ТРАНСФОРМЕРЫ VS CNN: СРАВНЕНИЕ ПРОИЗВОДИТЕЛЬНОСТИ ДЛЯ ОБНАРУЖЕНИЯ ЭМОЦИЙ В ВИДЕОПОТОКАХ В РЕАЛЬНОМ ВРЕМЕНИ

Джире. Ю
Томский политехнический университет, ИШИТР, гр. 8ПМ41, e-mail: djire01@tpu.ru
Научный руководитель: Н. Дж. Франсис, ст. преподаватель ОИТ ИШИТР ТПУ,
e-mail: natzinadzhuanita@tpu.ru

Аннотация

В этой работе сравниваются два метода глубокого обучения – CNN и Трансформеры – для распознавания эмоций по лицу в видео в реальном времени. Используя набор данных FER2013, мы оценили модели по точности, скорости и затратам вычислительных ресурсов.

Ключевые слова: распознавание эмоций, CNN, Трансформер, реальное время, FER2013, ViT, TimeSformer.

Введение

Распознавание эмоций человека с помощью компьютерных систем стало важным во многих сферах: голосовые помощники, безопасность, интеллектуальные интерфейсы. Это добавляет эмоциональную составляющую во взаимодействие между человеком и машиной. С развитием глубокого обучения (deep learning) появились два основных подхода к обработке изображений: сверточные нейронные сети (CNN), которые широко используются [1], и трансформеры (Transformers) – более современный метод, способный учитывать общий контекст изображения [2]. В этом докладе сравниваются эти два подхода для задачи распознавания эмоций по лицу в видео в реальном времени. Оценка проводится в единой экспериментальной среде, чтобы выявить точность, преимущества и ограничения каждого метода в реальных условиях.

Методология

Данные

Для этого исследования использовался набор данных FER2013, который широко известен в области распознавания выражений лица [3]. Он содержит более 35 000 черно-белых изображений лиц размером 48×48 пикселей, разделенных на семь классов эмоций: злость, отвращение, страх, радость, грусть, удивление и нейтральное выражение. Изображения были разделены на три части – обучение, проверка и тестирование. Затем данные были увеличены с помощью таких методов, как поворот, масштабирование (zoom) и горизонтальное отражение (flip), чтобы модель могла лучше обобщать информацию.

Архитектура CNN: ResNet50

Для реализации модели CNN мы использовали ResNet50 – глубокую нейронную сеть, состоящую из 50 слоев с остаточными связями. Такая архитектура была разработана для решения проблемы исчезающего градиента в глубоких сетях [4].

Модель обучалась с использованием библиотеки TensorFlow и предварительно обученных весов из базы данных ImageNet. Этот подход называется transfer learning (обучение с переносом знаний). Он помогает ускорить обучение и улучшить результаты модели на конкретных наборах данных.

Для оптимизации была использована функция потерь – перекрестная энтропия (cross-entropy), которая часто применяется в задачах классификации. Она определяется по следующей формуле:

$$L = -\sum y_i \cdot \log(\hat{y}_i) \quad (1)$$

где y_i – это истинное значение (ground truth), а $\hat{y}_i$ – предсказание модели для класса i.

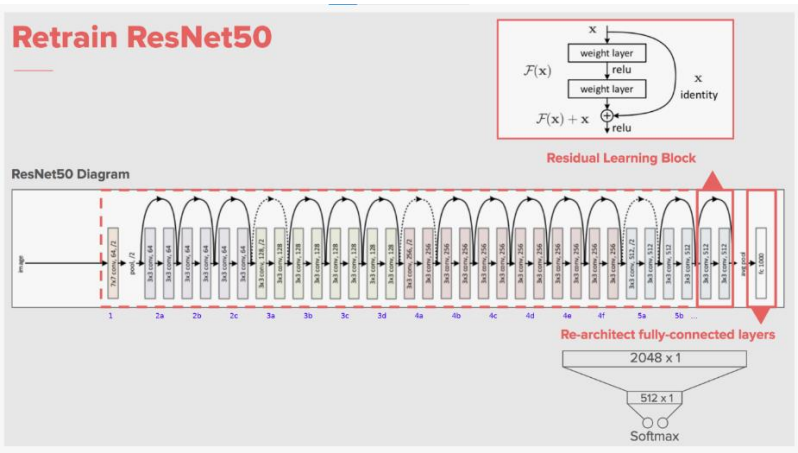

Рис. 1. Схема архитектуры ResNet50, используемой в нашей модели CNN

На рис. 1 показана полная структура ResNet50, включая остаточные блоки, последовательные сверточные слои и идентичные соединения. Также представлено, как были изменены финальные полносвязные (fully-connected) слои для выполнения задачи классификации выражений лиц по эмоциям.

Архитектура Трансформера: ViT

Для реализации модели Трансформера мы выбрали Vision Transformer (ViT), который был адаптирован с помощью библиотеки HuggingFace Transformers. ViT разбивает каждое изображение на небольшие блоки (патчи), которые затем линейно проецируются в пространство признаков перед применением механизма внимания, аналогичного тому, что используется в моделях обработки естественного языка, таких как BERT [5]. Перед обучением изображения были преобразованы в формат RGB и масштабированы до размера 224×224 пикселей в соответствии с требованиями модели ViT. В основе модели лежит механизм многоголового внимания, описываемый следующей формулой:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \qquad (2)$$

где:

- Q, K, V – это соответственно матрицы запросов (query), ключей (key) и значений (value), полученные из патчей изображения;
- d_k – это размерность ключей.
- произведение QK^T позволяет оценить степень сходства между патчами.
- функция softmax нормализует полученные значения, превращая их в веса внимания.

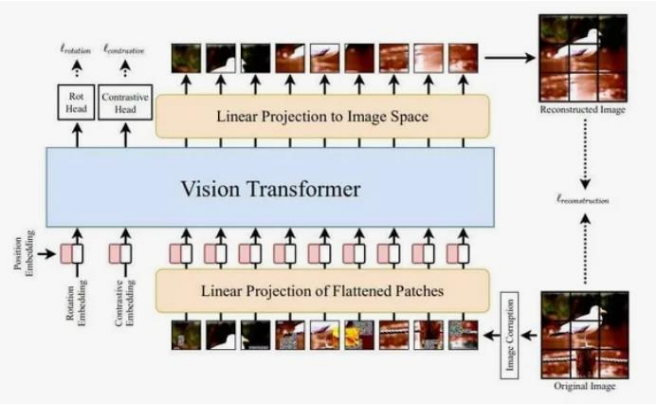

Рис. 2. Работа модели ViT (Vision Transformer)

На рис. 2 показано, как работает ViT. Сначала изображение делится на небольшие фрагменты (patches). Каждый фрагмент преобразуется в вектор с помощью линейного слоя. Затем к этим векторам добавляется позиционное кодирование (position embedding), после чего они обрабатываются основным блоком Transformer. Благодаря этому изображение анализируется целиком – каждый фрагмент учитывает общий контекст. Такая архитектура также позволяет восстанавливать изображение и использовать разные «головы» (например, contrastive или rotation heads), чтобы улучшить обучение без учителя (self-supervised learning).

Результаты и их обсуждение

Таблица 1. Точность моделей

Модель	Точность (%)
CNN (ResNet50)	57,7 %
ViT	≈ 42 %

Transformer чувствителен к качеству входных данных и требует больше примеров для хорошей генерализации.

Таблица 2. Время инференса

Модель	Фреймворк	Время инференса (мс/изображение)
CNN (ResNet50)	TensorFlow	~136,33 мс
ViT	PyTorch	~11,48 мс

ViT работает быстрее благодаря GPU-оптимизации PyTorch. CNN более стабилен при ограниченных ресурсах, но требует больше времени на обработку.

Визуальный анализ

Были подготовлены два типа визуализаций:

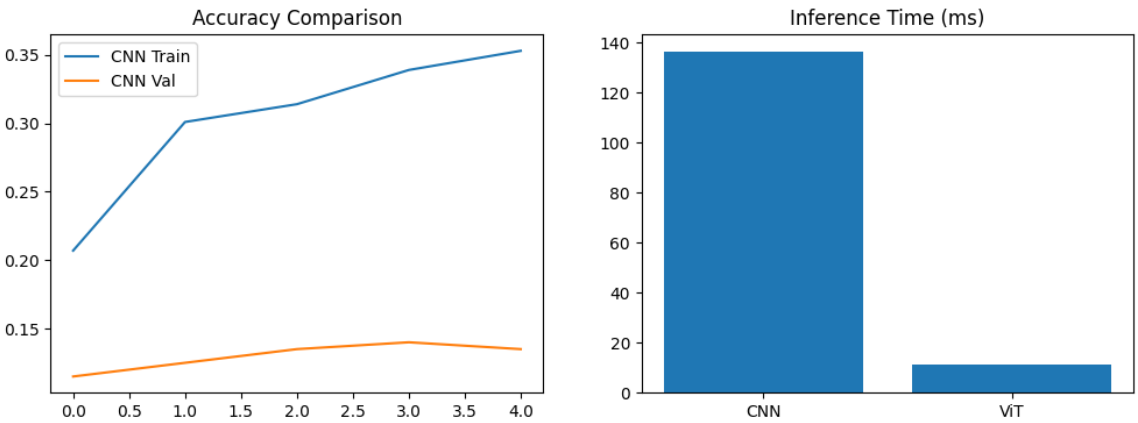

Рис. 3. График точности модели CNN (ResNet50) по эпохам и Столбчатая диаграмма, сравнивающая время инференса моделей

Визуальные результаты (Рис. 3.) для анализа эффективности моделей. График отображает изменение точности модели CNN (ResNet50) по эпохам, а столбчатая диаграмма показывает сравнение времени инференса между моделями CNN и ViT. Эти визуализации позволяют лучше понять различия в производительности моделей в реальных условиях.

Заключение

Это исследование подчеркивает сильные и слабые стороны двух популярных подходов для распознавания эмоций в реальном времени. CNN-модели работают быстро, занимают

немного памяти и хорошо подходят для встроенных систем. Модели Transformer лучше справляются с обработкой сложных визуальных данных, но требуют больше ресурсов. В будущем стоит изучить **гибридные модели**, которые будут сочетать скорость CNN и точность Transformer. Также было бы интересно протестировать такие архитектуры на видео из реальных условий, например, видеозвонков или онлайн-взаимодействий.

Список использованной литературы

1. Howard A.G., Zhu M., Chen B. и др. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications: preprint. – URL: arxiv.org/abs/1704.04861 (дата обращения: 10.04.2025).

2. Dosovitskiy A., Beyer L., Kolesnikov A. и др. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale: preprint. – URL: arxiv.org/abs/2010.11929 (дата обращения: 10.04.2025).

3. Mollahosseini A., Hasani B., Mahoor M.H. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild // IEEE Transactions on Affective Computing. – 2017. – Vol. 10, № 1. – P. 18-31.

4. He K., Zhang X., Ren S., Sun J. Deep Residual Learning for Image Recognition // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2016. – P. 770–778.

5. Bertasius G., Wang H., Torresani L. Is Space-Time Attention All You Need for Video Understanding? // Proceedings of the 38th International Conference on Machine Learning (ICML). – 2021. – P. 813–824.