УДК 004.85

МОДЕЛИРОВАНИЕ ИЗОБРАЖЕНИЯ ЛИЦА С ИСПОЛЬЗОВАНИЕМ РЕЧЕВОГО СИГНАЛА ПРИ ПОМОЩИ НЕЙРОННЫХ СЕТЕЙ ДЛЯ СИСТЕМ ВИДЕОКОНФЕРЕНЦСВЯЗИ

Шубина Е. А. (бакалавр, Университет ИТМО)
Научный руководитель – Беляев Е.А.
(к.т.н., вед. н.с., Университет ИТМО)

В данной работе рассматривается архитектура нейронной сети для генерации недостающих изображений в последовательности видеокадров с использованием звуковой дорожки.

Введение. В современных системах видеоконференцсвязи довольно часто возникают ситуации, когда пропускной способности канала связи не хватает для устойчивой передачи видеоданных. В такой ситуации передатчик либо вынужден существенно увеличить коэффициент сжатия видеоданных, что приводит к ухудшению качества передаваемой видеоинформации, либо к прекращению передачи видеоданных (передаётся только звук). Кроме этого, в канале связи могут возникать потери пакетов, из-за чего возникают дополнительные искажения или "замирания" при воспроизведении.

Основная часть. В настоящей работе предлагается замещать видео сгенерированным изображением лица человека, движения губ которого повторяют произносимый текст. Таким образом задача заключается в генерации изображения с определенными параметрами на основе звукового канала.

Поскольку данная задача является задачей генерации объекта, то для ее решения была выбрана архитектура нейронной сети GAN (Generative adversarial network). Архитектура представляет из себя две нейронные сети - генератор и дискриминатор. Задача генератора состоит в создании объекта похожего на объекты из тестового множества. Задача дискриминатора - в определении, принадлежит ли полученный на вход объект к тестовому множеству. Таким образом генератор и дискриминатор играют в антагонистическую игру.

Перед обучением нейронной сети необходимо обработать данные, для извлечения признаков. При этом звуковая дорожка и последовательность видеокадров обрабатываются отдельно друг от друга. Звуковая дорожка преобразуется в спектрограмму, которая разделяется на фрагменты. Число фрагментов совпадает с количеством видеокадров. Видеодорожка обрабатывается с помощью технологии распознавания лиц, представленных в библиотеке dlib. Выделяются 64 точки кадра, соответствующие контуру лица, губ, глаз и носа. В данной работе предлагается использовать только положение губ на кадре, так как положение носа и глаз оказывает существенно меньше влияния на формирование звуков. Полученные координаты и спектрограммы объединяются в один массив, который затем передается на вход нейронной сети.

Так как работать предстоит с изображениями и сигналами то было решено использовать в генераторе и дискриминаторе сверточные слои. Они позволяют отфильтровать изображение, оставив необходимые нам признаки и сфокусироваться на них. Так как обработка видео это задача, которая требует больших временных затрат, то для данных нейронных сетей была выбрана архитектура U-Net. Эта архитектура позволяет добиться одинаковых результатов с обычной сверточной сетью, используя меньшее количество обучающих данных. В этом случае нейронная сеть состоит из нескольких уровней. На каждом уровне присутствуют от одного до трех сверточных слоев, со стандартным ядром размера 3x3. Перва половина нейронной сети осуществляет сужение выходных данных с увеличением количества фильтров на каждом слое. Это позволяет настроить нейронную сеть так, чтобы уделять внимание наиболее важным признакам. Вторая часть нейронной сети работает наоборот, расширяя изображение, с уменьшением количества фильтров. В качестве функции активации используется функция ReLU, а в качестве метода обучения - метод обратного

распространения ошибки, с помощью стохастического градиентного спуска. Аудиоданные обрабатываются схожим образом, проходя несколько сверточных слоев с увеличением количества фильтров. Данная архитектура с небольшими различиями в слоях нормализации, параметрах сверток и функциях активации применялась и для генератора, и для дискриминатора. Так как в качестве метода обучения использовался метод обратного распространения ошибки, то для предотвращения ошибки исчезающих градиентов было решено использовать остаточные слои (Residual layers), то есть слои, в которых значение слоя подается на вход не только следующему слою, но и нескольким после.

Выводы. Полученные в ходе исследования результаты показывают, что с помощью аудиоданных возможно генерировать достаточно точное положение губ на изображении.

Шубина Е. А. (автор) Подпись

Беляев Е.А. (научный руководитель) Подпись