Audio Focused Multimodal Learning

A Seminar Report

Submitted in partial fulfillment of requirements for the degree of

Master of Technology

by

Shubham Nemani
203050011

under the guidance of

Prof. Preethi Jyothi
and
Prof. Ganesh Ramakrishnan

Department of Computer Science and Engineering

Indian Institute of Technology, Bombay

June 11, 2021

Acknowledgement

In the process of learning about Audio Focused Multimodal Learning in my seminar, **Prof. Preethi Jyothi** and **Prof. Ganesh Ramakrishnan** has played a vital role in guiding me the correct path to proceed on. Also, I would like to thank my seniors **Abhishek Thakur** and **Jayaprakash Akula** who helped in suggesting papers and how to dive into the field of Multi-Modal Video Analysis. I would also like to thank my fellow mates **Pranjal Saini** and **Jatin Lamba** for all the healthy discussions we had in the process. The constant support of my friends and families has greatly contributed to the success of this work.

Abstract

In Multi-modal video analysis, we focus on action recognition in video using audio as preview. In simple words, it tries to find what different activities are going on in video clips based on a predefined set of classes. The task is non-trivial as there are various complex issues in multimodal alignment. There are various approaches to solve this problem like knowledge distillation, contrastive learning, and curriculum learning approach. This report provides an overview of some of these approaches.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Multi-modal Learning

Our experience of the world is multimodal — we see objects, hear sounds, feel the texture, smell odors and taste flavors and then come up to a decision. Multimodal learning suggests that when a number of our senses — visual, auditory, kinesthetic — are being engaged in the processing of information, we understand and remember more. By combining these modes, learners can combine information from different sources [9].

A first fundamental step is learning how to represent inputs and summarizing the data in a way that expresses the multiple modalities. The heterogeneity of multimodal data makes it challenging to construct such representations.

A second step is to addresses how to translate data from one modality to another. Not only is the data heterogeneous, but the relationship between modalities is often open-ended or subjective.

Features need to be extracted from individual sources of information by building models that best suit the type of data. The next step is to combine information from two or more modalities to perform a prediction.

1.2 Self-supervised Learning

Given a task and enough labels, supervised learning can solve it really well. Good performance usually requires a decent amount of labels, but collecting manual labels is expensive and hard to be scaled up. Considering the amount of unlabelled data is substantially more than a limited number of human curated labelled datasets, it is kinda wasteful not to use them [10].

What if we can get labels for free for unlabelled data and train unsupervised dataset in a supervised manner? We can achieve this by framing a supervised learning task in a special form to predict only a subset of information using the rest. In this way, all the information needed, both inputs and labels, has been provided. This is known as self-supervised learning.

Self-supervised learning empowers us to exploit a variety of labels that come with the data for free.The self-supervised task, also known as pretext task, guides us to a supervised loss function. However, we usually don't care about the final performance of this invented task. Rather we are interested in the learned intermediate representation with the expectation that this representation can carry good semantic or structural meanings and can be beneficial to a variety of practical downstream tasks.

Chapter 2

Background

2.1 Knowledge Distillation

Knowledge distillation refers to the idea of model compression by teaching a smaller network, step by step, exactly what to do using a bigger already trained network. The 'soft labels' refer to the output feature maps by the bigger network after every convolution layer. The smaller network is then trained to learn the exact behavior of the bigger network by trying to replicate it's outputs at every level (not just the final loss) [11].

As seen in Fig [2.1], the highly complex teacher network is first trained separately using the complete dataset. This step requires high computational performance and thus can only be done offline (on high performing GPUs). While designing a student network, a correspondence needs to be established between intermediate outputs of the student network and the teacher network.

Pass the data through the teacher network to get all intermediate outputs and then apply data augmentation (if any) to the same. Now use the outputs from the teacher network and the correspondence relation to backpropagate error in the student network, so that the student network can learn to replicate the behavior of the teacher network.

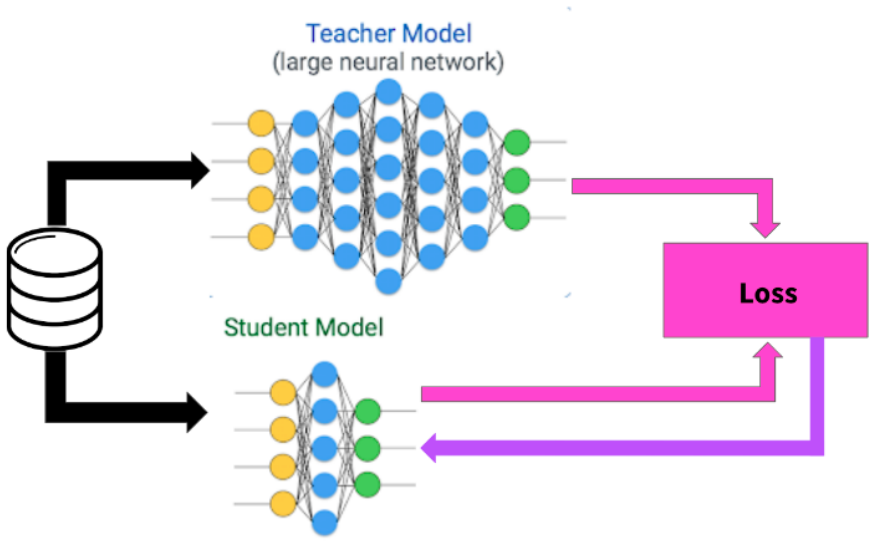

Figure 2.1: Knowledge Distillation

Chapter 3

Self-supervised Multi-modal Learning

It basically focuses on learning video representations using different modalities like visual, audio, text etc. in a self supervised manner.

3.1 Basic Architecture Framework

Video contains two modalities i.e. visual modality in frames and audio modality in audio. Frames are fed into a visual encoder to get visual features. Similarly audio is fed into audio encoder to extract audio features. Now these visual and audio features are jointly aligned in shared space using contrastive learning where i^{th} visual and audio features are more similar in comparision to j^{th} features. Figure [3.1] shows the basic architecture followed for self supervised multimodal learning.

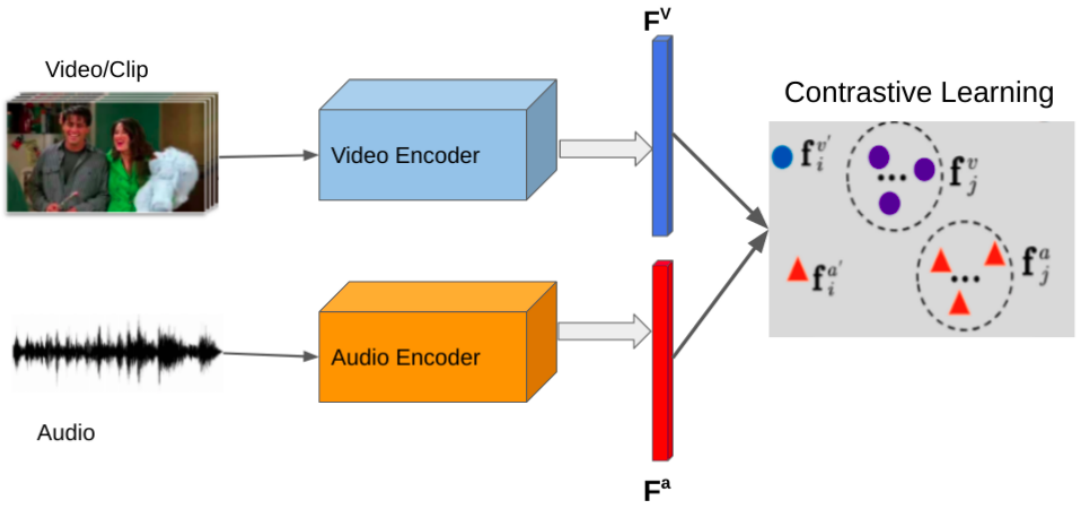

Figure 3.1: Multi-modal Learning Architecture

3.2 Listen To Look

Our goal is to perform accurate and efficient action recognition in long untrimmed videos [1]. Given a long untrimmed video V, the goal of video classification is to classify V into a predefined set of C classes. Because V can be very long, it is often intractable to process all the video frames together through a single deep network due to memory constraints. So the video is divided into N clips $\{V_1, V_2, ..., V_N\}$ which are taken

at a fixed hop size across the entire video. The final video-level prediction is obtained by aggregating the clip-level predictions of all N clips.

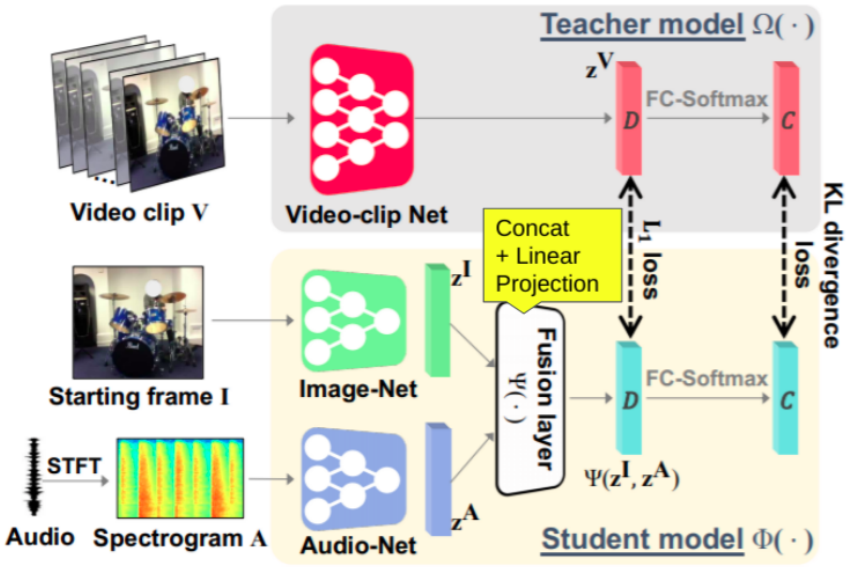

Figure 3.2: Listen to Look Architecture

3.2.1 Problem with previous approaches

But this is highly inefficient at two levels: (1) clip-level— within each short clip V, temporally close frames are visually similar, and (2) video-level—across all the clips in V, often only a few clips contain the key moments for recognizing the action.

3.2.2 Problem Formulation

Each video clip V is accompanied by an audio segment A. The starting frame I among the F frames within the short clip V usually contains most of the appearance cues already, while the audio segment A contains rich contextual temporal information. The idea is to replace the powerful but expensive clip-level classifier $\Omega(.)$ that takes F frames as input with an efficient image-audio classifier $\Phi(.)$ that only takes the starting frame I and its accompanying audio segment A as input, while preserving the clip-level information as much as possible.

$$\Omega(V_j) \approx \Phi(I_j, A_j), j \in \{1, 2, ..., N\}$$

3.2.3 Clip Level Preview

As shown in Fig. [3.2], the clip-based model takes a video clip V of F frames as input and based on that video volume generates a clip descriptor z^V of dimensionality D. A fully-connected layer is used to make predictions among the C classes in Kinetics. For the student model, using a two-stream network: the image stream takes the first frame I of the clip as input and extracts an image descriptor z^I ; the audio stream takes the audio spectrogram A as input and extracts an audio feature vector z^A. z^I and z^A are concatenated to generate an image-audio feature vector of dimensionality D using a fusion network $\Psi(.)$ that consists of two fully-connected layers. A final fully-connected layer is used to produce a C-class prediction like the teacher model.

3.2.4 Training Objective

The teacher model $\Omega(.)$ returns a softmax distribution over C classification labels. These predictions are used as soft targets for training weights associated with the student network $\Phi(.)$ using the following objective Eq. [3.1]:

$$\mathcal{L}_{KL} = - \sum_{\{(\mathbf{V},\mathbf{I},\mathbf{A})\}} \sum_{c} \Omega_c(\mathbf{V}) \log \Phi_c(\mathbf{I}, \mathbf{A}) \tag{3.1}$$

where $\Omega_c(V)$ and $\Phi_c(I, A)$ are the softmax scores of class c for the teacher model and the student model, respectively. We further impose an $\mathcal{L}_1$ loss on the clip descriptor z^V and the image-audio feature to regularize the learning process Eq. [3.2]:

$$\mathcal{L}_1 = \sum_{\{(\mathbf{z^V},\mathbf{z^I},\mathbf{z^A})\}} \left\| \mathbf{z^V} - \Psi\left(\mathbf{z^I}, \mathbf{z^A}\right) \right\|_1 \tag{3.2}$$

The final learning objective is a combination of these two losses Eq. [3.3]:

$$\mathcal{L} = \mathcal{L}_1 + \lambda * \mathcal{L}_{KL} \tag{3.3}$$

where is the weight for the KL divergence loss. The training is done over the image and audio student networks (producing representations z^I and z^A, respectively) and the fusion model $\Psi(.)$ with respect to a fixed teacher video-clip model.

Chapter 4

Contrastive Learning

Contrastive learning is a machine learning technique used to learn the general features of a dataset without labels by teaching the model which data points are similar or different [8].

In essence, contrastive learning allows our machine learning model to do the same thing Fig. [4.1]. It looks at which pairs of data points are "similar" and "different" in order to learn higher-level features about the data, before even having a task such as classification or segmentation.

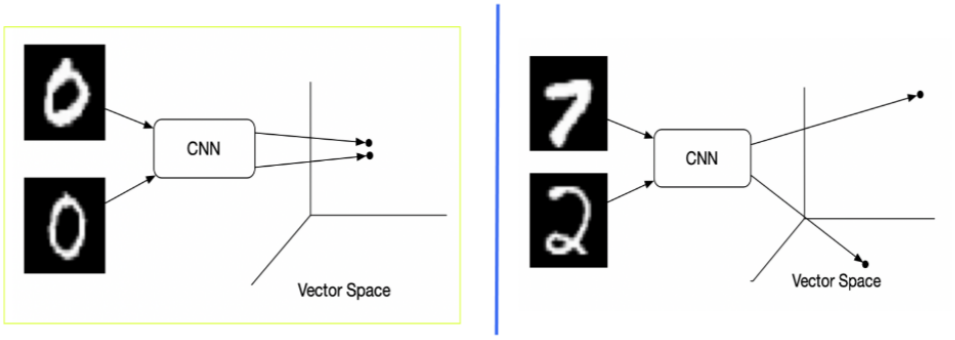

Figure 4.1: Contrastive Learning

4.1 Problem Formulation

The dataset consists of N video audio pairs, D = $\{(V_1, A_1), (V_2, A_2), ..., (V_N, A_N)\}$. As seen in the figure, each video V_i is fed into a video encoder which gives visual features v_i. Corresponding audio A_i is fed into audio encoder to extract audio features a_i.

Now contrastive learning tries to make v_i and a_i closer to each other rather than other j^{th} visual and audio features according to following loss Eq. [4.1]:

$$L_{ct} = -\log \frac{\exp\left(v_i^T.\bar{a}_i/\tau\right)}{\sum_{j=1}^{K+1}\exp\left(v_i^T.\bar{a}_j/\tau\right)} - \log \frac{\exp\left(a_i^T.\bar{v}_i/\tau\right)}{\sum_{j=1}^{K+1}\exp\left(a_i^T.\bar{v}_j/\tau\right)} \tag{4.1}$$

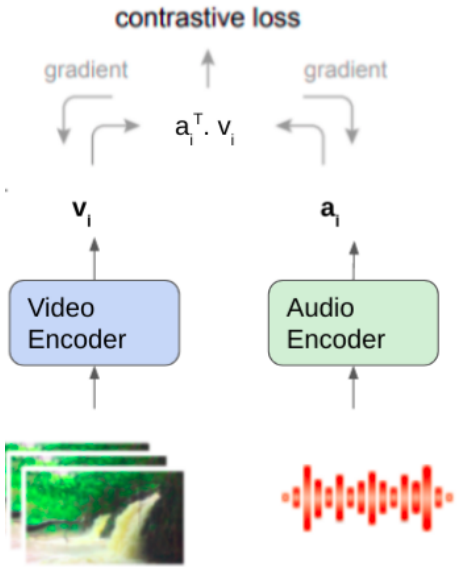

Figure 4.2: End to End Contrastive Learning

4.2 End to End Contrastvie Learning

The end-to-end update by back-propagation is a natural mechanism Fig. [4.2]. It uses samples in the current mini-batch. So the number of negatives is coupled with the mini-batch size, limited by the GPU memory size. It is also challenged by large mini-batch optimization. Some recent methods are based on pretext tasks driven by local positions, where the dictionary size can be made larger by multiple positions. But these pretext tasks may require special network designs such as patchifying the input or customizing the receptive field size, which may complicate the transfer of these networks to downstream tasks.

4.3 Memory Bank Approach

Another mechanism is the memory bank approach. A memory bank consists of the representations of all samples in the dataset. There are two separate memory banks for visual and audio features. The negatives can be sampled from memory bank for each minibatch without backpropagation, this reduces the computational complexity [7].

However, the representation of a sample in the memory bank was updated when it was last seen, so the sampled keys are essentially about the encoders at multiple different steps all over the past epoch and thus are less consistent. A momentum update is adopted on the memory bank according to equations Eq. [4.2] and Eq. [4.3]:

$$\bar{v_i} = (\lambda_v) * \bar{v_i} + (1 - \lambda_v) * v_i \tag{4.2}$$

$$\bar{a_i} = (\lambda_a) * \bar{a_i} + (1 - \lambda_a) * a_i \tag{4.3}$$

where λ_v and λ_a are momentum factor to be learned for visual and audio respectiveTrainingly.

4.4 Noise Contrastive Estimation

Like negative sampling, this is a technique for efficient learning when the number of output classes is large [4]. It can be seen like instance discrimination task where each video of dataset represent a different class, hence we have large number of classes. NCE can be used as unimodal (as in Eq. [4.4] and Eq. [4.5])as well as cross modal, Eq. [4.6]:

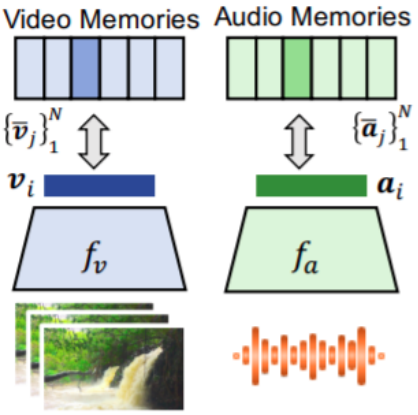

Figure 4.3: Unimodal NCE

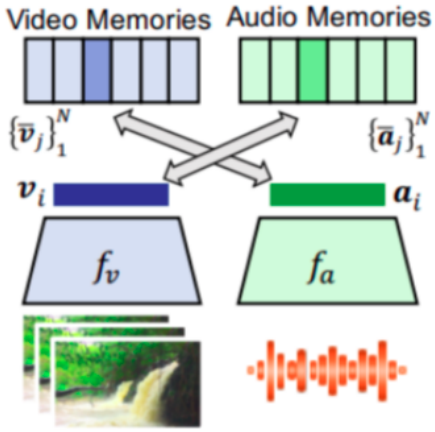

Figure 4.4: Crossmodal NCE

$$L_{visual}^{uni} = -\log \frac{\exp\left(v_i^T . \bar{v}_i/\tau\right)}{\sum_{j=1}^{K+1} \exp\left(v_i^T . \bar{v}_j/\tau\right)} \tag{4.4}$$

$$L_{audio}^{uni} = -\log \frac{\exp\left(a_i^T . \bar{a}_i/\tau\right)}{\sum_{j=1}^{K+1} \exp\left(a_i^T . \bar{a}_j/\tau\right)} \tag{4.5}$$

$$L_{audio-visual}^{cross} = -\log \frac{\exp\left(v_i^T . \bar{a}_i/\tau\right)}{\sum_{j=1}^{K+1} \exp\left(v_i^T . \bar{a}_j/\tau\right)} - \log \frac{\exp\left(a_i^T . \bar{v}_i/\tau\right)}{\sum_{j=1}^{K+1} \exp\left(a_i^T . \bar{v}_j/\tau\right)} \tag{4.6}$$

Chapter 5

Self Supervised Curriculum Learning

While most recent works focus on capturing the shared associations between the audio and visual modalities, they rarely consider multiple audio and video pairs at once and pay little attention to exploiting the valuable information within each modality [5].

Tackling this problem using a self-supervised audio-visual modality transfer framework termed SSCL to explore more coherent knowledge from a teacher network to a student network, where contrastive learning is leveraged to capture the correspondence between audio and visual information.

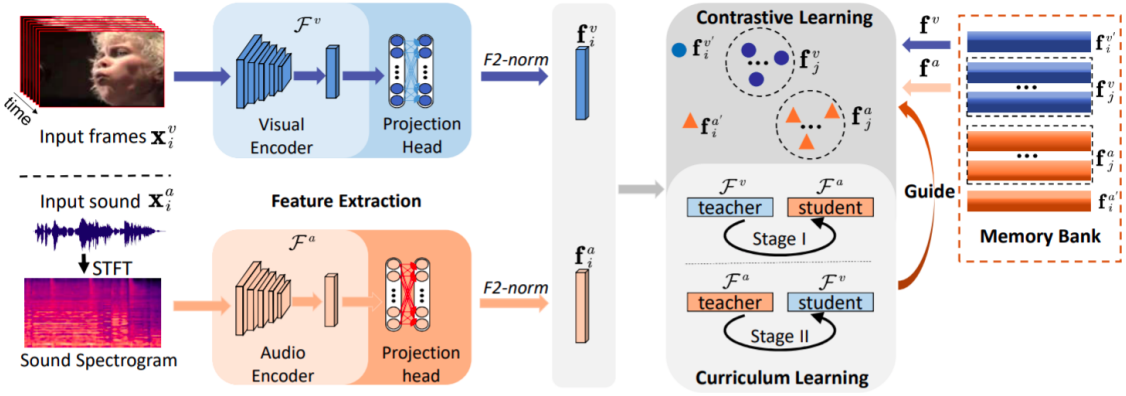

Figure 5.1: Curriculum Learning

5.1 Contribution

A two-stage curriculum learning process to reason about multiple single-modality instances and distill cross-modal correction information. This process not only improves the overall distillation performance but also regularizes the teacher and student model to generalize on noisy and complex scenarios.

5.2 Problem Formulation

The dataset consists of N video audio pairs, D = $\{(V_1, A_1), (V_2, A_2), ..., (V_N, A_N)\}$. As seen in the figure, each video V_i is fed into a video encoder which gives visual features v_i. Corresponding audio A_i is fed into audio encoder to extract audio features a_i.

5.3 Curriculum Learning

The curriculum learning strategy consists of two stages. From fig [5.1], in the stage-I, the student model i.e. audio encoder is fixed and only the parameters of the teacher model i.e. visual encoder are updated with a self instance discriminator Eq. [4.4], and then the teacher and student models are jointly trained according to Eq. [5.1]. While in the stage-II, exchange the role of teacher and student model i.e. now audio encoder will work as teacher and visual encoder as student. Audio encoder trained according to Eq. [4.5] and then both teacher and student are jointly trained using Eq. [5.2].

$$L_{audio}^{cross} = -\log \frac{\exp\left(a_i^T . \bar{v}_i/\tau\right)}{\sum_{j=1}^{K+1} \exp\left(a_i^T . \bar{v}_j/\tau\right)} \tag{5.1}$$

$$L_{video}^{cross} = -\log \frac{\exp\left(v_i^T . \bar{a}_i/\tau\right)}{\sum_{j=1}^{K+1} \exp\left(v_i^T . \bar{a}_j/\tau\right)} \tag{5.2}$$

Chapter 6

Compositional Contrastive Learning

The goal is to distill audio-visual knowledge learnt from heterogeneous audio and image modalities for video representation learning, while the cross-modal content may be semantically unrelated, the propose is to transfer knowledge across heterogeneous modalities, even though these data modalities may not be semantically correlated. Rather than directly aligning the representations of different modalities, audio, image, and video representations are composed across modalities to uncover richer multi-modal knowledge [2].

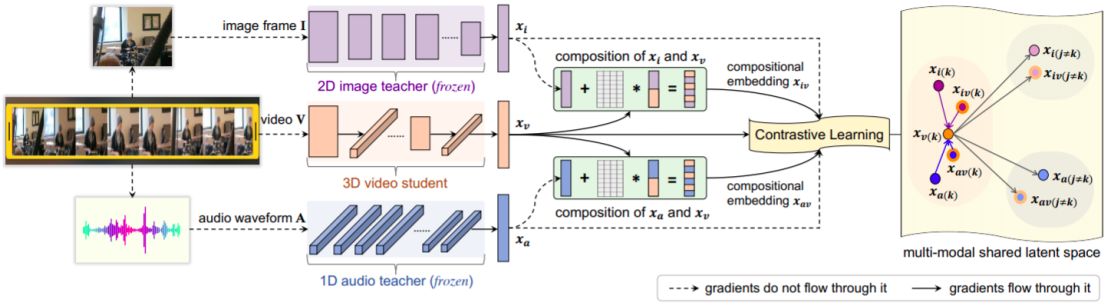

Figure 6.1: Compositional Contrastive Learning

6.1 Unimodal Representations of Audio and Vision

Given a dataset of N videos $D = \{V_i, y_i\}, i = \{1...N\}$ each video belongs to one of K video categories. As shown in Fig. [6.1], using 2 teacher networks, an image and audio teacher network. An image is randomly sampled from video and image features x_i are extracted using image teacher which is a 2D CNN network. Similarly, audio features x_a are extracted from audio teacher which is a 1D CNN network. Also, video is fed into a 3D CNN student encoder to extract video embedding x_v.

6.2 Compositional Multi-Modal Representations

The student and teacher embeddings may be semantically unaligned, to bridge the possible semantic gap and domain gap across modalities, the audio and image teacher embeddings are rectified by composing the teacher and student embeddings and constraining the compositional embeddings with task objective to close the possible semantic gap.

The teacher embeddings are recitfied using the following composition function $\mathcal{F}(.)$, which learns a

residual $f_\theta(.)$ that fuses two modalities by normalisation, concatenation and a linear projection:

$$\mathcal{F}_{av}\left(x_a, x_v\right) = x_{av} = x_a + f_{\theta_{av}}\left(x_a, x_v\right)$$
$$\mathcal{F}_{iv}\left(x_i, x_v\right) = x_{iv} = x_i + f_{\theta_{iv}}\left(x_i, x_v\right)$$

where x_{av}, x_{iv} are the compositional embeddings.

6.3 Classification Loss

The cross modal semantic gap is reduced by forcing the compositional embeddings to be classified into same class as that of video.

The classification loss is given by Eq. [6.1]:

$$\mathcal{L}_{cls} = \mathcal{L}_{ce}^{v}\left(x_v, k\right) + \mathcal{L}_{ce}^{av}\left(x_{av}, k\right) + \mathcal{L}_{ce}^{iv}\left(x_{iv}, k\right) \tag{6.1}$$

where $\mathcal{L}_{ce}^{m}\left(x_m, k\right)$ is negative log likelihood of classifying features x_m in class k.

6.4 Knowledge Distillation Loss

Given the unimodal embeddings and the compositional embeddings, we propose to distill the knowledge by pulling together the positive pairs while pushing away the negative pairs across modalities. The positive pairs could include the images, audios, and videos from the same video class k.

Specifically, for a triplet of audio, video, and their compositional embeddings extracted from the video V_i , the contrastive loss can be formed in every pair among them to reinforce their correspondence in the shared embedding space. Formally, a contrastive loss $\mathcal{L}_{ct}$ between a pair of audio and video embeddings $x_v(i), x_a(i)$ can be derived as below:

$$\mathcal{L}_{ct} = -\log \frac{\exp\left(\Phi\left(x_v(i), x_a(i)\right)/\tau\right)}{\sum_{j=1}^{B} \exp\left(\Phi\left(x_v(j), x_a(j)\right)/\tau\right)} = -\log p_{av}(i) \tag{6.2}$$

where Φ is a cosine similarity scoring function, τ is the temperature, $p_{av}(i)$ is the probability of assigning the video embedding $x_v(i)$ to its paired audio embedding $x_a(i)$ against the whole mini-batch of audio embeddings $x_a(j)$.

Therefore, we formulate a multi-class noise contrastive estimation (NCE) loss that brings the class label k into the loss formulation:

$$\mathcal{L}_{nce}\left(x_v, x_a\right) = -\frac{1}{B_p}\sum_{j=k}\log p_{av}(j) - \frac{1}{B_n}\sum_{j\neq k}\log\left(1 - p_{av}(j)\right) \tag{6.3}$$

where B_p, B_n are the number of positive pairs (from class k) and negative pairs (not from class k) for the video embedding x_v (labelled as k). Hence, the audio and image distillation loss is given by:

$$\mathcal{L}_a\left(x_v, x_a, x_{av}\right) = \lambda\mathcal{L}_{nce}\left(x_v, x_a\right) + (1-\lambda)\mathcal{L}_{nce}\left(x_v, x_{av}\right) \tag{6.4}$$

$$\mathcal{L}_i\left(x_v, x_i, x_{iv}\right) = \lambda\mathcal{L}_{nce}\left(x_v, x_i\right) + (1-\lambda)\mathcal{L}_{nce}\left(x_v, x_{iv}\right) \tag{6.5}$$

Chapter 7

Results

These are the results of accuracy for action recognition and sound recognition in video on different datasets.

1. **Listen to Look**

   - **Action Recognition**

Approach	Backbone	UCF101	ActivityNet
ListenToLook	ResNet-152	73.5	35.5
ListenToLook	R(2+1)D-152	82.5	47.0

Table 7.1: Action Recognition Benchmarks for Listen To Look

2. **Curriculum learning**

   - **Action Recognition**

Approach	Clip Size	UCF101	HMDB51
SSCL-stage-I	16x112x112	81.4	47.7
SSCL-stage-II	16x112x112	82.6	49.9
SSCL-stage-II	16x224x224	84.3	54.1
SSCL-stage-II	32x224x224	87.1	57.6

Table 7.2: Action Recognition Benchmarks for Curriculum Learning

   - **Sound Recognition**

Approach	Backbone	ESC-50	DCASE
SSCL-stage-I	2D-ResNet10	85.8	91.0
SSCL-stage-II	2D-ResNet10	88.3	93.0

Table 7.3: Sound Recognition Benchmarks for Curriculum Learning

3. **Compositional Learning**

   - **Action Recognition**

Approach	**UCF51**	**ActivityNet**
CCL-Audio Only	64.9	36.5
CCL-Image Only	69.1	46.3
CCL-Audio and Image	70.0	47.3

Table 7.4: Video Recognition Benchmarks for Compositional Contrastive Learning

Chapter 8

Conclusion and Future Work

We have discussed the importance of multimodal learning. We have also focused on efficiency for learning video representations, that can be used in wide variety of downstream tasks. How cross modal knowledge distillation helps in learning better features. We have explored both supervised and self supervised approaches for multimodal learning.

We can extend the idea of using self-supervised objectives in cases where we have the audio and the transcript. One can explore the idea of cross lingual setting, where audio and the transcript belong to different languages.

Bibliography

[1] R. Gao, T. Oh, K. Grauman, L. Torresani. "Listen to Look: Action Recognition by Previewing Audio". In CVPR, 2020.

[2] Yanbei Chen , Yongqin Xian , A. Sophia Koepke , Ying Shan, Zeynep Akata. "Distilling Audio-Visual Knowledge by Compositional Contrastive Learning". In CVPR, 2021.

[3] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, Andrew Zisserman. "Self-Supervised Learning of Audio-Visual Objects from Video". In CVPR, 2021

[4] Pedro Morgado, Nuno Vasconcelos, Ishan Misra. Audio-Visual Instance Discrimination with Cross-Modal Agreement. In CVPR-2021

[5] Jingran Zhang, Xing Xu, Fumin Shen, Huimin Lu, Xin Liu, Heng Tao Shen. Enhancing Audio-Visual Association with Self-Supervised Curriculum Learning. In AAAI-21.

[6] Ruohan Gao, Kristen Grauman. VISUALVOICE: Audio-Visual Speech Separation with Cross-Modal Consistency. In CVPR-2021.

[7] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. In CVPR-2020.

[8] https://towardsdatascience.com/understanding-contrastive-learning-d5b19fd96607

[9] https://lilianweng.github.io/lil-log/2019/11/10/self-supervised-learning.html.

[10] https://heartbeat.fritz.ai/introduction-to-multimodal-deep-learning-630b259f9291

[11] https://towardsdatascience.com/knowledge-distillation-simplified-dd4973dbc764