ПРИМЕНЕНИЕ НЕЙРОННЫХ СЕТЕЙ В ЗАДАЧАХ ПОИСКА МЕЛОДИИ ПО НАПЕВУ

Каминский А.В., магистрант

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Петровский Н.А. – канд. техн. наук, доцент

Рассматривается задача поиска музыкальных произведений по напетому фрагменту мелодии (QBSH). Обсуждаются основные сложности, связанные с вариативностью вокального исполнения и неточностями интонации. Предлагается метод на основе сверточной нейронной сети с контрастивной функцией потерь, использующий хроматические представления аудиосигнала и MIDI-данных. Анализируются результаты экспериментов, демонстрирующие эффективность предложенного подхода для сопоставления напетых мелодий с оригинальными композициями.

Задача поиска мелодии по напеву (Query-by-Singing-Humming, QBSH) является одной из ключевых задач в области музыкального информационного поиска. Пользователь напевает фрагмент мелодии, после чего система должна найти соответствующую композицию в базе данных. Основными трудностями решения задачи являются различия в высоте, а также вариации темпа исполнения.

Основные методы решения задач QBSH состоят в сравнении двух последовательностей признаков. Так, основным способом решения задачи является алгоритм динамического выравнивания по времени (dynamic time warping, DTW). Однако основным недостатком данного подхода является высокая вычислительная сложность ($O(N^2)$).

В данной работе предлагается метод, основанный на сверточной нейронной сети, которая обучается сопоставлять мелодические фрагменты аудио и MIDI в общем пространстве эмбеддингов [1].

Для проведения исследования был взят HumTrans датасет, содержащий midi-представления мелодий, а также соответствующие этим мелодиям напевы пользователей [1].

Для представления мелодии используется хроматическое представление (chroma features), которое отображает энергию сигнала по 12 классам высот, что соответствует количеству нот в одной музыкальной октаве. Хроматическое представление мелодии представлено на рисунке 1.

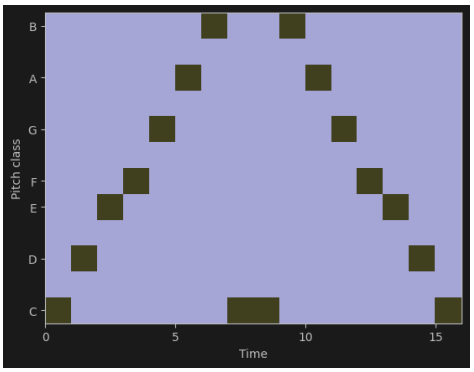

Рисунок 1 – Хроматическое представление мелодии

Поскольку длина мелодий может различаться, применяется разбиение последовательности на перекрывающиеся окна. Размер окна равен 100 с шагом перекрытия 50. Если длина последовательности меньше размера окна, применяется дополнение последовательности нулями (padding).

В предложенной модели для извлечения признаков используется сверточная нейронная сеть ResNet18, являющаяся одной из наиболее известных архитектур глубокого обучения для обработки изображений.

После вычисления эмбеддинга применяется контрастивная функция ошибки (InfoNCE Loss) [2]. Функция потерь используется для обучения сети формировать пространство эмбеддингов, в котором фрагменты одной и той же мелодии располагаются близко друг к другу, а фрагменты различных мелодий - на большем расстоянии. В процессе обучения эмбеддинги группируются по принадлежности к одному и тому же временному окну мелодии: такие эмбеддинги считаются положительными примерами. Одновременно все эмбеддинги, относящиеся к другим мелодиям, рассматриваются как отрицательные примеры. Функция потерь оптимизируется таким образом, чтобы увеличивать сходство между положительными парами и уменьшать сходство между отрицательными:

$$l = -\sum_{k=1}^{K} \sum_{z_k^i, z_k^j \in Z_k} \log \frac{\exp\left(\frac{sim(z_k^i, z_k^j)}{\tau}\right)}{\sum_{z_l \notin Z_k} \exp\left(\frac{sim(z_k^i, z_l)}{\tau}\right)}, \quad (1)$$

где K – количество групп в батче, Z_k – множество эмбеддингов одной группы, z_k^i – эмбеддинг из группы, z_l – эмбеддинг, не принадлежащий группе, $sim(x, y)$ – функция сходства (косинусное сходство), τ – температурный коэффициент, в данной работе равен 0.2.

Обучение производится с использованием оптимизатора Adam. Параметры обучения представлены в таблице 1.

Таблица 1 – Параметры обучения

№ п/п	Параметр	Значение
1	Размер батча	16
2	Размерность выходного эмбеддинга	256
3	Скорость обучения	0.002
4	Количество эпох обучения	100

График обучения нейронной сети представлен на рисунке 3.

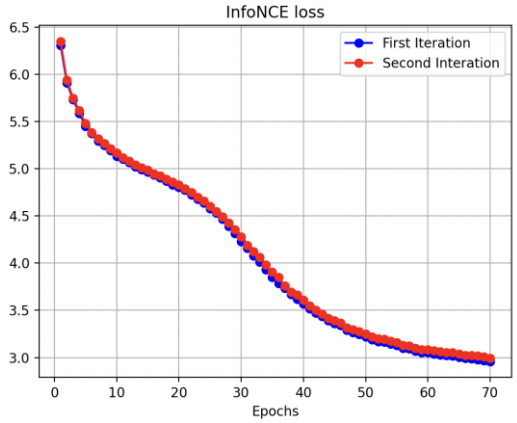

Рисунок 3 – График обучения нейронной сети

Для оценки качества поиска использовались две метрики: точность (accuracy), отражающая долю мелодий из тестовой выборки, которые были корректно распознаны нейронной сетью, и средний обратный ранг (Mean Reciprocal Rank, MRR), характеризующий позицию правильной мелодии в списке всех проранжированных кандидатов. Средний обратный ранг рассчитывается по формуле 2.

$$MRR = \frac{1}{U} \sum_{i=1}^{U} \frac{1}{rank_i}, \quad (2)$$

В результате проведённых экспериментов нейронная сеть продемонстрировала результаты, представленные в таблице 2.

Таблица 2 – Результаты обучения нейронной сети

№ п/п	Параметр	Значение
1	Accuracy	0.82
2	MRR	0.52

Точность 0.82 означает, что 82 % мелодий из тестовой выборки были корректно распознаны системой. Значение среднего обратного ранга 0.52 указывает на то, что правильная мелодия в среднем находится среди первых позиций в ранжированном списке результатов поиска.

Таким образом, в работе предложен метод решения задачи QBSH на основе сверточной нейронной сети. Использование модифицированной архитектуры ResNet18 позволяет извлекать компактные эмбеддинги мелодических фрагментов.

Список использованных источников:

1. A semi-supervised deep learning approach to dataset collection for query-by-humming task / A. Amatov, D. Lamanov, M. Titov, I. Vovk, I. Makarov, M. Kudinov // https://arxiv.org/pdf/2312.01092, 2023.

2. InfoNCE: Identifying the Gap Between Theory and Practice / E. Rusak, P. Reizinger // https://arxiv.org/pdf/2312.01092, 2025.