Отчет

Летняя школа по ИИ лето с AIRI 2025

Кудрявцев Василий, Ольга Волкова, Анна Томкевич, Игорь Соловьев

1 Abstract

В последние годы системы распознавания лиц людей достигли значительных успехов. Однако автоматическая верификация морд животных, в частности, кошек и собак, по-прежнему остается малоизученной областью. Основная сложность заключается в сборе качественных датасетов для этой задачи, что затрудняет получения высоких показателей метрик.

В данной работе предлагается система верификации морд кошек и собак, использующая как визуальную, так и текстовую информацию. Проведённые эксперименты демонстрируют высокую эффективность предложенного подхода. Полученные результаты по метрикам EER 2.84% подтверждают перспективность выбранного метода.

2 Introduction

Системы распознавания лиц являются широко используемыми биометрическими технологиями, которые в последние годы достигли значительных успехов. Например, модель ArcFace демонстрирует высокие показатели точности в задаче распознавания лиц человека. Однако системы верификации лиц животных остаются недостаточно изученной и недооценённой областью в компьютерном зрении.

Существуют отдельные исследования, посвящённые верификации лиц кошек и собак. Так, в одной из работ предлагается использовать уникальные особенности носа собаки для её идентификации. Однако данный подход требует близкого фотографирования животного, что не всегда возможно или удобно, так как некоторые люди боятся подходить слишком близко к животным. Кроме того, в популярных онлайн-базах данных, часто размещают фотографии потерянных собак, где нос животного не виден, что ограничивает применимость этого метода.

В связи с этим мы предлагаем новый подход, основанный на использовании модели CLIP, которая объединяет визуальную и текстовую информацию. Текстовая информация обрабатывается с помощью модели e5-base-v2, что позволяет более эффективно учитывать контекст и дополнительные данные.

3 Сбор и подготовка данных

Для решения поставленной задачи был проведён тщательный сбор и анализ данных. В работе использовались следующие источники:

- **Dogs World**[1] — датасет, содержащий около 300 000 изображений собак с разметкой по идентичности, а также дополнительной информацией: порода, возраст (не для всех экземпляров), координаты расположения на изображении и кличка животного.
- **Labeled Cats in the Wild**[2] — датасет, содержащий изображения кошек с разметкой по идентичности.

Для тестирования модели дополнительно был собран собственный датасет из открытых источников и приютов, а также использованы все данные, предоставленные организаторами.

4 Очистка данных

В ходе анализа датасетов было выявлено, что примерно 15% изображений содержат более одного животного. Для повышения качества данных была проведена автоматическая очистка с использованием модели **YOLOv12x**. Изображения, на которых детектировалось более одного животного, были удалены из выборки.

[1] https://www.kaggle.com/datasets/lextoumbourou/dogs-world
[2] https://www.kaggle.com/datasets/dseidli/lcwlabeled-cats-in-the-wild

5 Генерация текстовых описаний

Для обогащения данных текстовой информацией использовалась крупная языковая модель **Qwen 2.5 VL 72B**. С её помощью для каждого изображения формировалось подробное текстовое описание, включающее:

- тип животного;
- породу;
- категорию возраста: *baby*, *adult*, *senior*;
- детализированное описание внешних особенностей (окрас,пятна и др.).

Такой подход позволил создать структурированные и информативные аннотации для последующего обучения и тестирования моделей.

Однако из-за ограниченного объема времени мы не успели сгенерировать текстовое описание для всех изображений.

6 Сбор и подготовка данных

Для решения поставленной задачи был проведён тщательный сбор и анализ данных. В работе использовались следующие источники:

- **Dogs World**[3] — датасет, содержащий около 300 000 изображений собак с разметкой по идентичности, а также дополнительной информацией: порода, возраст (не для всех экземпляров), координаты расположения на изображении и кличка животного.
- **Labeled Cats in the Wild**[4] — датасет, содержащий изображения кошек с разметкой по идентичности.

Для тестирования модели дополнительно был собран собственный датасет из открытых источников и приютов, а также использованы все данные, предоставленные организаторами.

6.1 Очистка данных

В ходе анализа датасетов было выявлено, что примерно 15% изображений содержат более одного животного. Для повышения качества данных была проведена автоматическая очистка с использованием модели **YOLOv12x**. Изображения, на которых детектировалось более одного животного, были удалены из выборки.

6.2 Генерация текстовых описаний

Для обогащения данных текстовой информацией использовалась крупная языковая модель **Qwen 2.5 VL 72B**. С её помощью для каждого изображения формировалось подробное текстовое описание, включающее:

- название животного;
- породу (для кошек разметка по породе отсутствовала);
- категорию возраста: *baby*, *adult*, *senior*;
- детализированное описание внешних особенностей (окрас, характерные приметы и др.).

Такой подход позволил создать структурированные и информативные аннотации для последующего обучения и тестирования моделей.

[3] https://www.kaggle.com/datasets/lextoumbourou/dogs-world
[4] https://www.kaggle.com/datasets/dseidli/lcwlabeled-cats-in-the-wild

7 Архитектура решения

В качестве основы для решения задачи была выбрана модель **CLIP**, которую дообучали на собранных и размеченных датасетах. Такой подход позволил существенно улучшить результаты по сравнению с оригинальным решением.

В процессе обучения основной целью было формирование качественных эмбеддингов, хорошо разделяющих различные классы животных. Для этого использовались следующие методы:

- **Triplet Margin Loss** — лосс функция, способствующая максимальному разнесению эмбеддингов разных классов и минимизации расстояния между эмбеддингами одного класса.
- **Intra Pair Variance Loss** — дополнительная лосс функция, направленная на уменьшение внутриклассовой дисперсии эмбеддингов, что способствует их большей генерализации.

Во всех экспериментах использовались одинаковые параметры оптимизации:

- Оптимизатор: **Adam**
- learning rate: `1e-4`
- betas: `(0.9, 0.999)`

Такой подход обеспечил стабильность сравнения различных конфигураций и позволил объективно оценить влияние выбранных методов обучения на итоговое качество модели.

8 Эксперименты

Для оценки эффективности различных подходов мы провели серию экспериментов с моделью **CLIP**. В процессе обучения были заморожены все слои, кроме последних слоёв vision encoder, что позволило адаптировать модель под специфику задачи, сохраняя при этом преимущества предобученных представлений.

8.1 Экспериментальные сетапы

В ходе экспериментов были протестированы следующие конфигурации:

1. **CLIP-GmP-ViT-L-14 + Triplet Loss + Intra Pair Variance Loss**
2. **CLIP-GmP-ViT-L-14 + Triplet Loss**
3. **CLIP-GmP-ViT-L-14 + Triplet Loss + Intra Pair Variance Loss + SAM**
4. **CLIP-GmP-ViT-L-14 + Triplet Loss + Intra Pair Variance Loss + e5-small**
5. **ResNet50**
6. **CLIP ViT-B/32**

8.2 Метрики оценки

Для объективного сравнения моделей использовались следующие метрики:

- **EER (Equal Error Rate)** — стандартная метрика для задач верификации, отражающая точку, в которой доли ложных допусков (FAR) и ложных отказов (FRR) равны: $FAR(x) = FRR(x) = EER$.

8.3 Результаты

Анализ полученных результатов показывает, что добавление функции потерь *Intra Pair Variance Loss* способствует улучшению обобщающей способности модели по сравнению с базовыми конфигурациями.

Table 1: Результаты экспериментов по метрике EER

Конфигурация	**EER, %**
CLIP-GmP-ViT-L-14 + Triplet Loss + Intra Pair Variance Loss	2.90
CLIP-GmP-ViT-L-14 + Triplet Loss	3.20
CLIP-GmP-ViT-L-14 + Triplet Loss + Intra Pair Variance Loss + SAM	2.84
CLIP-GmP-ViT-L-14 + Triplet Loss + Intra Pair Variance Loss + e5-base-v2	3.77
CLIP-GmP-ViT-L-14	16.71
ResNet50	21.36
CLIP ViT-B/32	17.71

9 Веб-приложение

Для демонстрации работы сервиса была реализована веб-серверная архитектура, включающая следующие компоненты:

- **FastAPI** — для построения REST API;
- **S3** — для хранения изображений;
- **Streamlit** — для создания пользовательского интерфейса;
- **Qdrant** — для хранения и поиск эмбеддингов;
- **PostgreSQL** — для хранения метаданных.

Пользователь может воспользоваться эндпоинтом `/search/nearest_pets`, загрузив изображение животного. В ответ сервис возвращает идентификаторы k наиболее похожих животных (`pets_id`) из базы данных.

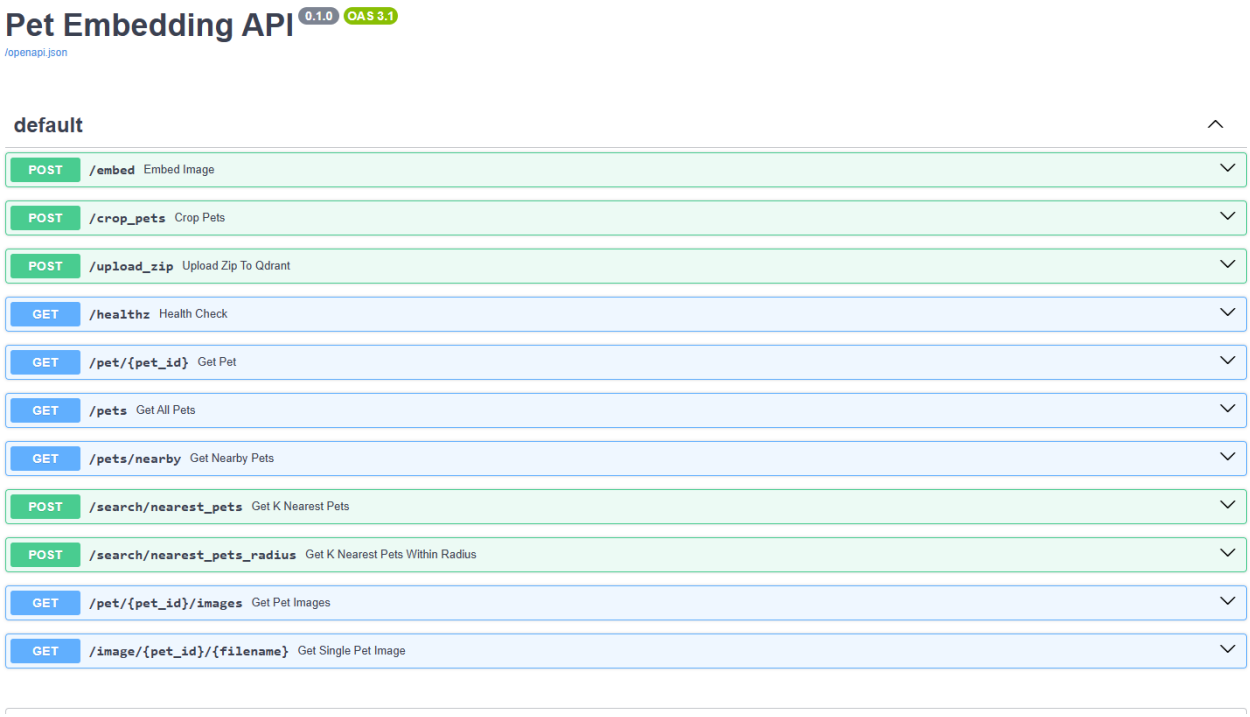

Figure 1: Основные эндпоинты сервиса