Создание алгоритма для создания музыкального (аудио) сопровождения к видеоклипам

Выполнил **Гамаюнов Никита**, с/б: 1032207777

Научный руководитель: **Карандашев Я.М.**

НЕЙРОКВАДРАТ

RUDN university

Задачи работы

- Исследовать возможности создания алгоритма для генерации звука по входному видео или фото
- Разработать подобный алгоритм
- Адаптировать все компоненты для работы при небольших вычислительных мощностях

Нейронная сеть

- Может быть представлена в виде функции:

$$N(x, \Theta): X \to Y$$

- Моделирует работу мозга при обработке информации

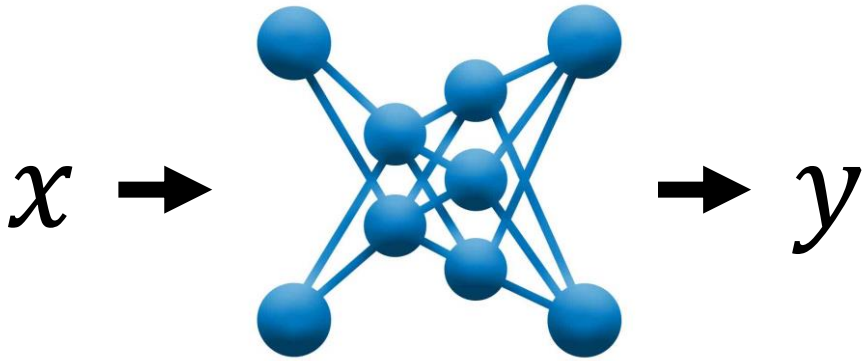
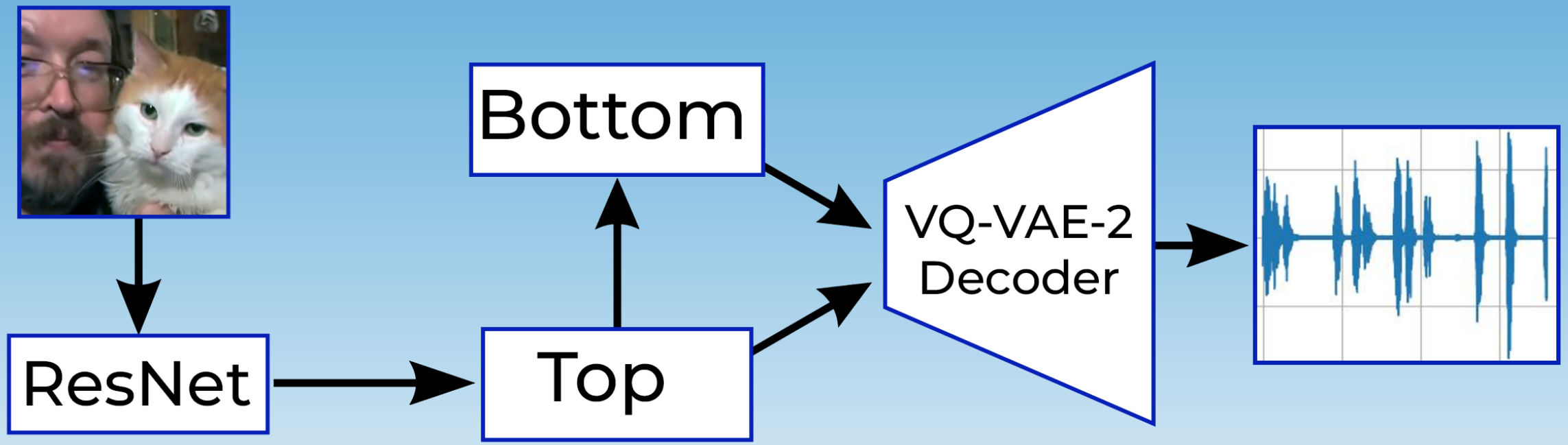


Схема алгоритма



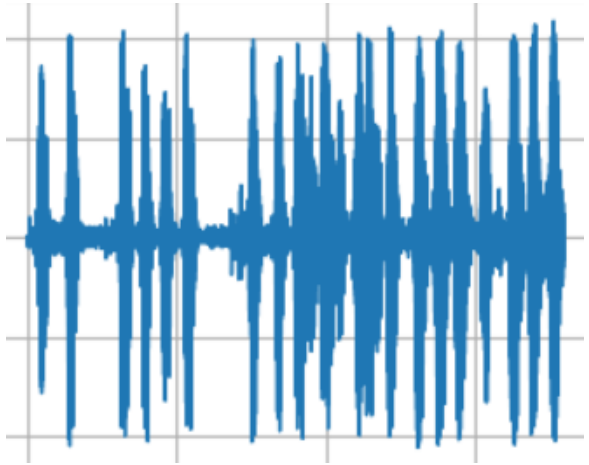
Датасеты

Используется два датасета, содержащие два класса, данные объединены в пары:

1. «фото – класс»

– кот

2. «звук – класс»



– кот

ResNet-50

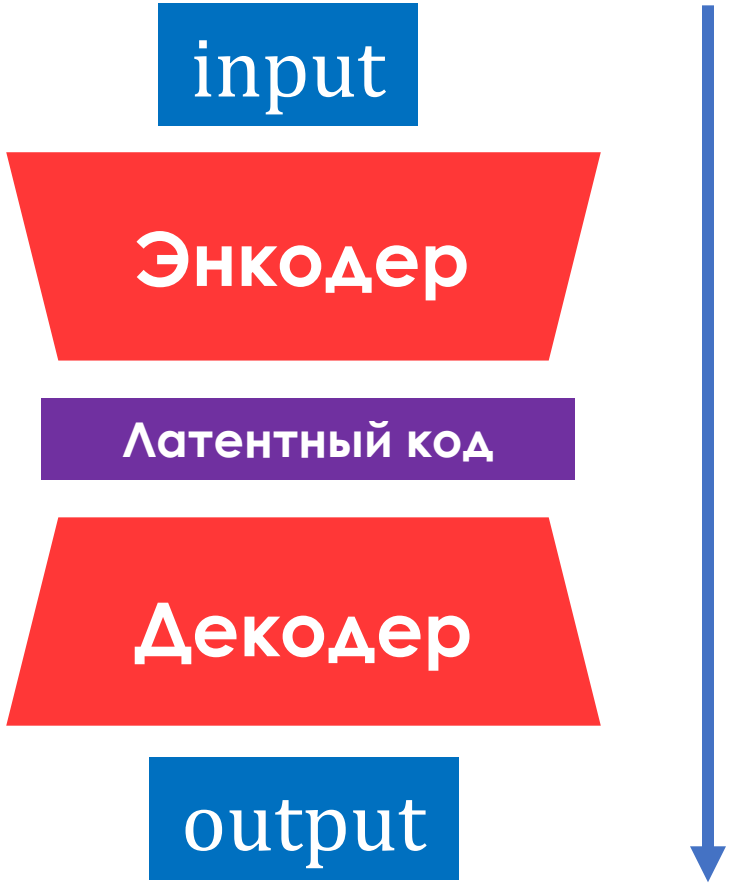
Мощная модель для классификации изображений

ResNet-50

$$\begin{cases} 0, \text{если кошка} \\ 1, \text{если собака} \end{cases}$$

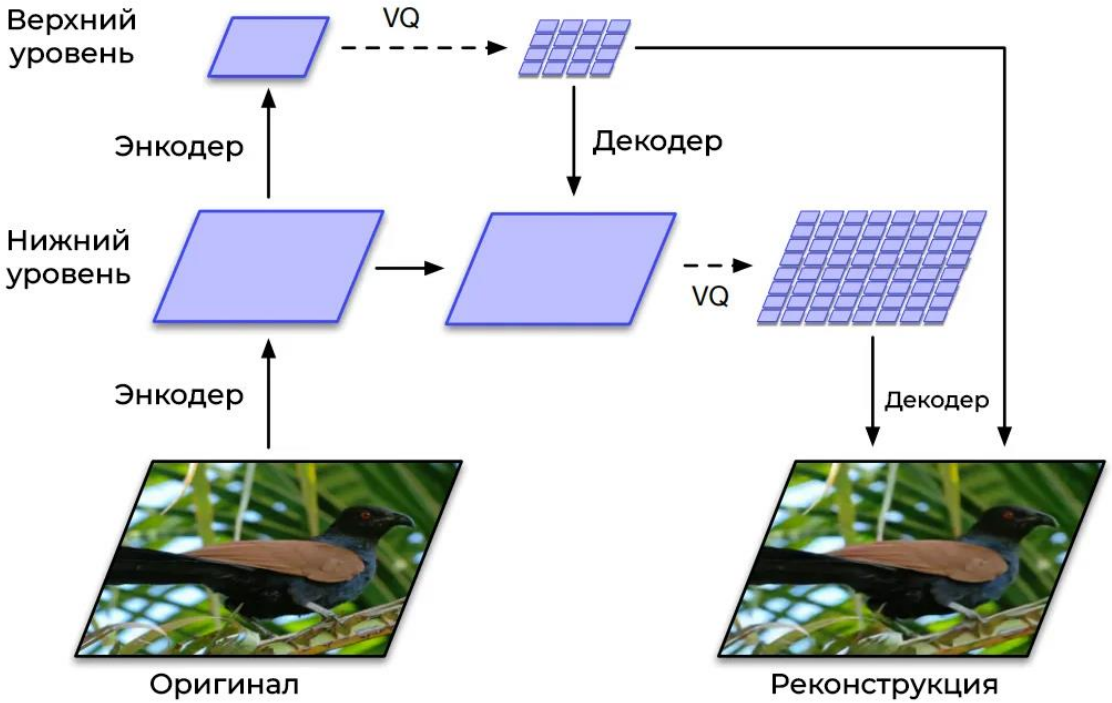
Автоэнкодер (AE)

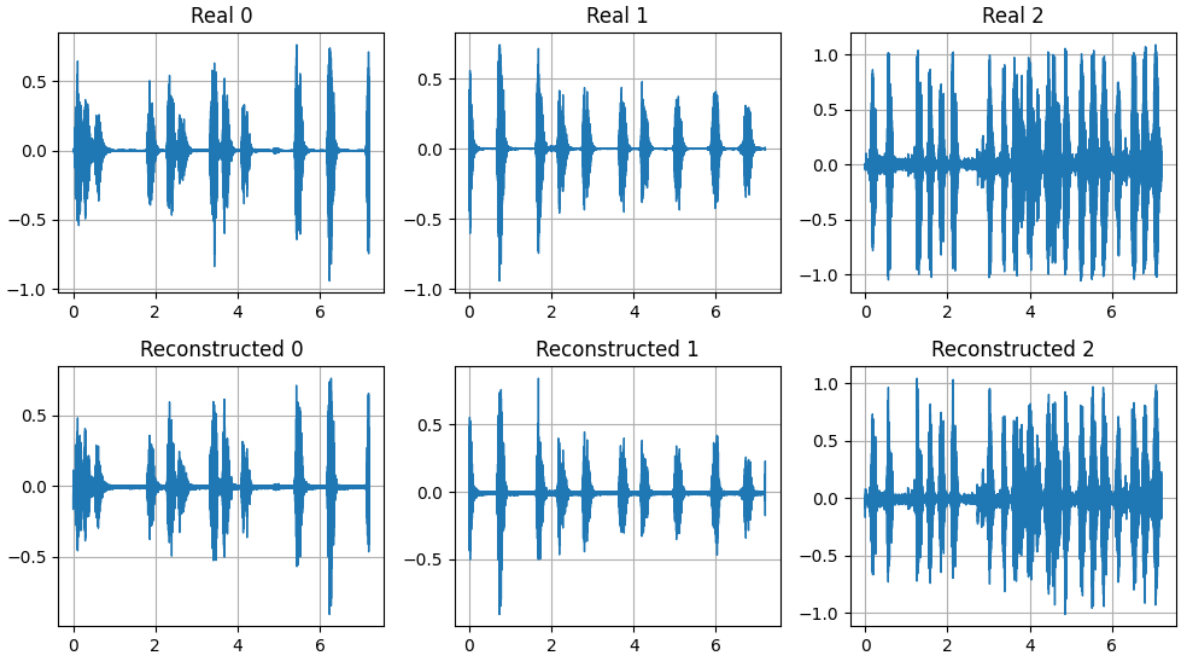
- Состоит из двух частей: **энкодера** и **декодера**
- Энкодер трансформирует входные данные в латентное представление
- Декодер восстанавливает данные в изначальный вид
- Задача – минимум потерь при кодировании и декодировании



VQ-VAE-2

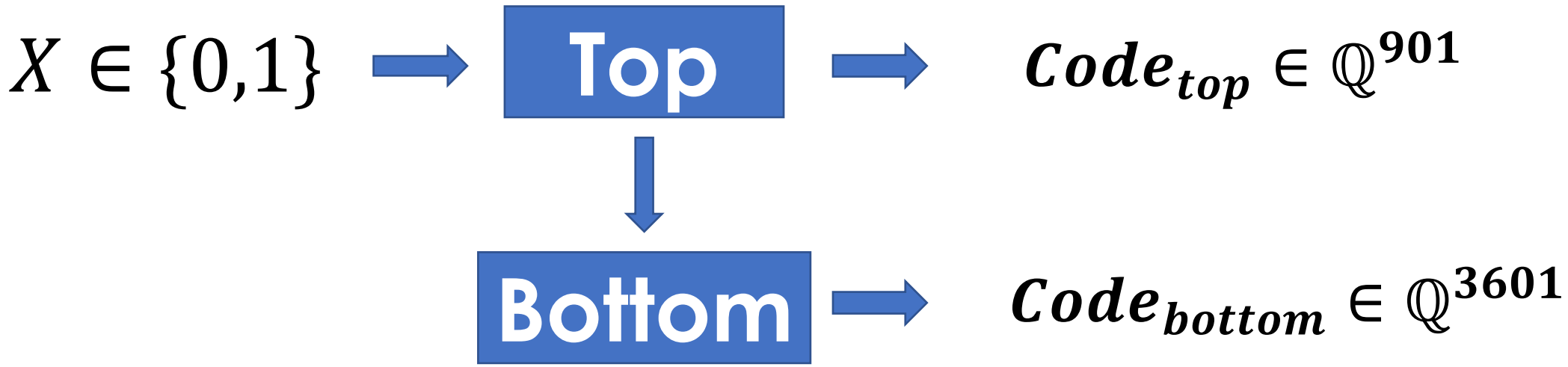
Использует два набора дискретных кодов





Модели Top и Bottom

Две модели для двух наборов кода VQ-VAE-2:



Внимание, трансформеры

$$\boldsymbol{Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_K}}\right)V}$$

$$Q = xW^Q$$
$$K = xW^K$$
$$V = xW^V$$

$$softmax(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{d_K} e^{z_j}}\ ,\ i = 1,2,\dots,d_K$$

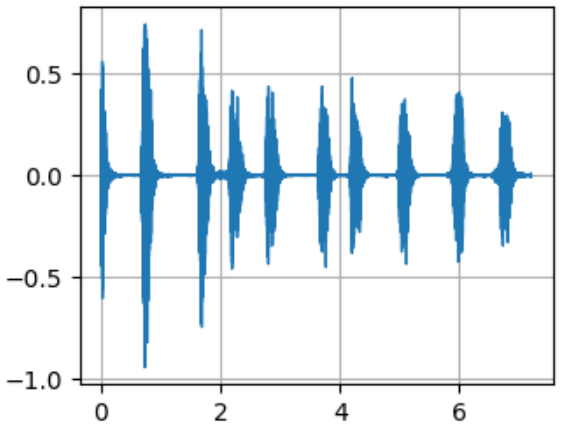
Инструменты реализации

- Язык программирования Python
- Фреймворк PyTorch, библиотеки numpy, matplotlib и torchaudio
- Графический ускоритель Nvidia Tesla P100
- Платформа Kaggle

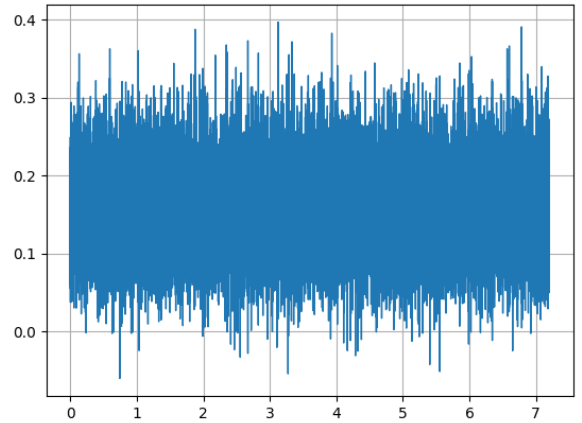
Модель	Количество параметров	Шагов обучения	Время обучения	Accuracy	FAD	Perplexity
VQ-VAE-2	580 641	10 920	3,4 часа	–	32883	–
Top	5 341 084	10 080	10,3 часов	–	–	6.217
Bottom	5 341 084	10 080		–	–	13.3276
ResNet-50 (пред-обученная)	23 512 130	До 60×10^4	Не приведено	0.51	–	–
ResNet-50 (дообучение)	23 512 130	160	2,06 минут	0.98	–	–

Результаты

- Хоть результаты генерации пока слабы, глядя на метрики, можно судить о том, что модели учатся и работают корректно
- В дальнейших исследованиях можно увеличить масштаб моделей и время обучения



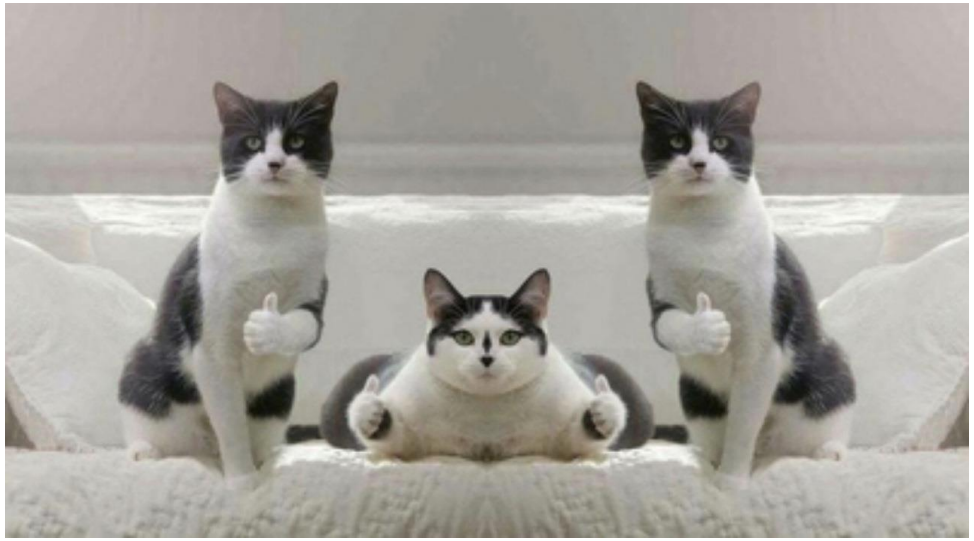
Лай собаки



Нейросетевой лай

Заключение

- Проведённое исследование подтвердило возможность создания алгоритма для генерации аудио по изображениям
- При работе на небольших мощностях приходится терять в качестве
- Алгоритм был построен из четырёх моделей, каждая из которых функционирует корректно



$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_K}}\right)V$$

Спасибо за *внимание!*

НЕЙРОКВАДРАТ

RUDN university