AV-SUPERB: Audio-visual Representations and How to Evaluate Them

Shang-Wen (Daniel) Li
FAIR
shangwel@meta.com

Yuan (Roger) Tseng
National Taiwan University
r11942082@ntu.edu.tw

Outline:

1. Why audio-visual representations & notable recent works
2. The AV-SUPERB benchmark (ICASSP 2024)
3. Some noteworthy findings
4. What’s next?

Computational systems should be able to perceive the world like we do, by combining auditory and visual inputs

- For audio event classification,

  audio events usually co-occur with visual actions

Computational systems should be able to perceive the world like we do, by combining auditory and visual inputs

- For audio event classification,

  audio events usually co-occur with visual actions

- For speech recognition,

  visual input provides more context to what is being said

Computational systems should be able to perceive the world like we do, by combining auditory and visual inputs

- For audio event classification,

  audio events usually co-occur with visual actions

- For speech recognition,

  visual input provides more context to what is being said

```
HYPOTHESIS: and then cut this piece on top of your shirt
```

ref: Gabeur et al., INTERSPEECH 2022,
AVATAR: Unconstrained Audiovisual Speech Recognition

Computational systems should be able to perceive the world like we do, by combining auditory and visual inputs

- For audio event classification,

  audio events usually co-occur with visual actions

- For speech recognition,

  visual input provides more context to what is being said

`HYPOTHESIS: and then cut the sleeves on top of your shirt`

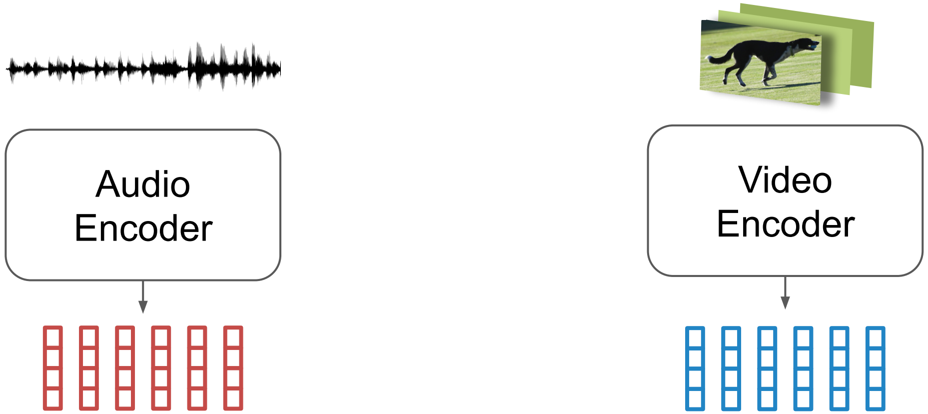
ref: Gabeur et al., INTERSPEECH 2022,
AVATAR: Unconstrained Audiovisual Speech Recognition

Computational systems should be able to perceive the world like we do, by combining auditory and visual inputs

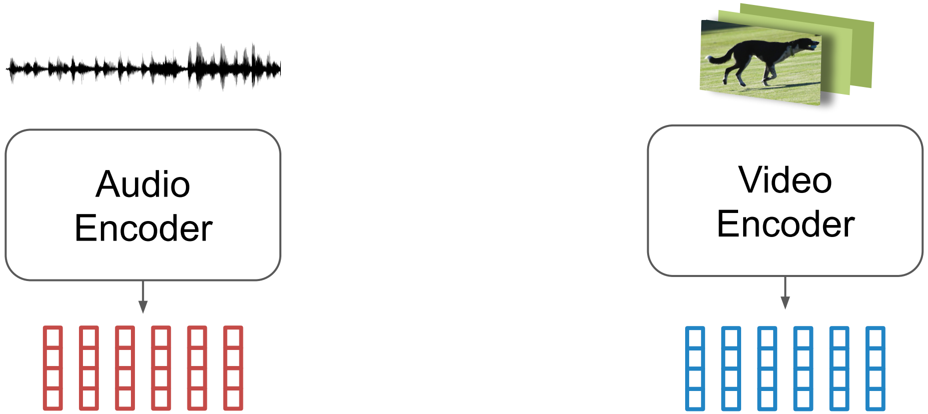
- For audio event classification,
  audio events usually co-occur with visual actions
- For speech recognition,
  visual input provides more context to what is being said

… which is why **audio-visual representation learning** is meaningful.

Contrastive Audio-visual Learning



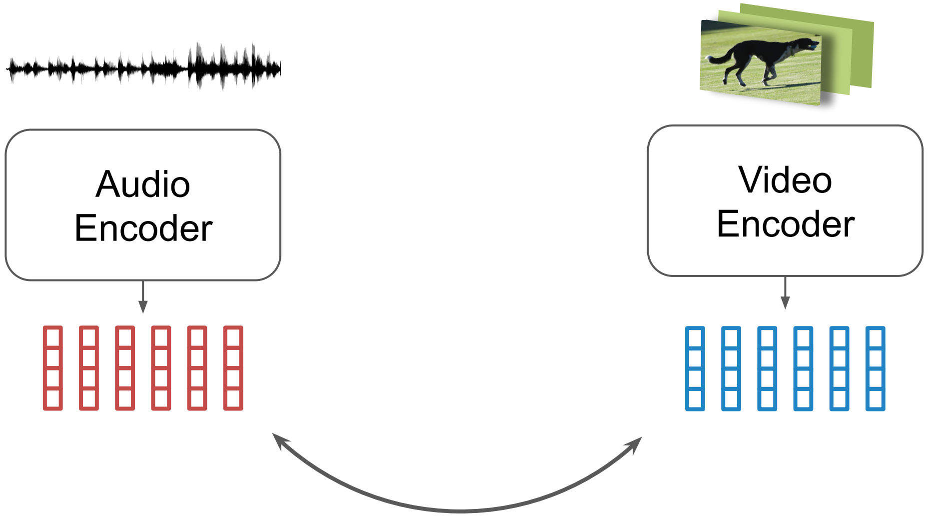
Contrastive Audio-visual Learning



continuous representation sequence (size T x D)

https://commons.wikimedia.org/wiki/File:Dog_Running_002a.jpg

Contrastive Audio-visual Learning

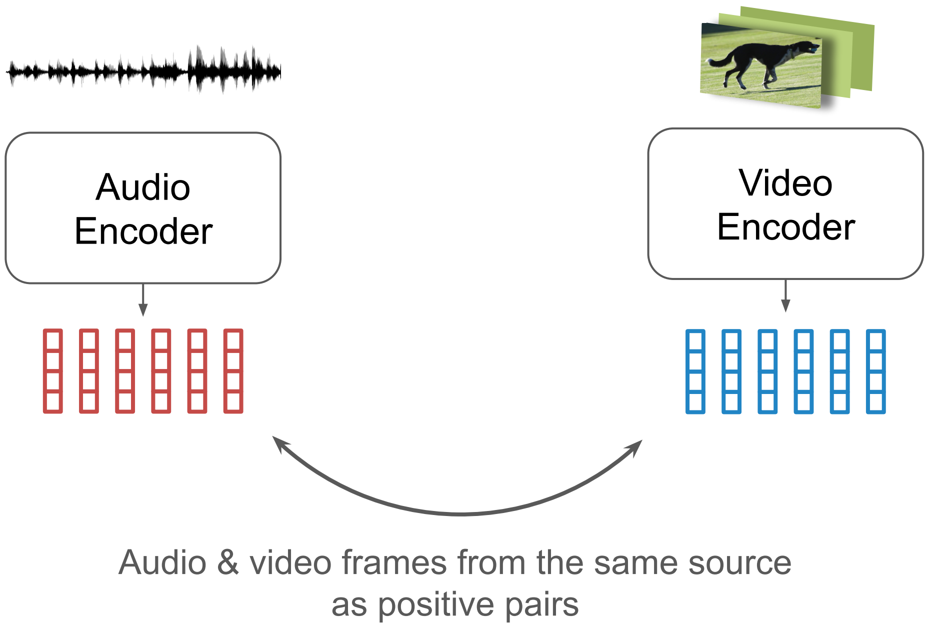


continuous representation sequence (size T x D)

Audio & video frames from the same source
as positive pairs

https://commons.wikimedia.org/wiki/File:Dog_Running_002a.jpg

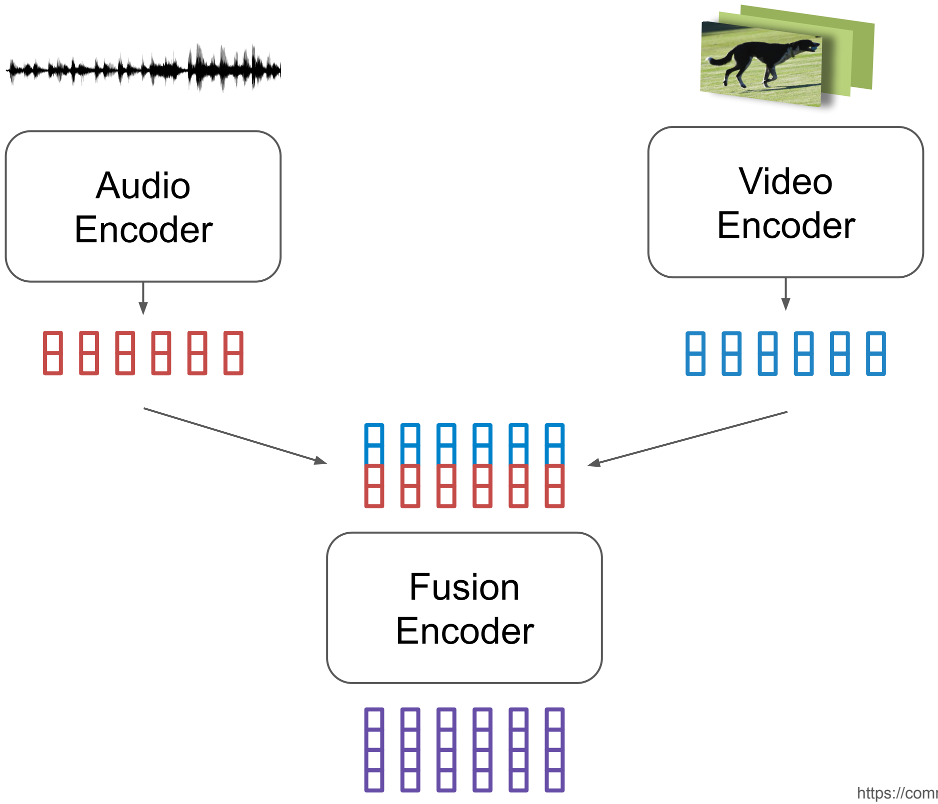
Contrastive Audio-visual Learning



notable works include: AVID-CMA and GDT in action recognition, and VisualVoice in speech separation

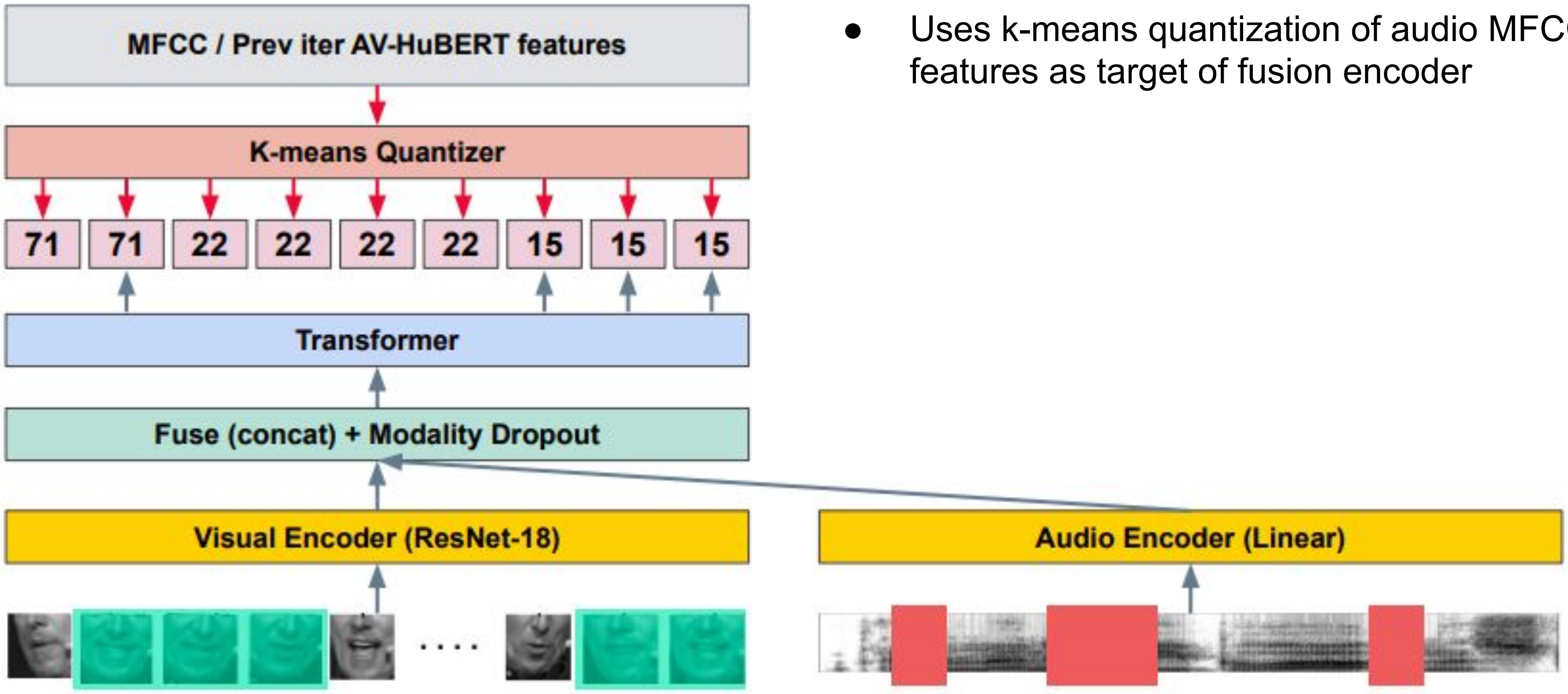
https://commons.wikimedia.org/wiki/File:Dog_Running_002a.jpg

Audio-visual Fusion



https://commons.wikimedia.org/wiki/File:Dog_Running_002a.jpg

AV-HuBERT

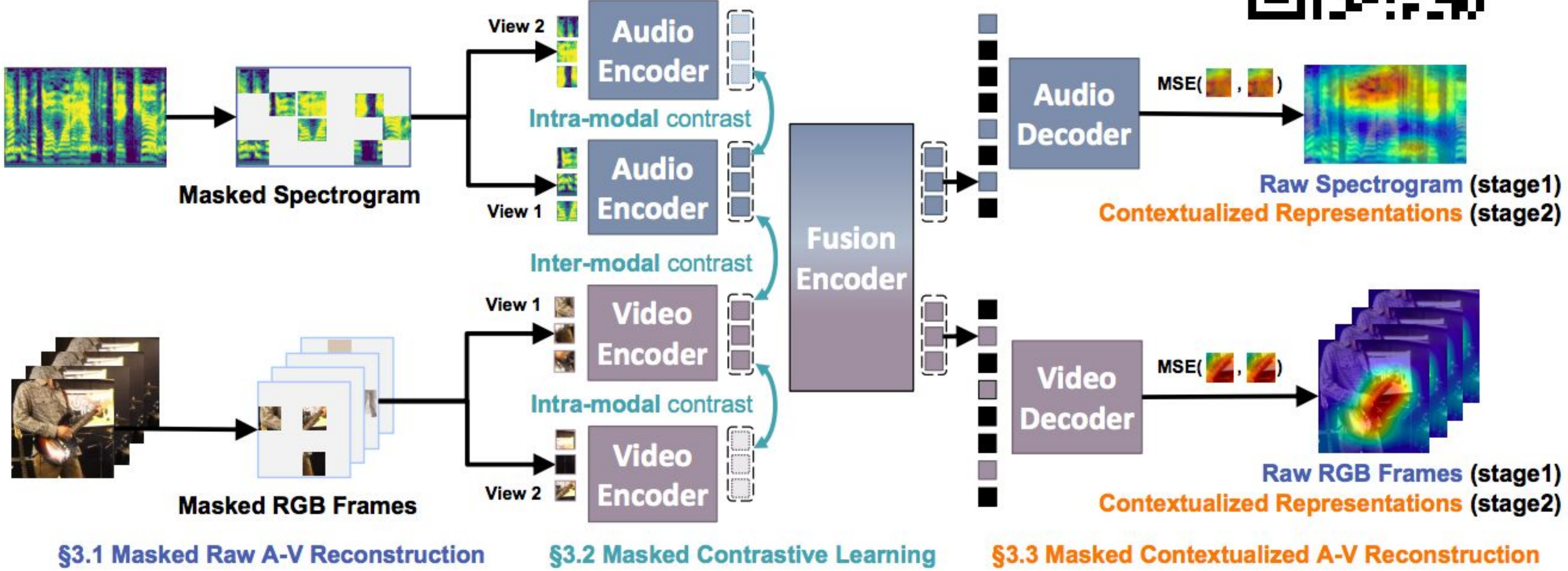


- Uses k-means quantization of audio MFCC features as target of fusion encoder

credit: Wei-Ning Hsu, https://aaai-sas-2022.github.io/static/media/Weining_Hsu_aaai2022_talk.7ac27a4c.pdf

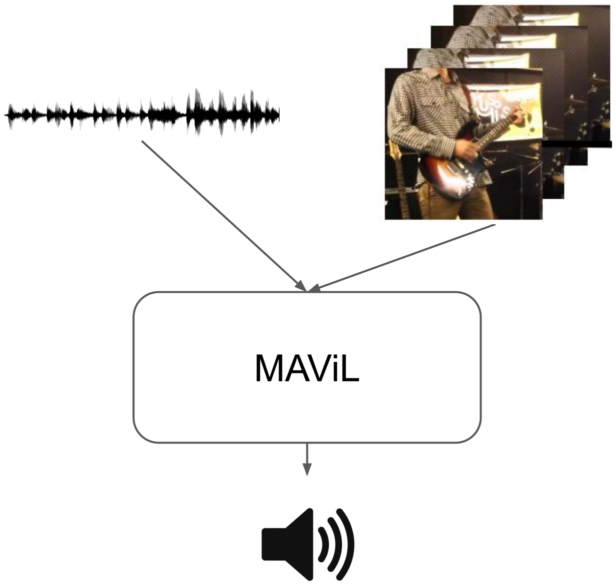
MAViL (NeurIPS 2023)

Paper link:

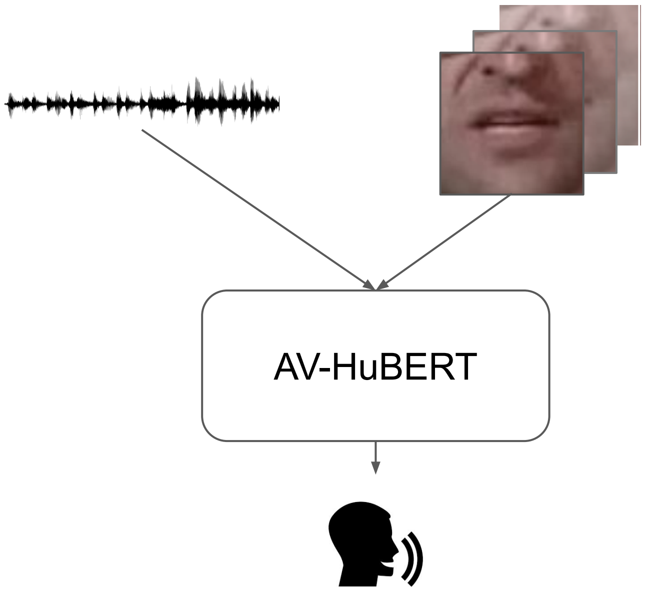


- new state-of-the-art on audio event classification & audio-to-video retrieval

Existing audio-visual models are designed for different tasks



Audio Event:
Electric Guitar



Speech Recognition:
“We already use artificial intelligence…”

References:
[1] MAVIL: https://arxiv.org/abs/2212.08071
[2] AV-HuBERT: https://arxiv.org/abs/2201.02184

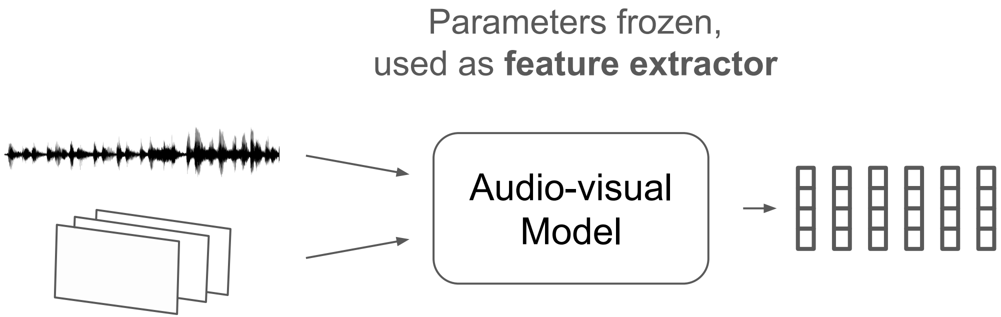
We can do all these tasks with one system:
Our brains!

→ How far are we from a model
that can similarly generalize?

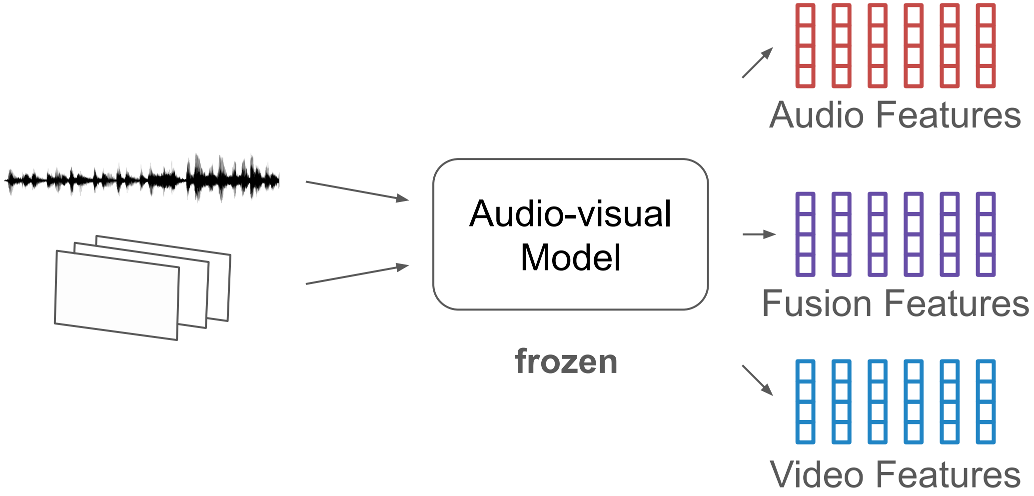
The AV-SUPERB benchmark:

Evaluation Protocol

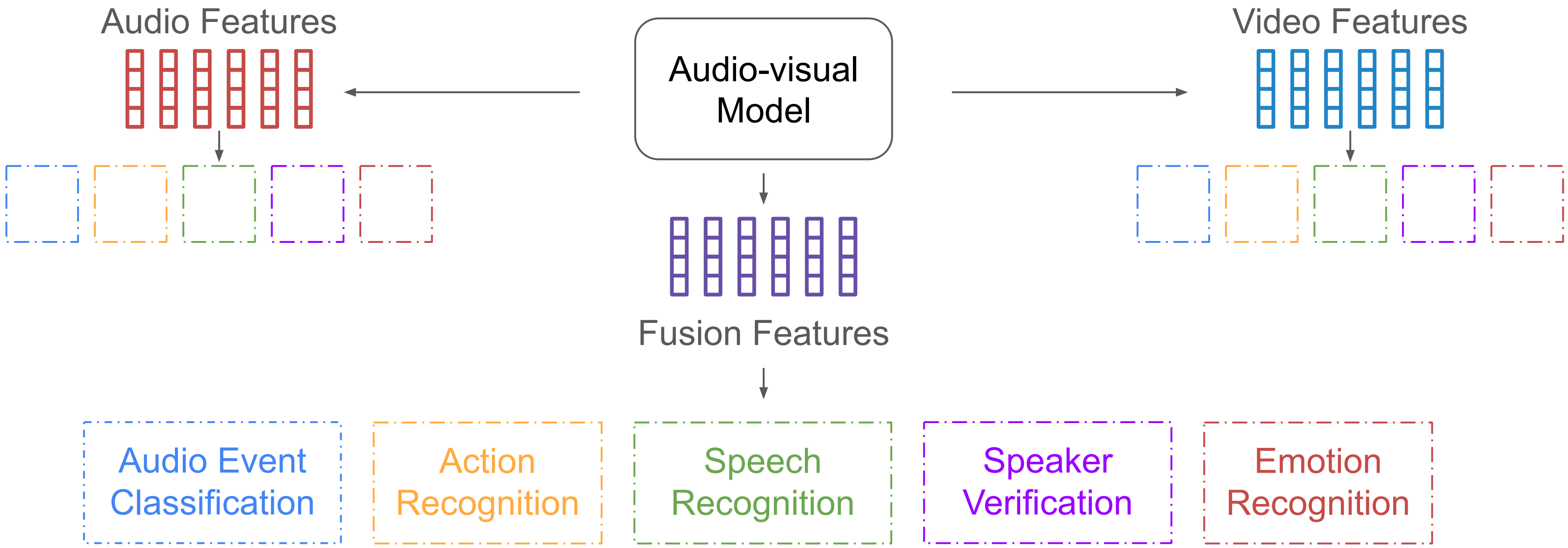
View representations as the output of feature extractors:



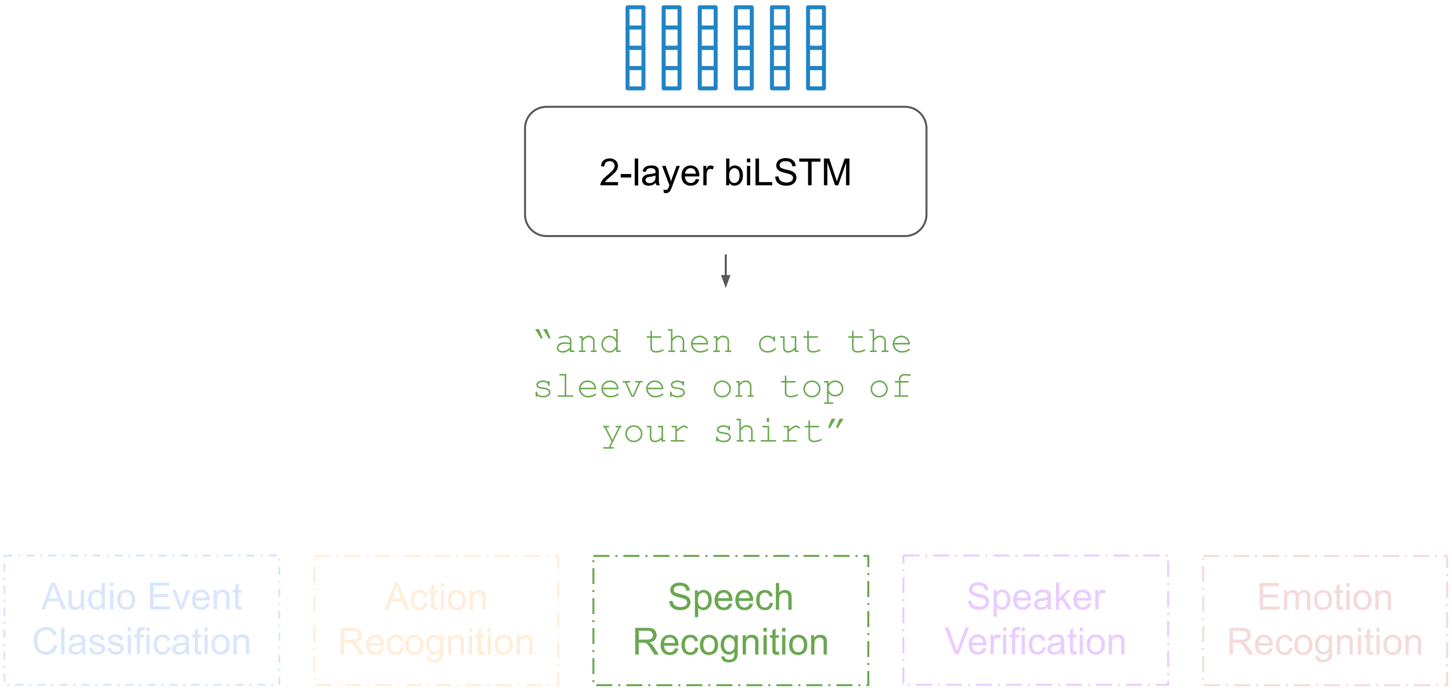
View representations as the output of feature extractors:



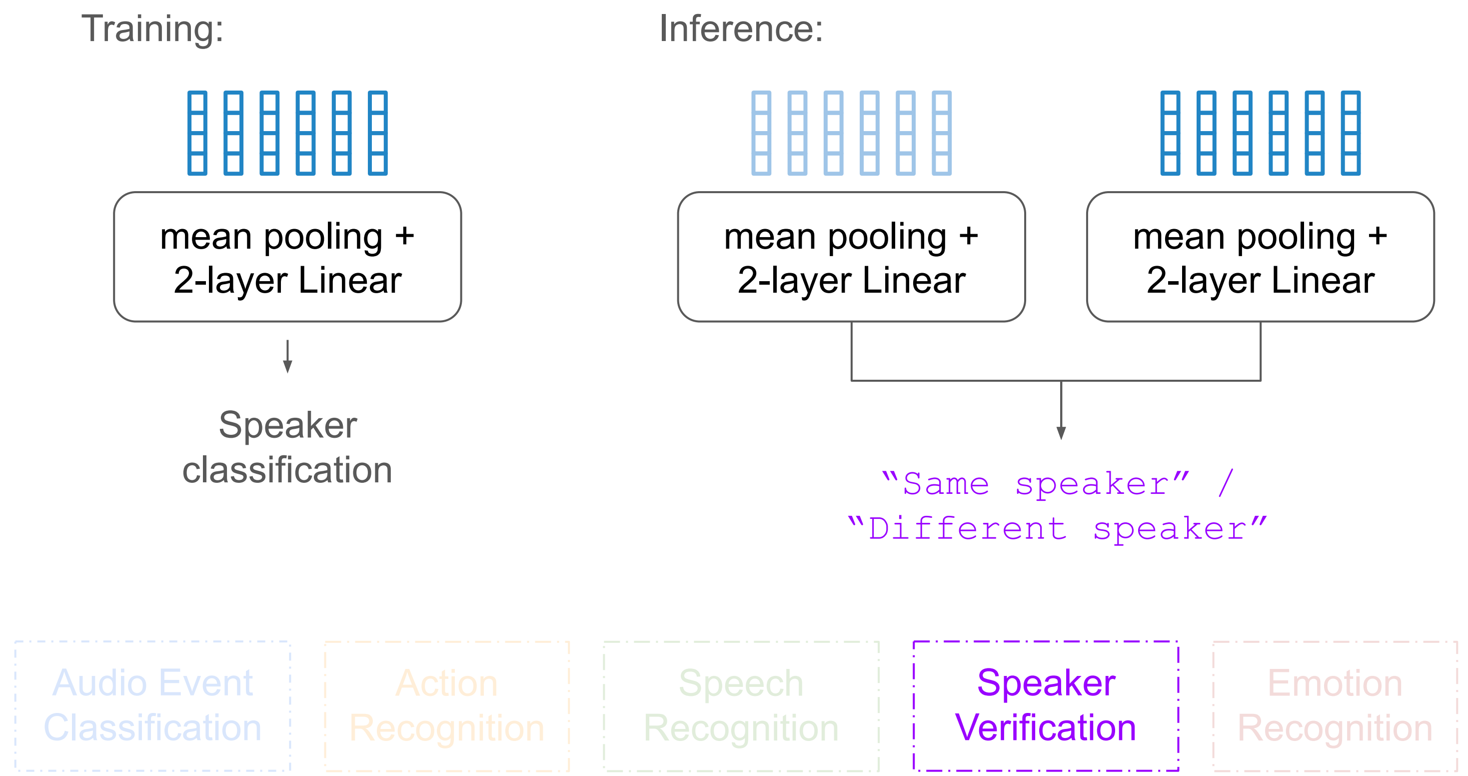
Each type of feature is evaluated on five tasks:



Train a small prediction head for each task:



Train a small prediction head for each task:



The AV-SUPERB benchmark:

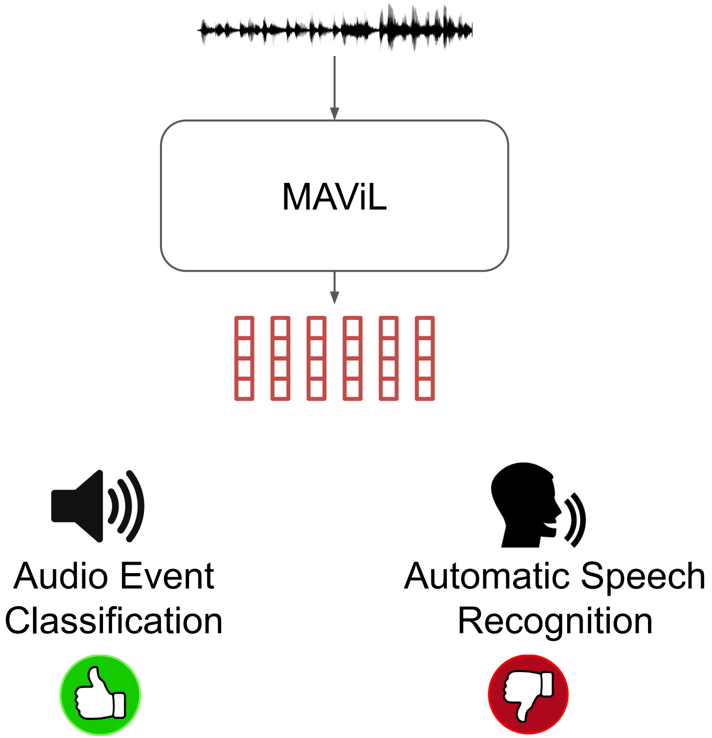
Some Noteworthy Findings

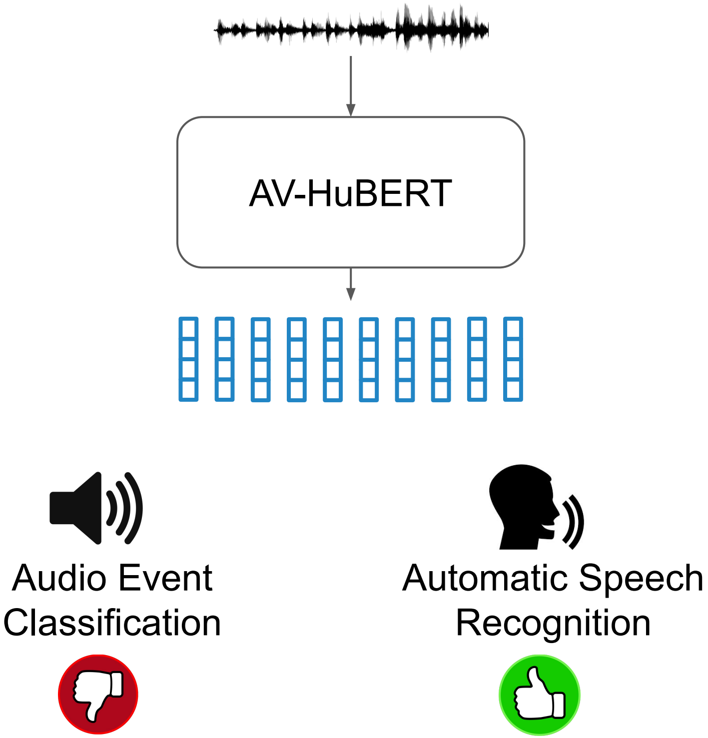
1. Existing pretrained SSL models do not generalize to all tasks

Representation Type	Params.	Overall Score	Audio-Visual				Speech-Visual		
			AEC		AR		ASR	ASV	ER
			AS-20K (mAP ↑)	VGGSound (Acc. ↑)	Kinetics-Sounds (Acc. ↑)	UCF101 (Acc. ↑)	LRS3-TED (CER ↓)	VoxCeleb2 (EER ↓)	IEMOCAP (Acc. ↑)
Audio-only									
FBANK	0	36.69	2.8	5.12	24.73	19.91	20.07	27.16	51.52
HuBERT	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	15.58	**62.14**
AV-HuBERT*	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	**14.45**	58.54
RepLAI	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	32.58	57.53
AVBERT	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	23.74	60.94
MAViL	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	20.71	59.46
Video-only									
HoG	0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	36.32	35.83
AV-HuBERT*	103M	33.48	2.4	5.90	24.73	37.55	**50.91**	**11.90**	26.59
RepLAI	15M	36.40	5.5	13.5	46.68	56.69	71.33	36.95	40.72
AVBERT	37M	47.69	11.5	28.73	62.67	77.42	72.29	20.00	**45.8**
MAViL	87M	49.70	**18.0**	**32.08**	**74.01**	**79.37**	74.03	24.58	43.03
Audio-visual fusion									
AV-HuBERT	103M	53.42	13.3	32.69	52.23	41.46	**2.75**	**9.46**	46.45
AVBERT	43M	54.85	22.9	44.54	71.31	71.76	70.12	18.31	**61.87**
MAViL	187M	62.36	**26.7**	**47.22**	**79.51**	**77.98**	30.18	19.67	54.94

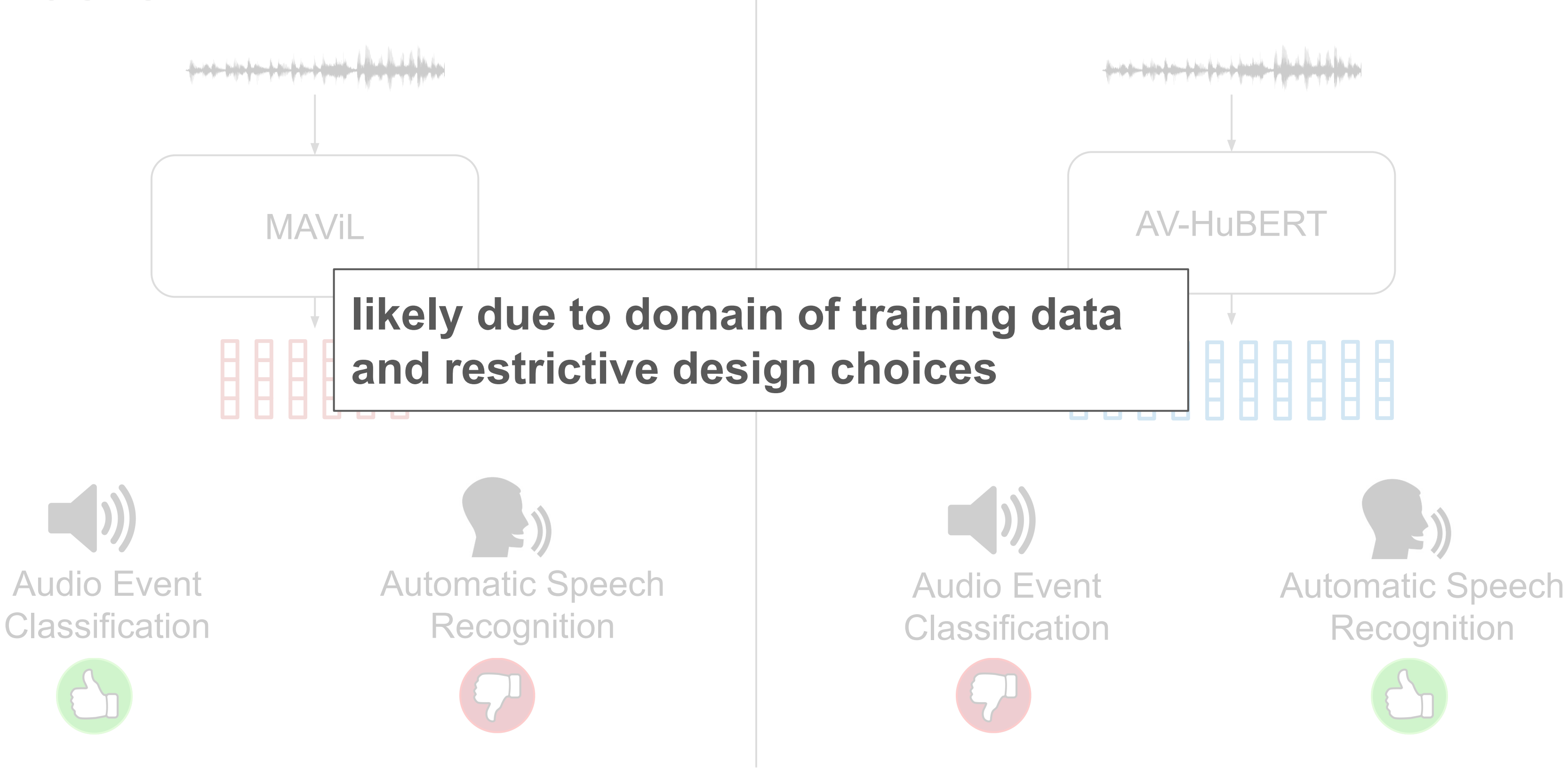
Evaluate 5 pretrained SSL models from different domains, with handcrafted features as baselines

1. Existing pretrained SSL models do not generalize to all tasks



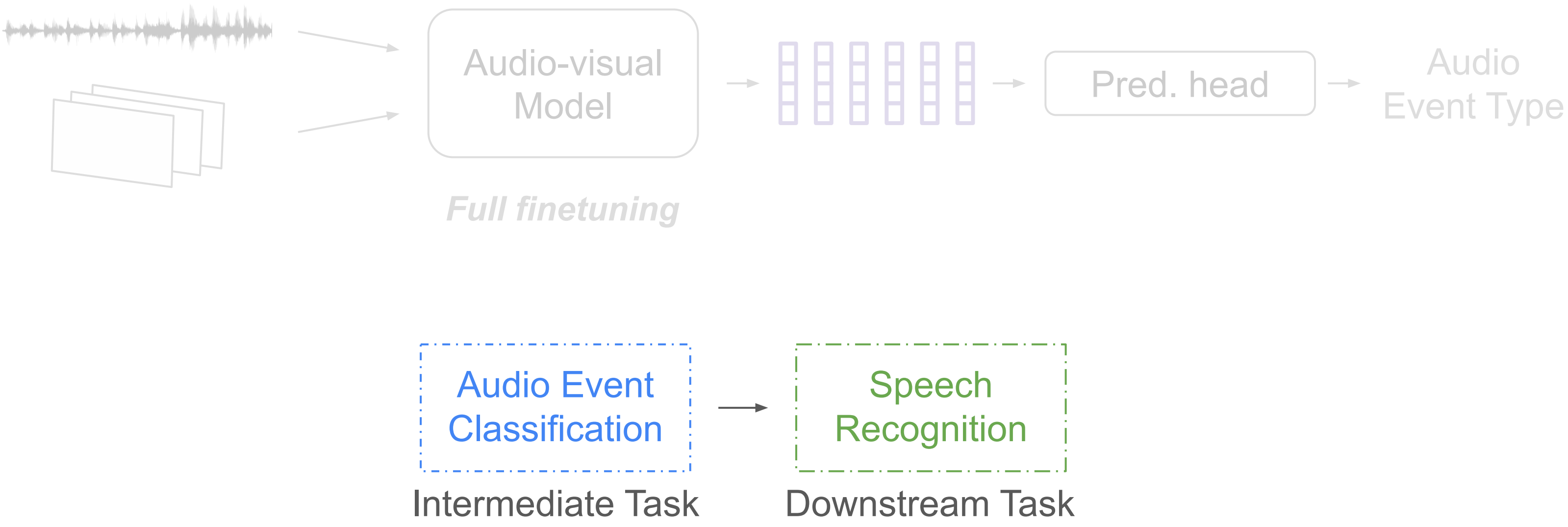


1. Existing pretrained SSL models do not generalize to all tasks

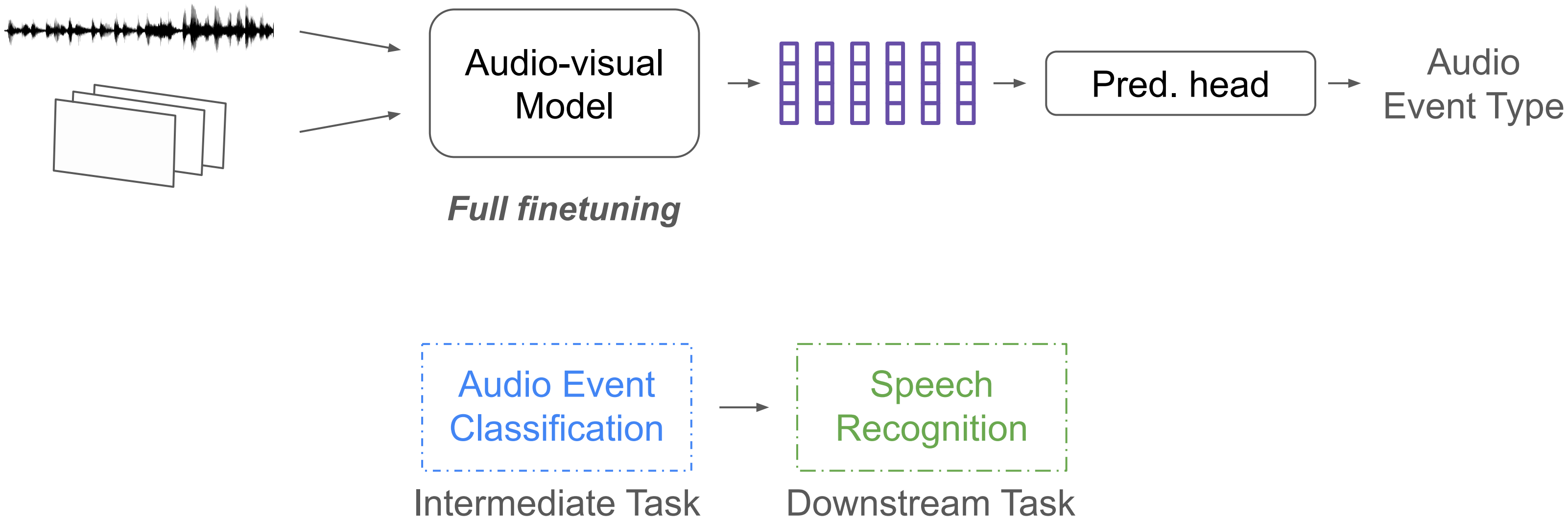


likely due to domain of training data and restrictive design choices

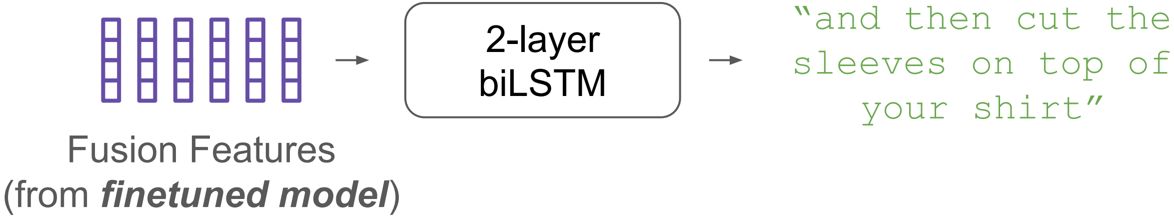
2. Intermediate-task finetuning can help but does not completely solve the problem

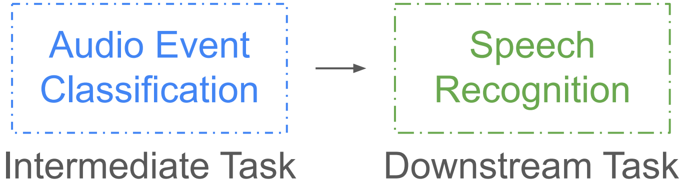


2. Intermediate-task finetuning can help but does not completely solve the problem



2. Intermediate-task finetuning can help but does not completely solve the problem



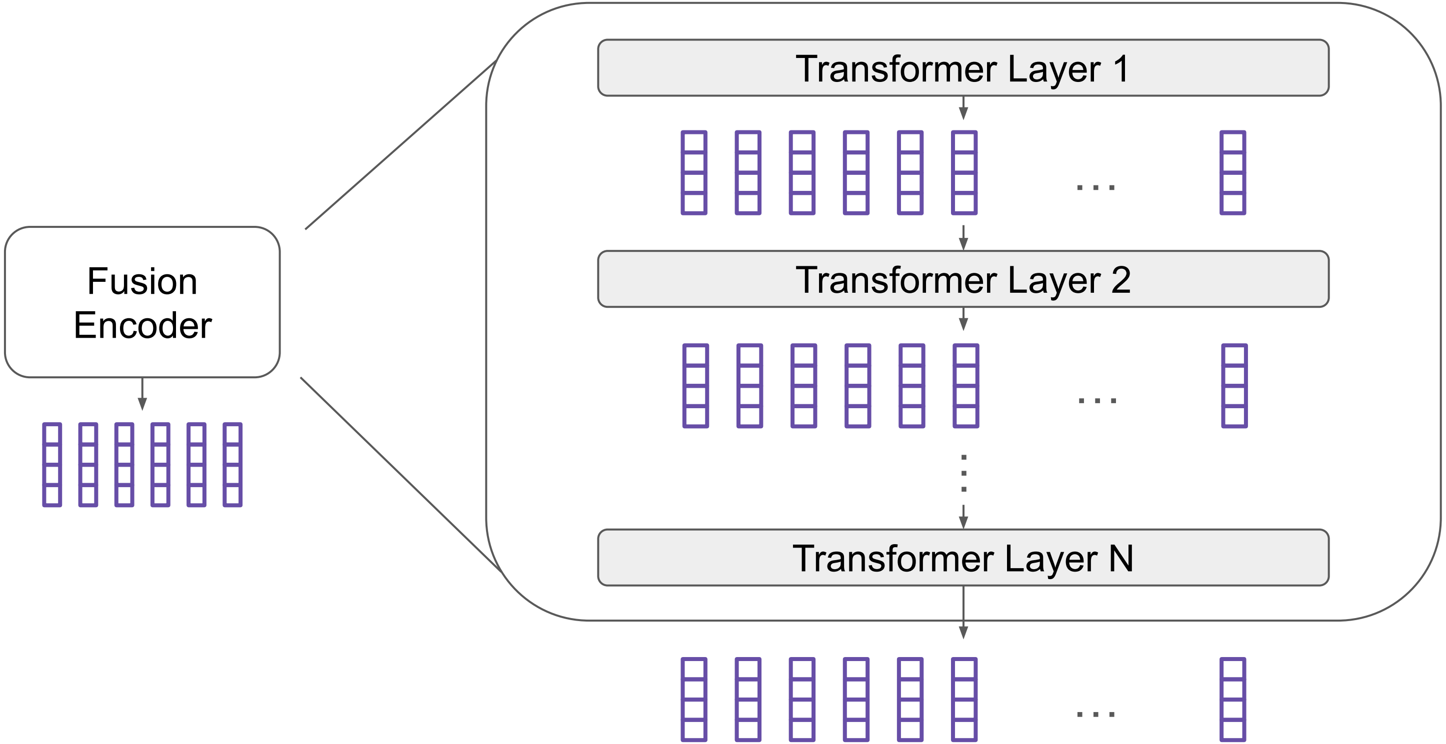


2. Intermediate-task finetuning can help but does not completely solve the problem

	Intermediate Task Fine-tuning Data	Audio-Visual				Speech-Visual		
		AEC		AR		ASR	ASV	ER
		AS-20K (mAP ↑)	VGGSound (Acc. ↑)	Kinetics-Sounds (Acc. ↑)	UCF101 (Acc. ↑)	LRS3-TED (CER ↓)	VoxCeleb2 (EER ↓)	IEMOCAP (Acc. ↑)
MAViL								
Audio	AudioSet-2M	28.3(+6.7)	44.79(+4.89)	62.93(+5.65)	50.10(+4.42)	23.99(+0.44)	21.77(-1.06)	58.17(-1.29)
Video		20.9(+2.9)	36.68(+4.58)	77.39(+3.38)	86.93(+7.56)	78.59(-4.56)	23.93(+0.65)	39.15(-3.88)
Fusion		39.1(+12.4)	55.94(+8.72)	84.93(+5.42)	88.07(+10.09)	30.65(-0.47)	18.61(+1.06)	46.35(-8.59)

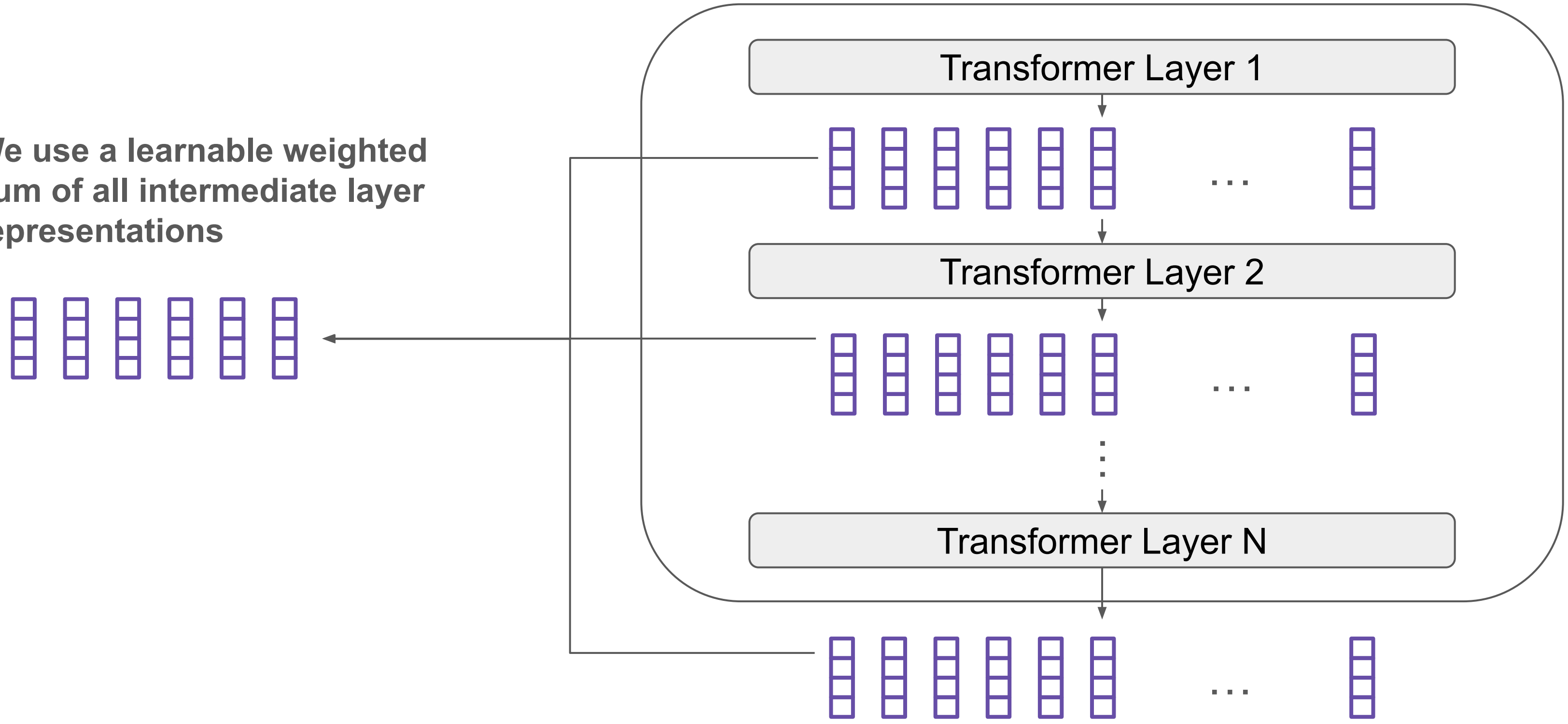
Ref: [our preprint]()

3. Representations from the last layer may be suboptimal



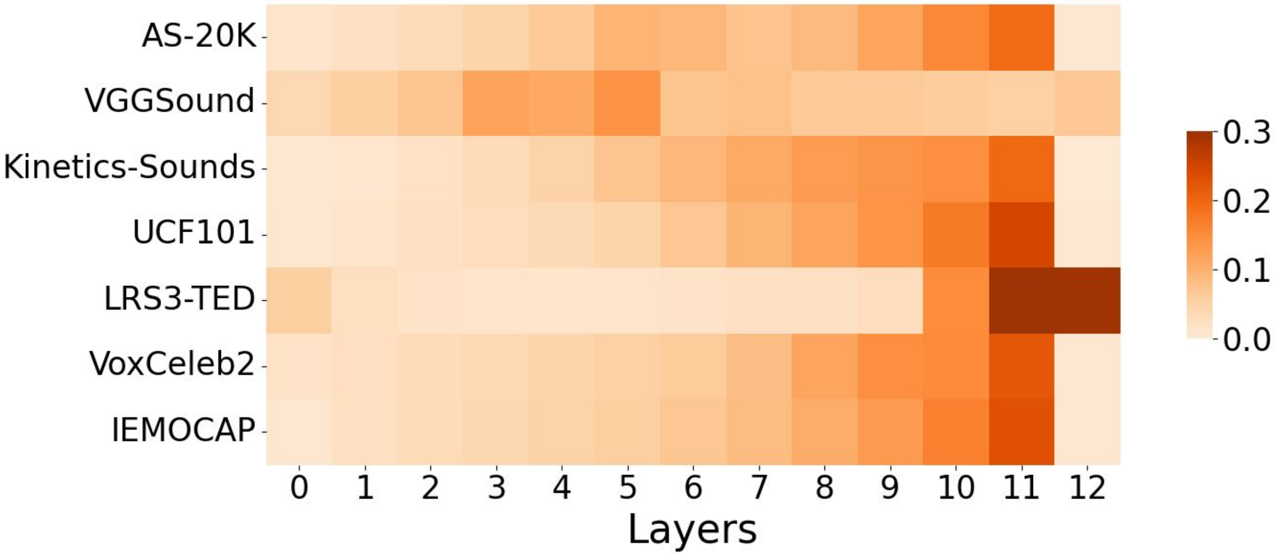
3. Representations from the last layer may be suboptimal

We use a learnable weighted sum of all intermediate layer representations



3. Representations from the last layer may be suboptimal

For AV-HuBERT fusion features, the peniultimate layer often contributes more



Heatmap of learned weights for each downstream dataset

What’s next?

- Including more realistic, useful, or fundamental tasks, such as retrieval, localisation, etc.
- Fairer comparision of models by unifying their training data, objective functions, or model architectures.

Check out our preprint on arXiv:

- Accepted to ICASSP 2024!
- In progress: an evaluation platform for researchers to benchmark new models

Paper link

Collaborators:


Yuan Tseng


Layne Berry


Yi-Ting Chen


I-Hsiang Chiu


Hsuan-Hao Lin


Max Liu


Puyuan Peng


Yi-Jen Shih


Hung-Yu Wang


Haibin Wu


Po-Yao Huang


Chun-Mao Lai


Shang-Wen Li


David Harwath


Yu Tsao


Shinji Watanabe


Abdelrahman Mohamed


Chi-Luen Feng


Hung-yi Lee

Thank you for listening!