Contrastive Alignment in Audio–Language Models: What It Misses and How to Fix It

2026. 04. 23.

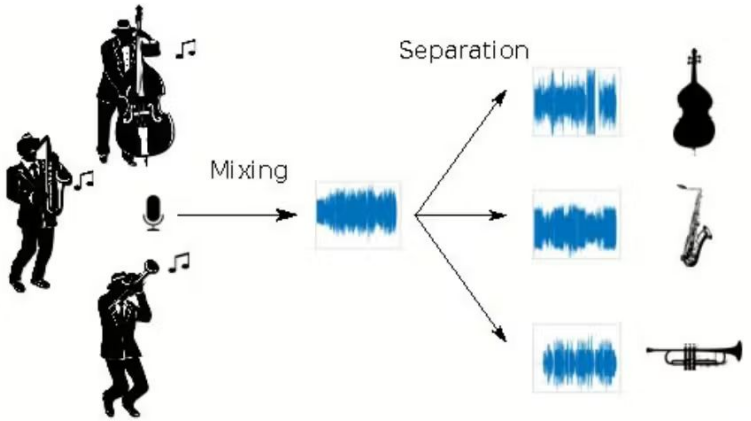
양 건

Natural Language Processing & Artificial Intelligence

KOREA UNIVERSITY 고려대학교 KOREA UNIVERSITY

Audio 분야의 근본 문제: 데이터 수집

- 소리를 인식·분리하는 모델을 만들려면 해당 소리의 학습 데이터가 충분해야 함
  - 하지만 실제 환경에서는 주변 소리 (노이즈) 가 섞이기에 원하는 소리만 녹음하기가 힘듦
    - 예) 목표: 비 내리는 소리 → 바람 소리도 섞여 들어감
      목표: 차 바퀴 굴러가는 소리 → 차 엔진 소리도 섞여 들어감
  - 특정 소리만 순수하게 담긴 데이터를 만드는 건 매우 비쌈
  - 깨끗한 단일 음원은 부족하지만, real-world audio 는 풍부함
    - 혼합음에서 원하는 소리를 분리해낼 수 있다면 데이터 문제 해결!
- Universal Sound Separation (USS)
  - 혼합음원에서 세상의 모든 소리를 다 분리해내는 모델 만들자는 목표



USS 의 학습 병목: 세밀한 annotation 비용

- USS 모델을 만들기 위해 “이상적으로" 필요한 데이터
  - Strong Label
    - 음원 속에 어떤 소리가, 몇 초부터 몇 초까지 등장하는지 모두 명시된 라벨

예) [tag: 개 짖는 소리, onset: 0.0s, offset: 1.0s]
[tag: 사람 뛰는 소리, onset: 1.2s, offset: 4.0s]

Strong Label 의 대안: Weak Label

- Weak Label
  - 클립 단위로 어떤 소리가 포함됐는지 tag 만 표시
  - 언제 등장하는지 (onset/offset) 는 표시하지 않음

예) [tags: 개 짖는 소리, 사람 뛰는 소리]

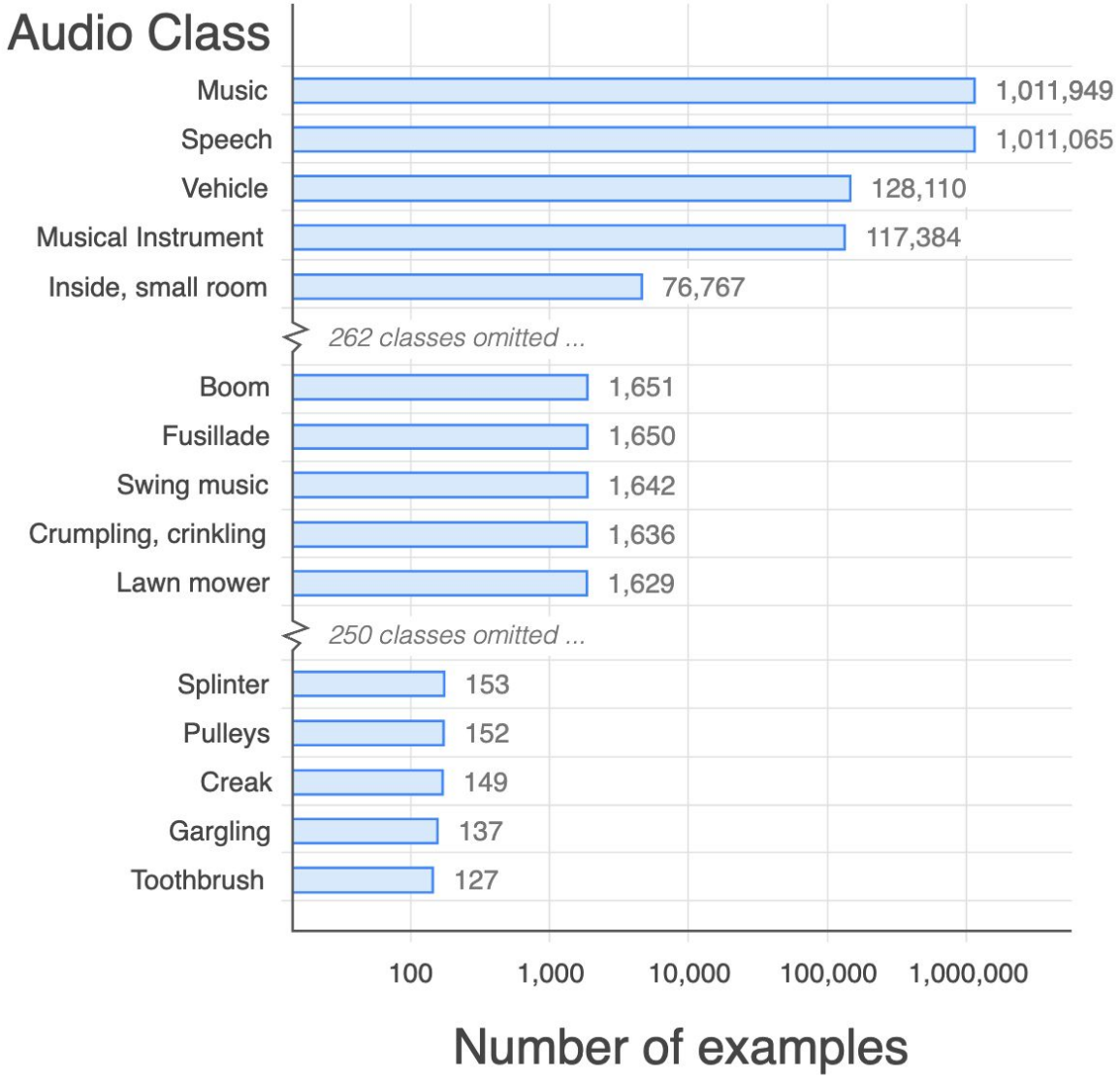
Weak Label 로 소리 분리 모델을 학습하는 방법

- weak label 에는 '언제 / 어디에' 정보가 없음
  - 그런데 소리 분리 모델은 혼합음에서 특정 구간을 뽑아내야 함

- Synthetic mixture 활용
  - Weak label 이 서로 다른 두 클립을 준비
    - 클립 A: tag "Dog"
    - 클립 B: tag "Piano"
  - 인위적으로 섞어 혼합음 M = A + B 생성
  - 모델에게 쿼리 "Dog" 와 혼합음 M 을 주고, A 를 복원하도록 학습

- 한계점
  - 개별 클립 A, B 도 real-world audio 라 노이즈가 섞인 데이터
    - 예) "Rain" 클립에는 바람 소리가 함께 섞임
  - Label-Signal 불일치가 잔존 → 모델 성능의 근본적 상한선으로 작용

AudioSet (Google Research, 2017)



2.1 million
annotated videos

5.8 thousand
hours of audio

527 classes
of annotated sounds

From USS to Query-Based Separation

- USS는 혼합 음원에서 세상의 모든 소리를 분리하려는 매우 큰 목표

- 세상 모든 소리를 분리하는 모델을 한 번에 만들기는 비현실적
  →모든 소리는 아니라도, 그 동안 모은 class 데이터 정도는 분리할 수 있는 모델을 먼저 만들자!

- Query-based Sound Separation (QSS)
  - pre-defined sound class 집합으로 학습
  - 쿼리로 class 를 받으면, 혼합음에서 해당 class 소리만 추출
  - Closed-Set 이라는 한계점 명확
    - 새로운 소리 추가? → 모델 재학습 필요
    - 쿼리는 pre-defined class 집합 안에서만 지정 가능하기에 표현력 부족

- Closed-set 이라는 한계를 어떻게 극복할까?

Closed-set을 넘어서: LASS

- Closed-set 제약을 깨려면? → 쿼리 자체를 open-vocabulary 로 만들자

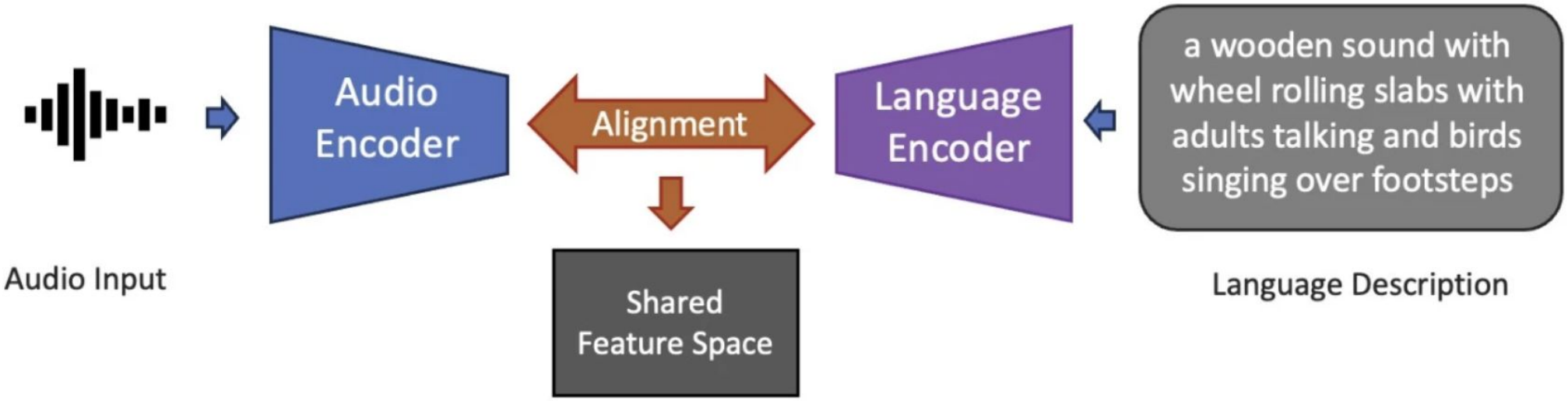
- Language-queried Audio Source Separation (LASS)
  - 쿼리를 free-form 자연어로 받음
    - 예: “개가 짖자 사람이 뛰어 달아나는 소리”

- QSS 와의 차이
  - pre-defined class 대신 자연어로 target sound를 지정할 수 있음
  - 학습 데이터에 없던 새로운 쿼리 사용 시에도 모델 재학습 불필요
  - 쿼리가 품을 수 있는 의미의 폭이 커짐 — 단일 class label 보다 더 유연하고 구체적인 기술이 가능함

- 중요한 전제
  - 자연어 쿼리와 오디오를 '충분히 잘' 정렬시킬 수 있어야 함

오디오와 자연어를 잇는 도구: CLAP

- CLAP (Contrastive Language-Audio Pretraining, 2023)
  - 대량의 audio-caption 쌍으로 학습된 오디오-텍스트 정렬 모델



Pre-CLAP vs. CLAP

	CLAP 이전 오디오 모델들	CLAP
학습 데이터 확보 비용	높음 (세밀한 정답 라벨이 필요함)	낮음 (대략적인 caption 만으로도 활용 가능)
표현 가능한 개념 범위	사전 정의된 label 만 가능	Open Vocabulary (자연어 기반 zero-shot 확장 가능)
의미 prior	label 간 의미 관계를 직접 활용하기 어려움	pretrained text encoder 를 통해 자연어 semantic prior 를 활용 가능
재사용성	힘듦 (task별 모델/데이터 설계 필요)	하나의 backbone 을 여러 ALM task 에 재사용 가능

ALM 전반으로 확장된 CLAP

- LASS

- Open-Vocabulary Sound Event Detection (OV-SED)
  - 학습 class 에 없던 소리도 semantic similarity 를 활용해 zero-shot 검출 시도
    - 예) 학습한 적 없는 새로운 class: "Fusillade" (뜻: 일제 사격)
      학습한 classes: "Explosion", "Gunshot/gunfire"
      → Fusillade 쿼리가 모델이 알고있는 class 임베딩과 가깝게 정렬되어 소리 검출 성공

- Audio-Text retrieval
  - 자연어로 오디오 찾기 / 오디오로 텍스트 찾기

- Text-to-Audio generation
  - CLAP 텍스트 임베딩을 diffusion 모델의 conditioning 으로 사용

- CLAPScore
  - 생성 오디오와 prompt 텍스트의 CLAP 임베딩 cosine similarity
  - Text-to-Audio 벤치마크의 표준 지표

Structural Limits of CLAP

- ALM 응용의 성능은 결국 CLAP 임베딩 공간의 품질에 의존
  - 그러나 CLAP 의 정렬 방식에는 몇 가지 구조적 한계가 존재

- 한계 1: 표현의 비대칭
  - 1차원적인 Text 와 다르게, Audio 신호는 시간 / 공간 (방향) / 주파수 가 반영되는 고차원 Modality
  - CLAP 은 두 Modality를 "같은 크기의 단일 벡터" 로 압축 (예. 10초 짜리 오디오 클립 ↔ 캡션)
    - 결과: 오디오 클립 내 세밀한 음향 속성은 임베딩에서 충분히 보존되기 어려움

- 한계 2: Hard Contrastive Objective 의 경직성
  - 대조 학습 (InfoNCE) 의 기본 목표
    - 현재 학습하는 (audio, text) pair 는 가깝게
    - 해당 batch 의 모든 나머지 pair 들은 “동등하게" 멀리 밀어냄
  - 문제점
    - 의미적으로 유사한 소리라도 서로 다른 pair이면 동일한 negative로 취급될 수 있음
    - 즉, semantic similarity의 정도를 직접 보존하는 학습은 아님

SMOOTHCLAP: SOFT-TARGET ENHANCED CONTRASTIVE LANGUAGE-AUDIO PRETRAINING FOR AFFECTIVE COMPUTING

Xin Jing[⋆] *Jiadong Wang*[⋆] *Andreas Triantafyllopoulos*[⋆] *Maurice Gerczuk*[⋆]
Shahin Amiriparian[†⋆] *Jun Luo*[†] *Björn Schuller*[⋆‡]

[⋆] CHI – Chair of Health Informatics, TUM University Hospital, Munich, Germany,
[†] Huawei, Netherlands, [‡]GLAM, Imperial College London, UK

ICASSP 2026

Motivation

- 감정 인식 domain에서 CLAP 의 한계
  - 감정 데이터는 class 사이의 경계가 본질적으로 모호함 (fuzzy boundary)
    → 예) 짜증 & 분노
  - CLAP 이 사용하는 InfoNCE 의 Hard Contrastive Objective 의 충돌
    - InfoNCE: matching pair 만 1, 나머지는 모두 "동등하게" 0 (one-hot target)
    - non-matching pairs 사이의 유사도 차이 는 학습 목표에 직접 반영되지 않음
      → 짜증 ↔ 분노 와 기쁨 ↔ 분노 의 거리가 동일하게 취급
  - CLAP embedding space 는
    감정 간 graded psychological structure 를 충분히 반영하지 못할 수 있음

Methods

- 핵심 아이디어: Hard Target → Soft Target
  - 기존 CLAP 은 batch 내 matching pair 만 정답으로 두는 one-hot identity target 을 사용
  - SmoothCLAP 은 각 sample 에 대해
    - pretrained audio encoder (wav2vec 2.0 large) 로 local audio feature 추출
    - pretrained text encoder (BERT) 로 text embedding 추출
  - 이후, batch 안에서
    audio-to-audio similarity / text-to-text similarity
    각각 계산하여 soft target distribution 을 생성
  - 두 similarity 정보를 결합해 non-matching pair 사이의 graded similarity 도 일부 반영
  - 기존 hard identity target 과 새로 만든 soft target 을 함께 사용하여 target 을 부드럽게 만듦
- Training objective 만 수정하며, inference pipeline 은 기존 CLAP 과 동일

Methods

- 배치 내 intra-modal similarity distribution

$$q_{ij}^{a2a} = \frac{\exp(\bar{\ell}_i^a \cdot \bar{\ell}_j^a / \tau_{a2a})}{\sum_{k=1}^{B} \exp(\bar{\ell}_i^a \cdot \bar{\ell}_k^a / \tau_{a2a})},$$

$$q_{ij}^{t2t} = \frac{\exp(e_i^t \cdot e_j^t / \tau_{t2t})}{\sum_{k=1}^{B} \exp(e_i^t \cdot e_k^t / \tau_{t2t})},$$

- 두 similarity 를 mix gamma ɣ 로 결합

$$q_{ij} = (1-\gamma) q_{ij}^{a2a} + \gamma q_{ij}^{t2t}$$

- 최종 soft target: hard identity 와 fusion factor β 로 결합

$$y_{ij} = (1-\beta)\delta_{ij} + \beta q_{ij}$$

- Loss: symmetric KL divergence (audio / text 축 모두)

$$\mathcal{L}_{\text{soft}} = \frac{1}{2B} \sum_{i=1}^{B} \Big[KL(y_i \,\|\, p_i^{a2t}) + KL(p_i^{a2t} \,\|\, y_i) + KL(y_i \,\|\, p_i^{t2a}) + KL(p_i^{t2a} \,\|\, y_i) \Big]$$

Results

- 학습: MSP-Podcast v1.9 (영어, 약 4.6 만 샘플)
- 평가: 8 개 affective computing task (영어 4 개 + 독일어 4 개)
- 지표: UAR (Unweighted Average Recall)

- SmoothCLAP의 전체 성능
  - 5 / 8 task 에서 최고 성능
  - 나머지 3개 중 2개에서 second-best
  - 대부분의 setting 에서 CLAP / ParaCLAP 대비 개선

Table 1. The Unweighted Average Recall (UAR) results on 8 affective computing tasks, comprising both English and German language-based datasets. **Bold** marks the best performance while underline represents the second best.

Dataset	Unweighted Average Recall (UAR)			
	CLAP	Pengi	ParaCLAP	SmoothCLAP
IEMOCAP (4cl/en)	.353	.345	.600	**.606**
RAVDESS (8cl/en)	.199	.148	**.228**	.175
CREMA-D (6cl/en)	.230	.245	.177	**.266**
TESS (7cl/en)	.232	.177	.170	**.275**
FAU Aibo (2cl/de)	.500	.470	526	**.555**
FAU Aibo (5cl/de)	**.211**	.185	.197	.204
ALC (2cl/de)	.511	.473	.537	**.541**
SLD (2cl/de)	.472	.485	**.507**	.496

Conclusion & Limitations

- 감정 인식 domain 에서,
  CLAP 의 hard one-hot target 을 soft target 으로 완화하는 것만으로도 generalization 이 개선될 수 있음
- 감정 domain 에 국한된 검증 — 범용 audio-language task 전체에 대한 일반화는 아직 미확인
- Soft target 의 품질이 외부 pretrained model (wav2vec 2.0) 에 의존

iKnow-audio: Integrating Knowledge Graphs with Audio-Language Models

Michel Olvera[1] **Changhong Wang**[1]
Paraskevas Stamatiadis[1] **Gaël Richard**[1] **Slim Essid**[2*]
[1]LTCI, Télécom Paris, Institut Polytechnique de Paris [2]NVIDIA
{olvera,changhong.wang}@telecom-paris.fr

EMNLP 2025

Motivation

- CLAP 의 또다른 문제: Acoustic Ambiguity
  - 음향적으로 유사한 소리들을 CLAP 이 잘 구분하지 못함
  - sound event 간 co-occurrence / hierarchy / causality 반영이 약함

- 실제 소리 이해에는 배경지식과 관계 정보 가 필요
  - 소리는 독립 label 이 아니라 상황 속 관계망 안에서 해석되어야 함

Motivation

- 폴리(Foley) 음향효과
  - 영화 후반 작업 시 소리를 스튜디오에서 도구를 활용해 인위적으로 재현하거나 창조하는 기술

Motivation

- 기존 audio dataset / taxonomy 는 sound 를 주로 독립된 category 로 다루며,
  sound 간의 structured semantic relation 을 충분히 담지 못함

- Research Question:

  CLAP embedding 만으로 부족한 부분을,
  Knowledge Graph 기반 external knowledge 로 보완할 수 있을까?

Methods

- 두 개의 핵심 구성 요소
  - Audio-centric Knowledge Graph (AKG): 소리들 사이의 ontological 관계를 인코딩한 KG
  - CLAP-KG: AKG 로 CLAP 의 top-k 예측을 refine 하는 추론 파이프라인

Methods

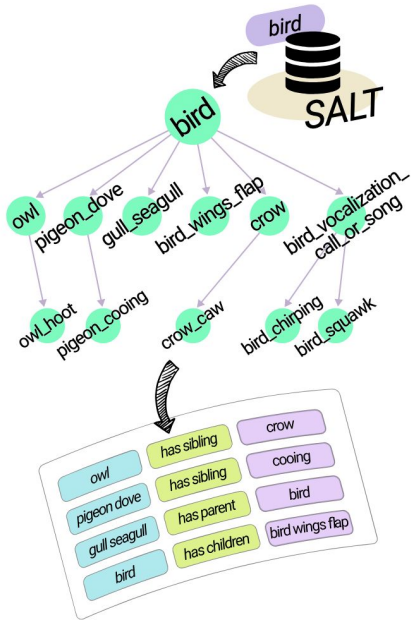


Figure 3: Generation of knowledge triples from SALT.

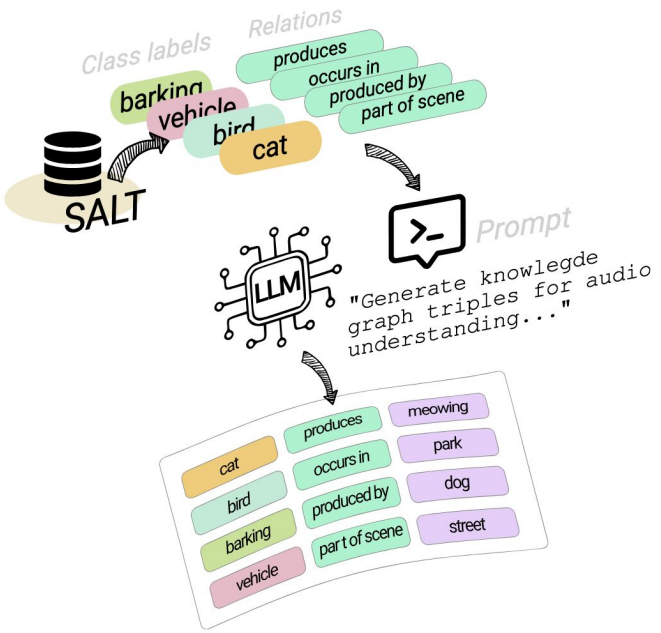


Figure 4: Generation of knowledge triples from LLMs.

- AKG: 구축 과정
  - SALT (Standardized Audio Label Taxonomy) 의 sound event label 을 기반
  - 두 가지 방식으로 triple 생성
    - SALT 의 hierarchical 구조에서 직접 추출
    - LLM (Mistral-7B) 프롬프팅으로 관계성 triple 생성
  - LLM 기반 검증 + 수동 curation 으로 필터링 → 20,387 개의 고품질 triple
  - 9 개의 relation 카테고리 (semantic, causal, taxonomic, spatial 등)

Methods

- KGE (Knowledge Graph Embedding) 학습
  - AKG 의 triple (h, r, t) 를 벡터 공간에 임베딩
  - 실험 결과 RotatE 가 최고 성능 → backbone 으로 채택

- Step 1: CLAP 의 초기 top-k 예측 (유사도 순 top-k 후보)

$$\hat{c} = \arg\max_{c \in C} \mathrm{sim}\left(\phi_{\mathrm{A}}(a), \phi_{\mathrm{T}}(c)\right),$$

$$C_k = \{\hat{c}^{(1)}, \ldots, \hat{c}^{(k)}\}, \quad \text{ranked by similarity.}$$

$$\phi_{\mathrm{KG}}(h, r, t) = -\left\|\mathbf{h} \circ \mathbf{r} - \mathbf{t}\right\|_2$$

- Step 2: KGE 로 각 후보의 semantic neighbors 예측
  - RotatE score
  - 각 후보 ĉ 에 대해 가장 plausible 한 tail top-m 개 선택

$$\mathcal{T}_c = \{(\hat{c}, r, t) \in \mathcal{T} \mid r \in \mathcal{R}_q\}$$

$$p_{\hat{c}, t_i^*} = \mathsf{concat}(\hat{c}, t_i^*)$$

- Step 3: Enriched prompt 생성 및 scoring

$$s(p) = \mathrm{sim}\left(\phi_{\mathrm{A}}(a), \phi_{\mathrm{T}}(p)\right)$$

Methods

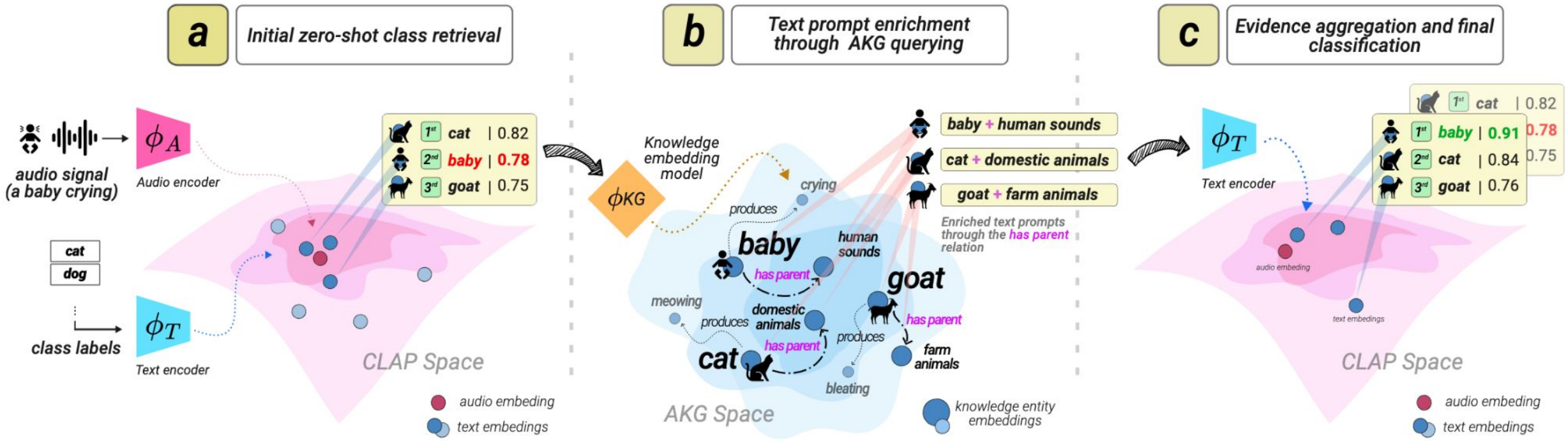
- Step 4: Log-Sum-Exp Aggregation
  - 원래 CLAP score 와 enriched prompt score 를 합산
  - 단순 평균 대신 log-sum-exp 를 쓰는 이유:
    → soft-max 방식으로 가장 강한 증거를 우선시하되, 나머지도 반영

$$\tilde{s}(\hat{c}) = \log\left(\exp(s(\hat{c})) + \sum_{p \in P_{\hat{c}}} \exp(s(p))\right)$$

- Step 5: 최종 재정렬

$$c^* = \arg\max_{\hat{c} \in C_k} \tilde{s}(\hat{c})$$

Methods


a
Initial zero-shot class retrieval
b
Text prompt enrichment through AKG querying
c
Evidence aggregation and final classification
audio signal (a baby crying)
Audio encoder
cat
dog
class labels
Text encoder
1st cat | 0.82
2nd baby | 0.78
3rd goat | 0.75
CLAP Space
audio embeding
text embeddings
Knowledge embedding model
ϕKG
crying
produces
baby
human sounds
has parent
meowing
produces
domestic animals
has parent
cat
goat
has parent
produces
farm animals
bleating
AKG Space
baby + human sounds
cat + domestic animals
goat + farm animals
Enriched text prompts through the has parent relation
knowledge entity embeddings
Text encoder
1st cat | 0.82
1st baby | 0.91
2nd cat | 0.84
3rd goat | 0.76
audio embeding
text embeddings
CLAP Space
audio embeding
text embeddings

Results

Metric	ESC50			US8K			TUT2017			FSD50K			AudioSet			DCASE17-T4		
	CLAP	+KG-agg	+KG	CLAP	+KG-agg	+KG	CLAP	+KG-agg	+KG	CLAP	+KG-agg	+KG	CLAP	+KG-agg	+KG	CLAP	+KG-agg	+KG
Hit@1	93.2	93.5	**95.4**	82.5	**84.5**	**85.9**	37.8	**49.3**	**47.9**	61.1	**63.6**	**64.0**	18.4	**19.6**	**19.9**	37.7	**43.0**	**45.9**
Hit@3	98.8	99.1	99.2	96.6	95.6	96.9	74.9	**82.1**	**83.3**	82.8	80.7	**84.2**	33.1	31.2	**34.4**	77.3	76.4	**78.5**
Hit@5	99.5	99.5	99.5	98.8	96.3	98.8	91.3	82.8	91.3	88.9	81.1	88.9	41.1	31.5	41.1	91.2	79.5	91.2
MRR	95.9	96.2	**97.2**	89.6	90.1	**91.5**	57.7	**64.3**	**65.4**	72.2	71.7	**74.3**	26.5	25.0	**27.7**	57.3	**58.5**	**63.1**

Table 2: Retrieval results (%) in terms of hit@1, hit@3, hit@5, and MRR on the six benchmark datasets. Each dataset has three sub-columns: CLAP (baseline), +KG-agg (CLAP-KG w/o aggregation), and +KG (CLAP-KG). Performance improvement larger than 1% over CLAP is in **bold**, and improvement of 1% or less is underlined.

- full model (+KG) 이 baseline CLAP 대비 전 dataset 에서 Hit@1 개선
- MRR 역시 모든 dataset 에서 상승

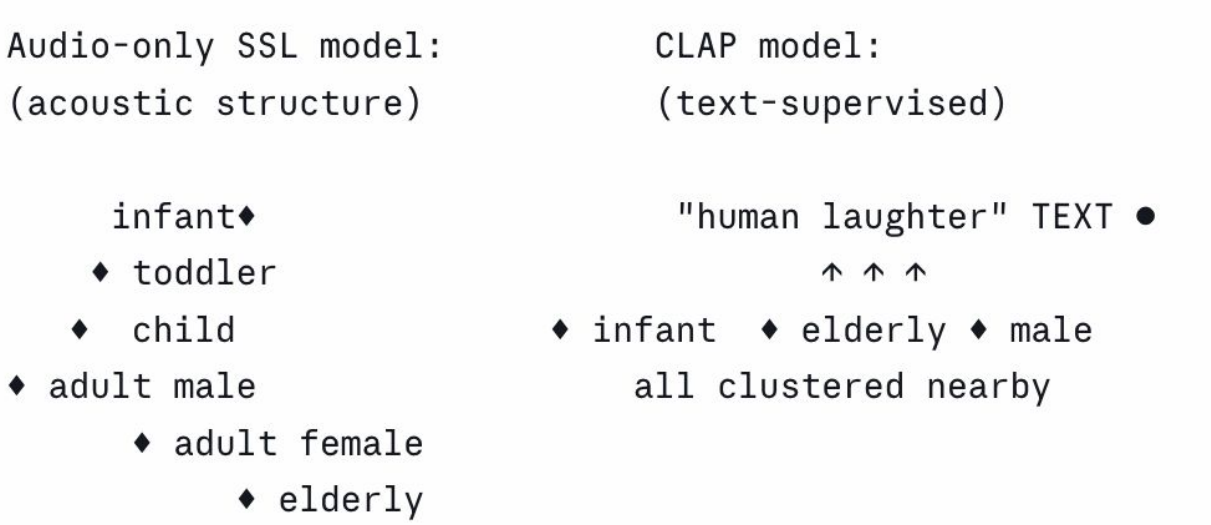
Conclusion

- Knowledge Graph 는 acoustically ambiguous class disambiguation 에 도움
- 특히 hierarchy / scene relation / contextual relation 이 성능 향상에 기여

Future Directions

- 의미적 유사 vs. 신호적 유사
  - CLAP 은 "text" 를 anchor 로 삼아 오디오를 정렬
    - 같은 text label 로 묶인 오디오들은 그 text 임베딩 근처로 수렴
    - 다른 text label 에 묶인 오디오들은 서로 멀어짐

  - 음원의 신호적 유사도와 무관하게 semantics 기준으로 정렬됨
    - 믹서기 & 잔디깎기 기계
      - 모두 회전 모터 소음 계열 — acoustic signal 은 매우 유사
      - 그러나 text label 이 다르기에 CLAP 임베딩 공간에서는 분리됨



Thank you