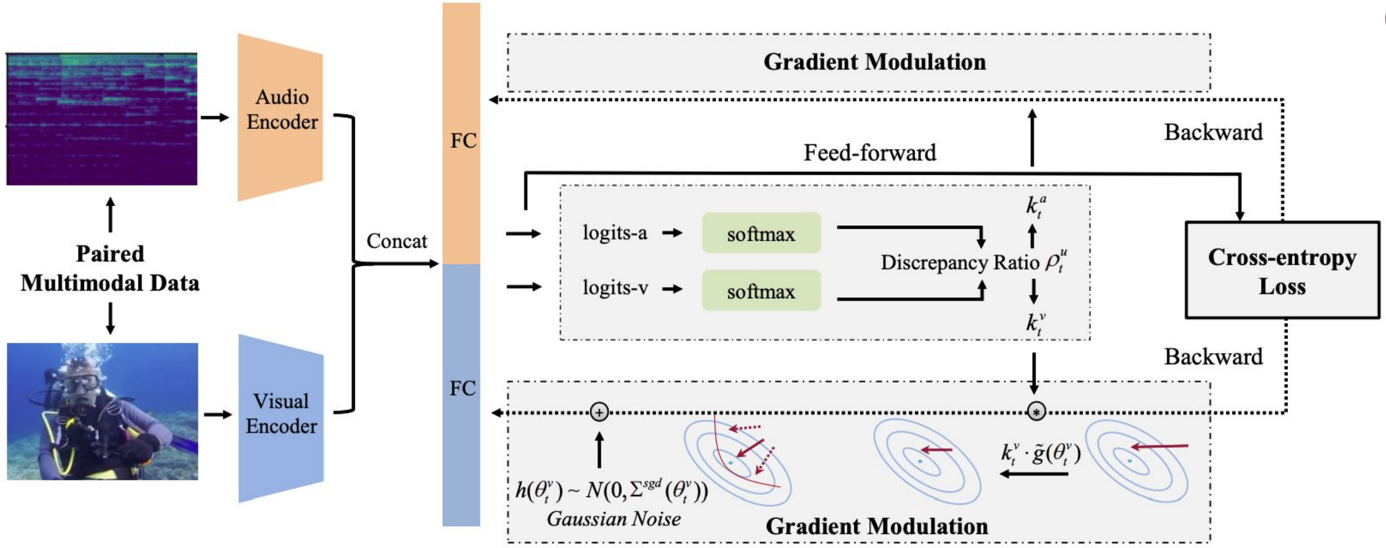


Figure 2. The pipeline of the On-the-fly Gradient Modulation with Generalization Enhancement strategy.

Component 1:

On-the-Fly Gradient Modulation (OGM)

Determines the relative balance for learning each modality.

Component 2:

Generalization Enhancement (GE)

Adds Gaussian noise to gradients to increase generalizability

On-the-fly Gradient Modulation (OGM)

weights parameters data weights parameters data

Eq. 2) $f(x_i) = W^a \cdot \underbrace{\varphi^a(\theta^a, x_i^a)}_{\text{audio encoder}} + W^v \cdot \underbrace{\varphi^v(\theta^v, x_i^v)}_{\text{image encoder}} + b.$

Full network

learning rate loss function

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

General equation for gradient descent

gradient

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

For stochastic gradient descent (SGD)

On-the-fly Gradient Modulation (OGM)

Eq. 2) $f(x_i) = \overset{\text{weights}}{W^a} \cdot \underbrace{\varphi^a(\overset{\text{parameters}}{\theta^a}, \overset{\text{data}}{x_i^a})}_{\text{audio encoder}} + \overset{\text{weights}}{W^v} \cdot \underbrace{\varphi^v(\overset{\text{parameters}}{\theta^v}, \overset{\text{data}}{x_i^v})}_{\text{image encoder}} + b.$

Full network

Eq. 6) $\theta_{t+1}^u = \theta_t^u - \overset{\text{learning rate}}{\eta} \nabla_{\theta^u} \overset{\text{loss function}}{L}(\theta_t^u).$

General equation for gradient descent

Eq. 7) $\theta_{t+1}^u = \theta_t^u - \eta \overset{\text{gradient}}{\tilde{g}}(\theta_t^u),$

For stochastic gradient descent (SGD)

On-the-fly Gradient Modulation (OGM)

Eq. 2) $$f(x_i) = \underset{\text{weights}}{W^a} \cdot \underbrace{\varphi^a(\underset{\text{parameters}}{\theta^a}, \underset{\text{data}}{x_i^a})}_{\text{audio encoder}} + \underset{\text{weights}}{W^v} \cdot \underbrace{\varphi^v(\underset{\text{parameters}}{\theta^v}, \underset{\text{data}}{x_i^v})}_{\text{image encoder}} + b.$$

Full network

Eq. 6) $$\theta_{t+1}^u = \theta_t^u - \underset{\text{learning rate}}{\eta} \nabla_{\theta^u} \underset{\text{loss function}}{L}(\theta_t^u).$$

General equation for gradient descent

Eq. 7) $$\theta_{t+1}^u = \theta_t^u - \eta \underset{\text{gradient}}{\tilde{g}}(\theta_t^u),$$

For stochastic gradient descent (SGD)

On-the-fly Gradient Modulation (OGM)

Eq. 8)

$$s_i^a = \underbrace{\sum_{k=1}^{M}}_{\text{over classes}} 1_{k=y_i} \cdot \text{softmax}(W_t^a \cdot \varphi_t^a(\theta^a, x_i^a) + \frac{b}{2})_k,$$ audio *performance*

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

Approximation of performance

Eq. 9) $$\rho_t^v = \frac{\overbrace{\sum_{i \in B_t} s_i^v}^{\text{over minibatch}}}{\sum_{i \in B_t} s_i^a}.$$

discrepancy ratio

On-the-fly Gradient Modulation (OGM)

Eq. 8)

$$s_i^a = \overbrace{\sum_{k=1}^{M}}^{\text{over classes}} 1_{k=y_i} \cdot \text{softmax}(W_t^a \cdot \varphi_t^a(\theta^a, x_i^a) + \frac{b}{2})_k,$$ audio performance

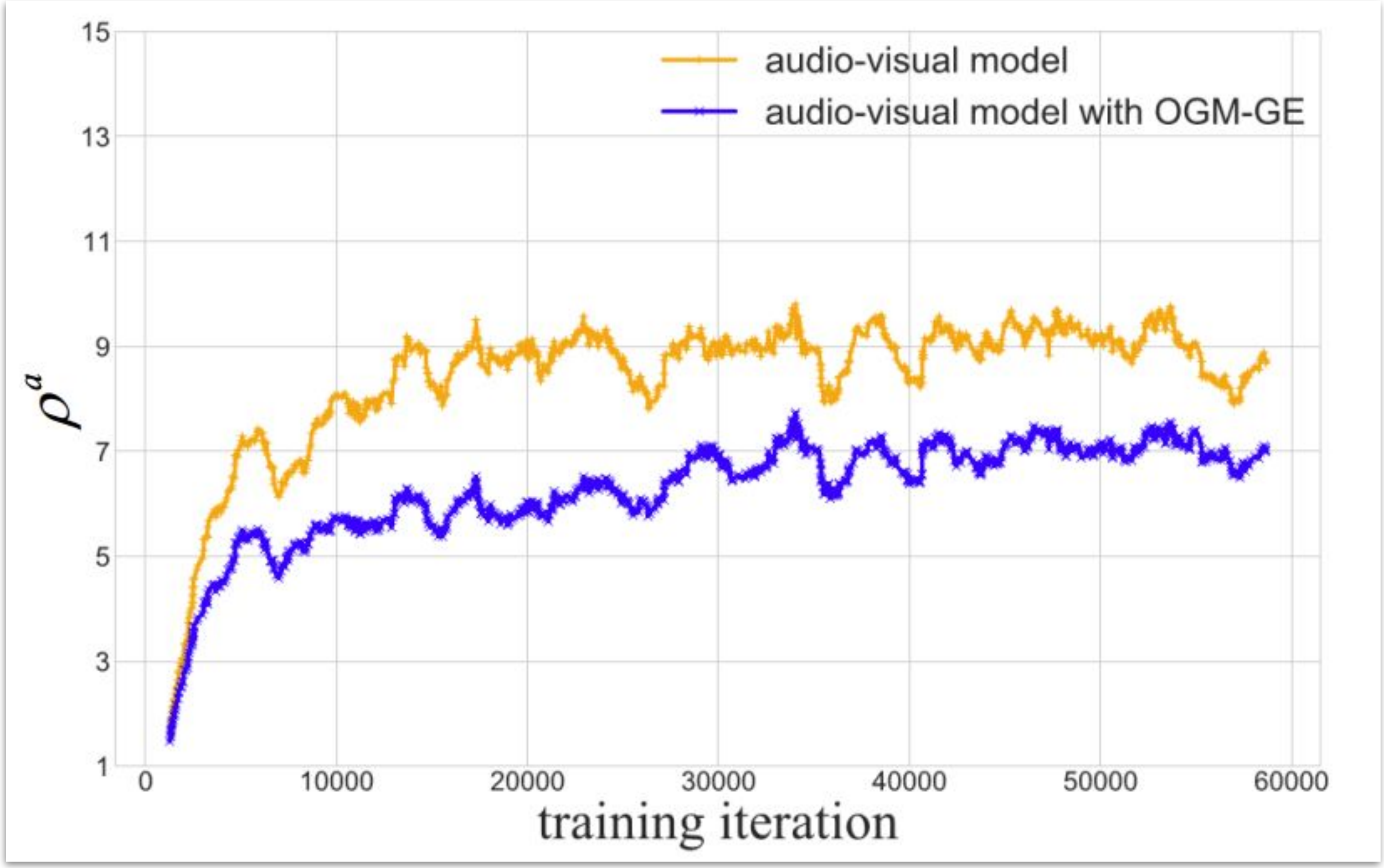
$$s_i^v = \sum_{k=1}^{M} 1_{k=y_i} \cdot \text{softmax}(W_t^v \cdot \varphi_t^v(\theta^v, x_i^v) + \frac{b}{2})_k,$$ image performance

Approximation of performance

Eq. 9) $$\rho_t^v = \frac{\overbrace{\sum_{i \in B_t} s_i^v}^{\text{over minibatch}}}{\sum_{i \in B_t} s_i^a}.$$

discrepancy ratio

On-the-fly Gradient Modulation (OGM)



OGM-GE creates a noticeable drop in discrepancy ratio.

Eq. 9) $$\rho_t^v = \frac{\overbrace{\sum_{i \in B_t} s_i^v}^{\text{over minibatch}}}{\sum_{i \in B_t} s_i^a}.$$

discrepancy ratio

On-the-fly Gradient Modulation (OGM)

Eq. 10) $$k_t^u = \begin{cases} 1 - \underbrace{\tanh(\alpha \cdot \rho_t^u)}_{\text{'penalty' term}} & \rho_t^u > 1 \\ 1 & \text{others,} \end{cases}$$

Without OGM

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

With OGM

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

Note: the minimum possible penalty is:
$1 - tanh(\alpha)$

On-the-fly Gradient Modulation (OGM)

Eq. 10) $k_t^u = \begin{cases} 1 - \overbrace{\tanh(\alpha \cdot \rho_t^u)}^{\text{'penalty' term}} & \rho_t^u > 1 \\ 1 & \text{others,} \end{cases}$

Without OGM

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

Note: the minimum possible penalty is:
$1 - tanh(\alpha)$

With OGM

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

Methodology (GE)

Generalization Enhancement (GE)

The gradient follows a normal distribution
as shown by the Central Limit Theorem

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

More SGD noise leads to better generalization.

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

SGD noise

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

Generalization Enhancement (GE)

The gradient follows a normal distribution
as shown by the Central Limit Theorem

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

More SGD noise leads to better generalization.

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

SGD noise

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

Generalization Enhancement (GE)

added noise

$$h(\theta_t^u) \sim \mathcal{N}(0, \Sigma^{sgd}(\theta_t^u))$$

Complete Equation

Eq. 16) $\theta_{t+1}^u = \theta_t^u - \eta(k_t^u \tilde{g}(\theta_t^u) + h(\theta_t^u))$

The goal of GE is to replace lost SGD noise.

Eq. 14) $\xi_t \sim \mathcal{N}(0, \Sigma^{sgd}(\theta_t^u)).$

regular SGD

Eq. 17) $\xi_t'' \sim \mathcal{N}(0, ((k_t^u)^2 + 1)\Sigma^{sgd}(\theta_t^u)).$

With both OGM and GE

Eq. 15) $\xi_t' \sim \mathcal{N}(0, (k_t^u)^2 \cdot \Sigma^{sgd}(\theta_t^u)),$

With only OGM

Generalization Enhancement (GE)

added noise

$$h(\theta_t^u) \sim \mathcal{N}(0, \Sigma^{sgd}(\theta_t^u))$$

Complete Equation

Eq. 16) $\theta_{t+1}^u = \theta_t^u - \eta(k_t^u \tilde{g}(\theta_t^u) + h(\theta_t^u))$

The goal of GE is to replace lost SGD noise.

Eq. 14) $\xi_t \sim \mathcal{N}(0, \Sigma^{sgd}(\theta_t^u)).$

regular SGD

Eq. 17) $\xi_t'' \sim \mathcal{N}(0, ((k_t^u)^2 + 1)\Sigma^{sgd}(\theta_t^u)).$

With both OGM and GE

Eq. 15) $\xi_t' \sim \mathcal{N}(0, (k_t^u)^2 \cdot \Sigma^{sgd}(\theta_t^u)),$

With only OGM

OGM-GE Algorithm

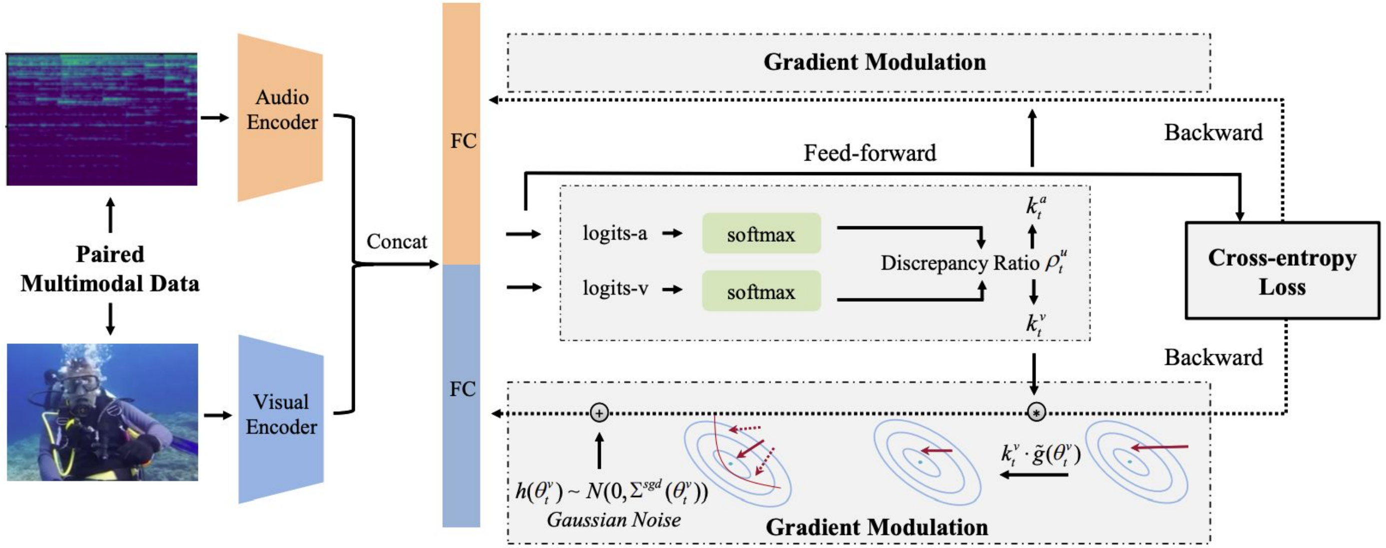


Figure 2. The pipeline of the On-the-fly Gradient Modulation with Generalization Enhancement strategy.

Algorithm 1 Multimodal learning with OGM-GE strategy

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

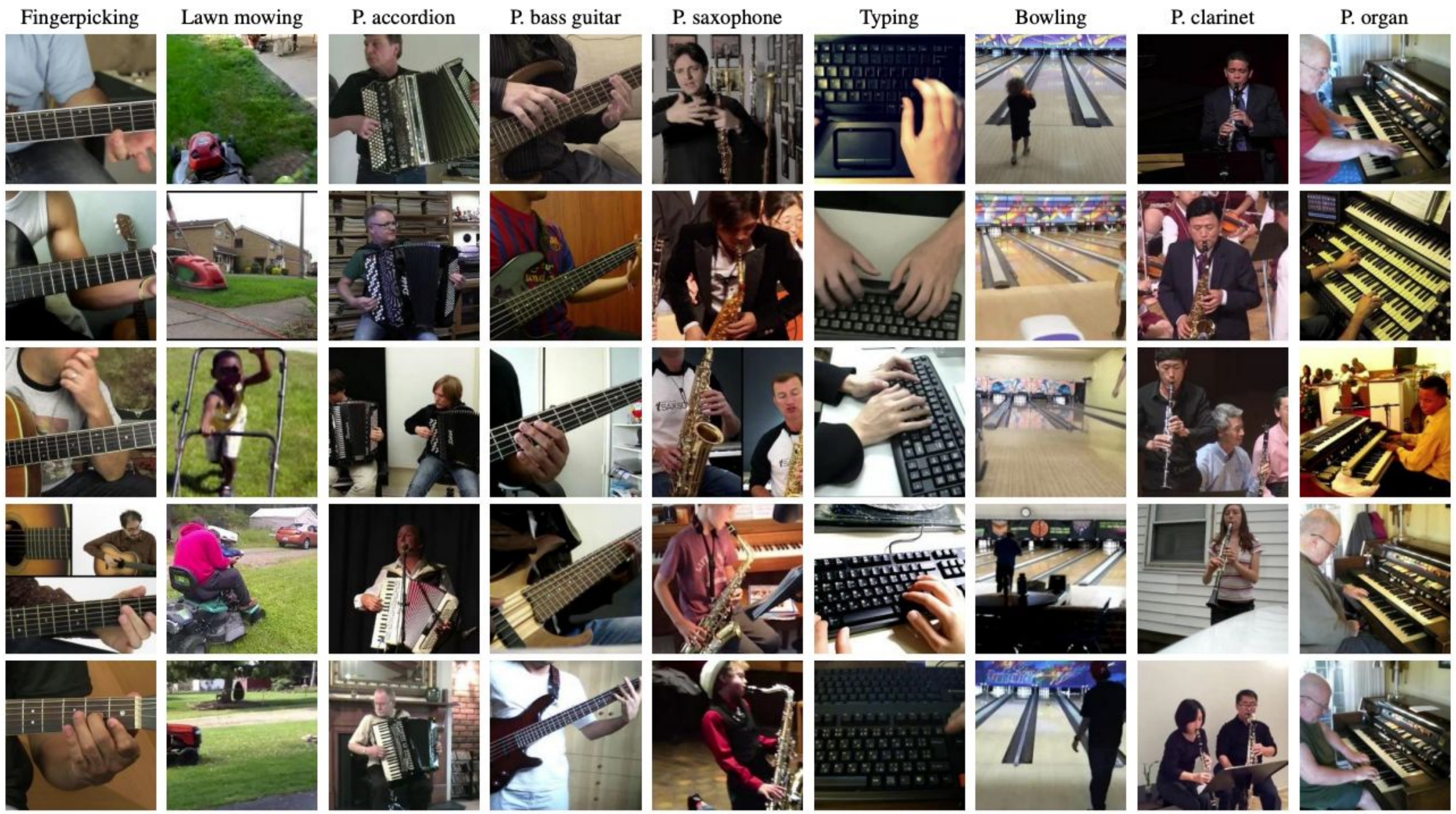
for $t = 0, \cdots, T-1$ **do**

- 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

Experimental Setup

9 of 34 class in Kinetics-Sounds



Datasets

Multi-modal categorization

1. CREMA-D
2. Kinetics-Sounds
3. VGGSound

Audio-Visual Localization

4. AVE

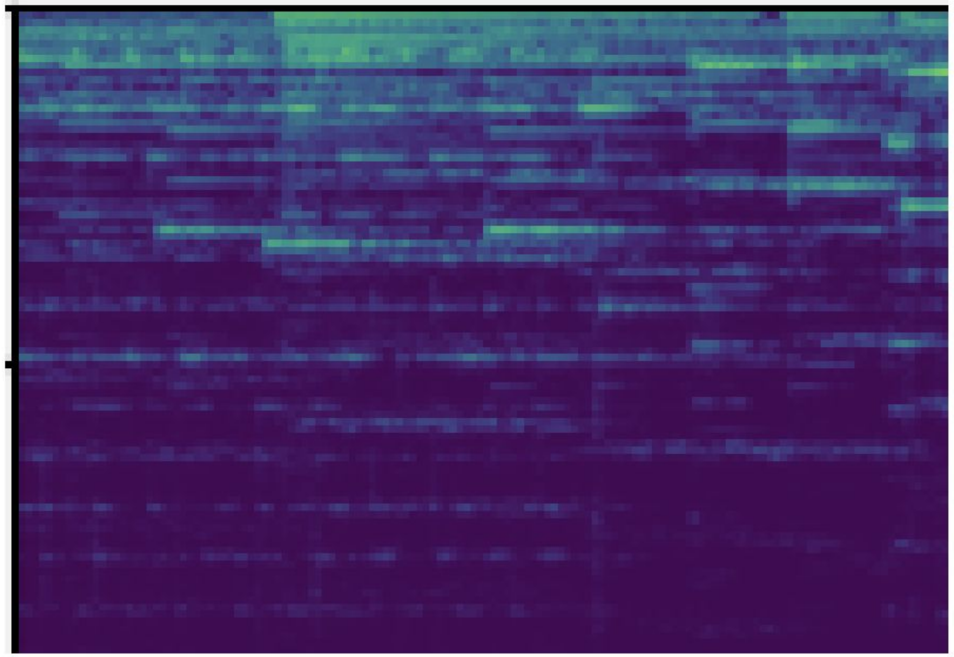
Experimental Settings

Visual Encoder

- ResNet18-based
- 3 frames per video
- Temporal Pooling

Audio Encoder

- Transformed to spectrogram
- ResNet18-based
- Input channels set to 1



Results

How does OGM-GE compare to **conventional fusion methods**?

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

OGM-GE consistently improves performance of baseline methods.

† indicates that OGM-GE was applied

How does OGM-GE compare to **other modulation strategies**?

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**

All methods make progress, but OGM-GE achieves the highest performance.

Can OGM-GE be **combined** with existing methods?

All multimodal approaches evaluated show improvements with OGM-GE

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

OGM-GE is not limited to disconnected encoders

PSP is an example using co-attention.

Results for the CREMA-D dataset

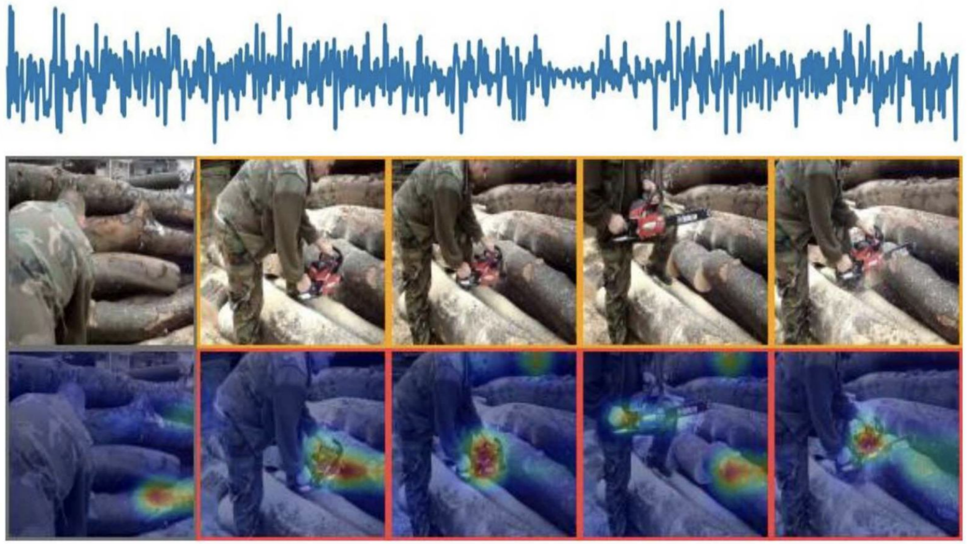
Method	Acc
I-vector [15]	53.6
X-vector [30]	55.6
MWTSM [12]	54.1
I-vector†	55.3
X-vector†	57.1
MWTSM†	**58.0**

† indicates that OGM-GE was applied

Can OGM-GE be applied to **other tasks**?

OGM-GE can also work on audio-visual event localization (AVE).

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**



Ablation Study

OGM-GE still improves performance when used with an Adam optimizer.

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**

Evaluation of learning rates and batch sizes.

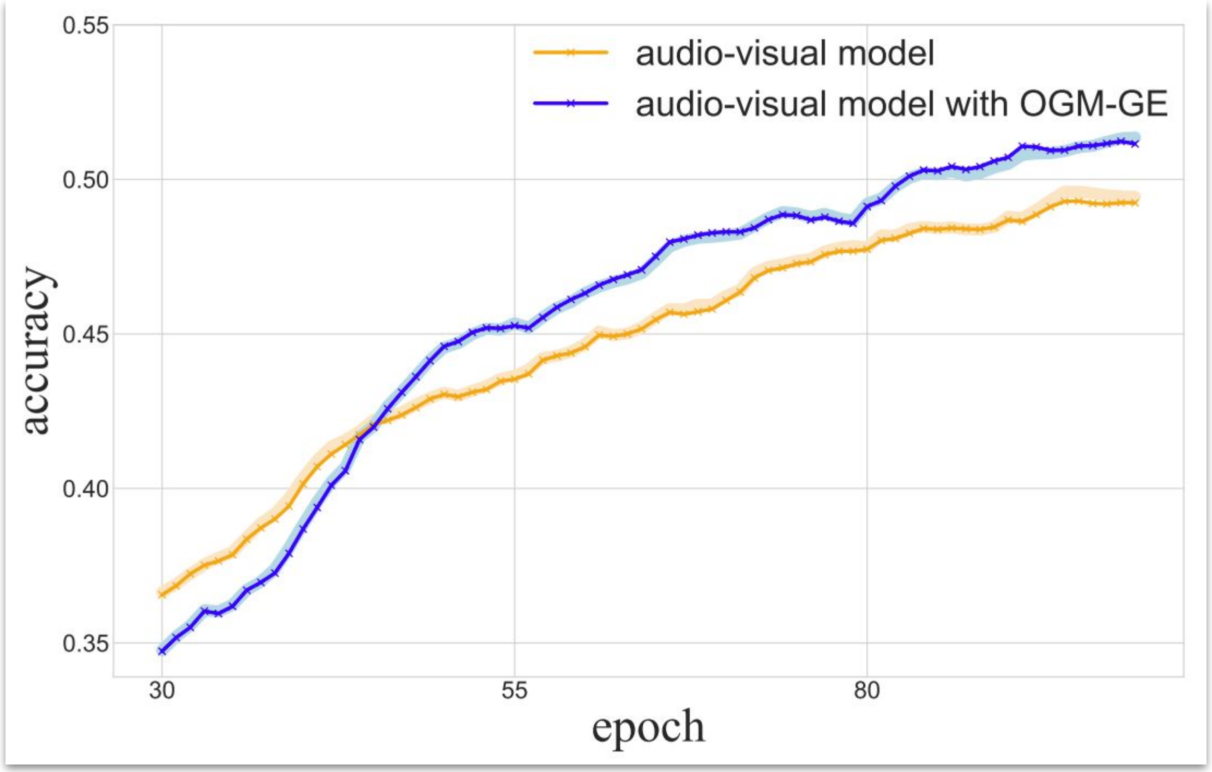
Settings	CREMA-D	VGGSound
(b=64, lr=1e-4)	50.4	48.3
(b=64, lr=5e-4)	51.0	48.7
(b=64, lr=1e-3)	51.8	49.1
(b= 64, lr=1e-3)	51.8	49.1
(b=128, lr=1e-3)	50.2	48.8
(b=256, lr=1e-3)	48.6	47.7
(b= 64, lr=1e-3) w/ GE	60.2	50.3

Strengths & Weaknesses

Strengths

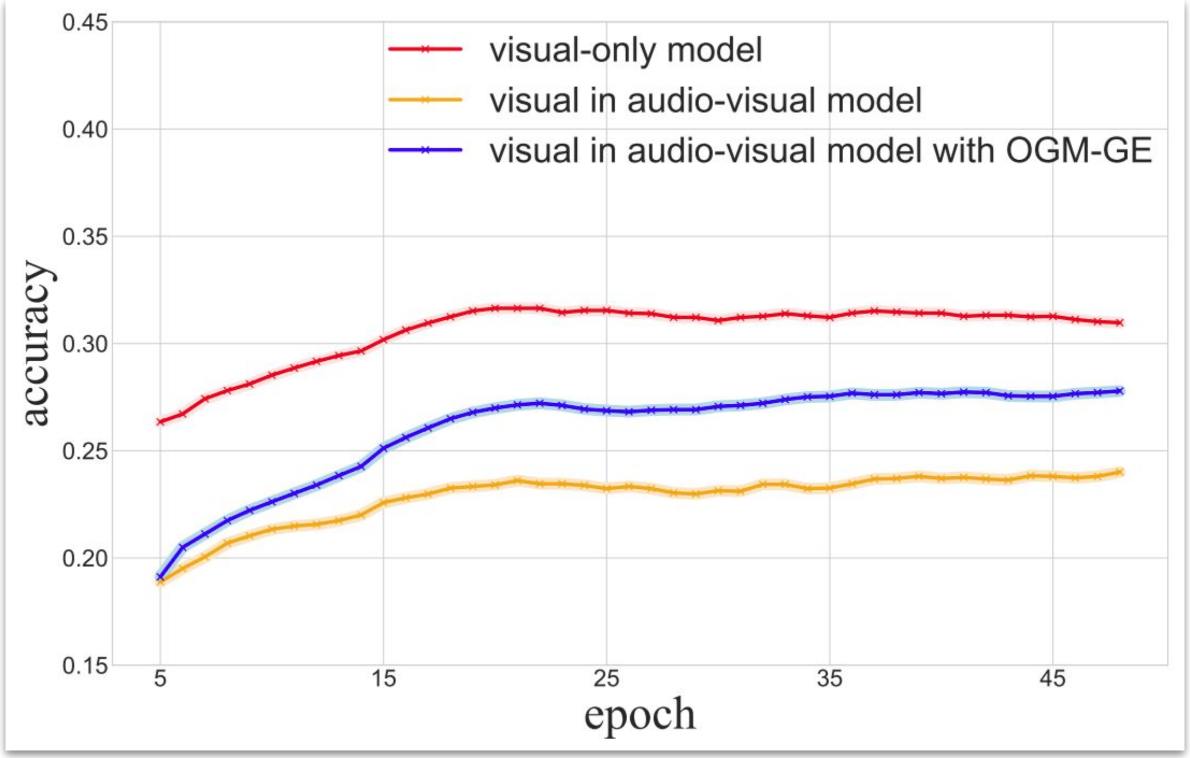
- No limited by modality
  - Applicable to text, full video, etc.
- The authors address the potential issue of interconnected encoders by evaluating PSP.
- Limited computational requirements, and fairly easy to implement.
- Consistently improves baseline results.

Evaluation of VGGSound dataset



Weaknesses

Evaluation of VGGSound dataset



- Algorithm seems to require a categorical output
  - Event localization is implemented as *"fine-grained classification"*.
- Minimum penalty is greater than 0.
  - May be a problem for near-equal settings.
- No method prescribed for optimizing degree of modulation α.
  - Ranges from 0.1 to 0.8
- Individual components still fall short of unimodal training

Discussion

Discussion Questions

1. Are the gains produced by adding Gaussian noise dependent on the modulation? (i.e. would running only GE improve the results)

2. How would you optimize the degree of modulation α, or make the algorithm choose it dynamically?

3. The authors claim the method is not limited to models with separated encoders, what architectures might cause OGM-GE to fail?

References

https://arxiv.org/pdf/1705.08168.pdf

https://arxiv.org/pdf/1905.12681.pdf

https://arxiv.org/pdf/2203.15332.pdf

https://arxiv.org/pdf/1911.12667.pdf

https://visualqa.org/

https://en.wikipedia.org/wiki/Spectrogram#/media/File:Spectrogram_of_violin.png

https://openaccess.thecvf.com/content/CVPR2022/papers/Zellers_MERLOT_Reserve_Neural_Script_Knowledge_Through_Vision_and_Language_and_CVPR_2022_paper.pdf