Weakly Paired Associative Learning for Sound and Image Representations via Bimodal Associative Memory

Sangmin Lee[1] Hyung-Il Kim[2] Yong Man Ro[1*]
[1] Image and Video Systems Lab, KAIST [2] ETRI
{sangmin.lee,ymro}@kaist.ac.kr hikim@etri.re.kr

Abstract

Data representation learning without labels has attracted increasing attention due to its nature that does not require human annotation. Recently, representation learning has been extended to bimodal data, especially sound and image which are closely related to basic human senses. Existing sound and image representation learning methods necessarily require a large number of sound and image with corresponding pairs. Therefore, it is difficult to ensure the effectiveness of the methods in the weakly paired condition, which lacks paired bimodal data. In fact, according to human cognitive studies, the cognitive functions in the human brain for a certain modality can be enhanced by receiving other modalities, even not directly paired ones. Based on the observation, we propose a new problem to deal with the weakly paired condition: How to boost a certain modal representation even by using other unpaired modal data. To address the issue, we introduce a novel bimodal associative memory (BMA-Memory) with key-value switching. It enables to build sound-image association with small paired bimodal data and to boost the built association with the easily obtainable large amount of unpaired data. Through the proposed associative learning, it is possible to reinforce the representation of a certain modality (e.g., sound) even by using other unpaired modal data (e.g., images).

1. Introduction

Data representation learning without labels is to learn general features from unlabeled data by exploiting automatically generated supervisory signals within the data. Since it is highly time-consuming and labor-intensive for people to annotate large-scale data manually, such representation learning methods have received increasing attention in industry and research fields. In this context, representation learning has been applied to various areas such as computer vision [10, 15, 17], natural language processing [7, 12], and sound signal processing [4, 39].

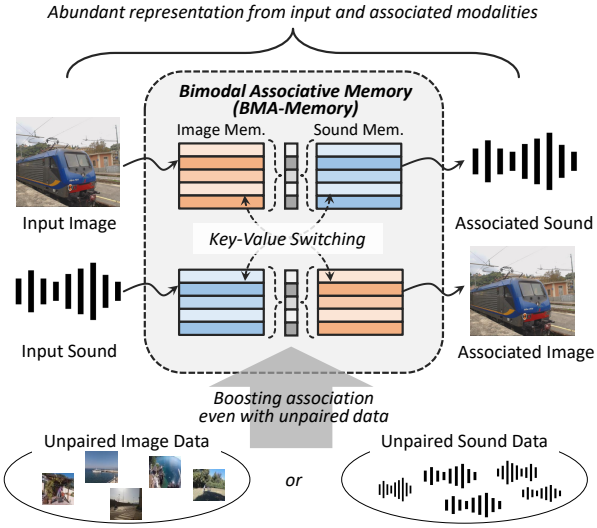

Figure 1. Concept of the proposed framework. The model can associate one modality (*e.g.*, sound) with a different modality (*e.g.*, image) through BMA-Memory to obtain abundant representations. Unpaired modality can boost the association between modalities.

Recently, as data samples are acquired in various multisensory environments, representation learning methods for bimodal data have been proposed. They aimed to learn feature representation from exploiting correspondence between bimodal data. In particular, many bimodal representation learning methods investigated the correspondence between auditory and vision which are closely related to basic human senses. These methods mainly attempted to learn bimodal representations from audio-video [26, 31] or sound-image [34, 38] data without labels. However, existing bimodal representation learning methods require a large number of data with corresponding pairs. Therefore, it is difficult to ensure the effectiveness of the methods in the weakly paired condition, which lacks paired bimodal data.

According to neurobiological studies, the cognitive functions related to a certain modality can be enhanced by re-

*Corresponding author

ceiving other modal stimuli in the human brain. There are several cases such as visual stimuli to multisensory cognition [41], auditory stimuli to visual cognition [5], and tactile stimuli to visual cognition [18]. It is possible because humans memorize multisensory modalities and associate them with each other in their brains. Bimodal cognitive functions are closely connected and influenced by each other.

Based on the observation, we introduce a new problem to deal with the weakly paired condition: *How to boost a certain modal representation even by using other unpaired modal data*, which has not been properly addressed in previous works. It is needed to devise such a method in order to extend and generalize bimodal representation learning as the human brain. In terms of sound-image data, we can expect to enhance the image representation even from unpaired sound data and vice versa. Based on this context, we focus on the representations of sound-image level rather than audio-video level because the weakly paired condition is more naturally observed in sound-image data. For example, we can acquire lots of animal images easily by web searching, whereas it is difficult to obtain animal sound data. In such weakly paired condition, it is worth to reinforce difficult-to-obtain modal (*e.g.*, sound) representation from other easy-to-obtain modal data (*e.g.*, image).

In this paper, we propose a novel bimodal associative memory (BMA-Memory) which enables to learn sound and image representations. BMA-Memory can store bimodal features in sound-image sub-memories and associate with one another naturally through a key-value switching scheme. Since another modality can be recalled through BMA-Memory, we can obtain abundant representation that includes both input and associated modalities from single modal input. Based on the memory, we introduce weakly paired associative learning to address weakly paired condition, which lacks paired data. BMA-Memory enables to build the sound-image association with small paired bimodal data and to boost the built association with the easily obtainable large amount of unpaired modal data. In unpaired associative learning, we construct pseudo bimodal pairs from unpaired data to enhance the bidirectional association. As a result, the representation of certain modality can be enhanced even by using other unpaired modal data. The concept of the proposed approach is shown in Figure 1.

The major contributions of the paper are as follows.

- We introduce a novel BMA-Memory with key-value switching to learn sound and image representations. It stores bimodal sound-image features and associates with one another. It enables to obtain abundant representations including both input and associated modalities even from single modal input.
- We propose weakly paired associative learning to address the weakly paired condition. It effectively enables to deal with boosting certain modal representation even by using other unpaired modal data in the weakly paired condition.

2. Related Work

2.1. Bimodal Representation Learning

Data representation learning without labels is to learn features from unlabeled data by using automatically generated supervisory signals within the data. To learn representations, pretext tasks are defined to train the model in self-supervised manners. Various pretext tasks have been investigated to utilize the structural properties of data. Such methods include rotation prediction [15], spatial context prediction [13], and jigsaw puzzle [33]. In recent years, representation learning methods with contrastive learning have shown remarkable effectiveness in learning image representations [10, 20]. The representations learned from pretext tasks are evaluated through other downstream tasks such as classification and retrieval.

Recently, representation learning methods have been extended to bimodal data as data samples are acquired in various multi-sensory environments [11, 22, 26, 31, 34, 38, 42, 44]. In particular, sound and vision which are closely related to basic human senses have been significantly investigated. These methods mainly attempted to learn bimodal representation from audio-video or sound-image data. In the case of audio-video data, Korbar *et al.* [26] proposed a representation learning method considering temporal synchronization of audio-video pairs. Alwassel *et al.* [1] introduced bimodal representation learning with audio-video clustering. In [32], a representation learning method for audio-video data was proposed by exploring bimodal agreement which groups together potentially paired multiple instances as positives. The attempts have been made to utilize bimodal correspondence between sound and image data as well. With sound-image data, Owens *et al.* [34] introduced representation learning which predicts the corresponding sounds from images. Senocak *et al.* [40] proposed an algorithm with learning sound-image representations for sound localization in images. In [38], the authors introduced enhanced representation learning from sound-image data with acoustic images by using knowledge distillation.

These existing methods require potentially paired bimodal data. To address the issue, we propose a novel method for learning sound-image representations from unpaired data with the consideration of the weakly paired condition. The strength of our algorithm lies in that it can enhance the feature representation of a certain modal input even by exploiting other unpaired modal data. Further, compared to bimodal semi-supervised works [3,6,9] which align different modalities in common space mainly for retrieval, our work is different in that the goal is to learn general features from unlabeled bimodal data in a self-supervised way.

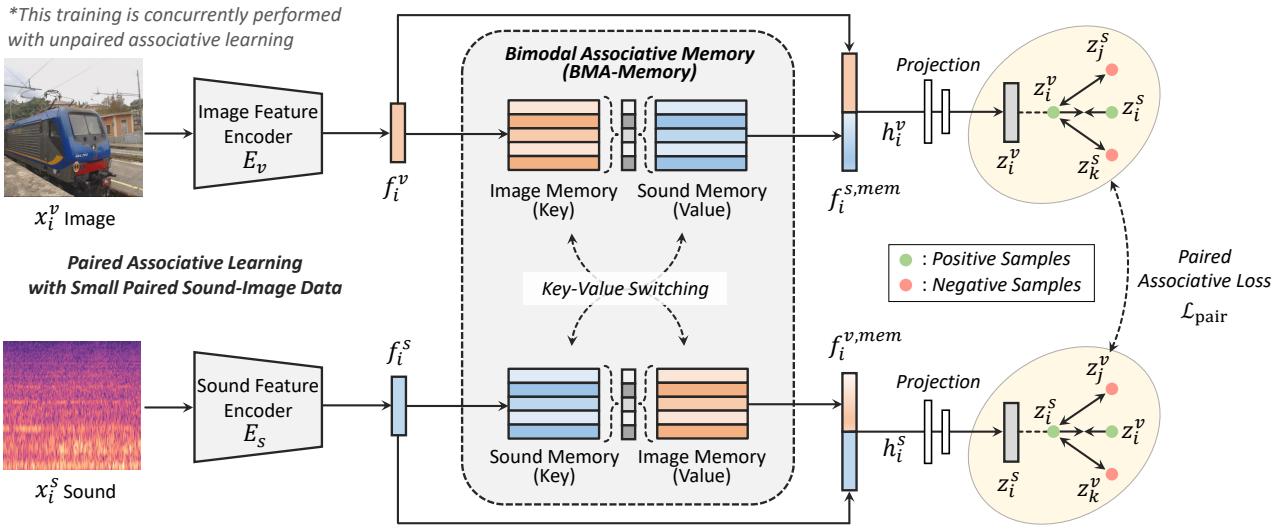

Figure 2. Proposed paired associative learning with BMA-Memory to exploit small paired sound-image data at training time. BMA-Memory includes image and sound sub-memories with a key-value switching scheme. During the training, paired images and sounds are used to make these memories associate with one another.

2.2. Memory-Augmented Network

A memory-augmented network indicates the neural network with external memory components which make it possible to read and write historical information. Memory-augmented networks have been proposed to solve various problems in the deep learning field. They were adopted in several tasks such as object tracking [14, 43], anomaly detection [16,36], predictive learning [19,28,29], and few-shot learning [8, 23, 45]. There exist methods that apply memory networks to self-supervised learning schemes. Lai *et al.* [27] introduced a self-supervised dense tracking model with the memory-augmented network which stores the information of past frames. Han *et al.* [19] proposed a predictive coding framework with the memory-augmented network. They tried to learn video representation by estimating the possible future states with the memory.

Compared to existing memory networks, we propose a novel BMA-Memory with key-value switching scheme that enables to naturally associate the sound with the image and vice-versa in a self-supervised manner. Based on BMA-Memory, we propose weakly paired associative learning for building and boosting the association between modalities.

3. Proposed Approach

Representation learning for bimodal data can be formulated as follows. Let x^v and x^s denote input image and sound data (*i.e.*, spectrogram), respectively. The goal is to jointly optimize two functions ($\mathcal{F}_v$, $\mathcal{F}_s$) for obtaining distinct sound and image representations (h^v, h^s) from (x^v, x^s) with self-supervised learning (*i.e.*, pretext learning). Note that $h^v = \mathcal{F}_v(x^v)$ and $h^s = \mathcal{F}_s(x^s)$. Then, the effectiveness of the representations is validated through downstream tasks (*e.g.*, image or sound classification).

3.1. Bimodal Associative Memory

Figure 2 shows weakly paired associative learning with a bimodal associative memory (BMA-Memory) for sound-image representations in the case of learning with small paired data. BMA-Memory is to store sound and image features and to link these modalities. The memory enables to recall the sound feature from image data and vice-versa. We can obtain more abundant representations by exploiting features of both input and recalled modalities.

Firstly, input image x^v and sound x^s become an image feature $f^v \in \mathbb{R}^c$ and a sound feature $f^s \in \mathbb{R}^c$ through each feature encoder (E_v, E_s), respectively. We adopt 2D-Conv architectures, ResNet-18 and ResNet-10 [21] for image and sound encoders, respectively. Note that input x^s has the form of the spectrogram image. The extracted image and sound features are used as memory queries to access an image memory M^v and a sound memory M^s which are sub-memories in BMA-Memory. The image and sound memories have matrix forms of $M^v = \{m_r^v\}_{r=1}^n \in \mathbb{R}^{n \times c}$ and $M^s = \{m_r^s\}_{r=1}^n \in \mathbb{R}^{n \times c}$, respectively with n slots and c channels. A row vector $m_r^v \in \mathbb{R}^c$ denotes the r-th memory item of M^v. BMA-Memory maps one to another modal space through key-value memory structure. It alleviates the domain gap from inconsistent distribution of different modalities. We introduce a key-value switching procedure to associate these memories in a self-supervised manner naturally. So for the image input, the image memory becomes the key while the sound memory becomes the value. The key-value memories are swapped in the

case of the sound input (see Figure 2). Addressing vectors $W^v = \{w_r^v\}_{r=1}^n \in \mathbb{R}^n$ and $W^s = \{w_r^s\}_{r=1}^n \in \mathbb{R}^n$ are obtained from key-memories M^v and M^s, respectively. Note that each addressing vector is used to access the components of each value-memory. The memory addressing scheme is shown in Figure 3. The addressing procedure in the case of image input feature f^v can be formulated as

$$w_r^v = \frac{\exp(d(f^v, m_r^v)/\tau_m)}{\sum_{r=1}^n \exp(d(f^v, m_r^v)/\tau_m)}, \quad (1)$$

$$d(f^v, m_r^v) = \frac{f^v \cdot m_r^v}{||f^v||\,||m_r^v||}, \quad (2)$$

where $d(\cdot,\cdot)$ indicates cosine similarity function, $\exp(\cdot)/\sum\exp(\cdot)$ denotes softmax function, and τ_m is a memory temperature. W^v is used to access the components of a value-memory to convert from image space to sound space. Note that value-memory indicates M^s for W^v. Each component w_r^v of W^v can be considered as an attention weight for the corresponding value-memory slot m_r^s. M^s outputs a sound memory feature $f^{s,mem} \in \mathbb{R}^c$ as follows

$$f^{s,mem} = \sum_{r=1}^n w_r^v m_r^s. \quad (3)$$

Finally, the image feature f^v and the sound memory feature $f^{s,mem}$ are concatenated to obtain target representation $h^v = [f^v;\ f^{s,mem}]$. h^v includes both input image and associated sound information. For input sound feature f^s, the overall addressing procedure is identical to that for the image feature f^v. Sound and image terms are just swapped. During the training phase, the weights of M^v and M^s are updated via backpropagation as [16, 28]. The objective loss is described in the next section.

3.2. Weakly Paired Associative Learning

3.2.1 Paired Associative Learning

We propose weakly paired associative learning that includes paired associative learning and unpaired associative learning. The proposed model is trained with small paired sound-image data as shown in Figure 2. The goal of paired associative learning is to build the link between image and sound memories in a self-supervised manner. With the i-th paired input image x_i^v and sound x_i^s, we can obtain target feature representations h_i^v and h_i^s, respectively. Then, they pass through the projection head that consists of 2-layer MLP to make z_i^v and z_i^s as [10]. z_i^v and z_i^s indicate projections which are actually used for pretext self-supervised learning. If z^v and z^s are from a pair (or the same clip), we consider them as a positive set (*e.g.*, z_i^v, z_i^s). Otherwise, we think of them as a negative set (*e.g.*, z_i^v, z_j^s). Making such a positive set distinctly close allows the memory to associate counterpart modality. The objective loss named paired associative loss $\mathcal{L}_{\texttt{pair}}$ has the variational form of noise contrastive

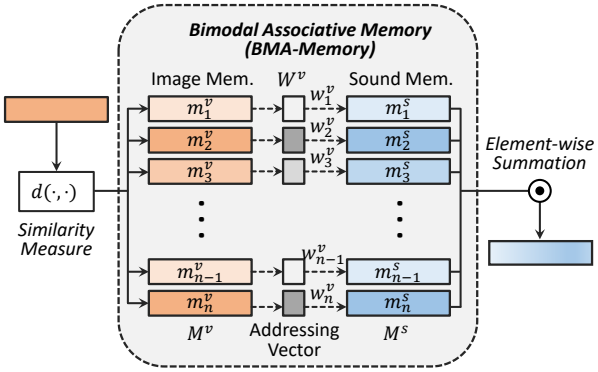

Figure 3. Detailed key-value addressing procedure of BMA-Memory in the case of image memory (key) to sound memory (value). Note that sound memory (key) to image memory (value) can be performed vice-versa.

loss [10, 20] and is applied to samples in a mini-batch (batch size=N). When there are N samples of z^v and N samples of z^s, except for itself and samples from a same clip, the rest samples can be considered as negative samples of a specific sample. We set $Z^p = \{z_i^p\}_{i=1}^{2N} = \{z_1^v, ..., z_N^v, z_1^s, ..., z_N^s\}$. The loss for paired associative learning is defined as

$$\mathcal{L}_{\texttt{pair}} = -\frac{1}{2N}\sum_{i=1}^{2N} \log \frac{\sum_{k^+} \exp(d(z_i^p, k^+)/\tau_l)}{\sum_{j=1}^{2N} \mathbb{1}_{[j\neq i]} \exp(d(z_i^p, z_j^p)/\tau_l)}, \quad (4)$$

where k^+ and τ_l indicate a positive sample from a sample clip (*e.g.*, z_i^v for z_i^s) and a loss temperature parameter, respectively. $\mathbb{1}_{[j\neq i]} \in \{0, 1\}$ represents an indicator function that has 1 iff $j \neq i$. By minimizing $\mathcal{L}_{\texttt{pair}}$, we can attract one another within the positive set and repel each other within the negative set. This makes it possible to recall the distinct sound modality from the image data and vice-versa. When we do not use the unpaired data, training is conducted by just minimizing $\mathcal{L}_{\texttt{pair}}$.

3.2.2 Unpaired Associative Learning

Further, the model can be trained with unpaired data to reinforce sound-image association which is built by the paired associative learning. To this end, we construct the pseudo bimodal pair including bidirectional memory associations (I→S, S→I). The strength of our algorithm lies in that it can enhance the feature representation of a certain modal input by exploiting other unpaired modal data. For example, we can reinforce the representation from input image data by additionally using just unpaired sound data, and vice-versa.

Figure 4 shows the case of exploiting the i-th unpaired sound x_i^s. Firstly, we get augmented sample $x_i'^s$ by applying augmentation algorithm to x_i^s. As an augmentation algorithm for $x_i'^s$, we adopt SpecAugment [35] which randomly masks the frequency and time bands of the spec-

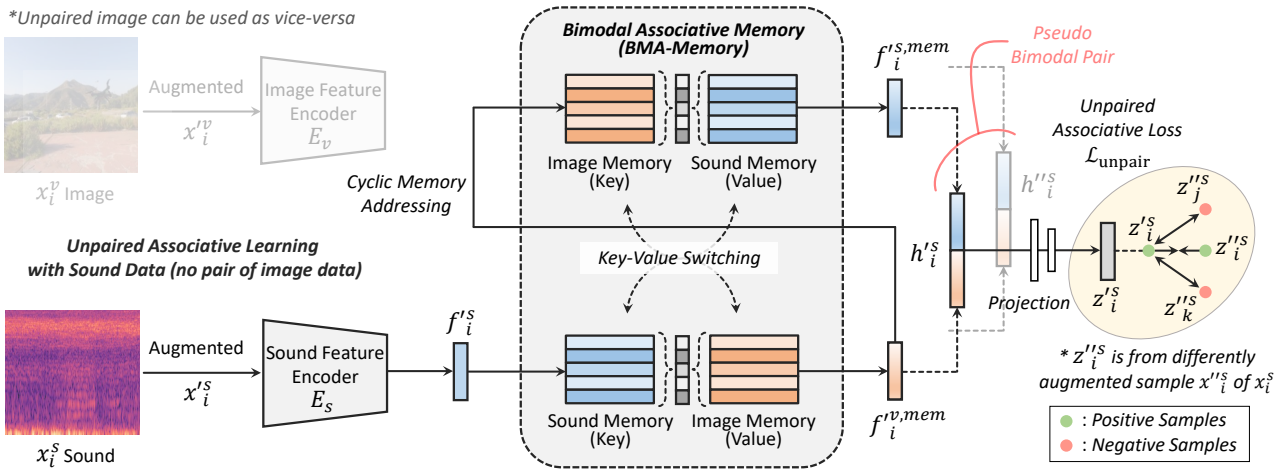

Figure 4. Proposed unpaired associative learning with BMA-Memory to exploit unpaired sound data at training time. During the training, unpaired images or sounds can be further exploited to enhance the association between image and sound sub-memories of BMA-Memory in a self-supervised way, which leads to boosting the representations. Through the unpaired associative learning, it is possible to enhance the representation of a certain modality by exploiting unpaired different modal data. As the case of unpaired sound data, unpaired image data also can be used in a similar way (memory accessing I→S, I→S→S→I).

trogram. $f_i^{\prime v,mem}$ is obtained from $x_i^{\prime s}$ through the sound-image (key-value) memories. Different from the paired associative learning, $f_i^{\prime v,mem}$ is then reused as a pseudo image memory query to pass through the image-sound (key-value) memories. Through this cyclic addressing, we obtain $h_i^{\prime s} = [f_i^{\prime v,mem};\ f_i^{\prime s,mem}]$ which is mainly composed of image and sound memory features. Making $h_i^{\prime s}$ discriminative can lead to boosting alignment of associations between image and sound memories because $h_i^{\prime s}$ contains bidirectional associations (I→S, S→I). Therefore, unpaired learning using sound can also reinforce the S→I as well as I→S associations, which can lead to the improvement of the image representation. To this end, we optimize the model with an unpaired associative loss $\mathcal{L}_{\texttt{unpair}}$. In terms of $\mathcal{L}_{\texttt{unpair}}$, we build the positive set as projections (*e.g.*, $z_i^{\prime s}$ and $z_i^{\prime\prime s}$) from a same sound sample (or a same clip). $z_i^{\prime\prime s}$ is the projection from $x_i^{\prime\prime s}$ which is differently random augmented sample of x_i^s. $z_i^{\prime\prime s}$ can be considered as the projection of pseudo bimodal pair. We set the negative set as projections (*e.g.*, $z_i^{\prime s}$ and $z_j^{\prime\prime s}$) from the different sound samples. $\mathcal{L}_{\texttt{unpair}}$ is also applied based on the units in a mini-batch. Similar to $\mathcal{L}_{\texttt{pair}}$, we can set $Z^{up} = \{z_i^{up}\}_{i=1}^{2N} = \{z_1^{\prime s}, ..., z_N^{\prime s}, z_1^{\prime\prime s}, ..., z_N^{\prime\prime s}\}$. In the case of unpaired sound data, the loss function for unpaired associative learning can be written as

$$\mathcal{L}_{\texttt{unpair}} = -\frac{1}{2N}\sum_{i=1}^{2N}\log\frac{\sum_{k^+}\exp(d(z_i^{up}, k^+)/\tau_l)}{\sum_{j=1}^{2N}\mathbb{1}_{[j\neq i]}\exp(d(z_i^{up}, z_j^{up})/\tau_l)}, \tag{5}$$

Unpaired associative learning proceeds simultaneously with the paired learning. Model training is conducted with a total loss $\mathcal{L} = \mathcal{L}_{\texttt{pair}} + \mathcal{L}_{\texttt{unpair}}$ with both paired and unpaired data. Paired data pass through the path of Figure 2, and unpaired data pass through the path of Figure 4. Then, they are optimized at once. Note that unpaired image x_i^v also can be used for unpaired associative learning. The procedure for this is just vice-versa for the case of unpaired sound. In this case, we adopt augmentations of [10] for images during the unpaired associative learning.

4. Experiments

4.1. Datasets

To validate the proposed method, we adopt public datasets which contain image and/or sound data. We use ACIVW [38] and Kinetics-400 [24] datasets for learning a self-supervised pretext task. For downstream tasks, ACIVW [38] and DCASE-2018 [30] datasets are utilized.

ACIVW. ACIVW [38] includes multimodal data with 5 hours of videos outdoors in the wild condition, obtained by an acoustic-optical camera. Raw signals are acquired from 128 microphones with a sampling frequency of 12.8kHz. Video frames are captured with 480 × 640 pixels and 12 frames per second. It also includes 36 × 48 × 512 multi-spectral acoustic images which have both spatial and auditory information. It consists of 10 classes with wild conditions: drone, shopping cart, traffic, train, boat, fountain, drill, razor, hair dryer, and vacuum cleaner. ACIVW is mainly utilized to investigate the correspondence between images and sounds. There are total 9k sound-image pairs. As [38], we used 70% of the dataset to train the model with the pretext task. The remains are used to validate the model with the downstream tasks.

Kinetics-400. This dataset [24] contains about 230k train-

Method	Training Data Types	Top-1 Accuracy
Supervised Learning	image	0.769
L^3 Vision Network* [2]	image + sound	0.544
Audio-Visual (H)* [38]	image + sound	0.667
Audio-Visual (H)* [38] (w/ Transfer Learning)	image + sound + acoustic image	0.732
Proposed Method*	image + sound + unpaired sound	**0.772**
AVID-CMA† [32]	image + sound	0.738
Proposed Method† (w/o Unpaired Associative Learning)	image + sound	0.745
Proposed Method†	image + sound + unpaired sound	**0.778**

Table 1. Performance results for image classification on ACIVW dataset. Except for the supervised model, all other models are train with ACIVW dataset in self-supervised ways. Unpaired sounds are from Kinetics-400 dataset. * and † indicates the accuracies obtained from KNN and linear evaluation protocol, respectively.

Method	Training Data Types	Top-1 Accuracy
Supervised Learning	sound	0.971
L^3 Audio Network* [2]	image + sound	0.361
HearNet* [38]	image + sound	0.757
HearNet* [38] (w/ Transfer Learning)	image + sound + acoustic image	0.795
Proposed Method*	image + sound + unpaired image	**0.936**
AVID-CMA† [32]	image + sound	0.902
Proposed Method† (w/o Unpaired Associative Learning)	image + sound	0.931
Proposed Method†	image + sound + unpaired image	**0.956**

Table 2. Performance results for sound classification on ACIVW dataset. Except for the supervised model, all other models are train with ACIVW dataset in self-supervised ways. Unpaired images are from Kinetics-400 dataset. * and † indicates the accuracies obtained from KNN and linear evaluation protocol, respectively.

ing videos with 400 classes such as riding a bike, salsa dancing, dunking basketball, and playing trumpet. Each clip lasts about 10sec and they are taken from different YouTube videos. Thus, it covers a large range of image and sound variations. Since the videos are obtained from YouTube, they have variable frame rates and resolutions. We sample a frame image and a sound in the middle of each clip. The dataset is used to train the model with the pretext task, especially in the unpaired associative learning.

DCASE-2018. It is the 2018 version of Detection and Classification of Acoustic Scenes (DCASE) [30]. The dataset includes sound recordings from six European cities with ten different acoustic scenes: airport, bus, metro, metro station, park, public square, shopping mall, street (pedestrian), street (traffic), and tram. The recordings are obtained with a 48kHz sampling rate. We use DCASE-2018 dataset to validate the model with the sound-based downstream task.

4.2. Implementation Details

Each image is normalized to the intensity of [0, 1] and resized to 224 $\times$ 224 pixels. Raw sound signals with 2sec are preprocessed to the form of log mel-spectrogram with 150 $\times$ 200. We adopt ResNet-18 [21] as our image encoder according to [38] and similarly ResNet-10 as the sound encoder since both image and sound have spatial information. Memory slot size n is fixed as 1,000. Memory and loss temperature parameters (τ_m, τ_l) are both set as 0.1 for all experiments according to [10]. The projected feature that is used to train the model with pretext tasks has 128-dimensional latent space as [10]. All of the proposed models are trained by the Adam optimizer [25] with a learning rate of 0.0002 and a batch size of 256. The experiments are conducted on a server system with TITAN RTX GPUs. We implement the model in PyTorch [37].

4.3. Evaluation on Downstream Tasks

To evaluate the representation quality, we follow a linear evaluation protocol [10] which is mainly adopted in the representation learning domain. First, a linear classifier is trained on top of the target representation with our frozen network. Then, we evaluate the test accuracy through the representation with the linear classifier to check the representation quality. If we want to evaluate the image representation power of our model, the representation $h^v = [f^v;\ f^{s,mem}]$ from the image input is used for image downstream tasks. $h^s = [f^{v,mem};\ f^s]$ from the sound input is used for sound downstream tasks (See Figure 2). Further, we also utilize k-nearest neighbor (KNN) to evaluate the representations.

Image Recognition. Table 1 shows the performance comparison results on ACIVW dataset in terms of classification with image data. All models are trained with the training set of ACIVW dataset. For evaluation of the image representation, we utilize the feature h^v obtained from image inputs (see Figure 2). We compare our method with existing sound-image works [2, 38] and applicable audio-video work [32]. Except for the supervised model, the rest of the models are trained in self-supervised manners. The supervised model is trained with label information of ACIVW dataset. Note that the supervised model and our model have the same ResNet-18 backbone architecture as [38]. As shown in the table, the proposed method outperforms other methods. The proposed model surpasses '*Audio-Visual (H)*' model with additional paired acoustic images which include both image and sound information. Note that unpaired sounds are more easily obtainable compared to the acoustic image pairs. By utilizing unpaired sound from Kinetics-400, the proposed method achieves the result better than the

Method	Training Data Types	Retrieval Accuracy				
		Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
L^3 Audio Network [2]	image + sound	0.097	0.119	0.242	0.267	0.301
HearNet [38]	image + sound	0.289	0.344	0.424	0.480	0.614
DualCamNet [38]	image + sound + acoustic image	0.334	0.370	0.429	0.482	0.624
Proposed Method (w/o Unpaired Associative Learning)	image + sound	0.498	0.541	0.588	0.614	0.705
Proposed Method	image + sound + unpaired image	**0.522**	**0.553**	**0.612**	**0.669**	**0.766**

Table 3. Performance results for bimodal retrieval (sound to image retrieval) on ACIVW dataset. All models are trained with ACIVW dataset in self-supervised manners.

Method	Training Data Types	Top-1 Accuracy
Supervised Learning	sound	0.595
L^3 Audio Network* [2]	image + sound	0.323
HearNet* [38]	image + sound	0.354
HearNet* [38] (w/ Transfer Learning)	image + sound + acoustic image	0.376
Proposed Method*	image + sound + unpaired image	**0.420**
AVID-CMA† [32]	image + sound	0.421
Proposed Method† (w/o Unpaired Associative Learning)	image + sound	0.538
Proposed Method†	image + sound + unpaired image	**0.562**

Table 4. Performance comparison results for sound classification on DCASE-2018 dataset in a zero-shot setting. All models are trained with ACIVW and tested on DCASE-2018 except for the supervised model. * and † indicates the accuracies obtained from KNN and linear evaluation protocol, respectively.

other methods and beyond the supervised model in terms of image recognition. The results show that the unpaired associative learning enhances the image representation even by using unpaired sound data from the different dataset.

Sound Recognition. The performance comparison results for sound classification on ACIVW dataset are shown in Table 2. Similar to the previous results, the models are trained with ACIVW dataset in self-supervised ways except for the supervised model. To evaluate the sound representation quality, we exploit the feature h^s obtained from sound inputs (see Figure 2). As shown in the table, the proposed method surpasses the other self-supervised methods. The unpaired associative learning with unpaired image data reinforces the sound recognition performance, which means that the sound representation is enhanced even by exploiting unpaired image data. Note that unpaired images are from Kinetics-400 dataset. As a result, the final model achieves the competitive performance compared to the supervised model.

Zero-Shot Sound Recognition. Further, we conduct the experiment on DCASE-2018 for sound classification to validate the generalizability in zero-shot setting. In this experiment, the models are trained with ACIVW dataset in a self-supervised way and validated on DCASE-2018 dataset. Note that the supervised model is trained with DCASE-2018 dataset with label information. As shown in Table 4, the proposed method shows better performances compared to the other methods. In particular, when unpaired images are additionally used, the proposed method achieves comparable performance to the supervised model. These results indicate the obtained feature representation is generalizable well to the different dataset.

Bimodal Retrieval. We additionally perform bimodal retrieval to verify how well the sound and image representations are associated with each other in a self-supervised manner. We select one sound sample and find the corresponding images which are close to the sound sample. It is correct if the sound and the retrieved image have the same class. Note that the bimodal retrieval is conducted based on image projection z^v and sound projection z^s which are used for matching at training time. Retrieved images are ordered according to the distance between the features of the sound and image. Experiments are performed on ACIVW dataset. As shown in Table 3, our method outperforms the other self-supervised methods in terms of bimodal retrieval for all conditions. The rank indicates how many candidates to retrieve. In particular, when unpaired associative learning is performed by using unpaired images, it shows better retrieval performances. It can be seen that the learning scheme with unpaired modal data strengthens the association between image and sound modalities.

4.4. Compensation for Paired Data Volume

Table 5 shows how effectively a decrease of the paired data volume can be compensated by using other unpaired modal data. As shown in the table, there is a considerable performance decrease in image classification when the paired data of ACIVW is reduced to 20% level (1k images). Interestingly, through the proposed unpaired associative learning with unpaired sounds (Kinetics-400), we

Paired Data Volume	Unpaired Associative Learning	Training Data Types	Top-1 Accuracy
100%	✗	image + sound	0.745
	✓	image + sound + unpaired sound	0.778
20%	✗	image + sound	0.693
	✓	image + sound + unpaired sound	0.749

Table 5. Performance evaluations in terms of image classification according to the amount of training paired data in ACIVW.

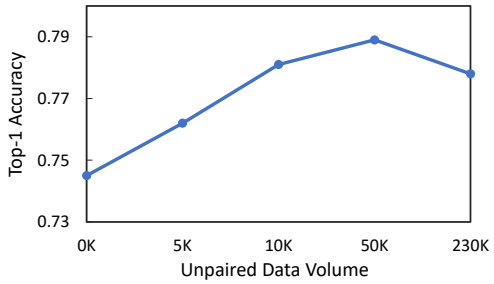

Figure 5. Performance evaluations in terms of image classification according to the amount of unpaired sound data in Kinetics-400.

achieve the performance that exceeds the performance using full paired data. In other words, with only 20% image data, we obtain a competitive image recognition performance compared to the 100% case by exploiting other unpaired modal data (*i.e.*, sound data). Further, it can be seen that the boosting effect from unpaired data is more significant when the amount of paired data is limited.

4.5. Effects of Unpaired Data Volume

Figure 5 shows accuracies for image classification on ACIVW according to the amount of unpaired sound data in Kinetics-400. 230k indicates the use of full unpaired sound data in Kinetics-400 while 0k indicates the model without unpaired associative learning. As shown in the figure, the best performance is not always achieved when all data is used unconditionally. Further higher performance can be achieved when a moderate amount of unpaired data is used. The highest performance was shown when 50k of unpaired data was used, and the performance decreased as the amount decreased after that. Note that there are 6k training sound-image pairs. When the volume imbalance of unpaired and paired data is severe, more unpaired data does not help to enhance the representation.

4.6. Effects of Memory Size

We perform the experiments to observe the effects of the memory size n on the representation learning performances. The memory size n indicates the number of slots in the image and sound sub-memories. We change n with an exponential scale (10, 50, 100, 500, 1,000, 5,000) for image classification on ACIVW dataset. Figure 6 shows the results. As the memory capacity increases, the performance tends to increase, which is saturated around n=500. Then it maintains the relatively stable values. Considering the exponential scale, the result represents the robustness to the setting of memory size around $n > 500$.

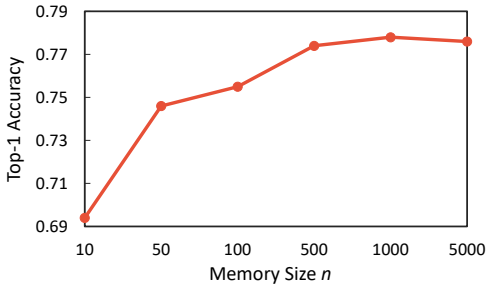

Figure 6. Effects of the memory size for image classification on ACIVW. Memory size is varied with exponential scale.

5. Discussion

A slight performance decrease is observed with extremely large unpaired data (See Figure 5). It seems that it is due to the severe imbalance of paired data volume and unpaired data volume because the training combination of unpaired data keeps changing even with the limited paired data. Effectively dealing with the extremely large amount of unpaired data can be investigated in further works.

6. Conclusion

The goal of the proposed work is to learn the sound-image representations even by exploiting unpaired modal data in weakly paired condition. To this end, we propose BMA-Memory with key-value switching to effectively store the sound-image features and associate one another modality in a self-supervised manner. Through BMA-Memory, we can obtain abundant representations which contain information of both input and associated modalities. Based on this memory, we devise weakly paired associative learning to build and boost the association between sound and image. It enables to enhance the representation of a certain modality even by using different modal data. As a result, the proposed method outperforms other sound-image representation learning methods. Further, we validate the effectiveness and practicality of the proposed method by conducting ablation studies and data volume analysis.

Acknowledgement. This work was supported by the IITP grant funded by the MSIT (No. 2020-0-00004).

References

[1] H. Alwassel, D. Mahajan, B. Korbar, L. Torresani, B. Ghanem, and D. Tran. Self-supervised learning by cross-modal audio-video clustering. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:9758–9770, 2020. 2

[2] R. Arandjelovic and A. Zisserman. Look, listen and learn. In *International Conference on Computer Vision (ICCV)*, pages 609–617, 2017. 6, 7

[3] Y. Aytar, L. Castrejon, C. Vondrick, H. Pirsiavash, and A. Torralba. Cross-modal scene networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(10):2303–2314, 2017. 2

[4] A. Baevski, S. Schneider, and M. Auli. vq-wav2vec: Self-supervised learning of discrete speech representations. In *International Conference on Learning Representations (ICLR)*, 2019. 1

[5] B. Barakat, A. R. Seitz, and L. Shams. Visual rhythm perception improves through auditory but not visual training. *Current Biology*, 25(2):R60–R61, 2015. 2

[6] A. K. Bhunia, P. N. Chowdhury, A. Sain, Y. Yang, T. Xiang, and Y. Song. More photos are all you need: Semi-supervised learning for fine-grained sketch based image retrieval. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4247–4256, 2021. 2

[7] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In *Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1877–1901, 2020. 1

[8] Q. Cai, Y. Pan, T. Yao, C. Yan, and T. Mei. Memory matching networks for one-shot image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4080–4088, 2018. 3

[9] L. Castrejon, Y. Aytar, C. Vondrick, H. Pirsiavash, and A. Torralba. Learning aligned cross-modal representations from weakly aligned data. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2940–2949, 2016. 2

[10] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, pages 1597–1607. PMLR, 2020. 1, 2, 4, 5, 6

[11] Y. Chen, Y. Xian, A. Koepke, Y. Shan, and Z. Akata. Distilling audio-visual knowledge by compositional contrastive learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7016–7025, 2021. 2

[12] J. Devlin, M. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019. 1

[13] C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. In *International Conference on Computer Vision (ICCV)*, pages 1422–1430, 2015. 2

[14] Z. Fu, Q. Liu, Z. Fu, and Y. Wang. Stmtrack: Template-free visual tracking with space-time memory networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13774–13783, 2021. 3

[15] S. Gidaris, P. Singh, and N. Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*, 2018. 1, 2

[16] D. Gong, L. Liu, V. Le, B. Saha, M.R. Mansour, S. Venkatesh, and A. Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *International Conference on Computer Vision (ICCV)*, pages 1705–1714, 2019. 3, 4

[17] J. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *Neural Information Processing Systems (NeurIPS)*, 2020. 1

[18] T. Guo, Y. Ren, Y. Yu, Y. Yu, Y. Hasegawa, Q. Wu, J. Yang, S. Takahashi, Y. Ejima, and J. Wu. Improving visual working memory with training on a tactile orientation sequence task in humans. *SAGE Open*, 11(3):21582440211031549, 2021. 2

[19] T. Han, W. Xie, and A. Zisserman. Memory-augmented dense predictive coding for video representation learning. In *European Conference on Computer Vision (ECCV)*, pages 312–329. Springer, 2020. 3

[20] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020. 2, 4

[21] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 3, 6

[22] C. Jia, Y. Yang, Y. Xia, Y. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning (ICML)*, 2021. 2

[23] Ł. Kaiser, O. Nachum, A. Roy, and S. Bengio. Learning to remember rare events. In *International Conference on Learning Representations (ICLR)*, 2017. 3

[24] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 5

[25] D.P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 6

[26] B. Korbar, D. Tran, and L. Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *Neural Information Processing Systems (NeurIPS)*, 2018. 1, 2

[27] Z. Lai, E. Lu, and W. Xie. Mast: A memory-augmented self-supervised tracker. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6479–6488, 2020. 3

[28] S. Lee, H. G. Kim, D. H. Choi, H. Kim, and Y. M. Ro. Video prediction recalling long-term motion context via memory

alignment learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3054–3063, 2021. 3, 4

[29] F. Marchetti, F. Becattini, L. Seidenari, and A. D. Bimbo. Mantra: Memory augmented networks for multiple trajectory prediction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7143–7152, 2020. 3

[30] A. Mesaros, T. Heittola, and T. Virtanen. A multi-device dataset for urban acoustic scene classification. In *Detection and Classification of Acoustic Scenes and Events Workshop (DCASEW)*, page 9, 2018. 5, 6

[31] P. Morgado, Y. Li, and N. Nvasconcelos. Learning representations from audio-visual spatial alignment. *Neural Information Processing Systems (NeurIPS)*, 33:4733–4744, 2020. 1, 2

[32] P. Morgado, N. Vasconcelos, and I. Misra. Audio-visual instance discrimination with cross-modal agreement. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12475–12486, 2021. 2, 6, 7

[33] M. Noroozi and P. Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision (ECCV)*, pages 69–84. Springer, 2016. 2

[34] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba. Learning sight from sound: Ambient sound provides supervision for visual learning. *International Journal of Computer Vision*, 126(10):1120–1137, 2018. 1, 2

[35] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le. Specaugment: A simple data augmentation method for automatic speech recognition. In *INTERSPEECH*, 2019. 4

[36] H. Park, J. Noh, and B. Ham. Learning memory-guided normality for anomaly detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14372–14381, 2020. 3

[37] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In *Neural Information Processing Systems Workshop (NeurIPSW)*, 2017. 6

[38] V. Sanguineti, P. Morerio, N. Pozzetti, D. Greco, M. Cristani, and V. Murino. Leveraging acoustic images for effective self-supervised audio representation learning. In *European Conference on Computer Vision (ECCV)*, pages 119–135. Springer, 2020. 1, 2, 5, 6, 7

[39] S. Schneider, A. Baevski, R. Collobert, and M. Auli. wav2vec: Unsupervised pre-training for speech recognition. In *INTERSPEECH*, 2019. 1

[40] A. Senocak, T. Oh, J. Kim, M. Yang, and I. S. Kweon. Learning to localize sound source in visual scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4358–4366, 2018. 2

[41] R. A. Stevenson, M. M. Wilson, A. R. Powers, and M. T. Wallace. The effects of visual training on multisensory temporal processing. *Experimental brain research*, 225(4):479–489, 2013. 2

[42] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid. Videobert: A joint model for video and language representation learning. In *International Conference on Computer Vision (ICCV)*, pages 7464–7473, 2019. 2

[43] T. Yang and A. B. Chan. Learning dynamic memory networks for object tracking. In *European Conference on Computer Vision (ECCV)*, pages 152–167, 2018. 3

[44] J. Zhang, X. Xu, F. Shen, H. Lu, X. Liu, and H. T. Shen. Enhancing audio-visual association with self-supervised curriculum learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 3351–3359, 2021. 2

[45] L. Zhu and Y. Yang. Inflated episodic memory with region self-attention for long-tailed visual recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4344–4353, 2020. 3