Multimodal Clustering Networks for Self-supervised Learning from Unlabeled Videos

Brian Chen [1] Andrew Rouditchenko[2] Kevin Duarte [3] Hilde Kuehne[4] Samuel Thomas[4,5]
Angie Boggust[2] Rameswar Panda[4,5] Brian Kingsbury[4,5] Rogerio Feris[4,5]
David Harwath[6] James Glass[2] Michael Picheny[7] Shih-Fu Chang[1]

[1]Columbia University, [2]MIT CSAIL, [3]University of Central Florida,
[4]IBM Research AI, [5]MIT-IBM Watson AI Lab, [6]UT Austin, [7]NYU-Courant CS & CDS,
{bc2754,sc250}@columbia.edu,{roudi,aboggust,glass}@mit.edu,kevin_duarte@knights.ucf.edu
{kuehne,rpanda}@ibm.com,{sthomas,rsferis,bedk}@us.ibm.com, harwath@cs.utexas.edu, map22@nyu.edu

Abstract

Multimodal self-supervised learning is getting more and more attention as it allows not only to train large networks without human supervision but also to search and retrieve data across various modalities. In this context, this paper proposes a self-supervised training framework that learns a common multimodal embedding space that, in addition to sharing representations across different modalities, enforces a grouping of semantically similar instances. To this end, we extend the concept of instance-level contrastive learning with a multimodal clustering step in the training pipeline to capture semantic similarities across modalities. The resulting embedding space enables retrieval of samples across all modalities, even from unseen datasets and different domains. To evaluate our approach, we train our model on the HowTo100M dataset and evaluate its zero-shot retrieval capabilities in two challenging domains, namely text-to-video retrieval, and temporal action localization, showing state-of-the-art results on four different datasets.

1. Introduction

To robustly learn visual events and concepts, humans seldom rely on visual inputs alone. Instead, a rich multimodal environment is utilized for understanding by combining multiple sensory signals along with various language representations. In this work, we focus on the problem of learning a joint embedding space across multiple modalities. Given that the features from different modalities are often not comparable, the goal is to learn the projections into a common space where features from different domains but with similar content are close to each other to allow for a direct retrieval across modalities.

To deal with multimodal data of this nature, several recent approaches use a contrastive loss to learn e.g. feature representations in a joint embedding space. The goal is to bring samples drawn from the same temporal instance closer to each other while keeping samples from different times apart. One problem arising from the contrastive loss is that this criterion does not consider the samples' semantic structure

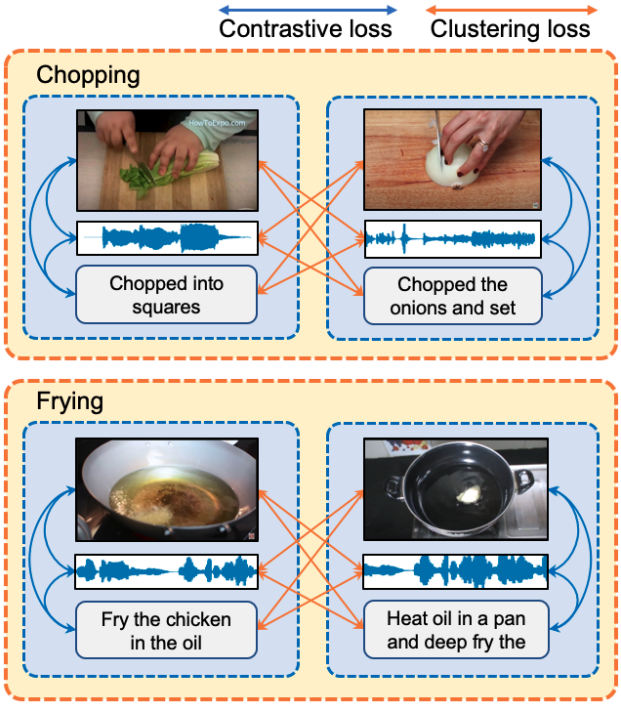

Figure 1: The Multimodal Clustering Network (MCN) combines a contrastive loss that learns feature representations to be close across different modalities such as video, audio, and text (blue box), with a clustering loss that draws instances that are semantically related together, e.g., scenes depicting the same semantic concept (e.g., chopping or frying) from different videos or different clips. (yellow box).

and similarity at different times: two samples are treated as a negative pair as long as they occur at different times regardless of their semantic similarity. This can have a considerable adverse impact on the learned representation. In a different formulation for learning representations, instead of comparing individual instances, clusters of instances are first created using a certain clustering algorithm [1]. This approach encourages samples semantically similar to each other (namely, samples in the same cluster) to be close in the embedding space. However, if we cluster features from

multi-modalities, those clusters would likely emerge only within the modalities separately, clustering audio instances with audio instances, visuals to visuals *etc*. Therefore, a mechanism that pulls the instances from different modalities together is crucial to cluster features from different modalities in a joint space. This leads to our proposed method *Multimodal Clustering Network* (MCN) that treats these two approaches as reciprocal information. Figure 1 provides a high-level overview of our approach.

To evaluate our proposed method, we address the challenging problem of zero-shot learning in two contexts: multimodal video retrieval and multimodal temporal action localization. We train our system on the HowTo100M dataset[9] and evaluate on various downstream tasks. MCN significantly outperforms the best text-to-video retrieval baseline over absolute 3% in recall and outperforms the temporal action localization baseline over 3.1% in recall, both in zero-shot settings.

2. Learning to Cluster Multimodal Data

To effectively construct a *joint representation space* from unlabeled narrated videos, we start with n narrated video clips. Each video clip is associated with its corresponding visual representation, audio representation and text narration. Given this input, the joint embedding space is learned, where the embeddings of video clips with semantically similar visual, audio, and text content are close to each other and apart when the content is dissimilar, as illustrated in Figure 1.

Using notation as in [10], let denote video $v \in \mathcal{V}$ as it's corresponding visual representation, let $a \in \mathcal{A}$ denote its corresponding audio and $t \in \mathcal{T}$, its matching text narration generated using an automatic speech recognition (ASR) system. Given a set of n tuples of associated video, audio and text narrations $\{(v_i, a_i, t_i)\}_{i=1}^{n} \in (\mathcal{V}\times\mathcal{A}\times\mathcal{T})^n$, as shown in Figure 2 (a), we first construct three parametrized mappings that derive embedding representations from the original video, audio and text signals. Transform $f : \mathcal{V} \rightarrow \mathbb{R}^d$ derives a d-dimensional embedding representation $f(v) \in \mathbb{R}^d$ from a video clip v, transforms $g : \mathcal{A} \rightarrow \mathbb{R}^d$ and $h : \mathcal{T} \rightarrow \mathbb{R}^d$, produce similar d-dimensional audio and text embeddings: $g(a) = z \in \mathbb{R}^d$ and $h(t) \in \mathbb{R}^d$. More details about model architectures are in Section 3.

Next, we introduce three loss functions to guide and properly situate these embeddings in the joint embedding space. The final model is trained to minimize sum of these losses.

$$L = L_{MMS} + L_{Cluster} + L_{Reconstruct} \tag{1}$$

2.1. Contrastive Loss for Learning Joint Spaces

To learn a joint space for the three modalities, we compute a contrastive loss on all pairs of modalities, $(v, t), (t, a), (a, v)$, as shown in Figure 2 (b). This loss maximizes the similarity between representations corresponding to any two modalities from the same instance (video clip) while minimizing the similarity of imposter pairs from the two modalities from one clip of video to another. In this work, we use the Masked Margin Softmax (MMS) function [6], which learned embedding vectors' dot product within a batch B. Features from each of the three modalities $\{V, A, T\}$ are assembled for each batch. The total contrastive loss L_{MMS} is the sum of pairwise losses using each of the three modalities:

$$L_{MMS} = L_{ta} + L_{vt} + L_{va} \tag{2}$$

where L_{ta}, L_{vt}, L_{va} represent the loss associated with pairwise modalities $(t, a), (v, t), (a, v)$ respectively. For a pair of modalities, for example the text and audio modalities, the individual loss L_{ta} is in turn given as:

$$L_{ta} = -\tfrac{1}{B}\sum_{i=1}^{B}\left[\left(\log\frac{e^{h(\mathbf{t}_i)\cdot g(\mathbf{a}_i)-\delta}}{e^{h(\mathbf{t}_i)\cdot g(\mathbf{a}_i)-\delta}+\sum\limits_{\substack{k=1\\k\neq i}}^{B} e^{h(\mathbf{t}_k^{imp})\cdot g(\mathbf{a}_i)}}\right)\right. \tag{3}$$
$$\left.+\left(\log\frac{e^{h(\mathbf{t}_i)\cdot g(\mathbf{a}_i)-\delta}}{e^{h(\mathbf{t}_i)\cdot g(\mathbf{a}_i)-\delta}+\sum\limits_{\substack{j=1\\j\neq i}}^{B} e^{h(\mathbf{t}_i)\cdot g(\mathbf{a}_j^{imp})}}\right)\right]$$

where a_j^{imp} represents imposter pairs from two modalities that are sampled from a batch but do not co-occur. By projecting all features to the same space and ensuring that the features across different modalities are comparable.

2.2. Clustering Multimodal Features

To ensure that representations of semantically related instances are close in the learned joint multimodal space, a clustering step is included as part of the training process.

Online K-means clustering. We applied standard clustering algorithm k-means that takes a set of vectors as input, in our case, the features M produced by the fused multimodal feature:

$$M = (f(\mathbf{v}) + g(\mathbf{a}) + h(\mathbf{t}))/3 \tag{4}$$

where we take the mean over features from three modalities to represent a multimodal instance. We cluster them into k distinct groups. More precisely, it outputs a $d \times k$ centroid matrix $C = \{\mu_1, .., \mu_k\}$ and the cluster assignments y_n of each multimodal instance n are defined by solving the following problem:

$$\min_{C\in\mathbb{R}^{d\times k}} \frac{1}{N}\sum_{n=1}^{N}\min_{y_n\in\{0,1\}^k} \|M_n - Cy_n\|_2^2 \tag{5}$$

We then acquire a centroid matrix C^* and a set of assignments $(y_n^*)_{n\leq N}$.

Semantic centroid learning. To learn the features closer to its multimodal semantic centroids. We proposed to use the

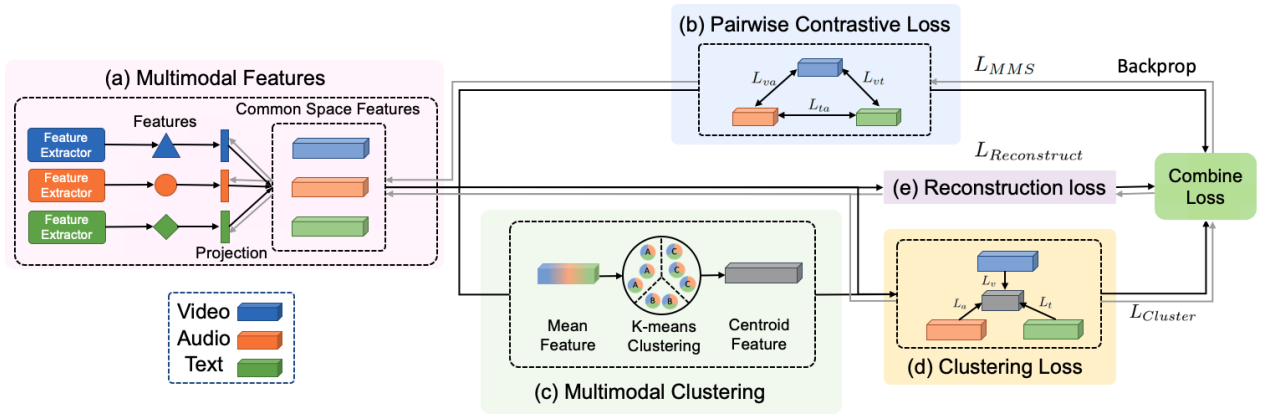

Figure 2: **Illustration of our proposed framework.** Our framework comprises four parts: (a) Extracting features from several modalities and projecting them into joint space. (b) Calculating contrastive loss pairwise to pull the features close across modalities. (c) Performing multimodal clustering across features from different domains in a batch. (d) Performing joint prediction across features to multimodal centroids to bring together semantically similar embeddings. (e) Reconstruction loss for regularization. Best viewed in color.

centroid as a contrastive loss reference target. This target pulls the features from three modalities closer to the centroid that is close to their multimodal instance feature M_n and pushes the features far away from the other centroid. For each modality, for example, the text modalities, the individual loss L_t is in turn given as:

$$L_t = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{e^{h(\mathbf{t}_i)\cdot\mu'-\delta}}{\sum_{k=1}^{K}e^{h(\mathbf{t}_i)\cdot\mu_k}} \tag{6}$$

where μ' is the nearest centroid for the multimodal instance feature M_i and μ'. We later sum over the loss from three modalities:

$$L_{Cluster} = L_v + L_a + L_t \tag{7}$$

In the end, the projected features learn to be closer to its centroid feature among the three and also learns to be closer in similar semantics.

Multimodal features reconstruction. We performed a reconstruction loss on top of the common space features from three modalities to stabilize the feature training during clustering. For each modality, for example, the visual modalities, the individual loss $L_{v'}$ is in turn given as:

$$L_{v'} = -\frac{1}{B}\sum_{i=1}^{B}\|f'(\mathbf{v}) - f(\mathbf{v})\|^2 \tag{8}$$

where $f'(\mathbf{v})$ represented the reconstructed features by feeding $\mathbf{v}$ into two linear layers as encoder and decoder. We then sum the loss over each modality:

$$L_{Reconstruct} = L_{v'} + L_{a'} + L_{t'} \tag{9}$$

3. Experiments

3.1. Implementation details

For the visual branch of the proposed MCN model we follow [9] and use 2D features from a ResNet-152 model [5], along with features from a ResNeXt-101 model [3]. For the audio branch of the network, we compute log-mel spectrograms and use a pre-trained DAVEnet model [4] to extract audio features. For the textual branch, the feature extraction process proposed in [9] is adopted to extract text representations: a GoogleNews pre-trained Word2vec model provides word embeddings., followed by a max-pooling over words in a given sentence to extract a sentence embedding.

3.2. Datasets

Training Dataset. Our models are trained on the HowTo100M [9] instructional video dataset, which contains 1.2M videos along with their corresponding audio that consists of speech and environmental sound and automatically generated speech transcriptions.

Downstream Datasets. For text-to-video retrieval, we evaluate our representations on the following two datasets. The **YouCook2** [11] dataset contains 3.5K cooking instruction video clips with text descriptions collected from YouTube. The **MSR-VTT** [2] dataset contains 10K human annotated video clip-caption pairs on various topics. For temporal action localization, the following two datasets were evaluated: The **CrossTask** [13] dataset contains 2.7K instructional videos that cover various topics with manual annotation for each frame. The **Mining Youtube** [8] dataset contains 250 cooking videos, 50 of each task, that are densely annotated.

Method	Mod	Model	TR	YouCook2 R@1	YouCook2 R@5	YouCook2 R@10	MSRVTT R@1	MSRVTT R@5	MSRVTT R@10
Random		-	-	0.03	0.15	0.3	0.01	0.05	0.1
Miech [9]	VT	R152+RX101	N	6.1	17.3	24.8	7.2	19.2	28.0
MIL-NCE* [10]	VT	R152+RX101	N	8.1	23.3	32.3	8.4	23.2	32.4
MCN (ours)	VAT	R152+RX101	N	**18.1**	35.5	45.2	**10.5**	**25.2**	**33.8**
MMV FAC [7]	VAT	TSM-50x2	Y	11.7	33.4	45.4	9.3	23.0	31.1
MIL-NCE [10]	VT	S3D-G	Y	15.1	**38.0**	**51.2**	9.9	24.0	32.4

Table 1: Comparison of text-to-video retrieval systems. Mod indicates modality used, where V: video, A: audio, T: text. TR indicates if a trainable backbone is used or not.

Method	Mod	Model	TR	CrossTask Recall	CrossTask IOD	CrossTask IOU	MYT Recall	MYT IOD	MYT IOU
Miech [9]	VT	R152+RX101	N	33.6	26.6	17.5	15.0	17.2	11.4
MIL-NCE* [10]	VT	R152+RX101	N	33.2	30.2	16.3	14.9	26.4	17.8
MCN (ours)	VAT	R152+RX101	N	35.1	**33.6**	**22.2**	**18.1**	**32.0**	**23.1**
ActBERT [12]	VT	R101+Res3D	N	37.1	-	-	-	-	-
ActBERT [12]	VT	+ Faster R-CNN	N	**41.4**	-	-	-	-	-
MIL-NCE [10]	VT	S3D-G	Y	40.5	-	-	-	-	-

Table 2: Evaluation of temporal action localization systems.

3.3. Downstream Tasks

Text-to-Video Retrieval. The goal of this task is to retrieve the matching video from a pool of videos, given its ground truth text query description. The model is tested on two video description datasets and evaluated on recall metrics: R@1, R@5, R@10. These evaluations are used to demonstrate the effectiveness of the contrastive loss and learned joint embedding space across three modalities.

Temporal action localization. The CrossTask [13] dataset considers the task of clip level action detection. The performance is reported as recall and computed as a ratio of the correctly predicted clips over the total number of clips in the video as used in [13]. The MiningYoutube [8] dataset considers the task of frame-level temporal action segmentation. Here, each test video is provided together with the respective actions and their ordering, including the background. The goal is to find the correct frame-wise segmentation of the video given the action order. More information of the metric and be found in the paper [8].

3.4. Comparison with State-of-the-art Methods

Zero-shot Video Retrieval. We first examine the results of the text-to-video retrieval task (Table 1). To allow comparability between different approaches, we use a fixed visual feature extraction backbone as described in [9] whenever possible. For the baseline MIL-NCE* [10], we apply their training strategy on the same visual feature set we use.On YouCook2, our model significantly outperforms prior works on the same architecture and shows even competitive results compared to models with trainable visual backbone (TR).

Zero-shot Action Localization. We examine the action localization tasks in Table 2. Given each frame in the video, we perform a zero-shot classification of the given labels and calculate the recall. For CrossTask, our method outperforms state-of-the-art approaches for self-supervised learning [10, 9] . We also evaluate our model on the MiningYoutube [13] temporal action localization benchmark. Our method outperforms state-of-the-art approaches for both self-supervised [10, 9] and weakly supervised [8] learning.

4. Conclusions

We have developed a novel self-supervised multimodal clustering network that learns a common embedding space by processing local (via a contrastive loss) and global (via a clustering loss) semantic relationships present in multimodal data. The multimodal clustering network is trained on a large corpus of narrated videos without any manual annotations. Our extensive experiments on multiple datasets show that creating a joint video-audio-language embedding space with a clustering loss is essential for self-supervised learning of good video representations. Our approach can be extended to more modalities such as optical flow or sentiment features and applied to other multimodal datasets for learning joint representation spaces without human annotation.

References

[1] M. Caron et al. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint*, 2020. 1

[2] J. X. et al. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*, 2016. 3

[3] K. Hara et al. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *CVPR*, 2018. 3

[4] D. Harwath et al. Jointly discovering visual objects and spoken words from raw sensory input. In *ECCV*, 2018. 3

[5] K. He et al. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[6] G. Ilharco et al. Large-scale representation learning from visually grounded untranscribed speech. In *CoNLL*, 2019. 2

[7] A. Jean-Baptiste et al. Self-supervised multimodal versatile networks. *arXiv preprint*, 2020. 4

[8] H. Kuehne et al. Mining youtube-a dataset for learning fine-grained action concepts from webly supervised video data. *arXiv preprint*, 2019. 3, 4

[9] A. Miech et al. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *CVPR*, 2019. 2, 3, 4

[10] A. Miech et al. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*, 2020. 2, 4

[11] L. Zhou, X. Chenliang, and J. J. Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018. 3

[12] L. Zhu et al. Actbert: Learning global-local video-text representations. In *CVPR*, 2020. 4

[13] D. Zhukov et al. Cross-task weakly supervised learning from instructional videos. In *CVPR*, 2019. 3, 4