Генерация MIDI бэнгеров с мульти-инструментальным выбором

Авторы работы: Охина Алина и Волков Иван

Цель проекта

Разработать модель для генерации MIDI-треков с возможностью выбора инструментов, сочетающую креативность и контроль пользователя

Задачи проекта

- Исследовать архитектуры нейросетей для генерации музыки (Transformers, RNN/LSTM, GAN)
- Собрать датасет MIDI-файлов с разметкой по инструментам (пианино, гитара, бас, скрипка и др.)
- Подготовить данные для обучения модели
- Обучить модель на мульти-инструментальных данных и оценить качество звучания
- Разработать интерфейс для интерактивного взаимодействия с пользователем

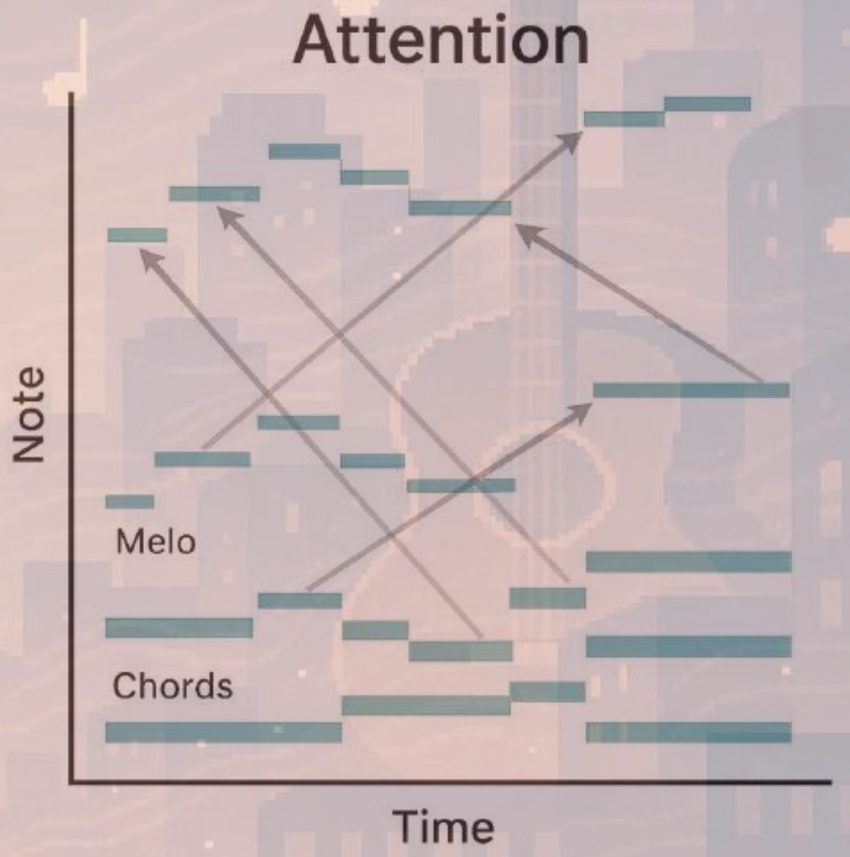
Существующие методы генерации музыки нейросетями

Метод	Пример работы	Преимущества	Недостатки
Рекуррентные сети (RNN/LSTM)	BachBot (генерация полифонической музыки в стиле Баха)	Хорошо улавливают временные зависимости за счёт рекуррентных связей Относительно просты в реализации для коротких последовательностей	Страдают от "забывания" длинных последовательностей Генерируют предсказуемые паттерны Медленная генерация
Generative Adversarial Networks (GAN)	MuseGAN — генерация многодорожечных MIDI-треков GANSynth (Google Magenta) — генерация сырого аудио в реальном времени	Параллельная генерация всего трека (не пошаговая) Могут имитировать сложные распределения данных	Ресурсоёмкость Нет гарантии, что сгенерированный трек будет гармоничным — возможны какофонии
Transformers	Music Transformer (Magenta) — генерация фортепианных композиций Allegro music transformer - мульти-инструментальный музыкальный трансформер	Лучше других справляются с аккордами и контрапунктом. Механизм внимания сохраняет контекст на всём протяжении трека	Высокая вычислительная сложность Риск переобучения

Обоснование выбора Transformer-модели

Мы выбрали архитектуру Transformer, потому что она:

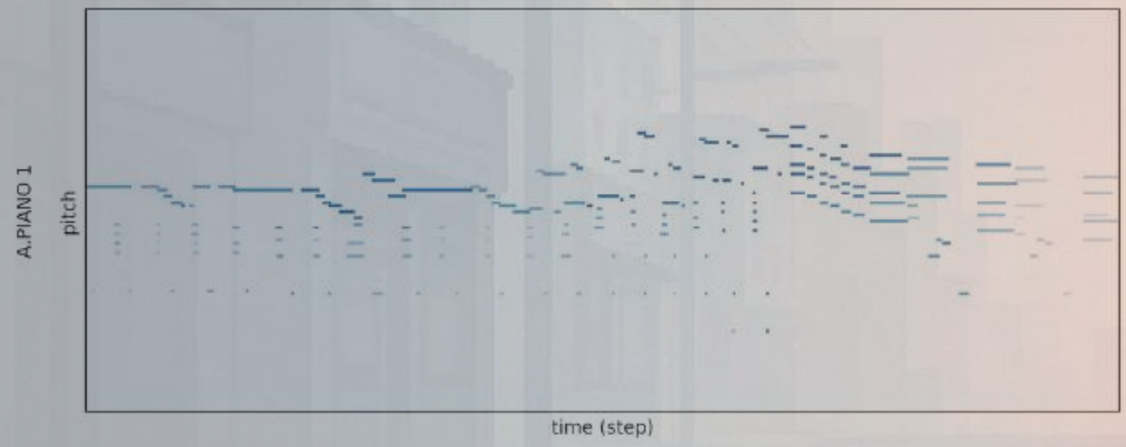
- Генерирует гармоничные треки за счёт анализа всей музыкальной структуры.
- Улавливает долгосрочные зависимости между аккордам и мелодией
- Даёт лучший quality/resource баланс по сравнению с RNN и GAN



Визуализация механизма внимания (attention)

Allegro transformer maker

- Гибкость генерации: Исходная модель поддерживает создание композиций с разными инструментами, что соответствует нашим требованиям.
- Трансформерная архитектура
- Модифицируемость: Архитектура позволила нам эффективно доработать модель под наши задачи



Пианоролл трека из датасета

Мы выбрали датасет Lakh MIDI, потому что он:

- Содержит разнообразные жанры
- Включает не только монофонические мелодии
- Предоставляет чистые MIDI-файлы с разделенными треками инструментов

Шаг 1: Подготовка MIDI-файлов

1. **Фильтрация**:
Оставили только файлы размером < 250 КБ
Удалили треки без чёткой разметки инструментов
Только один ведущий инструмент на трек

2. **Нормализация:**
Квантование времени - Все события привязаны к сетке 1/32 ноты
Фиксация длительностей - Длительности кратны 1/32
Нормализация громкости - громкость округляется с шагом 15 (8-127)
Группировка инструментов - 128 GM-патчей в 12 классов

3. **Токенизация**
Принцип кодирования следующий:

Каждое событие - 3 токена:
- Время (0-225)
- Длительность * 8 + громкость (256-511)
- Инструмент * 128 + нота (1280-2559)

Специальные токены:
- 3087 - START
- 3073 - Трек без ударных
- 3075 + N - Маркер инструмента

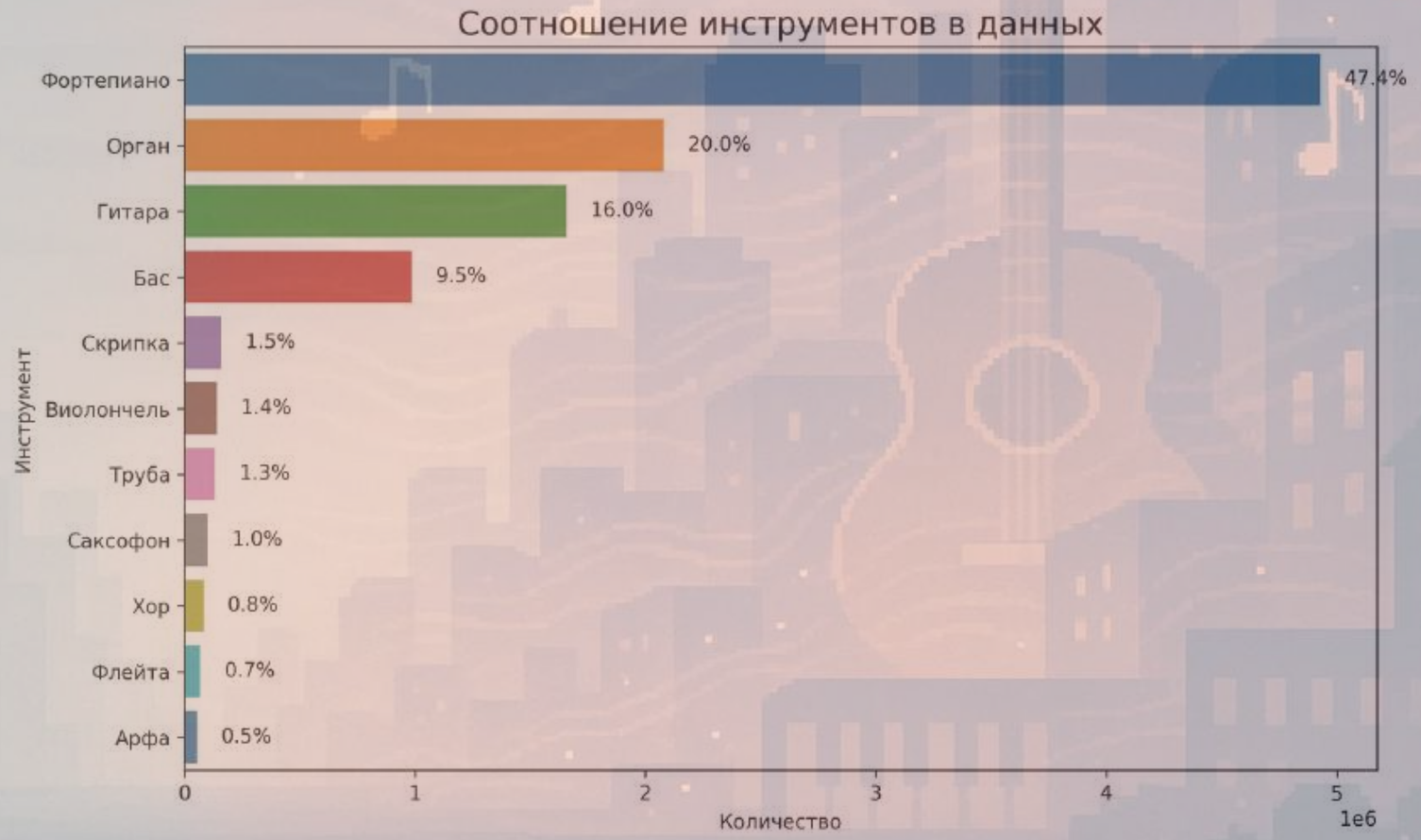
Итоговый датасет

Объем:

10,000+ MIDI-треков

Формат:

Токенизированные последовательности



Модель

1. Основные параметры

```
SEQ_LEN = 512
BATCH_SIZE = 64
NUM_EPOCHS = 5
LEARNING_RATE = 2e-4
```

2. Структура

```
TransformerWrapper(
    num_tokens=3088,        # Размер словаря
    max_seq_len=SEQ_LEN,    # Макс. длина последовательности
    attn_layers=Decoder(
        dim=512,            # Размерность эмбеддингов
        depth=12,           # Количество слоев
        heads=10,           # Головы внимания
        use_flash_attn=True  # Оптимизация внимания
    )
)
```

Особенности:

Авторегрессия:

- Предсказывает каждый токен на основе предыдущих
- Аналогично GPT для текста

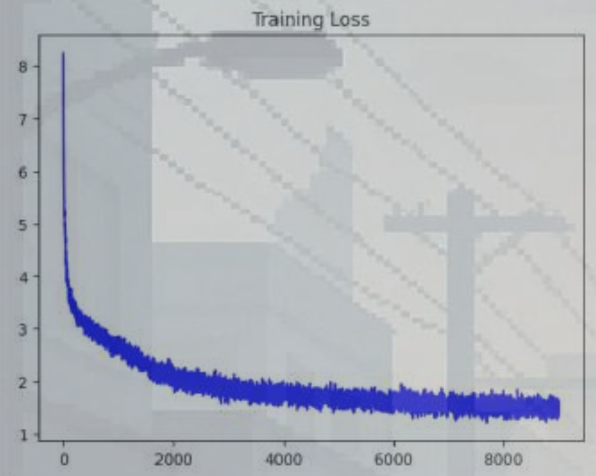
Оптимизации:

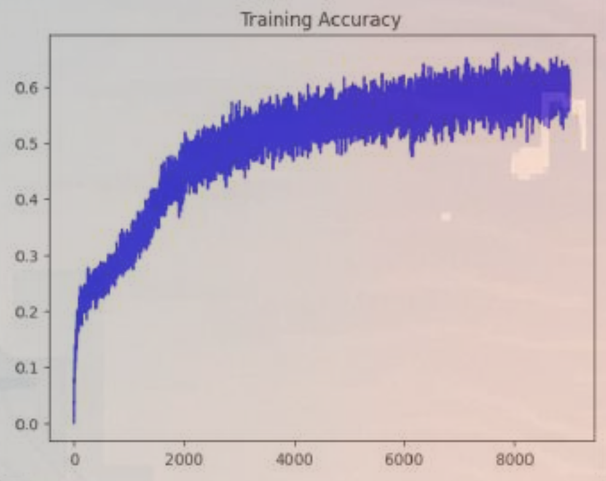
- Flash Attention — быстрые вычисления
- DataParallel — обучение на нескольких GPU

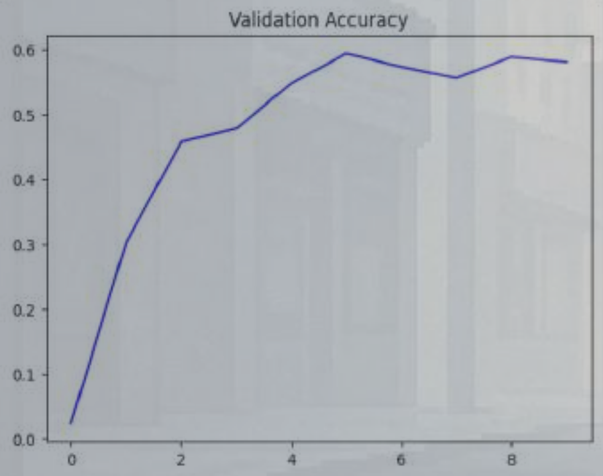
Слои:

- Эмбеддинги нот + позиции
- трансформер-блоки
- Нормализация и линейный слой

Оценка качества модели







Loss (Cross-Entropy)

- Главный индикатор обучения
- Чувствителен ко всем аспектам: высота ноты, инструмент, длительность

Accuracy

- Дополняет loss
- Показывает % идеально совпадающих предсказаний
- Особенно важен для контроля переобучения и сравнения разных архитектур

Разработка веб-интерфейса

Основные возможности:

- Выбор инструмента из 11 вариантов
- Возможность добавить барабанную дорожку к каждому инструменту

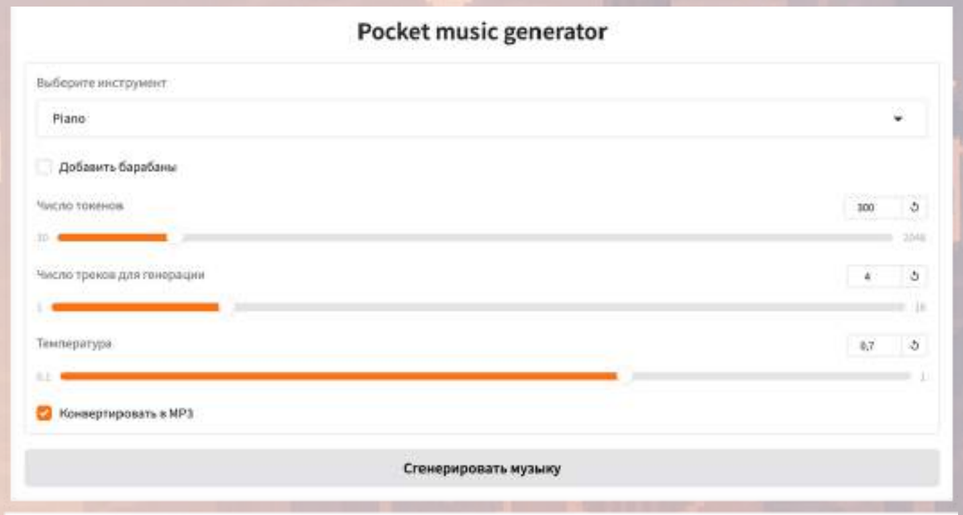
Гибкие параметры:

- Длина композиции
- Количество треков за раз

Температурный контроль креативности

Форматы вывода:

- MIDI-файл
- MP3 для мгновенного прослушивания




Спасибо за внимание!