Heterogeneous Scene Analysis via Self-supervised Audiovisual Learning

Di Hu[1], Zheng Wang[2], Haoyi Xiong[1], Dong Wang[2], Feiping Nie[2], and Dejing Dou[1]

[1]Big Data Lab, Baidu Research
[2]School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University

hdui831@mail.nwpu.edu.cn, zhengwangml@gmail.com, xionghaoyi@baidu.com, nwpuwangdong@gmail.com, feipingnie@gmail.com, doudejing@baidu.com

1. Introduction

Audiovisual concurrency provides potential cues for perceiving and understanding the outside world. Such concurrency comes from the simple phenomena of "Sound is produced by the oscillation of object", and exists through our daily life, such as the talking crowd, the barking dog, the roaring machine, etc. These inherent and pervasive correspondences provide us the reference to distinguish and correlate different audiovisual messages, then contribute to learning diversified visual appearances from their produced sounds, or perceiving various acoustic signals from their diversified sound-makers.

Previous work [6] has proved that machine intelligence is able to take advantage of the inherent audiovisual concurrency for possessing human-like audiovisual processing ability. However, the learning capacity of these self-supervised models is pervasively limited by the heterogeneous complexity of audiovisual scene, i.e., the scenes with different number of sound-sources, as shown in Fig. 1. Concretely, it is easy to align sound and its visual source in the simple scene with single sound, whereas more difficult for the complex one with multiple sounds as lack of one-to-one audiovisual alignment annotations. Besides, many works indiscriminately deal with these simple and complex audiovisual data via identical model, which could confuse the models when analyzing and aligning diversified audiovisual content without auxiliary annotations, even degrade the learning performance when more data is available for training [10, 3].

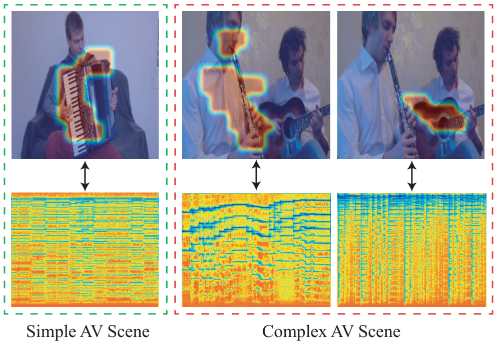


Figure 1. An illustration of the heterogeneous audiovisual (AV) scene complexity. The simple scene contains only one sound-source, while the complex scene contains multiple sound-sources.

To address these two challenges, we propose to study the heterogeneous audiovisual scene complexity, i.e., the number of contained sound-source for grading the heterogeneous scene into a set of audiovisual curriculum in different difficulty levels and perform differentiated audiovisual learning, from the easy ones to the hard ones. The core insight is that we can easily analyze and align the audiovisual content in the simple scenes with single sound, meanwhile it can also provide prior alignment knowledge for the learning in complex scenes. Then, we develop a flexible self-supervised learning model that could effortlessly target at the audiovisual scenes with different number of sound-sources, which can derive effective unimodal representations and infers the latent alignment between sound and sound-maker for both simple and complex scene. Last, a novel curriculum audiovisual learning strategy is proposed, where the difficulty level is determined by the number of sound-source in the scene. Experimental results on audiovisual sound localization and sound

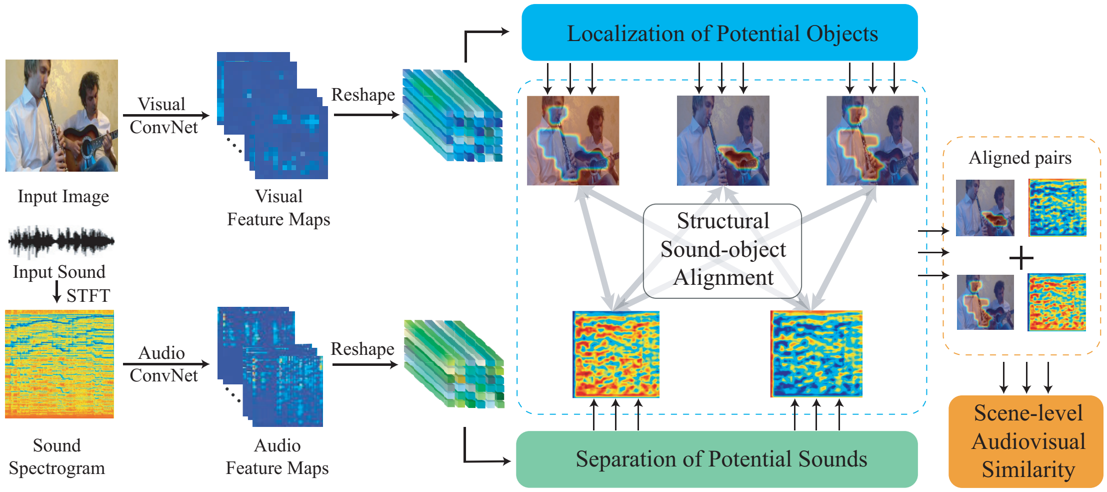


Figure 2. An overview of our proposed audiovisual learning model. Our model firstly represents pairwise audio and image as feature maps, then performs clustering over the reshaped maps to seek potential sound and objects, and learns to infer the alignment among these audiovisual contents. The whole model is optimized w.r.t. the scene-level similarity.

separation tasks show that such simple learning strategy not only makes our model much easier to train, but also improves the learning and alignment performance of audiovisual contents by comparison with the methods that utilize external visual knowledge, i.e., ImageNet-pretrained visual network [10] or detected sounding object representation [3]. An overview of our proposed audiovisual learning model is shown in Fig. 2

2. Approach

2.1. Audiovisual Learning Model

Inspired by clustering-based segmentation [2] and sound separation [5], we first propose to discover and disentangle the potential sounds and objects by analyzing and integrating their channel representations. Concretely, we propose to integrate these feature vectors in the channel space for each modality via soft K-means clustering [1]. The each of k center we obtained should correspond to certain modal component, such as specific object or sound. Meanwhile, the corresponding cluster assignment can be interpreted as a spatial-mask over feature map and indicates the location of sound-source in both modalities. Compared with the approximately discrete optimization in [6], our method is much simpler and more efficient.

Besides, for a given audiovisual scene, although the contained sounds and objects have been described as different clustering centers, it is still difficult to directly perform alignment between them only with supervision at the entire scene level. Therefore, for efficiently aligning multimodal representations, we also propose a structural alignment objective for maximizing the correlation of corrected sound-object pairs by using the pervasive concurrency of sound and sound-maker which helps to infer the latent alignment by comparing the matching degree of different sound-object pairs.

Last, by leveraging the audiovisual supervision, we employ the contrastive loss to train the audiovisual network and infer the latent alignment simultaneously, which encourages the audiovisual network to have higher matching confidence for the aligned sound-image pair than the mismatched ones.

2.2. Curriculum Learning

Usually, the audiovisual scenes in the wild contain different amounts of sound-sources, we find that directly performing audiovisual learning with these data will make the model very difficult to optimize and also lower the alignment performance. To settle this problem, we propose to train the audiovisual model step by step, which is about starting from simple scene then gradually increasing the difficulty level, where the number of sound-sources is considered as the reference for audiovisual scene complexity. Since the audiovisual scene complexity is crucial for curriculum training, we also propose an audiovisual counting model to estimate the number of sound-source in a given scene.

In practice, to effectively perform curriculum learning, all the audiovisual data have been sorted from simple to complex before training, according to the number of sound-sources in the scene. Meanwhile, the proposed audiovisual learning model can effortlessly be targeted to the scene in different complexity level, i.e., the cluster number is accordingly set to the number of sources for different learning stages. Based on these graded audiovisual data, we can train the audiovisual learning model in a curriculum fashion. More importantly, these models in different stages share the same audiovisual learning network.

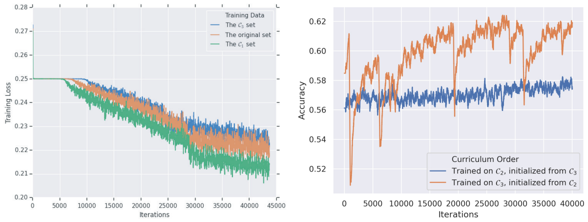
2.3. Audiovisual Perception

To effectively present proposed model's ability of object-level separation and alignment, we propose to perform sound source localization on visual scene. Moreover, to validate the effectiveness of inferred object representation further, we propose to perform sound separation based on visual guidance. The representative audiovisual separation network in [3] is adopted, and a variant of U-Net [7] is used to perform sound-source separation, similar to [3, 10].

3. Experiments

3.1. Curriculum Learning Evaluation

We aim to have an insight into how the curriculum strategy influences the audiovisual learning performance on **AudioSet-Balanced** Audioset [4] which are divided into different curriculums according to the number of contained sound-sources, e.g., the first curriculum $\mathcal{C}_1$ consists of videos with single sound-source. Finally, all the 19,443 valid video clips are divided into 9,239/7,098/2,685/421 $\mathcal{C}_1/\mathcal{C}_2/\mathcal{C}_3/\mathcal{C}_4$ curriculum clips. Concretely, to evaluate the effects of different audiovisual complexities, the original set and the curriculum set of $\mathcal{C}_1$ and $\mathcal{C}_3$ are selected for training the audiovisual network, respectively. From Fig. 3(a), we can obvious that the network trained with the simple curriculum of $\mathcal{C}_1$ enjoys the fastest convergence and lowest training loss, while the one trained with $\mathcal{C}_3$ suffers from the worst performance. In Fig. 3(b), we show the training accuracy obtained from two learning settings, one is firstly trained on the $\mathcal{C}_3$ set then the $\mathcal{C}_2$ set and the other is $\mathcal{C}_2$ then $\mathcal{C}_3$ set. Surprisingly, the model with correct curriculum order enjoys great advantages compared with the incorrect one. Hence, we consider that curriculum learning indeed helps to accelerate and improve the auiovisual learning performance.



(a) Curriculum effects (b) Curriculum order effects

Figure 3. The effects of curriculum learning in terms of training loss and accuracy.

Here, we perform acoustic scene classification by viewing the trained audiovisual model as a feature extractor. The ESC-50 dataset is chosen for evaluation and we follow the same pre-processing and train/test split as [6]. From the results in Table 1, we can summarize these results into four points. Firstly, learning with simple curriculum can provide proper initialization (51.25). Secondly, sound-source alignment better utilizes the audiovisual concurrency than scene-level alignment, especially in the complex scene (56.75 vs. 47.25). Thirdly, direct video-level alignment in the complex scene may deteriorate the pre-trained network (47.25 vs. 51.25). This is probably because the chaotic audiovisual correlation could confuse the scene matching objective, but it was ignored before. Fourthly, we validate that the merits of curriculum learning are not from more training data (56.75 vs. 53.00) again.

Table 1. Acoustic scene classification result on ESC-50, where $\mathcal{C}_1(\mathcal{C}_2)$ means trained on $\mathcal{C}_1$ but initialized from $\mathcal{C}_2$ and similarly for $\mathcal{C}_2(\mathcal{C}_1)$.

Training Strategy	Accuracy↑
$\mathcal{C}_1$	51.25
$\mathcal{C}_2(\mathcal{C}_1)+$ Scene-level alignment	47.25
$\mathcal{C}_2(\mathcal{C}_1)+$ Sound-source-level alignment	56.75
$\mathcal{C}_1(\mathcal{C}_2)+$ Sound-source-level alignment	53.00

3.2. Audiovisual Sound Localization

Table 2. Quantitative localization results on SoundNet-Flickr dataset [8]. AUC is the area under the cIoU curve. † means the methods are trained on SoundNet-Flickr, while others are trained on AudioSet-Balanced. ‡ means the results are based on the predicted results of Poisson regression model.

Methods	cIoU@0.5↑	AUC↑
Random	12.0	32.3
†Attention[8]	43.6	44.9
†DMC[6]	41.6	45.2
†‡Ours	48.4	47.4
Ours-$\mathcal{C}_1$	50.0	49.2
Ours-$\mathcal{C}_1$(unrelated)	19.2	36.8
Ours-$\mathcal{C}_2$	46.0	45.7

In this task, we aim to visualize the object location where the sound is produced. To evaluate the effectiveness of curriculum learning, our models trained in different curriculum levels are also considered. As shown in Table. 2, our models outperform all the other methods by a large margin. Compared with DMC [6], our model better localizes the sounding objects via the simpler and more effective feature aggregation mechanism (41.6 vs. 48.4). Moreover, our proposed structural alignment contributes to better align different sound-sources, even faced with multi-source scenes (41.6 vs. 46.0). Secondly, besides the aligned visual center, we also evaluate the unaligned visual center. As expected, they suffer from a large decline in both metrics, which indicates that our model can exactly distinguish sound-maker from background and align it with the produced sound (50.0 vs. 19.2). Thirdly, our model trained with curriculum $\mathcal{C}_2$ is worse than the one with $\mathcal{C}_1$ (50.0 vs. 46.0). This is because the test videos are all single-source, the multi-source videos in $\mathcal{C}_2$ may mix up the alignment knowledge learned in $\mathcal{C}_1$.

Table 3. Sound separation results.

(a) All the methods are trained only with solo videos in MIT-MUSIC-solo test dataset.

Methods	SDR↑	SIR↑
NMF-MFCC[9]	0.92	5.68
AV-Mix-Sep[3]	3.16	6.74
Sound-of-Pixels[10]	7.30	11.90
Co-Separation[3]	7.38	13.7
Ours	6.59	10.10

(b) Performance gain after training with both solo and duet videos, compared with only using solo.

Methods	ΔSDR↑	ΔSIR↑
NMF-MFCC[9]	0	0
AV-Mix-Sep[3]	0.07	0.27
Sound-of-Pixels[10]	-1.25	-2.09
Co-Separation[3]	0.26	0.10
Ours	0.54	0.94

3.3. Sound Separation

We evaluate the audiovisual sound separation performance on the MIT-MUSIC dataset. Table 3(a) infers that although the previous methods use additional visual knowledge for guiding the sound separation, e.g., ImageNet-pretrained visual model in Sound-of-Pixel [10] and finetuned instrument detector in Co-Separation [3], our model trained from scratch still shows comparable results in both SDR and SIR. Such phenomenon confirmed the quality of the sound-source localization results learned by completely self-supervised signal, which can provide effective visual representation of specific sound-maker. To validate the effectiveness of curriculum learning, our proposed model is further trained with the complex scene of duet videos. Table 3(b) shows the ablation results. After introducing more complex audiovisual data, the compared methods do not show obvious improvements when using identical models, even declines. In contrast, our differentiated audiovisual model can better utilize complex scenes to improve the ability of cross-modal perception by setting the number of sound-source to two, which attributes to merits of curriculum learning.

4. Conclusion

This paper proposes a self-supervised audiovisual learning model to perform targeted audiovisual content detection and structural alignment in the scene with heterogeneous complexity. A curriculum learning strategy is proposed to effectively train the model. We achieved noticeable audiovisual localization performance on object localization and sound separation.

References

[1] C. Bauckhage. Lecture notes on data science: Soft k-means clustering. Technical report, Technical Report, Univ. Bonn, DOI: 10.13140/RG. 2.1. 3582.6643. 2

[2] B. De Brabandere, D. Neven, and L. Van Gool. Semantic instance segmentation with a discriminative loss function. *arXiv preprint arXiv:1708.02551*, 2017. 2

[3] R. Gao and K. Grauman. Co-separating sounds of visual objects. *arXiv preprint arXiv:1904.07750*, 2019. 1, 2, 3, 4

[4] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2017. 3

[5] J. R. Hershey, C. Zhuo, J. L. Roux, and S. Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *IEEE International Conference on Acoustics*, 2016. 2

[6] D. Hu, F. Nie, and X. Li. Deep multimodal clustering for unsupervised audiovisual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9248–9257, 2019. 1, 2, 3

[7] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 3

[8] A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, and I. So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4358–4366, 2018. 3

[9] M. Spiertz and V. Gnann. Source-filter based clustering for monaural blind source separation. In *Proceedings of the 12th International Conference on Digital Audio Effects*, 2009. 4

[10] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba. The sound of pixels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 570–586, 2018. 1, 2, 3, 4