Weakly Paired Associative Learning for Sound and Image Representations via Bimodal Associative Memory - *Supplementary Material* -

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

1. Evaluation on Downstream Task with Fine-Tunning

Table 1 shows the performance results in terms of image classification with fine-tunning all weights. All models are trained with ACIVW dataset. 'Supervised Learning' indicates the model with supervised learning from the random weight initialization. The proposed methods indicate the models with supervised learning (*i.e.*, fine-tunning) from the weights obtained by the proposed self-supervised learning. They are firstly trained by ACIVW dataset without data labels in self-supervised manners. Then based on the learned weights, supervised learning is performed to optimize all the weights using the labels of the ACIVW dataset. As shown in the table, we can achieve the performance gain by initializing the weights with the proposed method. In particular, utilizing the other unpaired modal data can fairly contribute to the improvement of performance based on the fine-tunning.

Method	**Training Data Types**	**Top-1 Accuracy**
Supervised Learning	image	0.769
Proposed Method + Supervised Learning (w/o Unpaired Associative Learning)	image + sound	0.784
Proposed Method + Supervised Learning	image + sound + unpaired sound	**0.801**

Table 1: Performance results for image classification on ACIVW dataset according to fine-tunning the model from the randomly initialized weights and learned weights with the proposed method.

2. Effects of Loss Weight

Table 2 shows the image recognition performances according to the weight of losses on ACIVW. The performances are measured with linear evaluation protocol. Without heuristic loss weighting, naïve 1:1 setting can achieve fairly good performance.

Weight of Losses ($\mathcal{L}_{\texttt{pair}}$: $\mathcal{L}_{\texttt{unpair}}$)				
1:0.1	**1:0.5**	**1:1**	**1:2**	**1:10**
0.762	0.783	0.778	0.771	0.759

Table 2: Performance results according to the weight of losses for image recognition on ACIVW.

3. Effects of Training Data Types

Table 3, 4, and 5 show the performance results according to the combination of training data types. The performances are measured with linear evaluation protocol. I_A and S_A indicate image and sound from ACIVW while I_K and S_K indicate image and sound from Kinetics-400. Each training data type contributes to the image and sound recognition performances.

Method	**Training Data Types**	**Top-1 Accuracy**
Proposed Method† $(\mathcal{L}_{\texttt{unpair}}(I_A))$	image	0.708
Proposed Method† $(\mathcal{L}_{\texttt{pair}}(I_A + S_A))$	image + sound	0.745
Proposed Method† $(\mathcal{L}_{\texttt{pair}}(I_A + S_A) + \mathcal{L}_{\texttt{unpair}}(S_K))$	image + sound + unpaired sound	**0.778**

Table 3: Performance results according to the training data types for image classification on ACIVW dataset.

Method	**Training Data Types**	**Top-1 Accuracy**
Proposed Method† $(\mathcal{L}_{\texttt{unpair}}(S_A))$	sound	0.887
Proposed Method† $(\mathcal{L}_{\texttt{pair}}(I_A + S_A))$	image + sound	0.931
Proposed Method† $(\mathcal{L}_{\texttt{pair}}(I_A + S_A) + \mathcal{L}_{\texttt{unpair}}(I_K))$	image + sound + unpaired image	**0.956**

Table 4: Performance results according to the training data types for sound classification on ACIVW dataset.

Method	**Training Data Types**	**Top-1 Accuracy**
Proposed Method† $(\mathcal{L}_{\texttt{unpair}}(S_A))$	sound	0.468
Proposed Method† $(\mathcal{L}_{\texttt{pair}}(I_A + S_A))$	image + sound	0.538
Proposed Method† $(\mathcal{L}_{\texttt{pair}}(I_A + S_A) + \mathcal{L}_{\texttt{unpair}}(I_K))$	image + sound + unpaired image	**0.562**

Table 5: Performance results according to the training data types for sound classification on DCASE-2018 dataset in a zero-shot setting.

4. Network Structure Details

Image Feature Encoder E_v		
Layer	**Filter**	**Output Size** $(\boldsymbol{width}\times\boldsymbol{height}\times\boldsymbol{channel})$
Conv 1	7 × 7, 64 / stride 2	$(W/2)\times(H/2)\times 64$
Conv 2_x	3 × 3 max pool / stride 2 $\begin{bmatrix} 3\times 3, 64 \\ 3\times 3, 64 \end{bmatrix}\times 2$	$(W/4)\times(H/4)\times 64$
Conv 3_x	$\begin{bmatrix} 3\times 3, 128 \\ 3\times 3, 128 \end{bmatrix}\times 2$	$(W/8)\times(H/8)\times 128$
Conv 4_x	$\begin{bmatrix} 3\times 3, 256 \\ 3\times 3, 256 \end{bmatrix}\times 2$	$(W/16)\times(H/16)\times 256$
Conv 5_x	$\begin{bmatrix} 3\times 3, 512 \\ 3\times 3, 512 \end{bmatrix}\times 2$	$(W/32)\times(H/32)\times 512$
Avg Pooling	global avg pool	$1\times 1\times 512$

Table 6: Network structure details of the image feature encoder E_v, which has the form of ResNet-18. Suppose that the input image has the size of $\mathbb{R}^{W\times H\times C}$.

Sound Feature Encoder E_s		
Layer	**Filter**	**Output Size** $(\boldsymbol{width}\times\boldsymbol{height}\times\boldsymbol{channel})$
Conv 1	7 × 7, 64 / stride 2	$(W/2)\times(H/2)\times 64$
Conv 2_x	3 × 3 max pool / stride 2 $\begin{bmatrix} 3\times 3, 64 \\ 3\times 3, 64 \end{bmatrix}\times 1$	$(W/4)\times(H/4)\times 64$
Conv 3_x	$\begin{bmatrix} 3\times 3, 128 \\ 3\times 3, 128 \end{bmatrix}\times 1$	$(W/8)\times(H/8)\times 128$
Conv 4_x	$\begin{bmatrix} 3\times 3, 256 \\ 3\times 3, 256 \end{bmatrix}\times 1$	$(W/16)\times(H/16)\times 256$
Conv 5_x	$\begin{bmatrix} 3\times 3, 512 \\ 3\times 3, 512 \end{bmatrix}\times 1$	$(W/32)\times(H/32)\times 512$
Avg Pooling	global avg pool	$1\times 1\times 512$

Table 7: Network structure details of the sound feature encoder E_s, which has the form of ResNet-10. Suppose that the input sound spectrogram has the size of $\mathbb{R}^{W\times H\times C}$.

5. Qualitative Results for Bimodal Retrieval

Figure 1 shows the qualitative results of bimodal retrieval (*i.e.*, sound to image retrieval) on ACIVW dataset. The ✓ mark indicates the same class as the sound query while ✗ mark indicates the different class as the sound query. Note that sound queries and retrieved images are from the test set of ACIVW, which are not seen at training time. As shown in the figure, the model utilizing unpaired data can retrieve the images corresponding to the sound better, which means the unpaired data contribute to the bimodal retrieval. The results represents that upaired data can boost the association between sound and image modalities. The example sounds are available in the multimedia appendix.

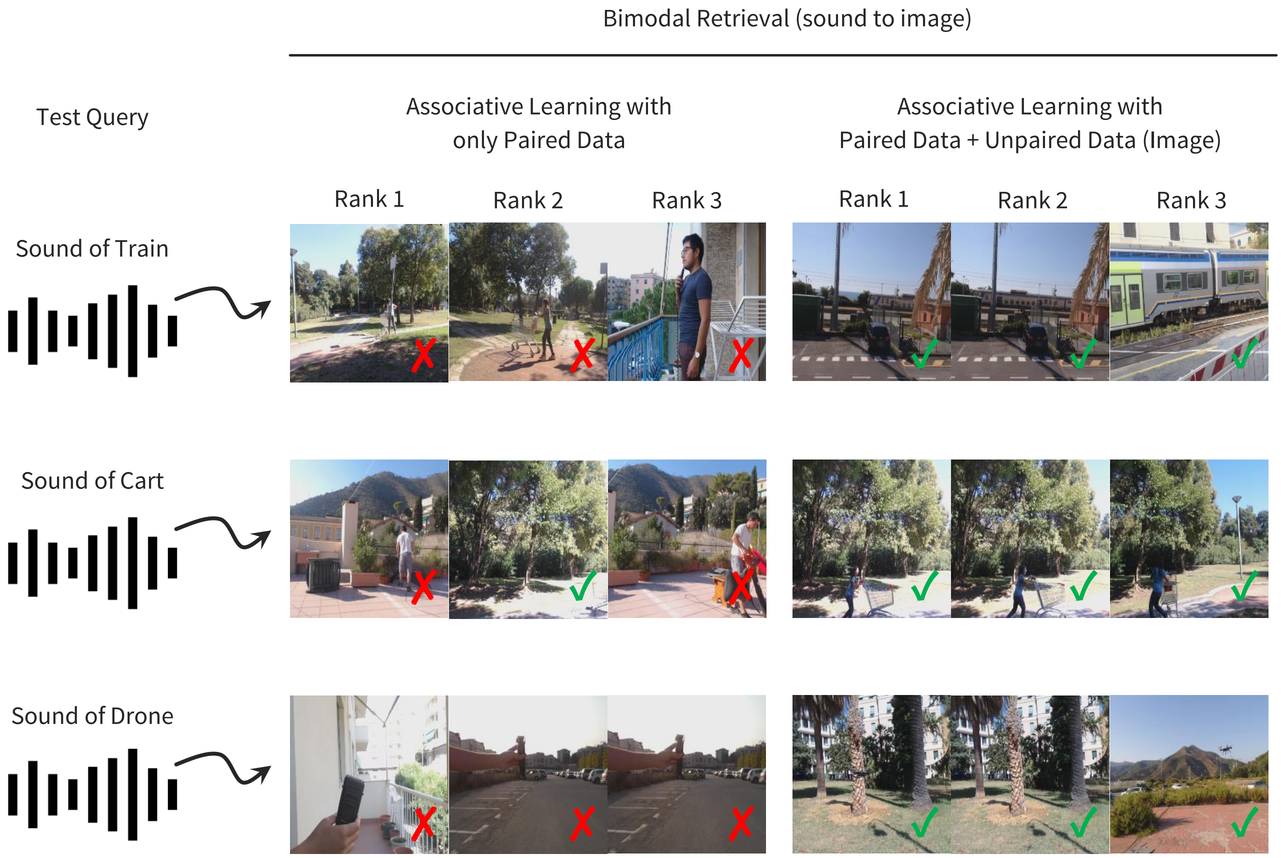
Bimodal Retrieval (sound to image)
Test Query
Associative Learning with only Paired Data
Associative Learning with Paired Data + Unpaired Data (Image)
Rank 1
Rank 2
Rank 3
Rank 1
Rank 2
Rank 3
Sound of Train
Sound of Cart
Sound of Drone
✗ ✗ ✗ ✓ ✓ ✓
✗ ✓ ✗ ✓ ✓ ✓
✗ ✗ ✗ ✓ ✓ ✓

Figure 1: Qualitative results for bimodal retrieval (sound to image retrieval) on ACIVW dataset. Models are trained with ACIVW dataset without labels.