Multimodal Few-Shot Learning for Micro Videos

Shray Mathur
UT Austin
shray@utexas.edu

Ishan Nigam
UT Austin
ishann@cs.utexas.edu

1. Problem Statement

Multi-modal few-shot learning for micro videos[1] is a challenging research problem that aims to recognize class labels in video clips with a limited number of examples, leveraging semantic knowledge from latent information. Such information may take the form of visual self-supervision, information dependent on the presence of one or more modalities (audio/ video) and novel architectures. However, most existing methods for few-shot learning have primarily focused on images, and few works have addressed this problem for videos (Fig.1). In this work, we propose to explore strategies towards proposing a novel framework for multi-modal few-shot learning that integrates audio and visual modalities to recognize class labels in micro videos.

We propose (Fig.2) to work with raw audio and video sequences and obtain their embedding representations. Our method incorporates a semantic embedding network that maps audio and visual features into a common semantic space, where few-shot recognition can be performed by leveraging class attributes. Additionally, we explore the projection of audio and visual features into a common embedding space, allowing our multimodal few-shot learning approach to work even if one modality is missing at test time. Sec.3 details the experiments we plan to carry out.

We evaluate our framework on the recently released 3MASSIV dataset [3], which comprises of 50,000 labeled and 100,000 unlabeled micro video clips, allowing us to explore the effectiveness of various supervised and unsupervised techniques. Since the benchmark also provides multiple labels per micro video, we plan to first attempt supervised few-shot classification and then move towards multi-label classification. The opportunity to explore both supervised and unsupervised methods for multi-modal sequential data presents a rich source of data for considering a number of experiments. We hope to pursue inquiries along several different directions in an attempt to perform novel research on multi-modal few-shot micro video classification.

[1]We adopt the terminology "micro video" to refer to video sequences that have a total duration of 30s or less.

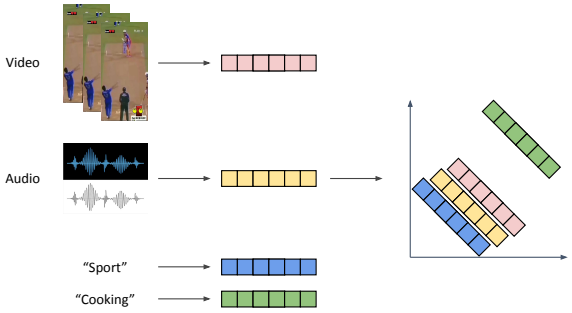

Figure 1. We study the problem of co-embedding (either or both) micro-video audio and video representations with class labels. We hope to explore this set-up under task ablations such as (1) low-shot learning, (2) cross-modal generalization (learning from one modality while testing on the other), and (3) missing modality @ test-time (which is common in real-world privacy-preserving scenarios). Sec. 3 elaborates upon the methods we plan to explore.

2. Related Work

2.1. Multi-Modal Co-Embedding Learning

Several prior works tackle the challenging task of few-shot learning for video classification using audio-visual inputs [7, 9, 12] by learning a shared embedding space for both modalities. A popular approach in these methods is to project the audio and visual features onto a common embedding space, which facilitates the learning of a mapping from the audio-visual input data to textual label embeddings. This enables the classification of samples from unseen classes, where at test time, the class with the closest word embedding to the predicted audio-visual output embedding is selected. [7, 9, 12] used temporally averaged features as inputs that were extracted from networks pretrained on video data. [8] propose a Temporal Cross-attention Framework (TCaF) which builds on [9] and additionally exploits temporal information by using temporal audio and visual data as inputs. We note that a significant element for our interest and subsequent exploration of co-embedding learning is Radford et al. [14].

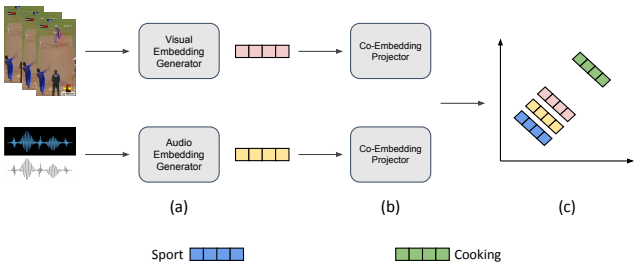

Figure 2. Proposed Method: (a) Raw audio and video sequences are transformed into embeddings using pre-trained networks. We intend to explore both pre-training these networks on unsupervised data as well as supervised fine-tuning on a subset of the labeled data. (b) The audio and video embeddings are projected into a common embedding space. This may be done using two-tower network architectures, but we intend to begin by exploring simple methods such as linear projections and shallow MLPs. (c) After projecting the audio and video embeddings into a common space (along with the labels) we intend to explore nearest neighbors, contrastive loss, and softmax classification to make the final prediction. NOTE: If time and resources permit, our final experiment will be to end-to-end learn (a), (b), and (c).

2.2. Self-supervised Visual Prior Learning

There is a vast and growing literature concerned with self-supervised visual learning [5, 6, 11]. Most notable of recent advances in this domain is the work by Chen et al. [1, 2], which proposes that the use of large-scale networks for pre-training and fine-tuning can enable surprising performance on few-shot visual tasks. He et al. [4] recently proposed to mask random patches of the input image and reconstruct the missing pixels as a form of self-supervision.

Both contrastive methods (such as SIMCLRv2 [2]) and generative methods (such as Masked Autoencoders [4]). Since we do not have the compute to explore self-supervised learning at scale, we intend to begin by using SIMCLRv2 pre-trained models and will consider Masked AutoEncoders based on time and compute constraints.

While a number of recent methods utilize attention for few-shot action recognition [10, 13, 15], we intend to focus on using self-supervised visual learning to improve our proposed method. Additionally, based on discussions with the Teaching Staff, we will attempt to generate synthetic data to improve our few-shot action classifiers.

3. Proposed Research

3.1. Data

The 3MASSIV dataset [3] consists of 50,000 expertly-annotated micro videos, with an average 20s duration, and an additional 100,000 unlabeled videos. The videos are drawn from the Moj social media platform and cover a wide range of popular short video trends such as pranks, fails, romance, and comedy. The micro video formats include self-shot videos, reaction videos, lip-syncs, among others. The dataset is multilingual, comprising videos in 11 different languages, and covers a diverse range of concepts, affective states, media types, and audio languages. The availability of the 3MASSIV dataset by Sharechat provides a valuable resource for researchers interested in developing machine learning models that can operate effectively in real-world social media platforms, where multi-modal content is prevalent.

Since we expect to be limited by compute resources, the entire 50,000 labeled video data will not be used for training. Instead, we will create a few-shot benchmark using approximately 5,000 video sequences.

3.2. Task

Few Shot Multi-Modal Learning Since it has not been as well-explored as few-shot image learning, the first task we focus on is few-shot multi-modal learning. Admittedly, we are also interested in working on this task due to its (lack of) demand on compute resources. Possible ablations include (1) unsupervised finetuning, and (2) transfer learning.

Cross-Modality Generalization As discussed with the Teaching Staff, cross-modal generalization is an interesting task from the perspective of privacy-preserving mobile device visual learning. We propose to explore: (1) learning from both modalities and testing using only one, and (2) learning from one modality and testing on the other.

3.3. Methods

Fig.2 summarizes our proposed methodology. In this section, we elaborate upon the details. (a) As a first step, we plan to directly use pre-trained audio and video networks to obtain embeddings for the raw sequential data. Next, we will explore the fine-tuning of these pre-trained embedding generators on unlabeled 3MASSIV data. (b) We plan to learn small (on the order of a few layers) transformers on the embedding sequences to obtain representations that are projected into a common embedding[2]. (c) The common embedding space will jointly optimize the audio and video representations with their class labels. We expect to experiment with a wide variety of contrastive learning methods (vanilla triplet loss ... → ... SIMCLRv2).

Implementation Notes **(1)** (a) and (b) may be jointly learned using two-tower network architectures. We will consider this step depending on time and resource constraints. **(2)** End-to-end learning (a), (b), and (c) seems quite intuitive, but we expect compute to be a limitation. **(3)** It is likely that (b) and (c) also need to be learnt jointly. Our current plan is to quickly implement this 3-part baseline and then merge pairs of the pipeline parts.

[2]This will be an interesting implementation exercise for our edification since neither of us has ever trained a transformer from scratch.

References

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[2] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020. 2

[3] Vikram Gupta, Trisha Mittal, Puneet Mathur, Vaibhav Mishra, Mayank Maheshwari, Aniket Bera, Debdoot Mukherjee, and Dinesh Manocha. 3massiv: multilingual, multimodal and multi-aspect dataset of social media short videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21064–21075, 2022. 1, 2

[4] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2

[5] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020. 2

[6] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):857–876, 2021. 2

[7] Pratik Mazumder, Pravendra Singh, Kranti Kumar Parida, and Vinay P Namboodiri. Avgzslnet: Audio-visual generalized zero-shot learning by reconstructing label features from multi-modal embeddings. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3090–3099, 2021. 1

[8] Otniel-Bogdan Mercea, Thomas Hummel, A Sophia Koepke, and Zeynep Akata. Temporal and cross-modal attention for audio-visual zero-shot learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XX*, pages 488–505. Springer, 2022. 1

[9] Otniel-Bogdan Mercea, Lukas Riesch, A Koepke, and Zeynep Akata. Audio-visual generalised zero-shot learning with cross-modal attention and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10553–10563, 2022. 1

[10] Ashish Mishra, Vinay Kumar Verma, M Shiva Krishna Reddy, S Arulkumar, Piyush Rai, and Anurag Mittal. A generative approach to zero-shot and few-shot action recognition. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 372–380. IEEE, 2018. 2

[11] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6707–6717, 2020. 2

[12] Kranti Parida, Neeraj Matiyali, Tanaya Guha, and Gaurav Sharma. Coordinated joint multimodal embeddings for generalized audio-visual zero-shot classification and retrieval of videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3251–3260, 2020. 1

[13] Toby Perrett, Alessandro Masullo, Tilo Burghardt, Majid Mirmehdi, and Dima Damen. Temporal-relational crosstransformers for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 475–484, 2021. 2

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[15] Hongguang Zhang, Li Zhang, Xiaojuan Qi, Hongdong Li, Philip HS Torr, and Piotr Koniusz. Few-shot action recognition with permutation-invariant attention. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 525–542. Springer, 2020. 2