1. Greetings everyone, my name is Dmitriy Volynov and today I would like to introduce a pioneer development which has a great potential to change a part of the music industry. What I’m going to show you is only a small part of what this technology can do in the future. This project was created together with my friend Nikita Pestretsov. So, let’s go.
2. I want to start with what a style transfer is all about. It is a technique where we take the artistic style of one image and apply it to another, creating a unique blend. Here is a combination of the content image, this parrot, and the style which creates specific patterns on the content.
3. For this stuff, people mainly use the neural network VGG-19. It is a network which has been pre-trained to extract features of various objects in the picture.
4. So, essentially, our idea was to try the same style transfer, but for music; And, if it’s possible, to create some crazy mixes like a popular rapper and a symphonic orchestra, or a serenade and a march. And what exactly do we need to do this?
5. Initial audio, presented as a sine wave, is not suitable for extracting crucial features of the whole melody. We need something different.
6. It’s better to use an audio spectrogram. It is a representation of audio frequencies, how they change over time. It’s usually used for audio neural networks. Another advantage is that we are able to convert it to an image. So, now we can try to feed this spectrogram to the VGG network, which we used to transfer the style to the parrot.
7. But here is a problem. The VGG model has been trained to extract features from specific objects like cars, animals, people, but not spectrograms. Therefore, we should write a custom network.
8. Our network consists of only one convolution layer, but it can properly extract 4 thousand features from the spectrogram.
9. Also we wrote two custom functions which calculate the losses. Loss functions measure how well the network’s predictions match the desired outputs. By calculating it the neural network can adjust its parameters to minimize this difference during training. The first function “Get content loss” is responsible for keeping a parrot as a result; for this stuff it uses Gram matrices. And the second one “Get style loss” calculates style generation over the content image.
10. So here you can see the process of training and how the spectrogram changes over time. Initially it is a noisy picture and loss is really huge, then it optimizes to be something like a spectrogram.
11. Now, when we know how to train it, we need to understand how to convert the result spectrogram back to audio. And here we use Griffin-Lim algorithm and Inverse Fourier transform which allows us to convert it back.
12. In conclusion, we think this technology would take a huge part of the modern music industry since high computational power. Here are some articles which helped us in accomplishing our goals. And also here is a link of the whole project so each of you can try to create something interesting.
13. Thank you for your attention!