Acoustics-Image Translation using Dual-GANs

Muhammad Sheheryar Naveed and Professor Juan Ye

The fast-technological advancement has grown even more rapidly than humans could've ever thought. The pace with which the Ai industry is moving is uncatchable due to the wide range of applications of deep learning. One specialized gizmo that adds to the computer vision industry is called dual-GANs. GAN (Generative Adversarial Networks) are considered backbone of Ai due to their enormous application in the field. Generally, GANs are used for data augmentation, however, the technique behind the idea has further opened a door of different experiments where deep learning can be applied.

GAN consist of a *discriminator* and *generator*. Both generator and discriminator are simple neural networks where discriminator tells whether the given input is true or false whereas a generator tries to fool the discriminator by making sure that the output it produces is too accurate that when fed to discriminator, discriminator specifies it as true. In this way both after several training sets achieve decent accuracies. In this way, given a single dataset, generator would be able to produce several datasets i.e. data augmentation. To delve deeper into the GAN technology, visit here.

Taking the GAN approach into action, another way to use generator and discriminator is to use two GANs together. While GANs are limited to only one domain, that is, GAN if fed with images would produce another set of images, dual-GANs can be used in dual domain, for instance, one GAN focuses on producing the real-time images whereas the other on producing the sketches. Hence, once dual-GAN is trained, the model would be in this case be able to produce the sketch of real time images when real-time images are fed and real-time images out of sketches when sketches are fed.

Dual GAN is not just limited to images but to any kind of data and the two domains can even be completely different from each other as far as the implementation is correct. Likewise, I used dual GAN where domain A was audio data and domain B was image data and thus did image to sound translation and vice versa.

Architecture: Let's call the GAN that converts sound to image the primal GAN and the GAN that converts image to sound, the secondary GAN. For the primal GAN the architecture of the discriminator was five conv2D layers with spectral norm along with 'leaky relu' function followed by the final linear layer. The same goes for the secondary discriminator's layers but with different filter sizes along the layers. For the primal generator (Image to sound translation), the architect was five two-dimensional deconvolutional layers and the same is for the secondary generator but there are two two-dimensional convolutional layers before that with 'leaky relu' as activation along with batch normalization. The image data was taken from a source called UrbanSound8K and fed as it is whereas for the audio data, it is too abstract for deep learning model to learn something out of it so it was first fed into soundnet tensorflow model which extracts features out of it and those were stored in .npy files and then those extracted features were fed into the model. Librosa was tried to do the same thing but it was too abstract, and the idea didn't work out much.

Application: This project has wide range of application and one such important application is the assistance of blind people to help them know of any obstacle in front of them.