Visually-aware Acoustic Event Detection using Heterogeneous Graphs

Amir Shirian[1], Krishna Somandepalli[2], Victor Sanchez[1], Tanaya Guha[3]

[1]University of Warwick, UK, [2]Google Research, USA, [3]The University of Glasgow, UK

INTERSPEECH 2022
Sep 18 - 22 • Incheon Korea

Motivation

Perception of auditory events is inherently multimodal, relying on both audio and visual cues.

Why modeling multi-modal data with a heterogeneous graph?

- Heterogeneous graphs are a compact, efficient, and scalable way to represent data involving multiple different entities and their relations.
- It explicitly captures the spatial and temporal relationships between the modalities.

Multimodal heterogeneous graphs lead to a closer coupling between concepts in multiple modalities, resulting in a significant performance improvement over various methods.

Contribution

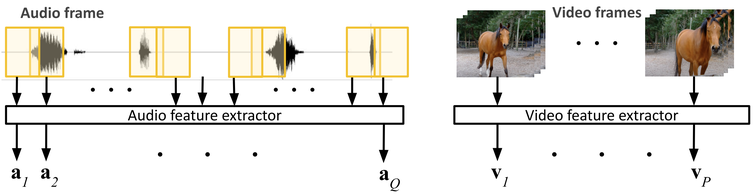
- Develop a graph construction method for converting an audiovisual clip to a multimodal heterogeneous graph.
- Propose a novel **heterogeneous graph neural network (HGNN)** that can capture modality-specific as well as complementary information between modalities.
- Leveraging heterogeneous graph modelling, we obtain the improved performance on **AudioSet** database for the task of acoustic event classification.

Problem Formulation

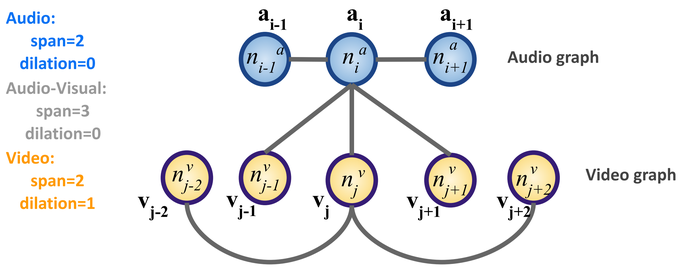
- Given:
  - Video clip heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, O, R)$
  - Where $\mathcal{V} \in \{\mathcal{V}_a, \mathcal{V}_v\}$, $\mathcal{E} = \{\mathcal{E}_{vv}, \mathcal{E}_{aa}, \mathcal{E}_{va}\}$, and $|O| + |R| > 2$
  - A graph specified by three adjacency matrices $\mathbf{A}_a$, $\mathbf{A}_v$, and $\mathbf{A}_{av}$
  - Each graph is associated with an acoustic label $\mathbf{y}_i$
- Goal:
  - We want to predict the acoustic event related to the audiovisual graph

Graph Construction

- It's a frame to node transformation.



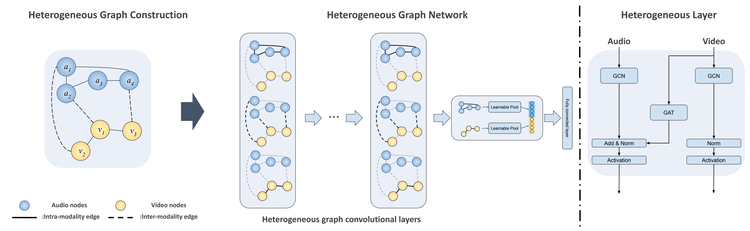
- Audio and video features were extracted from Q and P frames (short, overlapping segments).



- Each of these Q and P frames are associated with a heterogeneous node in a graph.
- Each modality has two specific hyperparameters: (i) ***span across time*** and (ii) ***dilation***.

Model

The overview of our proposed graph-based architecture



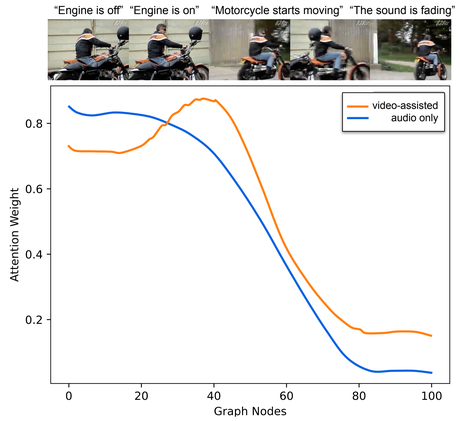
- Takes constructed heterogeneous graph as input.
- Produces node embedding with multiple HGNN layers
- Produces graph embedding with a multimodal pooling function.

Results

Compare with SOTA and graph baselines on AudioSet

Model	mAP	ROC-AUC	Params
Ours audio only	0.42±0.01	0.90±0.00	1.4M
Ours video only	0.15±0.02	0.75±0.01	1.5M
Ours both	**0.50** ±0.01	0.93±0.00	2.1M
Baselines			
ResNet-1D audio only	0.35±0.01	0.90±0.00	40.4M
ResNet-1D both	0.38 ±0.03	0.89 ±0.02	81.2M
LSTM audio only	0.40 ±0.00	0.90±0.00	0.8M
State-of-the-art			
DaiNet	0.25±0.07	-	1.8M
Spectrogram-VGG	0.26±0.01	-	6M
VATT	0.39±0.02	-	87M
SSL graph	0.42±0.02	-	218K
Wave-Logmel	0.43±0.04	-	81M
AST	0.44±0.00	-	88M

Qualitative result



Conclusions

- We transformed video clips into heterogeneous graphs by considering two hyperparameters in each modality.
- Proposed graph captures intra and inter modalities connections in both spatial and temporal domains.
- Our heterogeneous graph model produces higher or comparable performance to the state-of-the-art.

SCAN ME SCAN ME