中国人民大学
RENMIN UNIVERSITY OF CHINA

高瓴人工智能学院
Gaoling School of Artificial Intelligence

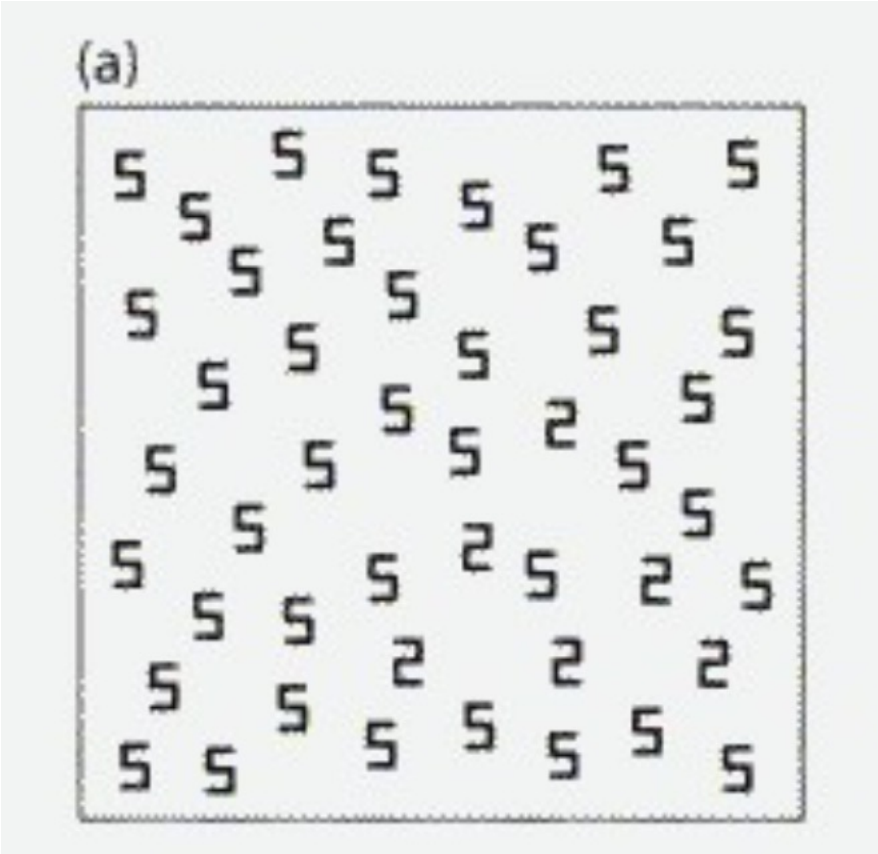
GeWu-Lab
Gaoling School of Artificial Intelligence
Renmin University of China

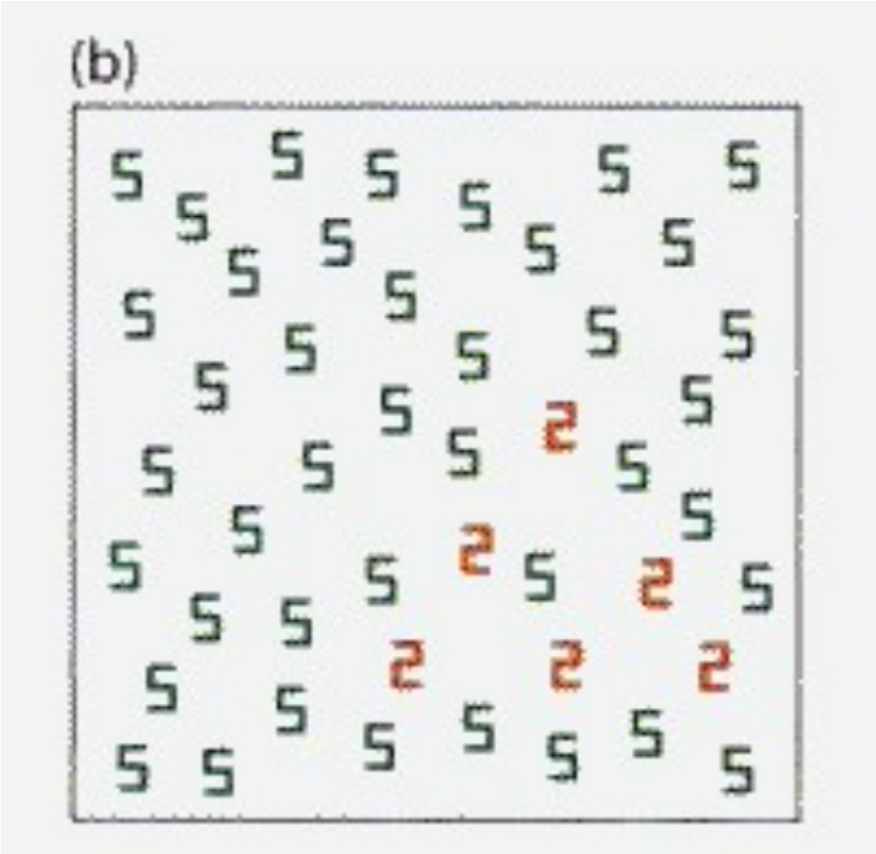
多模态场景的高效学习与理解方法探究

Di Hu（胡迪）

Email: dihu@ruc.edu.cn

2022-08-23


(a)


(b)

V. S. Ramachandran
拉马钱德兰

McGurk Effect (1976)


央视新闻
感动中国2021年度人物
江梦南
半岁时，江梦南双耳失聪，但她通过读唇语学会了“听”和“说”。凭借优秀的学习成绩，她成为家乡小镇上近年来唯一考上重点大学、最终到清华念博士的学生。江梦南的目标始终是明确的，那就是解决生命健康的难题。


江梦南
201.5万
22.5万
01:31

问题：不同模态是如何有效关联的？以及不同模态如何协同感知的？

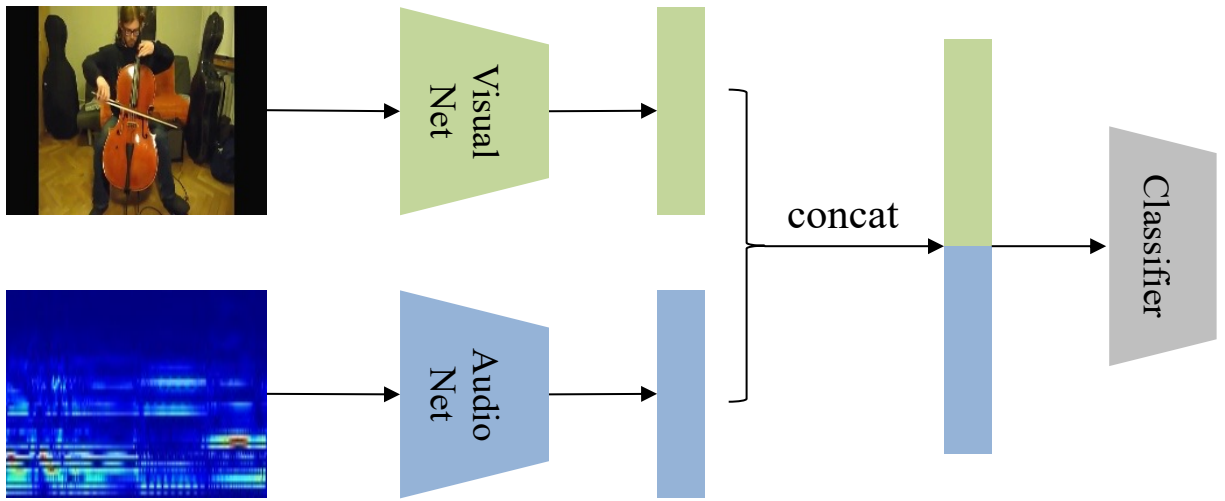
Cognitive **S**cience

+

Computer **S**cience

多模态学习机制探究 (CVPR'22 ORAL)

常用多模态学习框架

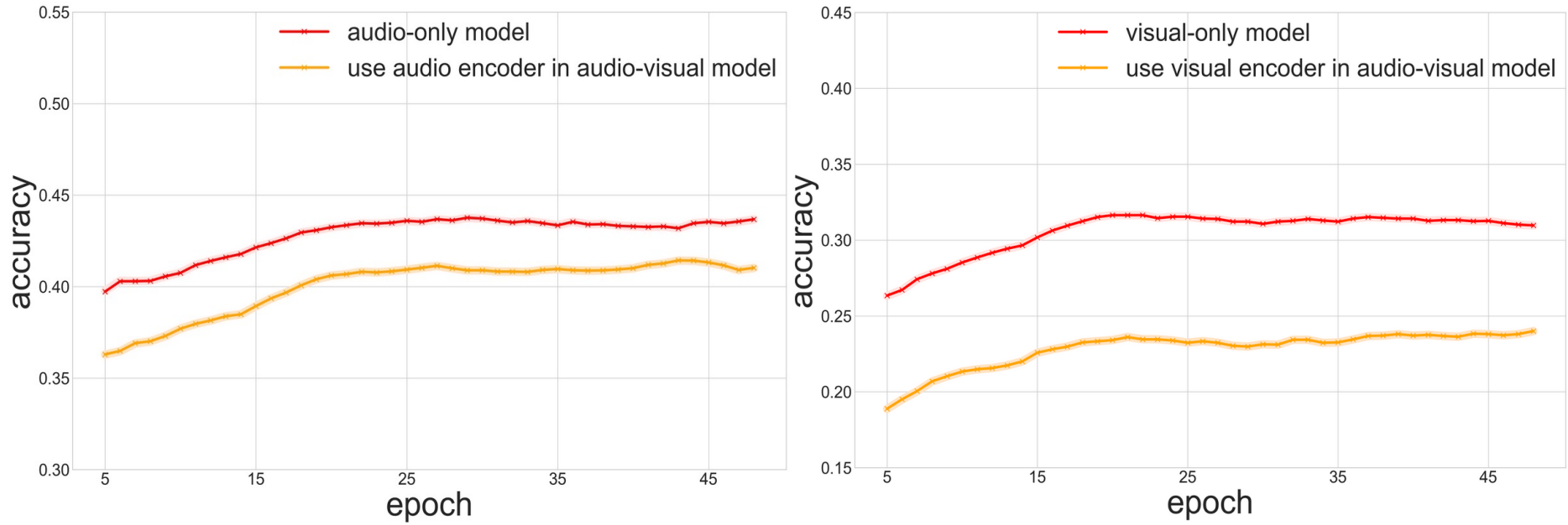


广泛使用的拼接框架

常用多模态学习框架

■ 单模态表征欠优化

- 以VGGSound数据集为例：

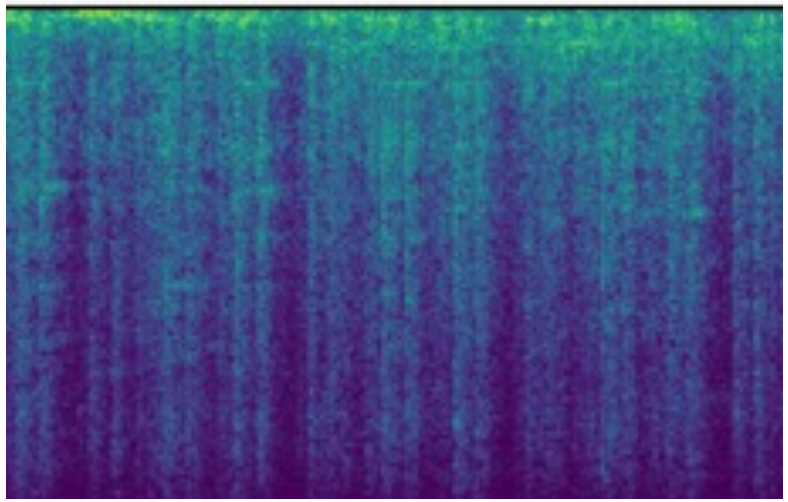


音频　▲ 单模态表征不同程度欠优化　视频

常用多模态学习框架

■ **Example：Bus**

Easy and clean visual data



Hard and noisy audio data

猜测：

- 优势模态贡献高
- 弱势模态贡献低

前馈-反向传播过程分析

前馈：

最终预测由两个模态贡献的加和构成

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

反向传播：

梯度大小最终由两个模态贡献的加和决定

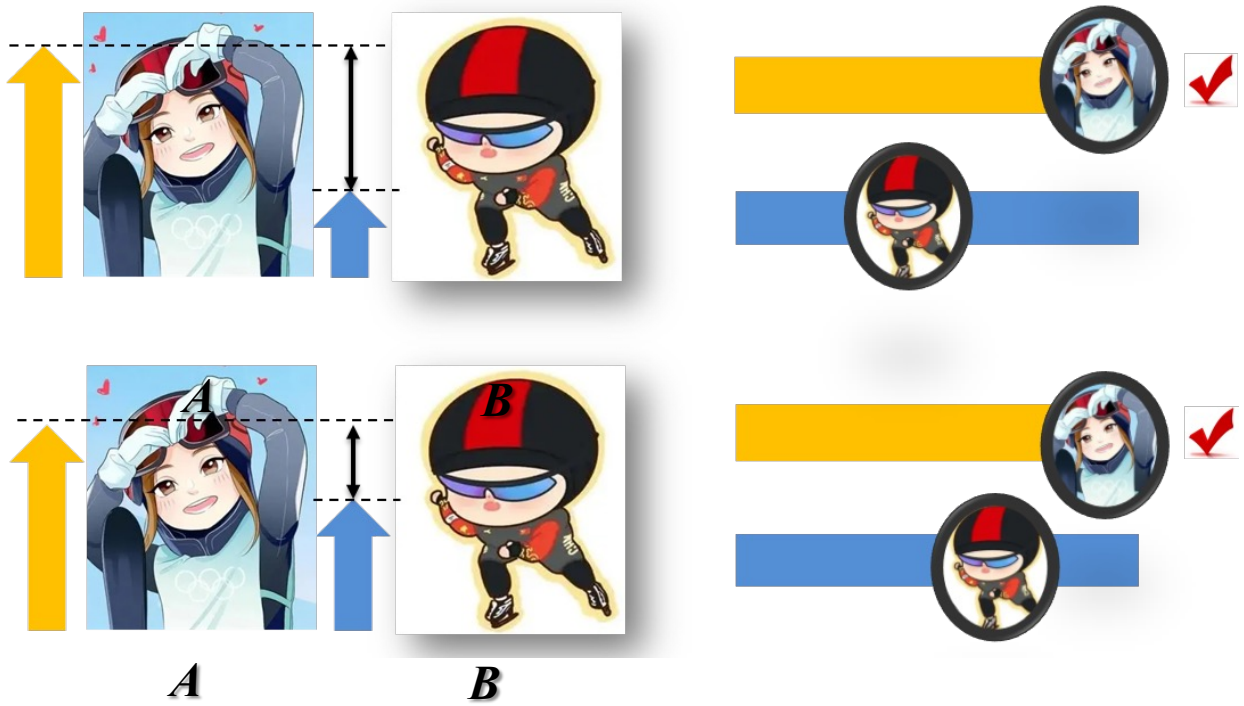
$$\frac{\partial L}{f(x_i)_c} = \frac{e^{(W^a \cdot \varphi_i^a + W^v \cdot \varphi_i^v + b)_c}}{\sum_{k=1}^{M} e^{(W^a \cdot \varphi_i^a + W^v \cdot \varphi_i^v + b)_k}} - 1_{c=y_i}$$

若某个模态表现更好，将会主导训练过程

模态间优化不平衡

编码器：$\varphi^u(\theta^u, \cdot)$　分类器：W and b　模态：$u \in \{a, v\}$

缓解不平衡训练

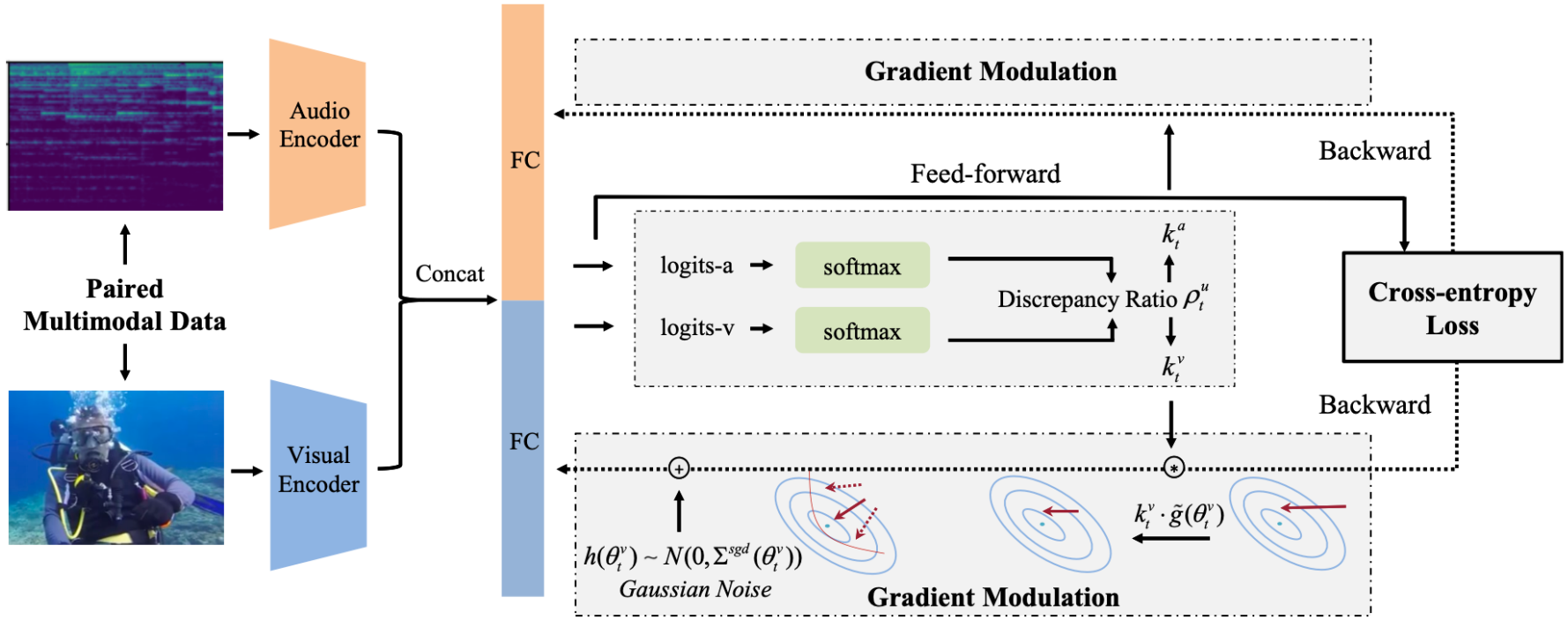


糟糕的团队合作

更优的团队合作

OGM-GE方法

- **On-the-fly Gradient Modulation (OGM) with Generalization Enhancement (GE)**



OGM-GE方法

■ On-the-fly Gradient Modulation (OGM)

动态梯度调整缓解不平衡：(以SGD优化为例)

调整基准:

单模态表现 s_i^u

模态间差异程度 ρ_t^u

$$\rho_t^a = \frac{\sum_{i\in B_t} s_i^a}{\sum_{i\in B_t} s_i^v},$$

→ **调整系数:**

超参数 α

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

→ **梯度调整:**

学习率 η, 梯度 $\tilde{g}$

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

$$\downarrow$$

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

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

OGM-GE方法

■ Generalization Enhancement (GE)

- SGD的泛化性可能受损：梯度噪声被削弱。
- 引入额外的随机噪声恢复泛化性。

梯度调整: ⟶ **噪声增强:**

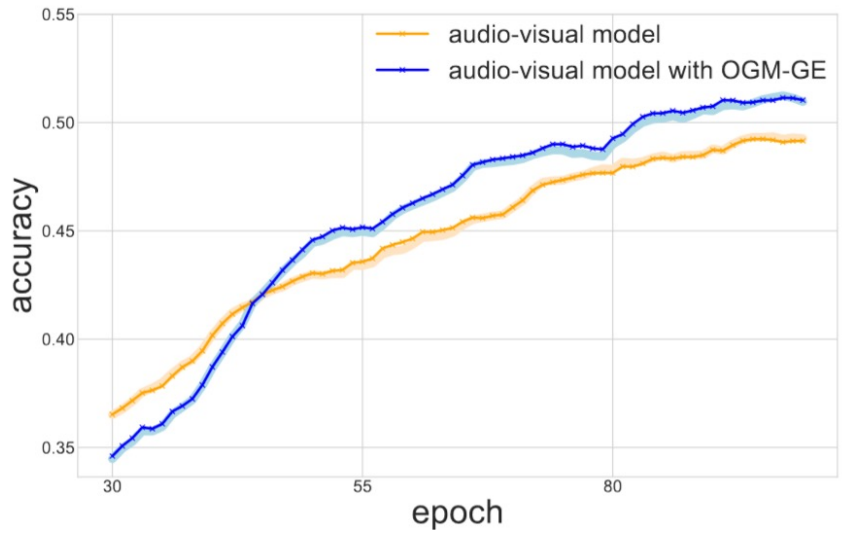
$$\theta^u_{t+1} = \theta^u_t - \eta \tilde{g}(\theta^u_t)$$

$$\theta^u_{t+1} = \theta^u_t - \eta \cdot k^u_t \tilde{g}(\theta^u_t)$$

$$\theta^u_{t+1} = \theta^u_t - \eta (k^u_t \tilde{g}(\theta^u_t) + h(\theta^u_t))$$

实验

■ 和基础融合方法的比较



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	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

† 表示在原方法的基础上增加OGM-GE方法。

实验

■ 和其他任务相关方法的比较

Dataset	KS	VGGSound
Method	Acc	Acc
TSN-AV	58.6	49.0
TSM-AV	60.3	48.8
TBN	60.8	49.4
PSP	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

CREMA-D dataset

Method	Acc
I-vector	53.6
X-vector	55.6
MWTSM	54.1
I-vector†	55.3
X-vector†	57.1
MWTSM†	**58.0**

† 表示在原方法的基础上增加OGM-GE方法。

实验

■ 和其他相关方法的比较

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

我们的方法既有效同时无需额外训练代价。

实验

■ 分类以外的应用

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

- 在视音事件定位模型中，场景更复杂，同时不同模态之间存在更复杂的交互。
- 我们的方法在更复杂的视听事件定位任务上保持有效性。

Part 01 小结

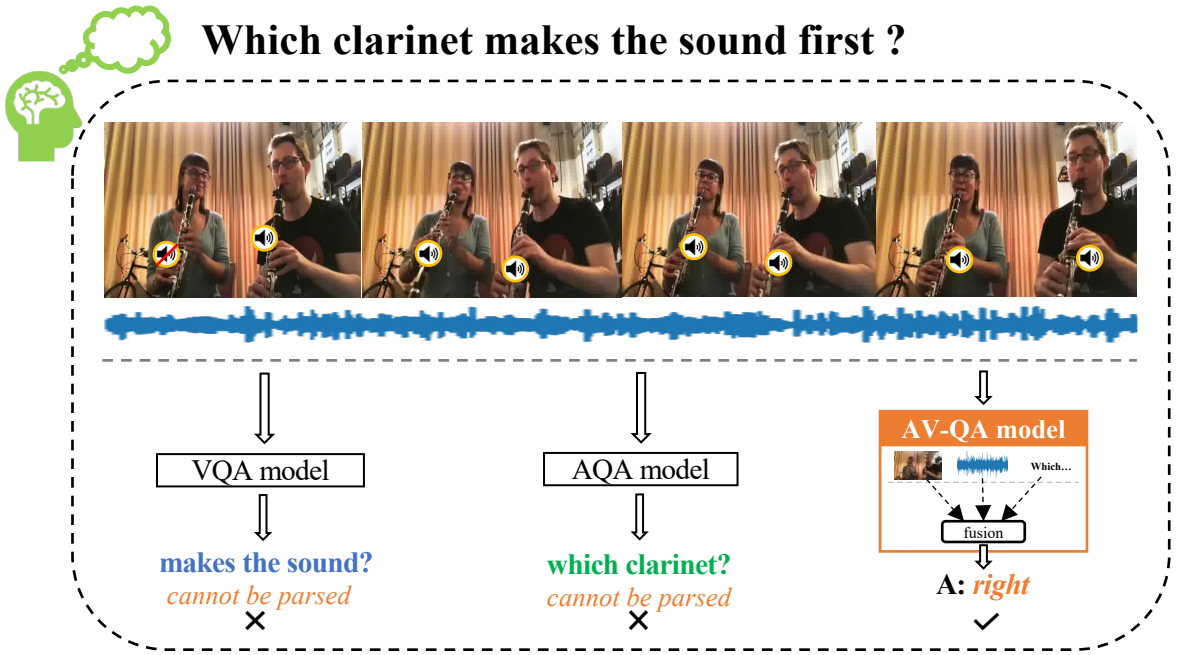
- 分析了多模态联合训练中的不平衡现象。
- 提出了OGM-GE方法，通过动态控制每个模态的优化过程缓解不平衡问题。
- OGM-GE方法不仅可以插入基础融合策略中，也可以插入现有的多模态模型中，具有良好的通用性。

视音场景分析
(CVPR'22 ORAL)

什么是AVQA任务？

■ Audio-Visual Question Answering (AVQA)

根据视频中不同物体的视觉对象、声音以及他们之间的关联回答问题。

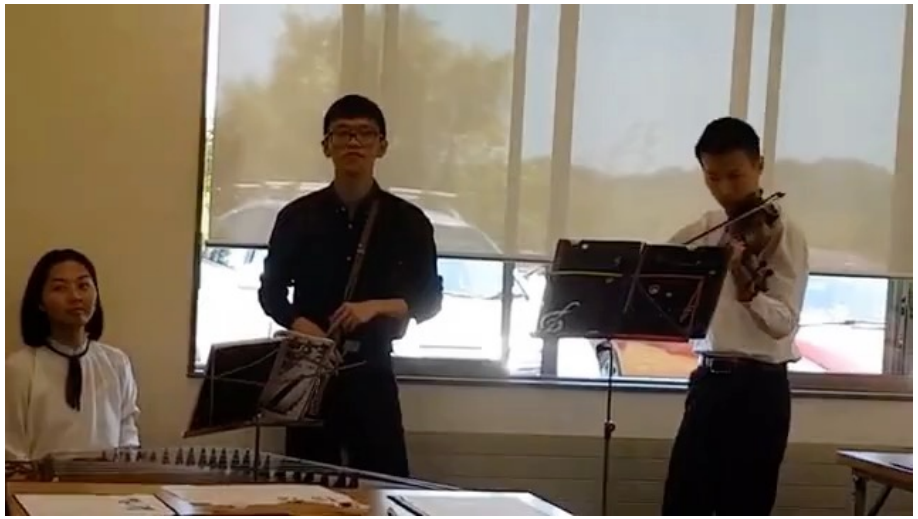


为了正确回答问题，需要有效的视听场景理解和时空推理。


MUSIC-AVQA DATASET
A large-scale audio-visual question answering dataset

MUSIC-AVQA数据集

■ Q-A Pair 样例



Question: What is the left instrument of the second sounding instrument?

Answer: guzheng

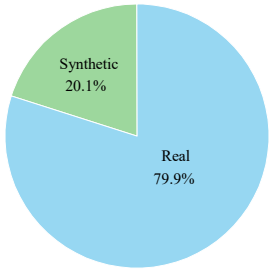
Audio-visual reasoning!

1) 声音：找出第二个发声的乐器: flute.
2) 视觉：推断出它左边的乐器

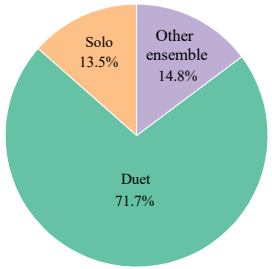
MUSIC-AVQA数据集

- 3 模态
- 9 问题类型
- 33 问题模版
- 22 乐器类型
- 9, 290 视频个数
- 超过 150 小时
- 45, 867 Q-A pair

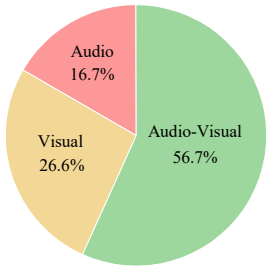
问题多样
场景动态
场景和问题复杂



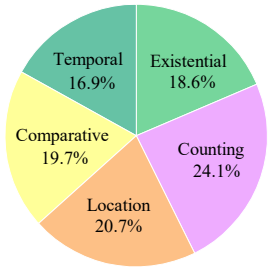
(a) Real and synthetic videos



(b) Different scene types



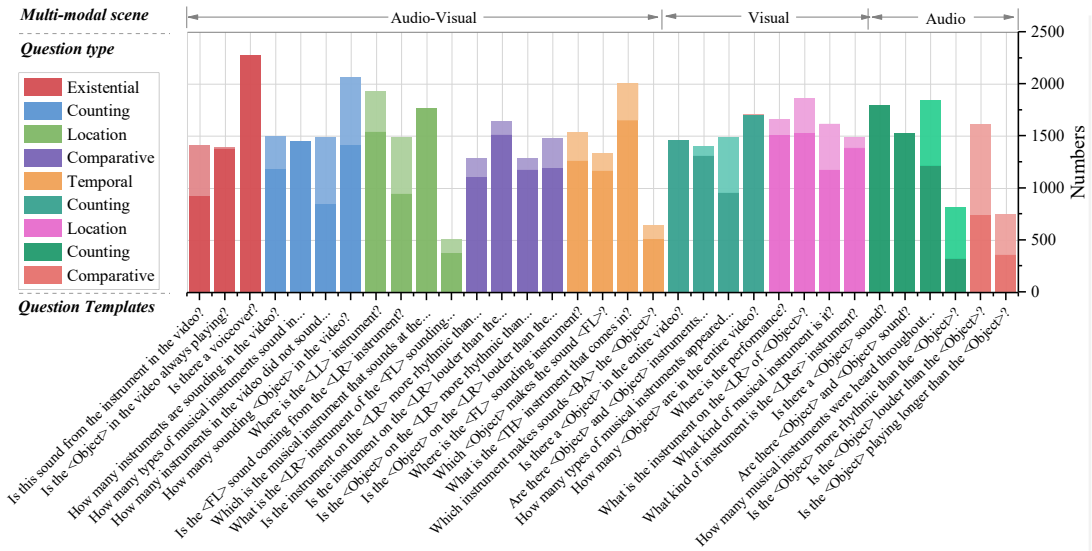
(c) Modalities QA pairs



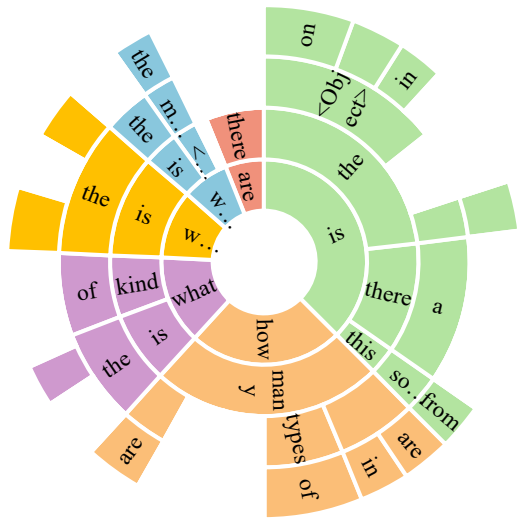
(d) Audio-visual questions

Formulas	Meaning
$\langle Object \rangle$	instrument name
$\langle LR \rangle$	left & right
$\langle LRer \rangle$	leftmost & rightmost
$\langle FL \rangle$	first & last
$\langle TH \rangle$	first, second, ...
$\langle LL \rangle$	Loudest and lowest
$\langle BA \rangle$	Before and after

(e) Question Formulas



(f) Distribution of question templates.



(g) Distribution of collected questions by their first four words.

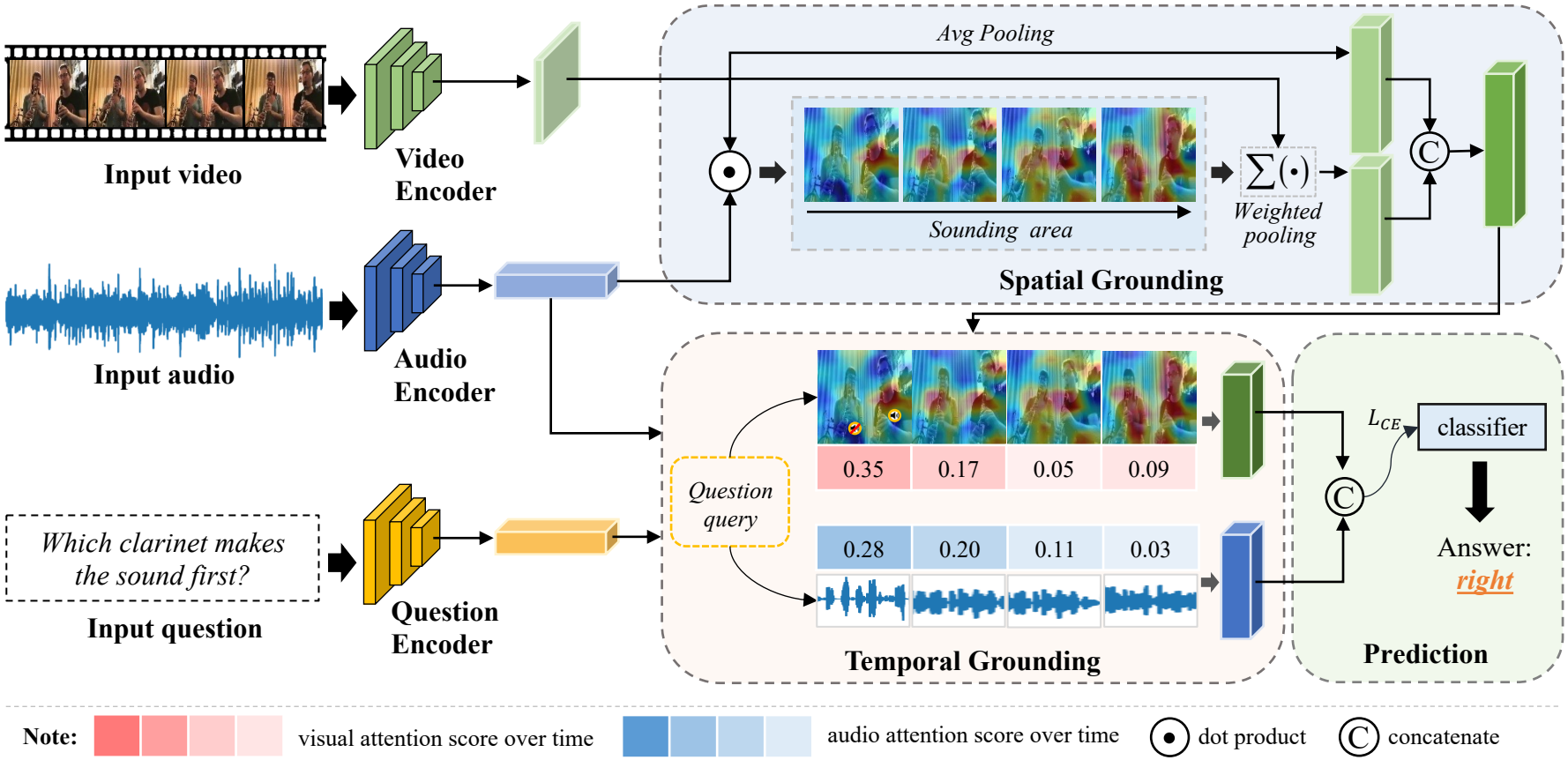
MUSIC-AVQA数据集

■ 和其他QA数据集的比较

Dataset	Origin	Main sound type	# Videos	Average video length	A Question	V Question	A-V Question				
							Existential	Location	Counting	Comparative	Temporal
ActivityNet-QA	ActivityNet	Background music	5.8K	180s	✗	✓	✗	✗	✗	✗	✗
TVQA	TV Show	Human speech	21.8K	60s/90s	✗	✓	✗	✗	✗	✗	✗
AVSD	Charades	Domestic sounds	8.5K	30s	✓	✓	✓	✗	✗	✗	✗
Pano-AVQA	Online	Visual object sound	5.4k	5s	✓	✓	✓	✓	✗	✗	✗
MUSIC-AVQA	YouTube	Visual object sound	9.3K	60s	✓	✓	✓	✓	✓	✓	✓

- 提供了涵盖音频、视觉和视听问题的Q-A pair，比其他数据集更全面。
- 由音乐表演场景组成，包含丰富的视听成分 (高质量的A-V交互，减少噪声（例如背景音乐）问题）。

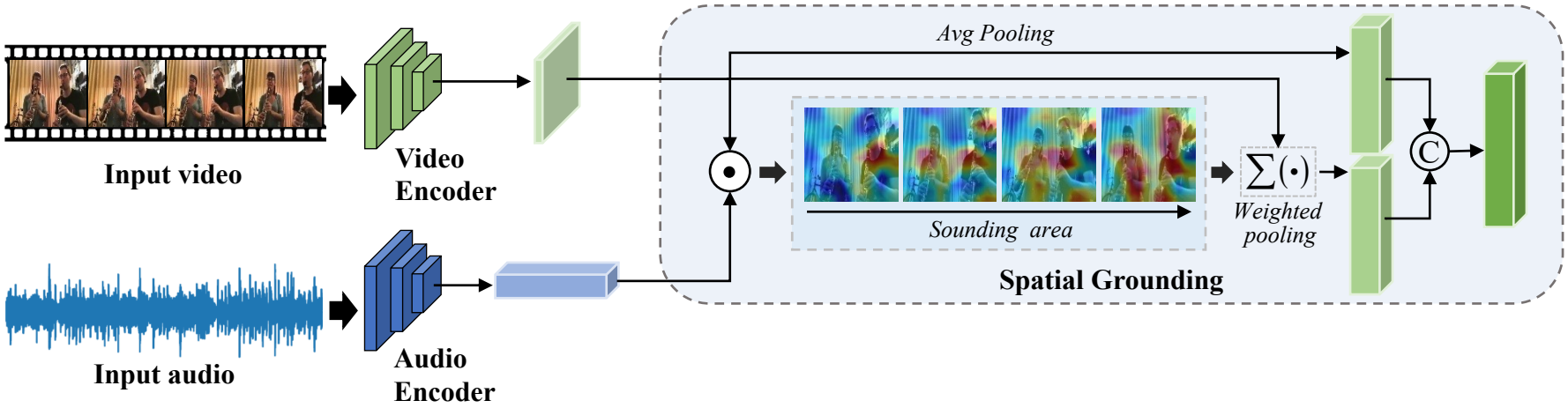
视音时空问答模型



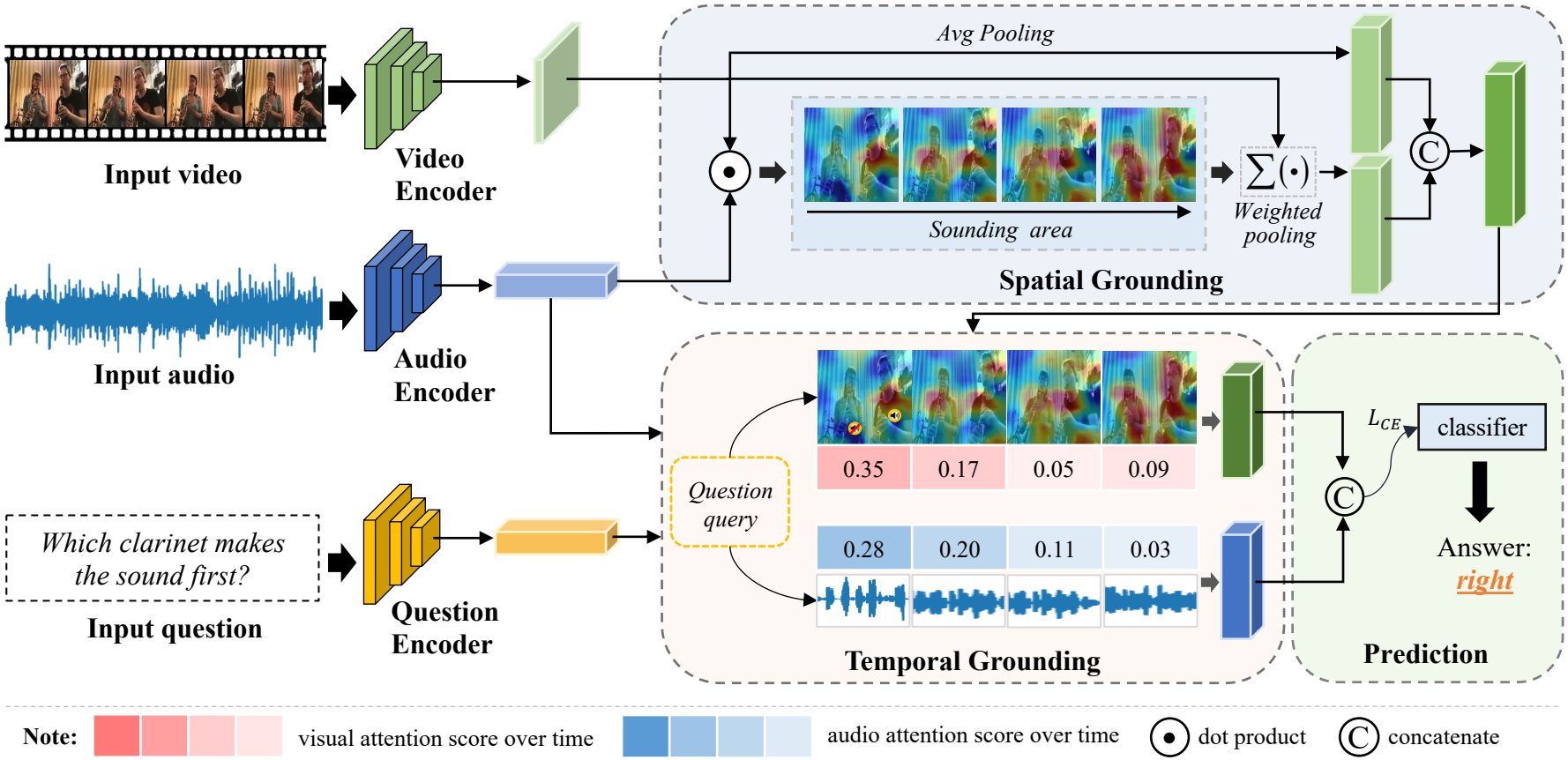
Spatial grounding 模块 → 声源定位

Temporal grounding 模块 → 捕捉时序关键片段

视音时空问答模型


Input video
Video Encoder
Input audio
Audio Encoder
Avg Pooling
Sounding area
Weighted pooling
Spatial Grounding

视音时空问答模型



Spatial grounding 模块 → 声源定位

Temporal grounding 模块 → 捕捉时序关键片段

实验

■ 结果与分析

- 多种模态结合促进QA任务
- 时空推理

Method	A Question	V Question	A-V Question	All
Q	65.19	44.42	55.15	54.09
A+Q	67.78	62.75	63.86	64.26
V+Q	68.76	67.28	63.23	65.28
AV+Q	70.67	69.72	65.84	67.72
AV+Q+TG	73.01	73.18	68.02	70.27
AV+Q+TG+SG	74.06	74.00	69.54	71.52

* TG: Temporal Grounding; SG: Spatial Grounding.

Ablation study on input modalities and the proposed modules.

多模态感知和有效的时空推理能够促进AVQA任务。

实验

■ 和其他QA方法的比较

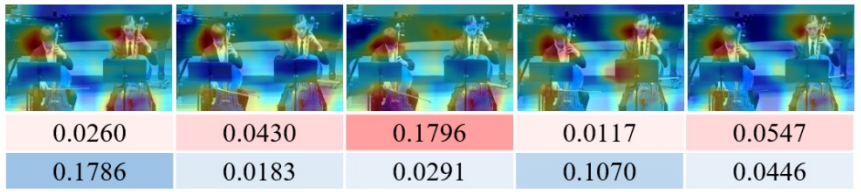
Task	Method	Audio Question			Visual Question			Audio-Visual Question						All
		Counting	Comparative	Avg.	Counting	Location	Avg.	Existential	Location	Counting	Comparative	Temporal	Avg.	Avg.
AudioQA	FCNLSTM	70.45	66.22	68.88	63.89	46.74	55.21	82.01	46.28	59.34	62.15	47.33	60.06	60.34
	CONVLSTM	74.07	**68.89**	72.15	67.47	54.56	60.94	**82.91**	50.81	63.03	60.27	51.58	62.24	63.65
VisualQA	GRU	72.21	66.89	70.24	67.72	70.11	68.93	81.71	59.44	62.64	61.88	60.07	65.18	67.07
	BiLSTM Attn	70.35	47.92	62.05	64.64	64.33	64.48	78.39	45.85	56.91	53.09	49.76	57.10	59.92
	HCAttn	70.25	54.91	64.57	64.05	66.37	65.22	79.10	49.51	59.97	55.25	56.43	60.19	62.30
	MCAN	77.50	55.24	69.25	71.56	70.93	71.24	80.40	54.48	64.91	57.22	47.57	61.58	65.49
VideoQA	PSAC	75.64	66.06	72.09	68.64	69.79	69.22	77.59	55.02	63.42	61.17	59.47	63.52	66.54
	HME	74.76	63.56	70.61	67.97	69.46	68.76	80.30	53.18	63.19	62..69	59.83	64.05	66.45
	HCRN	68.59	50.92	62.05	64.39	61.81	63.08	54.47	41.53	53.38	52.11	47.69	50.26	55.73
AVQA	AVSD	72.41	61.90	68.52	67.39	74.19	70.83	81.61	58.79	63.89	61.52	61.41	65.49	67.44
	Pano-AVQA	74.36	64.56	70.73	69.39	75.65	72.56	81.21	59.33	64.91	64.22	63.23	66.64	68.93
	Our method	**78.18**	67.05	**74.06**	**71.56**	**76.38**	**74.00**	81.81	**64.51**	**70.80**	**66.01**	**63.23**	**69.54**	**71.52**

AVQA results of different methods on the test set of MUSIC-AVQA. The top-2 results are highlighted.

- 我们的方法在绝大多数问题类型上都取得了显著的提升，
  尤其是在***Counting*** 和 ***Location*** 问题上。

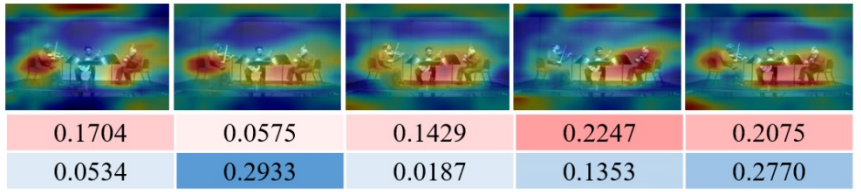
实验

■ 结果可视化



0.0260	0.0430	0.1796	0.0117	0.0547
0.1786	0.0183	0.0291	0.1070	0.0446

(a) Q: How many sounding cello in the video? **A:** two ✓



0.1704	0.0575	0.1429	0.2247	0.2075
0.0534	0.2933	0.0187	0.1353	0.2770

(b) Q: How many types of musical instruments sound in the video? **A:** two ✓

0.0056	0.2831	0.0538	0.1243	0.3722
0.0083	0.0638	0.1110	0.3016	0.0433

(c) Q: Where is the last sounding instrument? **A:** right ✓

0.1223	0.1086	0.0855	0.1293	0.1439
0.1609	0.0807	0.0237	0.0601	0.0333

(d) Q: Where is the first sounding instrument? **A:** left ✓

0.1562	0.1392	0.1148	0.0613	0.1748
0.2185	0.2693	0.0049	0.0411	0.0429

(e) Q: Is the first sound coming from the right instrument? **A:** yes ✓

0.2406	0.0239	0.0229	0.0024	0.0402
0.0676	0.1469	0.0120	0.0145	0.0359

(f) Q: What is the left instrument of the first sounding instrument? **A:** erhu ✗

Note: visual attention score over time audio attention score over time

Part 02 小结

- 介绍了AVQA任务，其目的是回答有关视频中不同视觉对象、声音及其关联的问题。
- 建立了一个大规模的MUSIC-AVQA数据集，其中包括45,867个Q-A pairs，涵盖了视音模态和多种不同的问题类型。
- 提出了视音时空问答模型解决细粒度的场景理解和时空推理。

问题：不同模态是如何有效关联的？以及不同模态如何协同感知的？

多模态学习机制研究

■ 通过动态梯度调控缓解不平衡的多模态联合训练

Balanced Multimodal Learning via On-the-fly Gradient Modulation，

CVPR 2022 ORAL

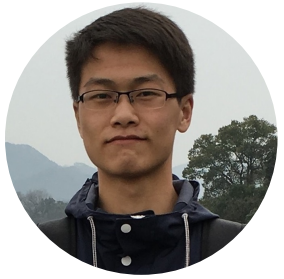
由中国人民大学高瓴人工智能学院GeWu实验室主导

GitHub Repo

Xiaokang Peng†
Renmin University of China

Yake Wei†
Renmin University of China

Andong Deng
Shanghai Jiaotong University



Dong Wang
Shanghai Artificial Intelligence Laboratory

Di Hu*
Renmin University of China

视音场景分析

■ 视音场景理解和时空推理

Learning to Answer Questions in Dynamic Audio-Visual Scenarios，

CVPR 2022 ORAL

由中国人民大学高瓴人工智能学院GeWu实验室主导

项目主页

Guangyao Li†
Renmin University of China

Yake Wei†
Renmin University of China

Yapeng Tian
University of Rochester

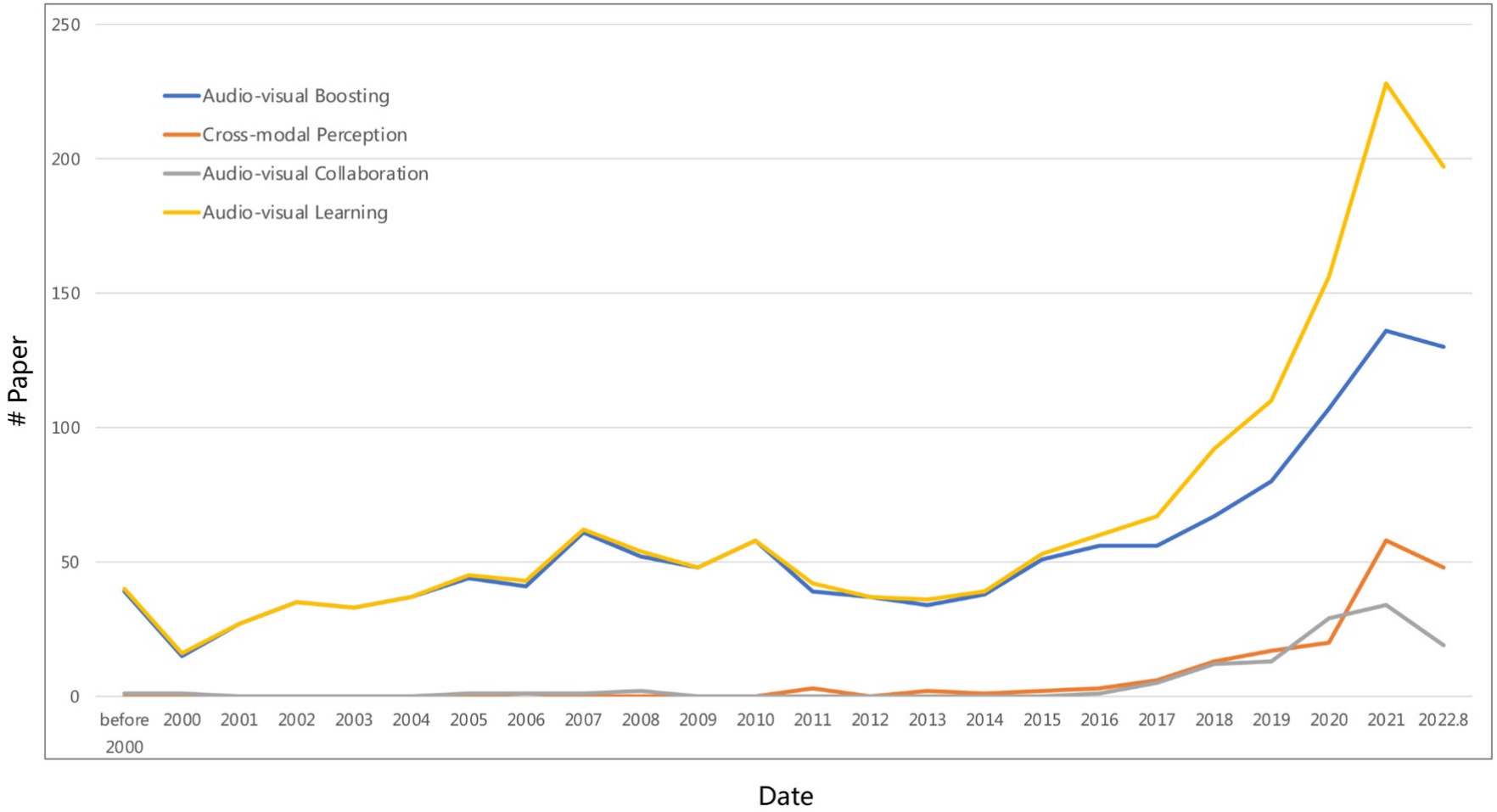
Chenliang Xu
University of Rochester

Ji-Rong Wen
Renmin University of China

Di Hu*
Renmin University of China

新问题：面向这种自然模态数据的多模态研究现在是什么进展呢？大家都在研究什么？

一张图


250
200
150
100
50
0
Paper
Audio-visual Boosting
Cross-modal Perception
Audio-visual Collaboration
Audio-visual Learning
before 2000
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022.8
Date

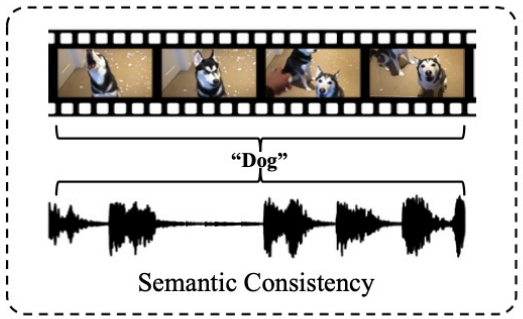
Learning in Audio-visual Context: A Review, Analysis, and New Perspective

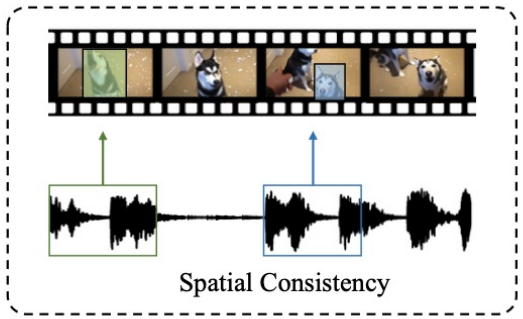
Yake Wei, Di Hu, Yapeng Tian, Xuelong Li, *Fellow, IEEE*

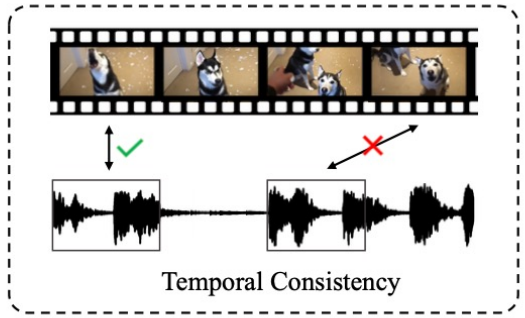
Abstract—Sight and hearing are two senses that play a vital role in human communication and scene understanding. To mimic human perception ability, audio-visual learning, aimed at developing computational approaches to learn from both audio and visual modalities, has been a flourishing field in recent years. A comprehensive survey that can systematically organize and analyze studies of the audio-visual field is expected. Starting from the analysis of audio-visual cognition foundations, we introduce several key findings that have inspired our computational studies. Then, we systematically review the recent audio-visual learning studies and divide them into three categories: audio-visual boosting, cross-modal perception and audio-visual collaboration. Through our analysis, we discover that, the consistency of audio-visual data across semantic, spatial and temporal support the above studies. To revisit the current development of the audio-visual learning field from a more macro view, we further propose a new perspective about audio-visual scene understanding, then discuss and analyze the feasible future direction of the audio-visual learning area. Overall, this survey reviews and outlooks the current audio-visual learning field from different aspects. We hope it can provide researchers with a better understanding of this area.

Index Terms—Sight, Hearing, Audio-visual learning, Audio-visual boosting, Cross-modal perception, Audio-visual collaboration, Survey

一种归类与一次探究







Audio-visual Boosting

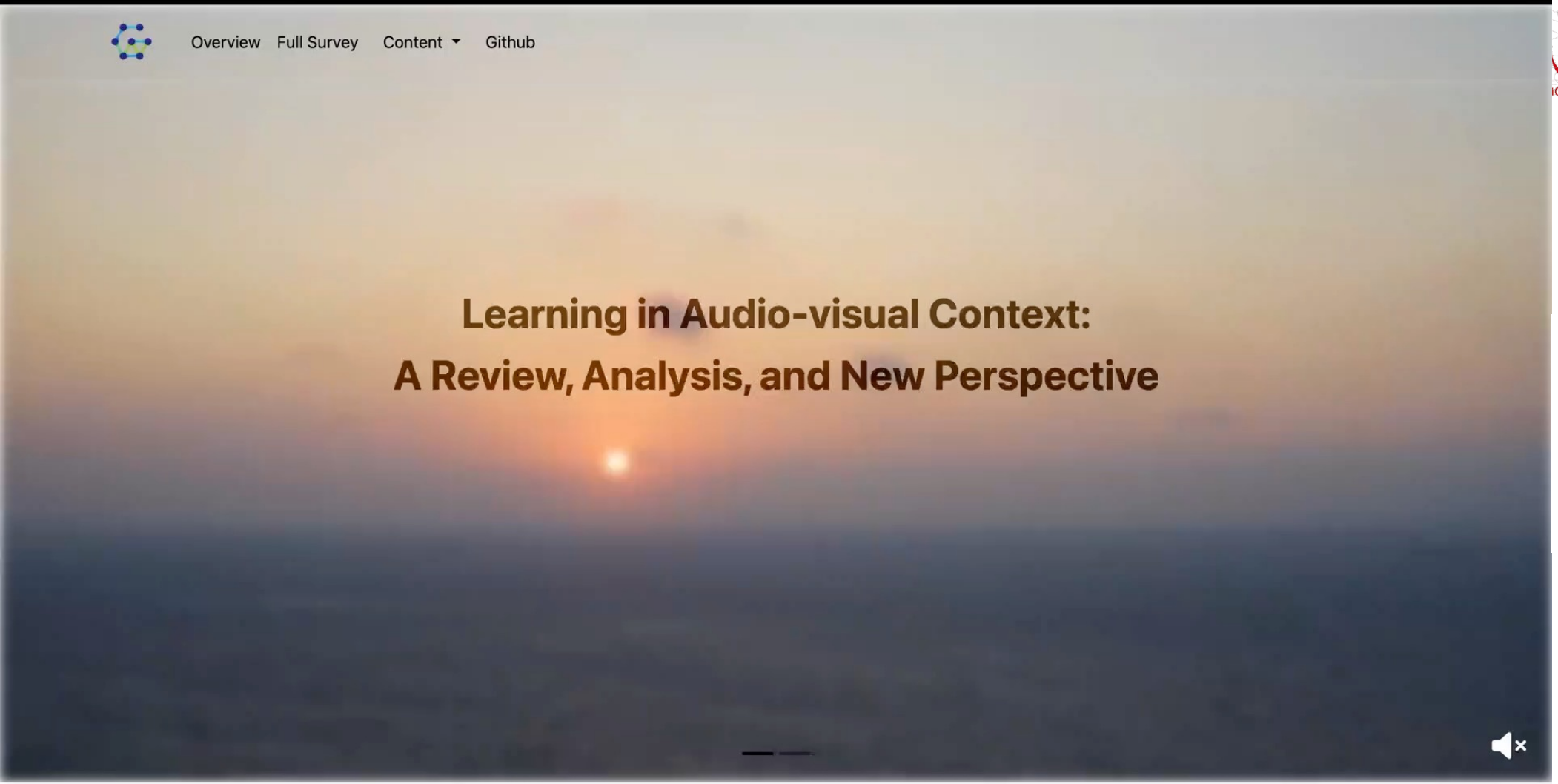
- Audio-visual Recognition
  - Speech Recognition
  - Speaker Recognition
  - Action Recognition
  - Emotion Recognition
- Uni-modal Enhancement
  - Speech Enhancement/Separation
  - Object Sound Separation
  - Face Super-resolution/Reconstruction

Cross-modal Perception

- Audio-visual Generation
  - Mono Sound Generation
  - Spatial Sound Generation
  - Video Generation
  - Depth Estimation
- Audio-visual Transfer Learning
- Cross-modal Retrieval

Audio-visual Collaboration

- Audio-visual Representation Learning
  - Audio-visual Consistency
  - Deep Clustering
  - Pre-training Model
- Audio-visual Localization
  - Sound Localization in Videos
  - Audio-visual Saliency Detection
  - Audio-visual Navigation
- Audio-visual Event Localization/Parsing
- Audio-visual Question Answering/Dialog



Sunrise field

Learning in Audio-visual Context: A Review, Analysis, and New Perspective

Yake Wei
Renmin University of China

Di Hu*
Renmin University of China

Yapeng Tian
University of Texas at Dallas

Xuelong Li
Northwestern Polytechnical University

* Indicates Corresponding Author

中國人民大學
RENMIN UNIVERSITY OF CHINA

高瓴人工智能学院
Gaoling School of Artificial Intelligence

GeWu-Lab
Gaoling School of Artificial Intelligence
Renmin University of China

Thank You for listening!

Di Hu (胡迪)

Email: dihu@ruc.edu.cn

2022-08-23