Cyclic Co-Learning of Sounding Object Visual Grounding and Sound Separation Supplementary Material

Yapeng Tian[1], Di Hu[2,3*], Chenliang Xu[1*]
[1]University of Rochester, [2] Gaoling School of Artificial Intelligence, Renmin University of China
[3] Beijing Key Laboratory of Big Data Management and Analysis Methods
{yapengtian,chenliang.xu}@rochester.edu, dihu@ruc.edu.cn

We include this appendix to describe more details about audio-visual sound separation network and our implementation. Moreover, we provide more evaluation results for silent sound separation.

Appendix

Audio-Visual Sound Separation Network

We adopt the same audio-visual separator as in [4], which consists of three modules: audio network, visual network, and audio-visual sound synthesizing network.

We use Time-Frequency representation of sound and project raw waveform to spectrogram with the STFT. The audio network transforms magnitude of STFT spectrogram S_m from the input audio mixture to a C-channel feature map $A_m \in \mathcal{R}^{C\times F\times T}$ with an U-Net [3] structure, where $C = 32$ in our experiments. We use an ResNet-18 [1] followed by a linear layer to predict a C-dimension object feature $f_{o_i^{(k)}} \in \mathcal{R}^C$ for each object $O_i^{(k)}$. The audio-visual sound synthesizing network takes A_m and $f_{o_i^{(k)}}$ as inputs and output a spectrogram mask $M_i^{(k)} \in \mathcal{R}^{F\times T}$ with a linearly transformed dot product of the audio and visual features. The separated sound spectrogram: $S_i^{(k)} = S_m \odot M_i^{(k)}$ can be obtained by masking the sound mixture, where $\odot$ is the element-wise multiplication operator. The waveform $s_i^{(k)}$ of the object can be reconstructed by the inverse short-time Fourier transform.

Implementation Details

We train our networks using PyTorch [2] library with 4 NVIDIA 1080Ti GPUs. The batch size and epoch number are set to 48 and 60, respectively. The learning rate is set to $1e-4$ and it will decrease by multiplying 0.1 at 30- and 50-th epochs, respectively. For the three-step training, we train 60 epochs for each step. We conduct ablation studies with several models: Grounding only, Random Obj, CoL, and CCoL, they are defined as follows:

Grounding Only: The Grounding only model is trained only with grounding losses: l_{grd_s};

Random Obj: This baseline randomly selects objects from detected object proposals to train a audio-visual sound separation model;

CoL: The co-learning model (CoL) jointly learns visual grounding and sound separation using the l_{col};

CCoL: The cyclic co-learning (CCoL) further strengthens the interaction between the two tasks optimized by l_{ccol}.

Additional Evaluation

To further validate silent sound separation performance, we compute the sound energy for separated sounds of silent objects. We empirically set a threshold: 20 to decide whether the separation is successful. The success rates for silent objects classification are 8.6% and 92.3% for SoP and CCoL, respectively. An interesting observation is that the success rate of our CCoL is even higher than the corresponding grounding accuracy, which demonstrates that our model tends to generate weak sounds for silent objects.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[2] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019.

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 2015.

*Corresponding authors.

[4] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *ECCV*, 2018.