Towards Audio to Scene Image Synthesis using Generative Adversarial Network

Chia-Hung, Wan
National Taiwan University
wjohn1483@gmail.com

Shun-Po, Chuang
National Taiwan University
alex82528@hotmail.com.tw

Hung-Yi, Lee
National Taiwan University
hungyilee@ntu.edu.tw

Abstract

Humans can imagine a scene from a sound. We want machines to do so by using conditional generative adversarial networks (GANs). By applying the techniques including spectral norm, projection discriminator and auxiliary classifier, compared with naive conditional GAN, the model can generate images with better quality in terms of both subjective and objective evaluations. Almost three-fourth of people agree that our model have the ability to generate images related to sounds. By inputting different volumes of the same sound, our model output different scales of changes based on the volumes, showing that our model truly knows the relationship between sounds and images to some extent.

1 Introduction

People now are trying to make machines work like humans. Researchers are attempting to teach machines to comprehend natural languages, to understand the content in images, etc. After understanding the content, we also want machines to describe what they see [1][2]. In addition, we also want machines to have the ability to imagine. In the task of text-to-image [3], machine can turn text descriptions into images. In this paper, we want machines to imagine the scenes by listening to sounds. We hope that when hearing sounds, machine can draw the object that is making sounds and the scene that the sound is made. For instance, after hearing the sparrows chirp, machine can draw a picture of sparrows with probably trees or grass as background.

The technology we use to learn an audio-to-image generator is based on GAN. In this paper, we fuse several advanced techniques of conditional GANs [4] including spectral normalization [5], hinge loss [6][7], projection discriminator [8] and auxiliary classifier [9] into one model. Machine learns the relationships between audio and visual information from watching videos. We create a dataset from SoundNet Dataset [10] by using pretrained image classification and sound classification models to apply data cleaning. After training, the audio-to-image generator can produce recognizable images, and the advanced techniques of conditional GAN achieve better Inception score [11][12] than the naive conditional GAN. In addition, we show that our model learns the relationship between sounds and images by inputting the same sound with different volume levels.

2 Dataset

In previous work [10], videos crawled from the webs are used to train a sound classification model, SoundNet, to classify where or what is in the sounds. Here we use the screenshots of videos and sound segment files in the dataset to train our audio-to-image models. Most of sound segment files in our dataset are around 30 seconds long, and we resize all the screenshots to size of 64*64. However, we found that the corpus for training SoundNet cannot be directly used to train audio-to-image models because there are some discrepancies between images and sounds. The screenshots and the sounds of videos can be unrelated. To relieve the difficulties of learning sound-image matching, we use an image classifier and a sound classifier to clean up the dataset automatically. We classified sounds in

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

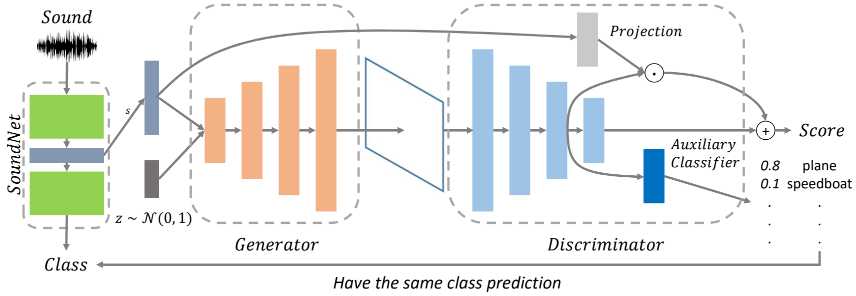


Figure 1: Model architecture with projection discriminator and auxiliary classifier.

those videos into categories by the pretrained sound classification model, SoundNet [10]. We also use Inception model[13], an image classification model, to classify the images. If the classification results for the image and sound are not the same, the sound-image pair would be discarded. Because the above data cleaning procedure is automatic, it cannot be perfect, but it remarkably improves the quality of the generation results.

Because some objects are very rare in the training data, to make the training of audio-to-image plausible, we chose nine classes with the most examples in the training data. The number of training examples for each class is listed in Table 1. The total number of sound-image pairs for training is 10701, and the total number of sound segments for testing is 248.

Class	Plane	Soccer	Piano	Baseball	Speedboat	Dam	Dog	Drum	Guitar
# of data	2803	2077	1899	1708	900	584	264	259	207

Table 1: Number of training data in different classes.

3 Approach

Due to the success of text-to-image synthesis [3], which utilized text embeddings as condition for generators to generate correlated images, our work is based on similar model architecture. Recently, there are some researches trying to improve the generation by limiting discriminator to be a function in 1-Lipschitz continuity [8][5][14] or utilizing another auxiliary classifier in discriminator [9]. We fuse these approaches into one model. Therefore, although the algorithm for GAN training is similar to text-to-image, the discriminator architecture and loss function used here are very different. The model architecture is illustrated in Figure 1.

3.1 Generator

The generator is shown in the left hand side of Figure 1. The input sound segment is first represented by a sequence of feature.Using SoundNet for feature extraction is illustrated in Figure 1. Then all the features in the sequence are averaged into a single vector s. The vector s is taken as the condition of the generator. Then, we concatenate a noise vector z sampled from normal distribution with our sound condition as the input to generator. Generator is the cascade of several transposed convolution layers with hyperbolic tangent function as the activation function in the last layer. The output of the generator is an image generated based on the input condition.

3.2 Discriminator

The discriminator is in the right hand side of Figure 1. The discriminator takes a pair of sound segment and image as input, and outputs a score. The architecture of discriminator is the cascade of several convolution layers with spectral normalization [5] in each layer. The convolution layers takes an image as input and outputs a scalar representing the quality of the image. The projection layer which is simply a linear transformation projects the sound vector into a latent representation [8]. Then by computing inner-product between projected vector and the output of one of the convolution layer, we obtain a similarity score representing the degree of match between the audio and image.

The final output of the discriminator is the addition of the similarity score and the scalar that solely comes from convolution layers. The final score represents not only the realness of images but also relevance between sounds and images. The discriminator learns to assign large score to the sound-image pairs in the training data, and low score to the sound and its generated image. While the generator tries to fool discriminator, it learns how to generate images which are relevant to input condition and looks like real photos.

In Figure 1, there is an auxiliary classifier. The classifier shares weights with the convolution layers in discriminator, and they are jointly learned. Because in the training data, the class of the sound segment and image pair can be obtained by SoundNet and Inception model, the classifier can learn to predict the class of an input image from the training data. The generator will learn to generate images that can be correctly classified by the auxiliary classifier.

4 Experiments

Our training procedure follows standard GAN training algorithm. Generator is composed of four deconvolution layers with ascending number of kernels. Discriminator is composed of four convolution layers and with linear function as activation function of final layer. To keep this adversarial training procedure in balance, more training steps are needed for generator to catch up discriminator. We train generator five times per each update of discriminator. The input dimension is 266 which consist of 256-dimension SoundNet feature and 10-dimension z sampled from normal distribution. The whole optimization process is based on Adam optimizer with learning rate 0.0002, and we train 300 epochs for all experiments. Generated images and corresponding audio files in this section can be found in `https://wjohn1483.github.io/audio_to_scene/index.html`.

4.1 Qualitative Results

Sampled images from generator by inputting the sounds not in training data are shown in Fig 2. Sounds belonging to some classes can generate relatively high quality images. For speedboat or plane, there are eye-catching objects in the generated images. The generator truly generates the images that are interpretable to some extent. Some classes of images get worse quality of images than others. This may be because the imbalance and variance in different classes of training data. The number of training examples may explains why some classes performed better than the others. We also found that for all the sounds classified into drums, they still have very high diversity. There are many kinds of drum and are played in variant places. It becomes an obstacle for model to generate image from the sound of such class.

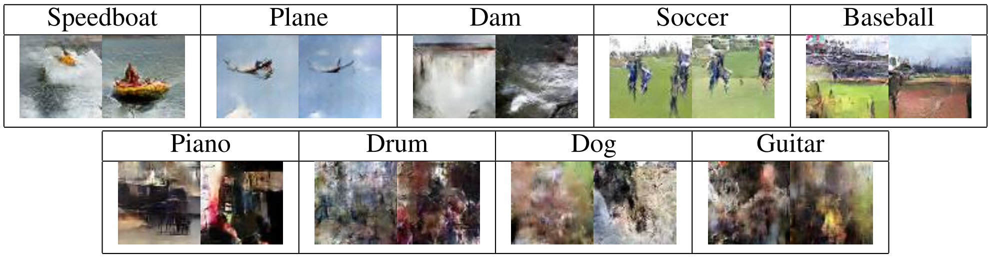


Figure 2: Samples from our model. Each image is generated from a sound segment. The labels are the classes predicted by SoundNet.

4.2 Sound Volume

To further investigate whether our model truly learns the relation between sound an vision, we tune the volume of sounds to observe the influences on generated images. We input those tuned sound features into our generator which was pretrained on standard volume scale. The images are shown in Table 2. The numbers on top indicates the scale of volume that we modified our sound files. In those images, we can see different scale of splashes. As the volume goes up, the scale of splashes become larger. We can see that our model truly learned the relation between characteristic of sound and image. In this case, the volume of sounds is reflect on splashes.

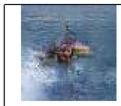
0.5 times	Original	2 times	3 times	0.5 times	Original	2 times	3 times
							

Table 2: Generated images by inputting different volumes of sounds. The numbers in the table is the relative loudness to the original sound. (Left 4 is the class of speedboat; Right 4 is the class of dam)

4.3 Ablation Study

The architecture of our model contains spectral normalization, hinge version loss, projection discriminator and auxiliary classifier. In this subsection, we want to know the influence of each part in our model. Table 3-(I) shows the Inception score of different types of model. Row (a) shows the upper bound of this task, which is obtained by inputting all the real images we have in training and testing data to calculate Inception score. The Inception score obtained in this way is 4.44, which is the highest score we can get. We can use this upper bound score as a criterion to measure the quality between generated images and real images. In both rows (b) and (c), we used the same network architecture as in [3], but we substitute sound embedding for sentence embedding. In row (b), we apply improved W-GAN[14] on original text to image architecture, which use gradient penalty to make sure discriminator is in 1-Lipschitz continuity.

The table shows that improved W-GAN cannot get good Inception score in this task. On the other hand, conditional GAN can perform better. By adding different tricks mentioned above, we can get improvements step by step. It shows that tricks do help our model to generate better images. Finally, with all the technologies, we can get 2.83 in Inception score, which performs relatively good compare to our upper bound.

Model	(I) Inception Score	(II) Human avg. score
(a) Upper bound	4.44 ± 1.91	-
(b) Improved WGAN	1.42 ± 0.13	-
(c) Conditional GAN	2.21 ± 0.38	-
(d) + Spectral Norm	2.45 ± 0.48	1.90
(e) + Hinge Loss	2.49 ± 0.51	2.74
(f) + Projection Discriminator	2.61 ± 0.41	3.16
(g) + Auxiliary Classifier	2.83 ± 0.53	3.70

Table 3: (I) Inception scores of different models. (II) Human scores on different models.

4.4 Human Evaluation

4.4.1 Evaluation on ablation study

In this section, we want to prove that the improvement of different models is not only shown on Inception score but also on human feeling. We ask ten people to help us evaluate our models. Our experimental setup is as follows, we sample some pairs of image and corresponding sounds in testing data. Then, let people listen to those testing sounds and rate from 1 to 5. If the generated image is unreal or uncorrelated to testing sound, people should rate this pair with lower score. On the contrary, if the generated image seems real enough and have high correlation with sound, this pair should get higher score. The results are shown in Table 3-(II). We can see that most people think the model with all tricks performed the best. Although those models get close scores in Inception score, they get scores which have at least 0.4 gap between different models.

4.4.2 Correlation between sounds and images

To measure the correlation between sounds and images, we ask people to choose the most correlated image from two different images after hearing a sound from testing data. These two images are conditioned on different class of sounds so that if our model can generate images related to given class, people will choose the corresponding image which is generated by inputting sound that they just listen to, rather than image generated by inputting sampled sound from other classes.

Our results shows that 73% people choose the image generated by the sound they hear, 11% people choose the image generated by sound sampled from other classes, and 16% people think both of the images cannot represent the sound they listen to. Most of the people think the images that our model generated are correlated to input sounds. It shows that our model has the ability to generate images related to given sounds.

5 Conclusion

In this paper, we introduce a novel task in which images are generated conditioned on sounds. Base on SoundNet dataset, we utilize image and sound classification results to build a relatively cleaner image-sound paired dataset. By applying different methods to our generative model, the model can generate images with better quality in terms of both subjective and objective evaluations.

References

[1] Subhashini Venugopalan, Marcus Rohrbach, Jeff Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence – video to text. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[2] Shun-Po Chuang, Chia-Hung Wan, Pang-Chi Huang, Chi-Yu Yang, and Hung-Yi Lee. Seeing and hearing too: Audio representation for video captioning. In *Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE*, pages 381–388. IEEE, 2017.

[3] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.

[4] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[5] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

[6] Jae Hyun Lim and Jong Chul Ye. Geometric gan. *arXiv preprint arXiv:1705.02894*, 2017.

[7] Dustin Tran, Rajesh Ranganath, and David M Blei. Deep and hierarchical implicit models. *arXiv preprint arXiv:1702.08896*, 2017.

[8] Takeru Miyato and Masanori Koyama. cgans with projection discriminator. *arXiv preprint arXiv:1802.05637*, 2018.

[9] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. *arXiv preprint arXiv:1610.09585*, 2016.

[10] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, 2016.

[11] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016.

[12] Shane Barratt and Rishi Sharma. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018.

[13] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[14] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.