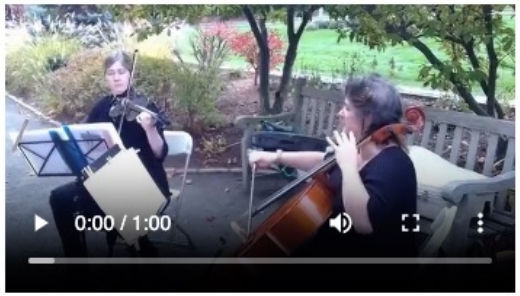
Simple Modality Alignment for Audio-Visual QA

Kyung Myung Ko
ko112@purdue.edu

Introduction

Can we understand the continuous scene without supervision of text?
=> Use contextualized representations to interact between the modalities!



Question: How many instruments are sounding in the video?
Answer: two
To answer the question, an AVQA model needs to first identify objects and sound sources in the video, and then count all sounding objects. Although there are three different sound sources in the audio modality, only two of them are visible. Rather than simply counting all audio and visual instances, exploiting audio-visual association is important for AVQA.

Different approach than the Proposal:

Contrastive Learning w/out pre-train -> Cross-attention contextualized alignment

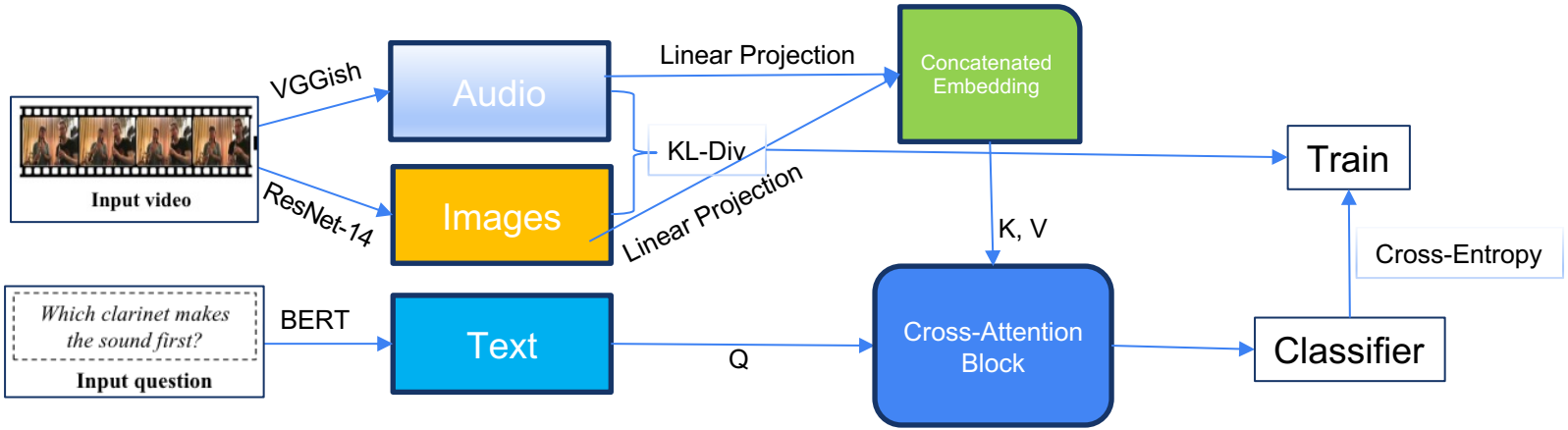
Learning to Answer Questions in Dynamic Audio-Visual Scenarios, Li et al, 2022 CVPR

Model

Initial Idea(contrastive learning) failed to implement due to limitation of accessing raw image/audio data. Performing perturbation at the encoded representation does not fit into the optimization category of maximizing mutual information.

Main Idea:

1. Cross-attention between the text representation & concatenated audio and visual representation
2. Maximize mutual information between audio & image representation by adding separate KL Div loss during training



Experiments and results

Trained with Adam Optimizer with learning rate 0.0001 for 40 epoch, using scheduler to drop by 0.1 every 10 epoch.
Utilized pre-layer norm transformer to stabilize the training.

Accuracy

	Audio			Visual			Audio-Visual						
	Count	Comp	Avg	Count	Location	Avg	Existential	Location	Count	Comp	Temporal	Avg	Total Avg
Baseline	**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**
Mine	64.44	60.17	62.31	55.51	57.75	56.63	**82.20**	40.72	44.00	62.98	40.34	54.05	57.67

F-1 Score

Accuracy			**0.57**	**9129**
Macro Avg	0.37	0.37	0.37	9129
Weighted Avg	0.55	0.55	0.55	9129

Analysis

Interesting Findings:

- Performance on temporal questions has inverse relationship to other questions based on my approach.
- Too much information from both modality may confuse the model to not answer audio-only and visual-only questions.

What I learned:

- Great amount of time spent to understand the attention mechanism with thorough experiments, pre-LN, post-LN, etc.

Future Tasks:

1. Other alignment methods, CCA, Soft-DTW for audio-visual alignment

2. Apply perturbation at the raw input level for more robustness