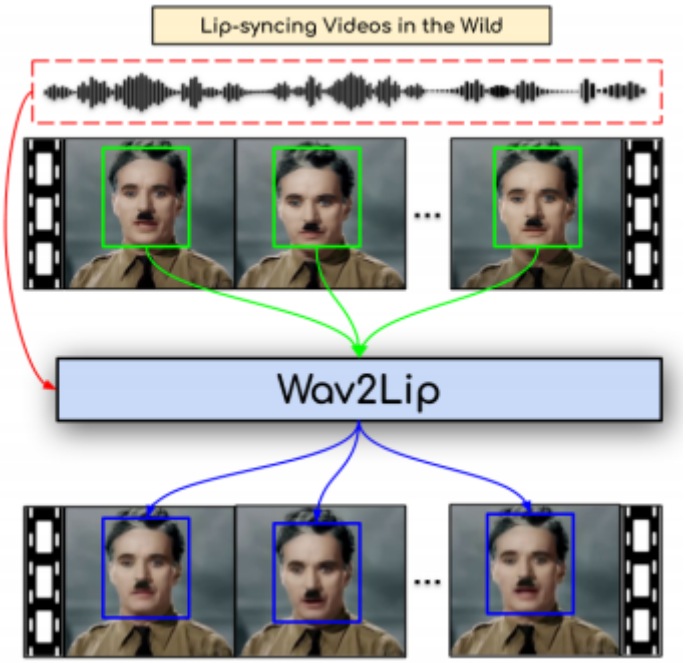
Всутпление

Wav2Lip - это генеративная сеть принимающая на вход видеоряд и аудио с записью человеческой речи. Финальная задача - максимально точно синхронизировать движения губ на видео с записью речи. У этой технологии может быть много применений, но самое, как нам кажется, очевидное - синхронизация губ при дубляже в фильмах.
У данное работы есть два аналога - это SyncNet и LipGAN, но оба они дают результаты значительно хуже. Wac2Lip является продолжением идеи LipGAN, основные изменения тут лежат в дискриминаторе, в то время как генератор модели остался практически неизменным.



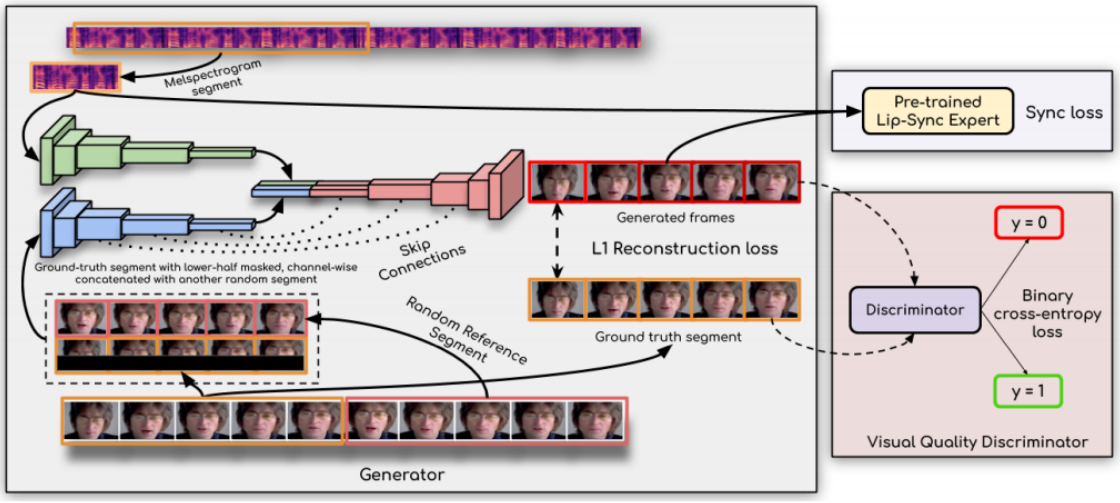
Генератор

Как уже было сказано выше, генератор Wav2Lip особо не выделяется на фоне конкурентов. На вход в генератор подаются отдельно мелспектрограмма записи речи и кадр из видеоряда. Мелспектрограмма прогоняется через speech энкодер, а изоражения соответственно через identity энкодер, после чего полученные feature map конкатенируются и подаются в face декодер, на выходе из которого уже будет готовый фрейм.
Оба энкодера представляют из себя обычные свертки, декодер - набор двумерных конволюций с добавлением ConvTransposed. В целом, архитектура без изменений была скопирована из LipGAN.

Дискриминатор

Основное же нововведение статьи заключается в дискриминаторе. Авторы замечают, что проблема LipGAN заключается именно в плохом дискриминаторе(его точности достигает лишь 54% в распознавание десинхронизации губ с речью). Корень проблемы заключается в том, что при обучении дискриминатора на результатах генератора, как это обычно делается в GAN, дискриминатор начинает концентрироваться на визуальных артефактах, чего ему абсолютно достаточно для обнаружения сгенерированного изображения, особенно в начале обучения, в то время как сама синхронизация мимики остается не у дел. Авторы предлагают использовать предобученный заранее на реальных данных дискриминатор, и штрафовать генератор за неверную синхронизацию уже им. Этот дискриминатор в статье называется Lip-Sync Expert, и по архитектуре схож с SyncNet.

Этого дискриминатора недостаточно, так как визуальные артефакты все еще остаются проблемой, поэтому при обучении также используется уже привычный нам дискриминатор для обнаружения этих артефактов обучаемый совместно с генератором, на его выходах. Собственно, тот же самый дискриминатор, что и в LipGAN.
В итоге добавление нового дискриминатора для синхронизации привело к существенному улучшению результатов.
Помимо описанных выше нововведений авторы также подают в дискриминатор несколько кадров, что логично, ведь контекст в аудио очень важен.



Обучение сети

Первым делом обучается дискриминатор синхронизации. Для его обучения используются цветные(в отличие от SyncNet) изображения. В качестве функции потерь используется косинусное расстояние и бинарная кросс энтропия. В итоге вероятности десинхронизации определяется как:

$$P_{sync} = \frac{v^{*}s}{max(||v\ ||_2 + ||s||_2, e)}$$

, где v и s - эмбеддинги видео и аудио соответственно. При обучении используется размер батча 64, окно в 5 кадров и оптимизатор Adam с шагом в 1e-3. Авторы статьи утверждают, что им удалось обучить данный Lip-Sync Expert до точности 91%, напомним, что точности дискриминатора в LipGAN составил 54%.

Далее обучается генератор и обычный дискриминатор. У дискриминатора лосс стандартный - $L_{disc} = E_{x`L_G}[log(D(x))] + L_{gen}$, где $L_{gen} = E_{x`L_g}[log(1 - D(x))]$. Лосс синхронизации

считается как $E_{sync} = \frac{1}{N}\sum - log(P^{i}_{sync})$. Также считается ошибка реконструкции, сверяющая

оригинальное видео с восстановленным с помощью генератора $L_{rec} = \frac{1}{N}\sum ||L_g - L_G||_1$.

Финальный лосс генератора выглядел следующим образом:

$$L_{total} = (1 - s_g - s_w)L_{recon} + s_w E_{sync} + s_g L_{gen}$$

sw и sg - веса лоссов, которые авторы статьи предлагают устанавливать как 0.03 и 0.07 соответственно. Обучение проводится с помощью оптимизатора Adam с шагом 1e-4

	LRW [8]			LRS2 [1]			LRS3 [3]		
Method	**LSE-D ↓**	**LSE-C ↑**	**FID ↓**	**LSE-D ↓**	**LSE-C ↑**	**FID ↓**	**LSE-D ↓**	**LSE-C ↑**	**FID ↓**
Speech2Vid [17]	13.14	1.762	11.15	14.23	1.587	12.32	13.97	1.681	11.91
LipGAN [18]	10.05	3.350	2.833	10.33	3.199	4.861	10.65	3.193	4.732
Wav2Lip (ours)	**6.512**	**7.490**	3.189	**6.386**	**7.789**	4.887	**6.652**	**7.887**	4.844
Wav2Lip + GAN (ours)	6.774	7.263	**2.475**	6.469	7.781	**4.446**	6.986	7.574	**4.350**
Real Videos	7.012	6.931	—	6.736	7.838	—	6.956	7.592	—

Что было сделано нами

Данные

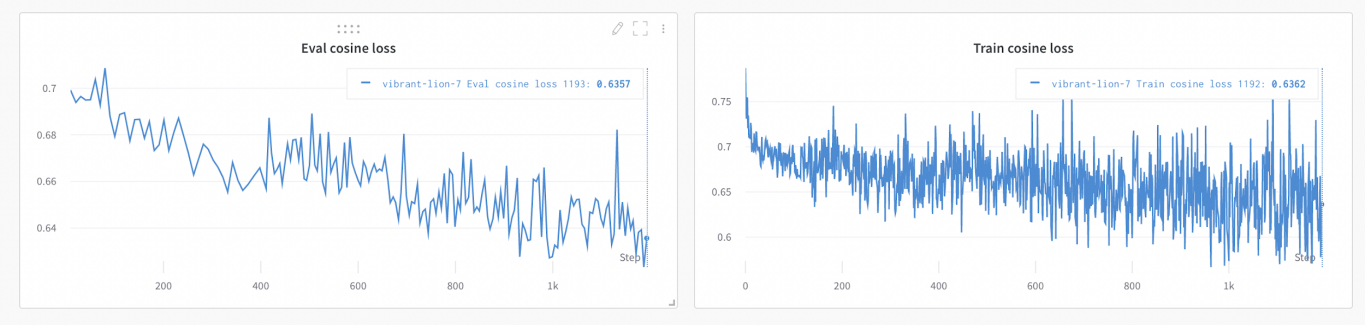
За основу взяли реализацию модели авторов. Они обучались на закрытых данных LRS2, при этом очень большого размера. Поэтому для обучения мы взяли открытый датасет VoxCeleb2 с ~1 миллионом видео, из которых мы взяли 50000 объектов.
На обучение ушло приблизительно 4 дня. Также много времени потратили на фикс багов и переписывание кода, т. к. изначально код оказался нерабочим.

Препроцесинг

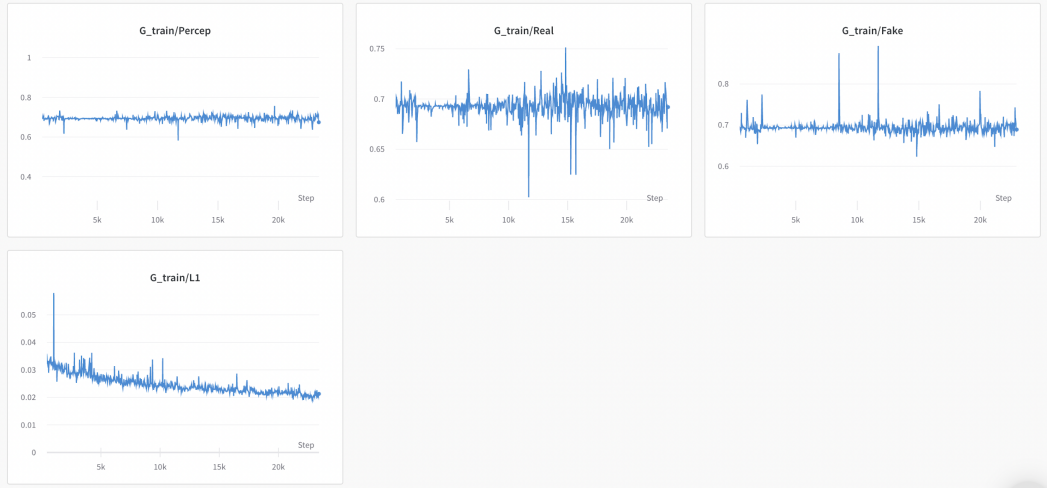
Был переписан модуль препроцессинга данных. Около дня ушло, чтобы разбить видео на фреймы с частотой 25fps и аудио (т. к. данных было очень много). Так много времени ушло, так как на каждом фрейме была использована модель face-alignment для нахождения лица (для нее использовали уже обученную модель)

Обучение

Также были переписаны модули обучения первого дискриминатора SyncNet, на обучение которого ушел ~1 день.
Графики обучения:



Модуль обучения Wav2lip тоже изменили, так опять же ничего не запускалось. На обучение ушло ~2 дня.
Графики обучения:



В дополнительные дни для обучения увеличили датасет до 100000 объектов на новом сервере с 4 видеокартами, но к сожалению визуально результаты лучше не стали.

Результаты

В целом результаты получились неплохие, но хуже, так как в оригинальной статье использовалось больше данных и большего разрешения.

Примеры видео будут лежать в папке results.

К сожалению веса модели пока не можем прислать, так как сервер, на котором обучались, после перезапуска пока не доступен. Веса постараемся прислать как только появится доступ к серверу.