Deep Cross-Modal Audio-Visual Generation

Lele Chen[1], **Sudhanshu Srivastava**[1], **Zhiyao Duan**[2] and **Chenliang Xu**[1]

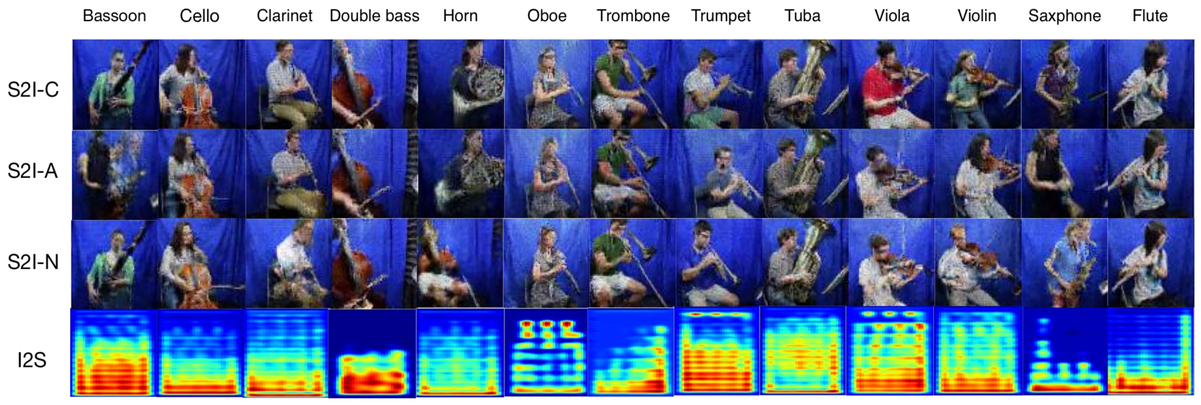
{lchen63, ssrivas6}@cs.rochester.edu, {zhiyao.duan, chenliang.xu}@rochester.edu

[1]: Department of Computer Science, [2]:Department of Electrical and Computer Engineering

Introduction

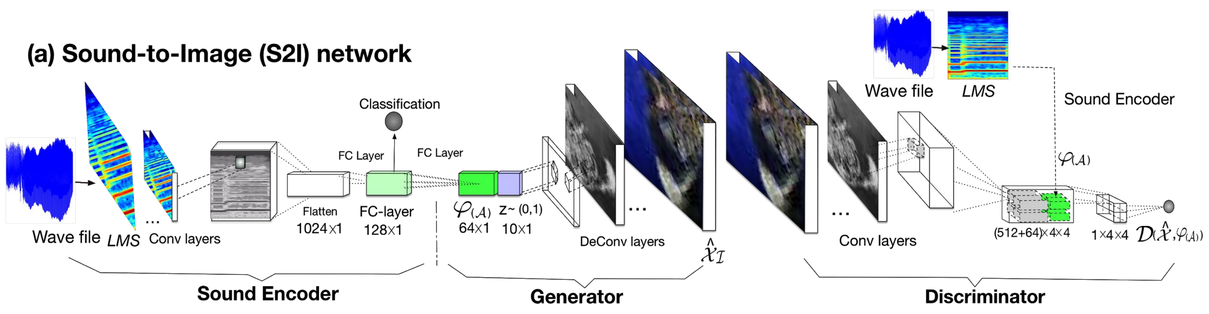
Motivation: Cross-modal audio-visual perception has been a long-lasting topic in neurology and psychology. However, the problem of computational cross-modal audio-visual generation has not been systematically studied in computer vision, audition or multimedia communities.

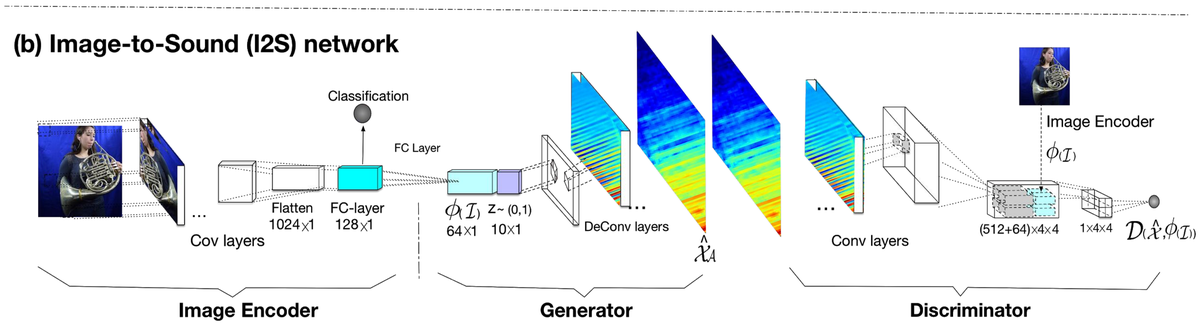
Objective: In this paper, we make the first attempt to solve this cross-modal generation problem leveraging the power of deep generative adversarial training. Our system is trained with pairs of visual and audios, which are typically contained in videos, and is able to generate one modality (image/sound) from the other modality (sound/image). We generate images in two scenarios, instrument-oriented and pose-oriented; we generate sound in log mel-spectrum (LMS).



Generated outputs using our cross-modal audio-visual generation models

Cross-Modal Generation Model





The overall diagram of (a)S2I GAN network and (b) I2S GAN network

S2I Generator: The S2I generator is denoted as: $G_{S\mapsto I}:\mathbb{R}^{|\varphi(A)|}\times\mathbb{R}^{Z}\mapsto\mathbb{R}^{I}$. The sound encoding vector of size 128 is first compressed to a vector of size 64 via a fully connected layer followed by a leaky ReLU, which is denoted as $\varphi(A)$. Then it is concatenated with a random noise vector $z\in\mathbb{R}^{Z}$. The generator takes this concatenated vector and produces a synthetic image $\hat{x}_I\leftarrow G_{S\mapsto I}(z,\varphi(A))$ of size 64x64x3.

S2I Discriminator: The S2I discriminator is denoted as: $D_{S\mapsto I}:\mathbb{R}^{I}\times\mathbb{R}^{|\varphi(A)|}\mapsto[0,1]$. It takes an image and a compressed sound encoding vector and produces a score for this pair being a genuine pair of image and sound.

Conditional GANs: $\min_G\max_D V(D,G)=\mathbb{E}_{x\sim p_{data}(x)}[\log D(x|y)]+\mathbb{E}_{x\sim p_z(z)}[\log(1-D(G(z|y)))]$. Here y is a condition-a label for example, $G(z|y)$ is the output sample generated by the Generator network given noise z and condition y, $D(x|y)$ is a score between 0 and 1 corresponding to how genuine the sample x is as an element satisfying condition y.

Sound Encoder and Image Encoder

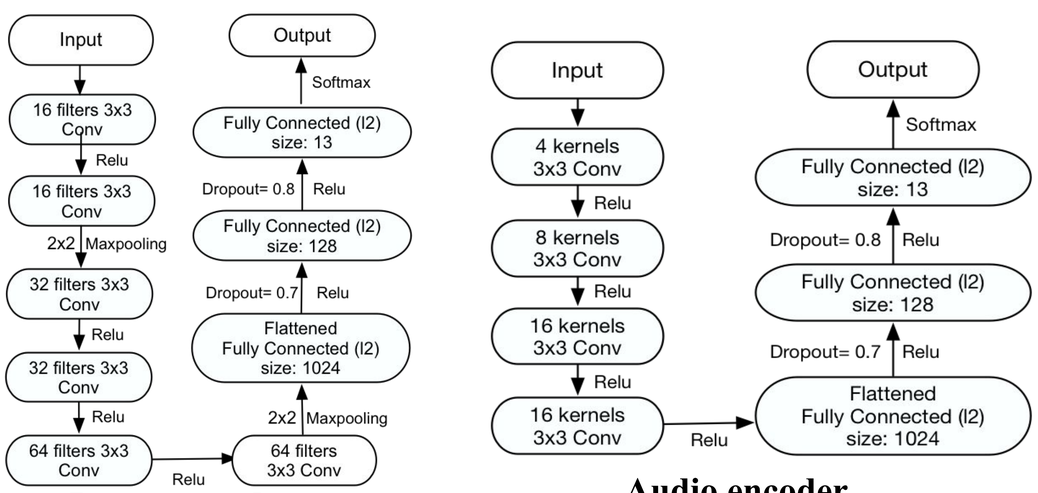


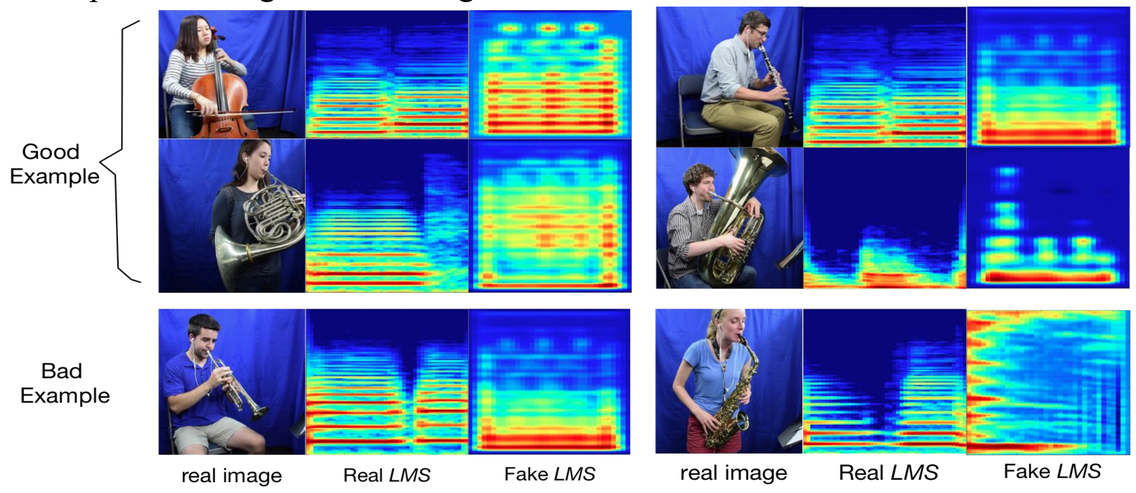
Image encoder

Audio encoder

Results

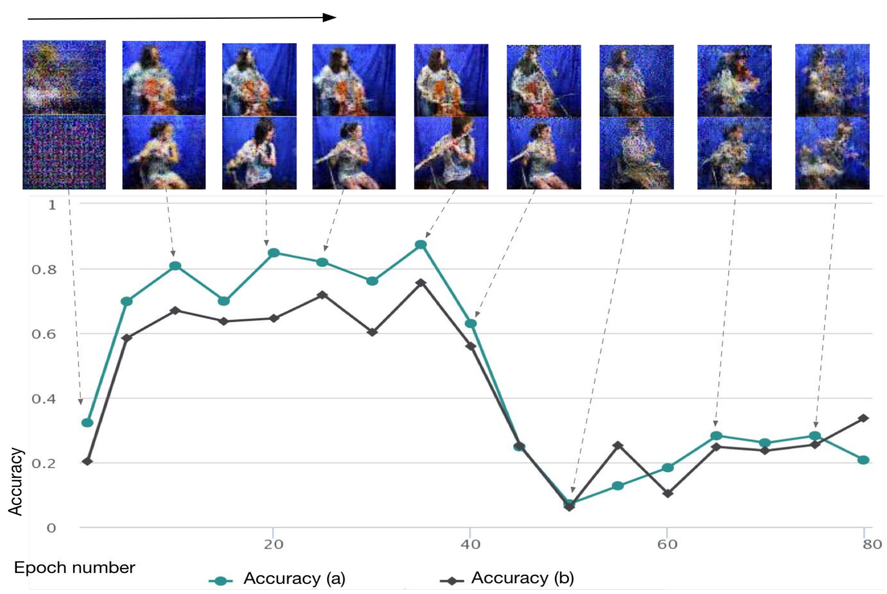
Generated images with different poses

S2I (Pose-oriented): Figure above shows output of pose-oriented model. We can see this model can generate a sequence of images. **I2S:** Figure below shows the mel-spectrum we generated using I2S network.



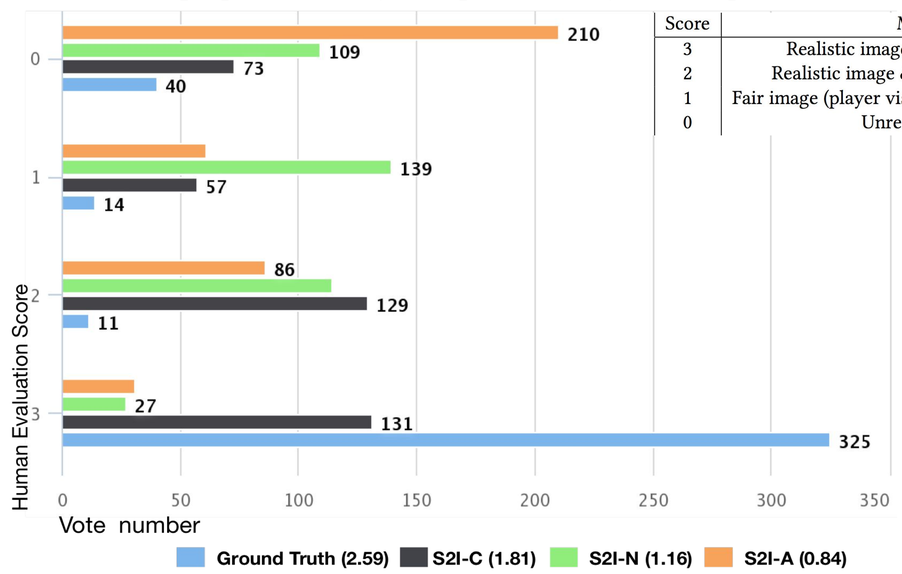
Generated audio LMS of different instruments

Evaluation



Classification accuracy on generated images

Classifier-based evaluation: We build a image verifier trained by real images (acc >95%). Figure above shows the relation between image quality and verification accuracy. **Human-based evaluation:** We also have human subjects evaluate our sound-to-image generation (see figure below). Score guideline is shown in the table.



Score	Meaning
3	Realistic image & match instrument
2	Realistic image & mismatch instrument
1	Fair image (player visible, instrument not visible)
0	Unrealistic image

Result of human evaluation on generated images