Separating the "Chirp" from the "Chat": Self-supervised Visual Grounding of Sound and Language

CVPR 2024

Hanyi Shin

Korea University

July 10, 2024

Overview

Overview

0. Abstract

Q. 특정 Sound에 상응하는 Image Segment?
(speech and sound prompted semantic segmentation)

- (sound) "A dog is barking"
- (image)

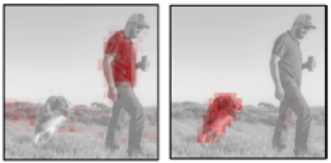


Figure 1: Inference Example01

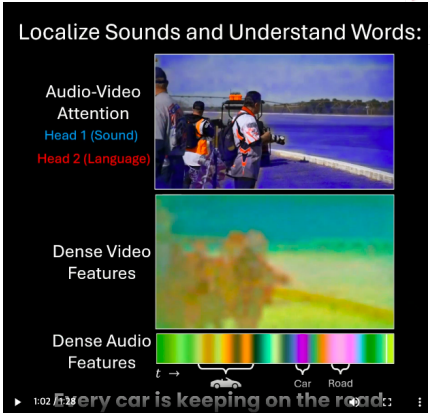


Figure 2: Inference Example 02

0. Abstract

- Algorithm Overview:
  How can we train audio-visual paired signal(a, v)?

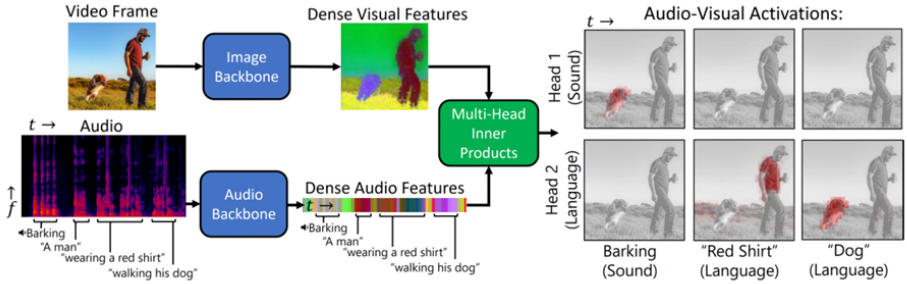


Figure 1. Visual overview of the DenseAV algorithm. Two modality-specific backbones featurize audio and visual signals. We introduce a novel generalization of multi-head attention to extract attention maps that discover and separate the "meaning" of spoken words and the sounds an object makes. DenseAV performs this localization and decomposition solely through observing paired stimuli such as videos.

Figure 3: Algorithm Architecture

Overview

1. Introduction

TASK : Audio-Video Multi-modal 연구의 해결 과제

1. 높은 해상도의 이미지 segmentation이 필요
2. audio-video 간의 multi-modal gap을 해결
3. 이전 연구(ImageBind)에서 test 및 inference 에 사용된 dataset 및 코드를 공개하지 않음
4. audio-visual signal 데이터 부족

1. Introduction

This paper's Objective :

- 고해상도 audio-visual signal(high-resolution AV correspondences) 을 학습할 수 있는 self-supervised architecture, 즉 DenseAV model을 발표.
- 이미지의 지엽적 특징에 기반한 유사도(A local-feature-based image similarity function)를 정의.(to improve a network's zero-shot localization ability)
- speech and sound prompted semantic segmentation를 평가하기 위한 새로운 evaluation dataset 제시.
- 단 하나의 contrastive supervision(w multi-head architecture)을 사용해서 audio signal을 입력받았을 때, 별도의 notation 없이 sound/language를 독립적으로 구분할 수 있음.

DenseAV의 Task는 크게 2가지로 구분

- Speech/Sound Prompted Image Segmentation
- Audio-Visual retrieval

Overview

2. Related Work

(1) Contrastive Learning

(left) A visual representation x에 대한 Contrastive Learning

(right) multi-modal representation (a, v)에 대한 Contrastive Learning

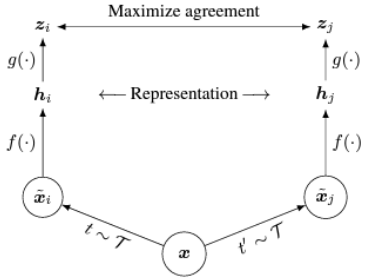


Figure 4: A simple Contrastive

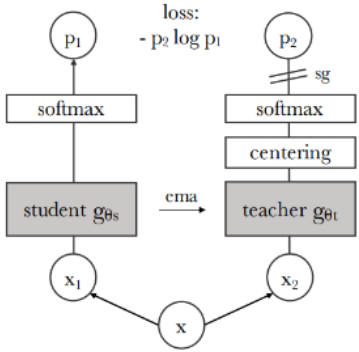
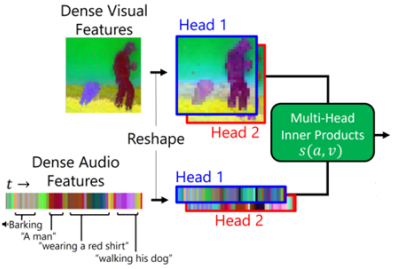


Figure 5: DiNO Vision Transformer

2. Related Work

(2) Audio-Visual(AV) Representation
audio-vision 간의 내적을 사용해 "global" 표현.
*previous) ImageBind 모델 (Local 특징 표현이 아쉽다는 이슈가 있었고, 이전 연구들은 고해상도도 지원하지 않음)

(3) Audio-Visual(AV) modality gap
audio-vision 간의 similarity function $s(a, v)$ 정의



Overview

*Idea)
audio-visual 사이에 뭔가 object가 있다면, 공통된 couple signal이 비슷할 것이다.

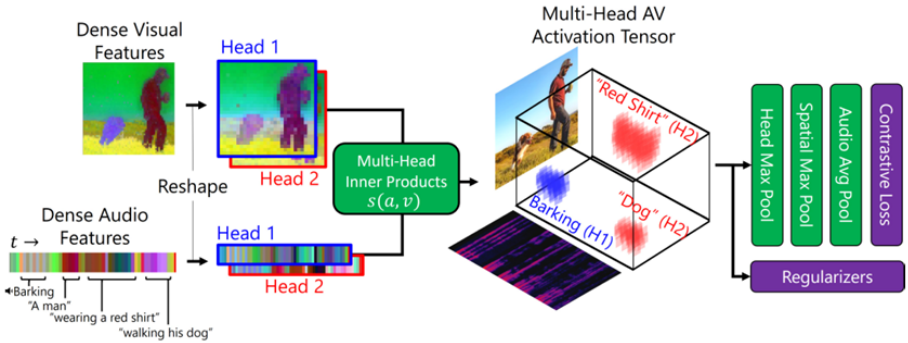


Figure 7: Multi-modal Contrastive Learning Framework

3.1. Multi-Headed Aggregation of Similarities

Aggregation Method

- Video Signal은 max-pooling / Audio Signal은 average-pooling 적용 (from "soft" average-pooling to "hard" max-pooling)

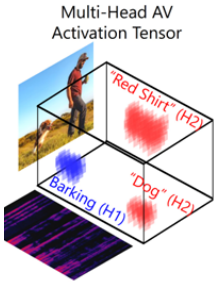


Figure 8: Inner Product

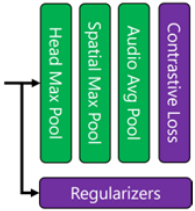


Figure 9: Max pooling

3.1. Multi-Headed Aggregation of Similarities

Audio-Visual 간의 유사성(Similarity) Score를 계산

- a local similarity volume:

$$s(a, v) \in \mathbb{R}^{kfthw} = \sum_{c=1}^{C} a[c, k, f, t] \cdot v[c, k, h, w]$$

- 위의 volume $s(a, v)$ 를 single score $S(a, v)$로 계산:

$$\mathcal{S}(a, v) = \frac{1}{FT} \sum_{f=1}^{F} \sum_{t=1}^{T} \max_{k,h,w}(s(a, v)[k, f, t, h, w])$$

* MISA loss 로 채택
a는 audio tensor, $a \in [Channel, K - heads, Frequency, Time]$
v는 video tensor, $v \in [Channel, K - heads, Height, Width]$

3.2. Loss

Visual-retrieval term에서는 $\mathcal{L}_{A\to V}$ 를 계산,
반대로 Audio-retrieval에서 사용하는 $\mathcal{L}_{V\to A}$는 $\mathcal{L}_{A\to V}$와 symmetric한 관계에 있다.

Contrastive Loss InfoNCE
similarity between "positive" pairs (a_b, v_b) 는 B 개
dissimilarity between "negative" pairs는 $B^2 - B$ 개 있다.

$$\mathcal{L}_{A\to V} = \frac{1}{2B}\sum_{b=1}^{B}\left(\log\frac{\exp(\gamma\mathcal{S}(a_b,v_b))}{\sum_{b'=1}^{B}\exp(\gamma\mathcal{S}(a_b,v_{b'}))}\right)$$
($\mathcal{L}_{V\to A}$ is symmetric)

3.3. Audio and Visual Featurizers

2개 modality-specific network를 구성하기 위한 backbone 모델 채택

- Audio는 HuBERT pretrained 모델 사용
  - - LibriSpeech datasets으로 훈련(self-supervision, no-label)
- Video는 DINO vision transformer Self-supervised 모델을 사용
  - - ImageNet datasets으로 훈련(no-label)

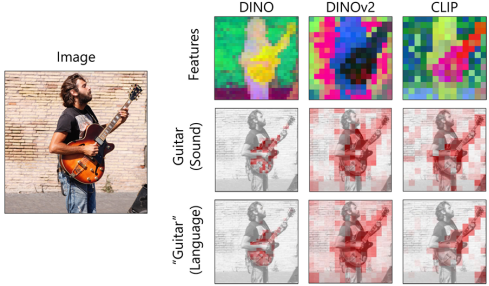


Figure 10: Comparison of Visual Backbone

3.3. Audio and Visual Featurizers

위의 architecture에 각각 layer-norm과 convolution을 추가한다.

- Audio Featurizer: HuBERT (+ layer norm + Convolution 3x3)
- Video Featurizer: DINO (+ layer norm + Convolution 1x1)

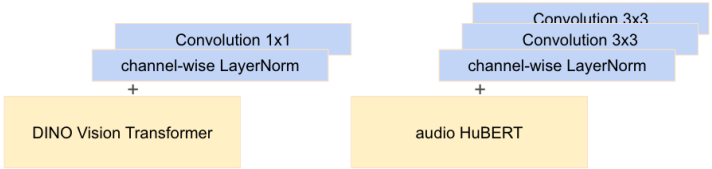


Figure 11: Aligners

Q. Then why?
학습을 더 안정적으로 진행하고, saturation을 방지
pre-trained model의 sensitive parameter 특성을 감안.

3.4. Regularizers

$\mathcal{L}_{\text{Dis}}$: *Disentanglement Regularizer*

B : Batch of positive a-v pairs
head $1(k = 1)$: distinguish the meaning of words
head $2(k = 2)$: capture the sounds objects produce
(*kth attention head)
$\mathcal{L}_{\text{Dis}}$ is "cross-term" generalization of the l^2 *regularizer*

$$\mathcal{L}_{Dis} = \text{Mean}\left(|s\left(a_b, v_b\right)[k=1] \circ s\left(a_b, v_b\right)[k=2]|\right)$$

3.4. Regularizers

$\mathcal{L}_{\text{Stability}}$: *Stability Regularizer*

- 안정적인 학습 진행(stable convergence)을 위함.
- dissimilarity보다 similarity에 집중하기 위함. (non-negative pressure를 주기 위해 아주 작은 값을 가짐)

$\mathcal{L}_{\text{Stability}}$ is standard regularizer like Total Variation

Total Loss function:

$$\mathcal{L} = \mathcal{L}_{A \to V} + \mathcal{L}_{V \to A} + \lambda_{\text{Dis}}\, \mathcal{L}_{\text{Dis}} + \mathcal{L}_{\text{Stability}}$$
$$(\lambda_{\text{Dis}} = 0.05)$$

3.5. Training I

training dataset : AudioSet, PlacesAudio 사용
ADE20K는 Task별로 다르게 Customized되어 사용됨.

dataset info)

Table 1: default

Dataset	No.Class	Image Size	Volume	explanation
AudioSet	632classes	10s length	5.8hours	training sound
PlacesAudio	205scenes	10s length	85GB	training speech
*ADE20K	478classes	256x256	3030pairs	evaluation data

3.5. Training II

1. Step : Data augmentation
   - 1.1 Image Augmentation)
     resize, color, random flip 등의 이미지 augmentation을 진행
     random sampling
   - 1.2 Audio process)
     sampling rate=16kHz로 맞추어서 re-sample
     10초 길이로 고정(Trimming or Padding)

training setting : 8개의 V100 GPUs 사용, batch size = 80

3.5. Training III

2. Step : Training
   2.1 Step1 : Warming up Aligners
   aligners 를 업데이트하기 위한 warming up 진행
   - backbone은 고정 : DINO and HuBERT backbone은 parameter 고정
   - 3,000 steps 동안 진행
   - aligners(norm + convolution)만 업데이트

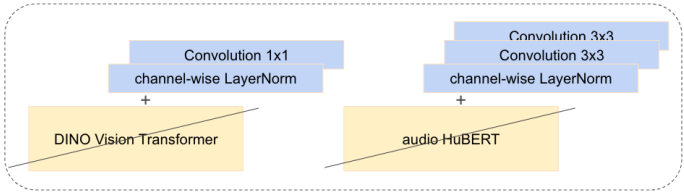


Figure 12: Warming up Aligners Backbone-Fixed

3.5. Training IV

2.2 Step2 : Full Training

- backbone+aligners 전체 파라미터 사용, fine tune all
- 800,000 steps 동안 진행
- DINO transformers(for visual)를 fine tune - Q, K, V layers를 업데이트
  Low rank adaptation(LoRA) 적용(rank=8)

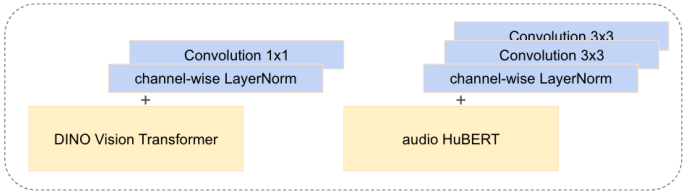


Figure 13: Full Training with Backbone

Overview

평가 metric:
mean average precision (mAP), mean intersection over union (mIoU) 사용

Evaluation dataset 제작

- Task1 : Speech Prompted Image Segmentation (ADE20K)
  - 각 object마다 10개 images씩 채택
  - paired speech signal은 "A picture of [object]"라는 문구의 TTS service를 이용해서 제작.
  - segment 부분이 전체 픽셀의 5퍼센트 미만일 경우, 데이터 제외
  - Total 478 classes, 3030개의 image-object pairs를 만듦.
- Task2 : Sound Prompted Image Segmentation(ADE20K+VGGSound)
  - ADE20K의 각 class마다 3개씩 VGGSound ontology 후보군 설정
  - cosine similarity $<.85$ sound 제외
  - Total 20개 classes, 106개의 image-object pairs를 만듦

Table 1.

Method	Speech Semseg. mAP	Speech Semseg. mIoU	Sound Semseg. mAP	Sound Semseg. mIoU
DAVENet [26]	32.2%	26.3%	16.8%	17.0%
CAVMAE [21]	27.2%	19.9%	26.0%	20.5%
ImageBind [20]	20.2%	19.7%	18.3%	18.1%
Ours	**48.7%**	**36.8%**	**32.7%**	**24.2%**

Table 1. **Speech and Sound prompted semantic segmentation.**

Method	Places Audio Retrieval I → A @1	I → A @5	I → A @10	A → I @1	A → I @5	A → I @10	AudioSet Retrieval I → A @1	I → A @5	I → A @10	A → I @1	A → I @5	A → I @10
[25]	12.1%	33.5%	46.3%	14.8%	40.3%	54.8%	-	-	-	-	-	-
[24]	13.0%	37.8%	54.2%	16.1%	40.4%	56.4%	-	-	-	-	-	-
DAVENet [26]	12.7%	37.5%	52.8%	20.0%	46.9%	60.4%	-	-	-	-	-	-
DAVENet* [26]	13.3%	38.3%	51.2%	20.5%	45.3%	57.2%	0.10%	0.70%	1.30%	0.10%	0.30%	1.20%
CAVMAE*[21]	36.7%	70.3%	81.7%	33.9%	65.7%	77.7%	22.8%	44.9%	55.7%	21.1%	41.7%	50.7%
ImageBind[20]	0.10%	0.50%	1.10%	0.10%	0.40%	1.10%	29.6%	55.4%	64.5%	31.8%	57.3%	66.5%
Ours	**65.3%**	**90.0%**	**94.2%**	**64.4%**	**89.4%**	**94.3%**	**35.1%**	**58.0%**	**68.2%**	**33.6%**	**59.3%**	**68.4%**

Table 5. Full cross modal retrieval results using the same setting of Table 2. We note DenseAV outperforms all baselines in all metrics and all datasets.

Figure 14: semantic segmentation results - expanded version

Table 2.

Method	Places Acc. @10		AudioSet Acc. @10	
	I → A	A → I	I → A	A → I
[25]*	46.3%	54.8%	-	-
[24]*	54.2%	56.4%	-	-
DAVENet [26]*	52.8%	60.4%	-	-
CAVMAE [21]	81.7%	77.7%	55.7%	50.7%
ImageBind [20]	1.10%	1.10%	64.5%	66.5%
Ours	**94.2%**	**94.3%**	**69.8%**	**68.1%**

Table 2. **Cross-modal retrieval using 1000 evaluation videos from the PlacesAudio and AudioSet validation datasets.**

Figure 15: Cross-modal retrieval results

Table 3.

Measure : prediction disentanglement score(near 100 = disentanglement)

- Pred.Dis : $\delta_{\text{pred}}\ (k, k') = \text{AP}\left(\left(\hat{\mathcal{S}}\left(a_b, v_b\right)_k\right)_1^B, \left(I\left[k'\right]_b\right)_1^B\right)$
- Act.Dis : $\delta_{act}\left(k, k'\right) = 1 - \frac{1}{\sum_{b'} I[k']_{b'}} \sum_{b=1}^{B} \hat{\mathcal{S}}\left(a_b, v_b\right)_k \cdot I\left[k'\right]_b$
- $I\left[k'\right]_b$ is an indicator variable of whether the signal (a,v) arises from the sound dataset

Table 3.

Method	Pred. Dis.	Act. Dis.
No $\mathcal{L}_{Dis}$, No Head Max Pool	64.1%	70.3%
No $\mathcal{L}_{Dis}$	**99.9%**	86.5%
Ours	**99.9%**	**91.2%**

Table 3. Quantitative ablation study of the impact of max-pooling attention heads and adding our disentanglement loss, $\mathcal{L}_{Dis}$. Intuitively, max-pooling attention heads allows each head to specialize on its own specific set of triggers. Our disentanglement loss further encourages the heads to operate independently and orthogonally.

Figure 16: max-pooling attention heads specialize on its own specific set of triggers

Table 4.

(left) VGGSound annotation's bounding boxes do not reward high-resolution, but good results
(right) additional noise-robustness experiments(w MUSAN dataset)

Method	cIoU	AUC
DAVENet	6.8%	21.2%
CAVMAE	7.9%	25.0%
ImageBind	3.4%	20.5%
Attention10K	18.5%	30.2%
AVObject	29.7%	35.7%
LVS	34.4%	38.2%
SLAVC	38.8%	38.8%
FNAC AVL	39.4%	39.4%
Ours	**40.6%**	**40.6%**

Figure 17: Performance on VGGSound Source localization.

Method	mAP	mIoU
DAVENet	31.8%	26.1%
CAVMAE	27.2%	23.8%
ImageBind	20.2%	19.7%
Ours	**48.1%**	**36.6%**

Figure 18: Speech Prompted Semantic Segmentation Noise Robustness

Overview

5. Conclusion

- one audio signal contrastive learning으로 sound에서 meaning of words를 알아낼 수 있다.
- multi-head attention 구조를 사용하여 high-resolution, semantically meaningful, AV aligned representation을 만들 수 있다.
- 이전 SOTA 모델과 비교했을 때도 cross-modal retrieval과 sound-prompted semantic task에서 모두 좋은 prediction 성능을 보였다.

References I

Antonio Torralba David Harwath and James Glass. Unsupervised learning of spoken language with visual context. In *Advances in Neural Information Processing Systems*, page 29, 2016.

D´ıdac Sur´ıs Galen Chuang Antonio Torralba David Harwath, Adria Recasens and James Glass. Jointly discovering visual objects and spoken words from raw sensory input. In *In Proceedings of the European conference on computer vision (ECCV)*, page 649–665, 2018.

Nicolas Papernot Nicholas Frosst and Geoffrey Hinton. Analyzing and improving representations with the soft nearest neighbor loss. In *In International conference on machine learning (PMLR)*, page 2012–2020, 2019.

Nicholas Frosst and Hinton (2019) David Harwath and Glass (2016) David Harwath and Glass (2018)