**Сравнение алгоритмов машинного и глубокого обучения в задаче классификации эмоций**

В.В. Никитин, А.М. Насибуллин, С.Л. Макаров

Университет ИТМО

[owlahikit@gmail.com](mailto:owlahikit@gmail.com) , [alan.nasibyllin@yandex.ru](mailto:alan.nasibyllin@yandex.ru) , [semon.makarov.2003@mail.ru](mailto:semon.makarov.2003@mail.ru)

Аннотация

В данной статье проводится сравнительный анализ эффективности алгоритмов машинного и глубокого обучения в задаче классификации эмоций. Область классификации эмоций имеет важное значение для понимания и анализа человеческого поведения и взаимодействия. Цель исследования состоит в определении наиболее эффективного подхода для классификации различных эмоциональных состояний на основе собранных данных.

В статье рассмотрены различные методы машинного и глубокого обучения, включая метод опорных векторов (SVM), случайные леса (Random Forest), feature extractor, сверточные нейронные сети и visual transformer (ViT). Эти методы анализируются с точки зрения их способности обрабатывать и классифицировать эмоциональные данные, а также сравниваются по показателям точности и надежности.

Результаты исследования могут быть полезными для разработчиков систем распознавания эмоций и применяются в различных областях, включая социальные медиа, медицинскую диагностику и образование. Проведенный анализ позволяет выявить преимущества и недостатки каждого метода и выбрать наиболее подходящую модель классификации эмоций для конкретной задачи.

.

**Ключевые слова**: распознавание эмоций на лицах, сверточные нейронные сети, машинное обучение, многоклассовая классификация, настройка гиперпараметров, искусственный интеллект, анализ данных, компьютерные науки.

**1. Вступление**

В современном информационном обществе огромное количество данных, сгенерированных и собранных каждый день, открывает новые возможности для применения алгоритмов машинного и глубокого обучения в различных областях. Одной из таких областей является классификация эмоций, которая имеет важное значение для определения и понимания человеческого поведения и взаимодействия. Классификация эмоций представляет собой сложную задачу, требующую точного определения и классификации различных эмоциональных состояний на основе предоставленных данных. В данной статье мы предлагаем сравнить и проанализировать эффективность алгоритмов машинного и глубокого обучения в решении задачи классификации эмоций.

Цель данного исследования заключается в оценке и сравнении различных методов классификации эмоций, а также в определении наиболее эффективного подхода для достижения высоких показателей точности и надежности. Мы будем рассматривать различные алгоритмы машинного и глубокого обучения, такие как метод опорных векторов (SVM), случайные леса (Random Forest), Catboost, экстракторы функции (Feature Extractor) HOG и Canny, сверточные нейронные сети (CNN), visual transformer (ViT), с акцентом на их возможности и ограничения в обработке и классификации эмоциональных данных. Изучение и сравнение этих алгоритмов позволит нам получить более глубокое понимание их преимуществ и недостатков, а также определить наиболее подходящую модель классификации. Ключевой метрикой будет являться точность (Accuracy) определения аффекта на фото из заранее подготовленного датасета фотографий лиц людей, поделенных на 7 классов: удивленный, испуганный, отвращенный, веселый, грустный, злой и нейтральный. Результаты данного исследования могут быть полезными для специалистов в области разработки систем распознавания эмоций, а также для различных практических применений, включая социальные медиа, медицинскую диагностику, образование и другие области, где классификация эмоций играет важную роль.

Структура статьи следующая: мы начнем с обзора предыдущих исследований по классификации эмоций и описания основных методов машинного и глубокого обучения. Выдвинем свои цели и задачи. Затем мы рассмотрим экспериментальную настройку и оценку вышеописанных алгоритмов в задаче классификации эмоций, а также изучим полученные результаты и проведем их анализ. В заключение мы представим выводы и рекомендации по выбору наиболее эффективного алгоритма для классификации эмоций.

**2. Обзор литературы**

В данном обзоре литературы рассмотрим 10 статей, связанных с проблемой обнаружения и классификации эмоций с использованием алгоритмов машинного обучения.

В статье "Emotion Recognition and Classification Using Machine Learning Techniques" исследуется применение различных методов машинного обучения для распознавания и классификации эмоций. Авторы описывают использование алгоритмов, таких как метод опорных векторов, случайный лес и наивный Байесовский классификатор. Исследование проведено на наборе данных, содержащем информацию о физиологических показателях и эмоциональном состоянии испытуемых.

Данная статья описывает один из методов распознавания выражений в видео, предлагая новый подход с использованием 3D глубокой нейронной сети. Основным результатом является предложение 3DInception-ResNet (3DIR) архитектуры, расширяющей 2D Inception-ResNet для эффективной обработки последовательностей изображений. Основные компоненты - 3DInception-ResNet, использование LSTM для учета временных зависимостей и интеграция лицевых ориентиров. Авторы предоставляют результаты экспериментов, демонстрируя превосходство предложенного метода над многими существующими методами в задачах "субъект-независимые" и "межбазовые". Особенно важно отметить улучшение в распознавании эмоций на базах данных FERA и DISFA. В целом, статья предоставляет обширное объяснение принятых решений и результатов экспериментов. Внедрение 3D глубоких нейронных сетей, интеграция LSTM и использование лицевых ориентиров делают предложенный метод более гибким и эффективным для решения сложных задач распознавания фациальных выражений в видео.

Статья "Deep Learning and Machine Learning based Facial Emotion Detection using CNN" исследует использование глубокого обучения для распознавания эмоций по лицевым выражениям. Авторы предлагают архитектуру сверточных нейронных сетей, включая использование сверточных слоев, слоев объединения и полносвязных слоев. Эксперименты на различных наборах данных показывают высокую точность распознавания эмоций.

В статье "Функция распознавания эмоций с использованием сверточных нейронных сетей с многоклассовой классификацией и байесовской оптимизацией для настройки гиперпараметров" авторы Локеш Беджагам и Решми Чакрадхара исследуют распознавание эмоций на лицах с использованием сверточных нейронных сетей. Они предлагают метод, который использует многоклассовую классификацию и байесовскую оптимизацию для настройки гиперпараметров модели. В работе они обсуждают важность распознавания эмоций для межличностных отношений и коммуникации. Модель применяется к датасету и сравниваются ее результаты с другими подходами к распознаванию эмоций. Полученные результаты свидетельствуют об эффективности предложенного метода. Работа представляет интерес для исследователей, занимающихся распознаванием эмоций и применением сверточных нейронных сетей в этой области.

В статье "A Survey on Emotion Detection using Machine Learning Techniques" авторы проводят обзор существующих методов обнаружения эмоций с использованием алгоритмов машинного обучения. Они анализируют различные подходы, такие как основанные на тексте, аудио и изображениях, а также использующие комбинированные методы. В статье представлена обширная классификация методов и обзор актуальных исследований в области обнаружения эмоций.

Довольно хорошая ознакомительная презентация по EmotiW Challenge – соревнованию по распознаванию эмоций в различных условиях. Оно проводится с 2013 года, включая аудио- и видеоанализ эмоций в реальных и неконтролируемых условиях, таких как движение головы, фоновая музыка, шум, и спонтанные проявления эмоций. Участники используют базы данных, такие как Acted Facial Expressions in the Wild, и современные методы, включая глубокое обучение. В презентации представлены базовые методы, бейзлайны, и статистика участия команд и принятых статей. Победителями соревнования в 2017 году стали команды, использующие методы, основанные на сверточных нейронных сетях и объединении мультимодальных данных для распознавания эмоций в реальных условиях. Приведены выводы, такие как преимущество методов глубокого обучения, сложности обнаружения лиц и их частей, предложены направления будущего исследования - использование больших объемов данных и работа с более спонтанными выражениями эмоций в группах.

В "Emotion Detection: A Machine Learning Project" автор проводит проект по обнаружению эмоций с использованием алгоритмов машинного обучения. В исследовании использовались различные методы обработки данных и классификации, включая использование модели BERT, нейронной сети с долгой краткосрочной памятью (LSTM) и градиентного бустинга. Результаты говорят о высокой эффективности предложенных методов в обнаружении и классификации эмоций.

В статье рассматривается проблема распознавания эмоций, выделяя два основных подхода: классический, основанный на классификации ключевых точек лица, и современный, использующий глубокие сверточные сети (CNN). В классическом подходе используются различные классификаторы, такие как SVM и Random Forest, а также дескрипторы, включая LBP, HOG, SIFT и LATCH. В современном подходе предлагается использование предварительно обученных сетей на ImageNet и их дальнейшее переобучение для распознавания эмоций. Рассматриваются также методы анализа последовательностей кадров, включая подачу высокоуровневых признаков от CNN в рекуррентные сети и применение 3D-CNN для анализа видео. На конкретные архитектуры нейросетей не делается упор, но подчеркивается возможность использования любой сети с базовой архитектурой, предварительно обученной на ImageNet.

В статье проведено исследование эмоционального распознавания с использованием трех моделей: аудио-SVM, CNN-LSTM и Video-C3D. В использованных данных выделяются AFEW6.0, SFEW и FER2013. Для каждой модели: аудио-SVM использовал OpenSMILE для извлечения аудио-фич и показал удовлетворительные результаты, однако мог столкнуться с ограничениями в обработке сложных аудио-паттернов; CNN-LSTM продемонстрировал повышенную эффективность в распознавании злости и счастья, но оказался менее точен в других эмоциях, вывод - необходимость дополнительных исследований и улучшений для более точного распознавания; Video-C3D включает две модели C3D-1 и C3D-2, которые обладают признаками переобучения, C3D-2, обученный с уменьшенным темпом, продемонстрировал повышенную стабильность. Inception-ResNet-v2 показал удовлетворительные результаты, с высокой уверенностью в предсказании счастья, но с ошибками в предсказаниях отвращения. Каждая модель имеет свои сильные и слабые стороны, подчеркивается важность комбинированных подходов для достижения оптимальных результатов.

Статья представляет решение задачи распознавания эмоций на видео EmotiW2017 с использованием ансамбля моделей. Основные методы включают в себя использование сверточных нейронных сетей (CNN), предварительно обученных на крупных наборах данных по распознаванию лиц, таких как VGGFace, а также проприетарные FR-Net-A, FR-Net-B и FR-Net-C. Эти нейросети подвергаются настройке на данных FER2013 и EmotiW2017. В дополнение к пространственным признакам, исследователи использовали аудио-признаки, вычисленные с использованием Opensmile. Эти признаки, вместе с пространственными признаками, были объединены в ансамбль моделей, включая линейные SVM. Эксперименты включали в себя аугментацию данных, как на уровне кадров, так и добавление спектральных признаков с использованием преобразования Фурье. В результате проведенных экспериментов удалось улучшить общую точность классификации на тестовом наборе данных до 60.03%. Выведен вывод о важности предварительного обучения на крупных наборах данных по распознаванию лиц для повышения точности распознавания эмоций.

Эти 10 статей представляют различные методы и подходы к классификации и обнаружению эмоций с использованием алгоритмов машинного обучения. Они подтверждают эффективность таких методов в решении проблемы и предлагают новые идеи для дальнейших исследований в этой области.

**3. Методология**

В качестве метрики было взято accuracy, как простая и понятная метрика для сравнения методов машинного обучения и нейронных сетей. Точность определения эмоций определялась для 7 классов эмоций:

# 1 : 'surprise',

# 2 : 'fear',

# 3 : 'disgust',

# 4 : 'happy',

# 5 : 'sad',

# 6 : 'angry',

# 7 : 'neutral'

Анализировался датасет RAF-DB (Real-world Affective Faces Database), который был разработан группой исследователей Wenhao Ding, Xilin Chen и Xiangtuo Chen. Датасет содержит более 30000 изображений, которые показывают лица людей, выражающих различные эмоции. Изображения размечены по 7 базовым эмоциям, таким как радость, грусть, злость и т.д. Показанные изображения в датасете включают разнообразные условия освещения, позы и вида лица. Изображения представлены в цветовом формате и имеют разрешение 256x256 пикселей.

Ссылка на датасет: <http://www.whdeng.cn/raf/model1.html>

Классы в тренировочной выборке имели дисбаланс по количеству изображений (см. рисунок 1).

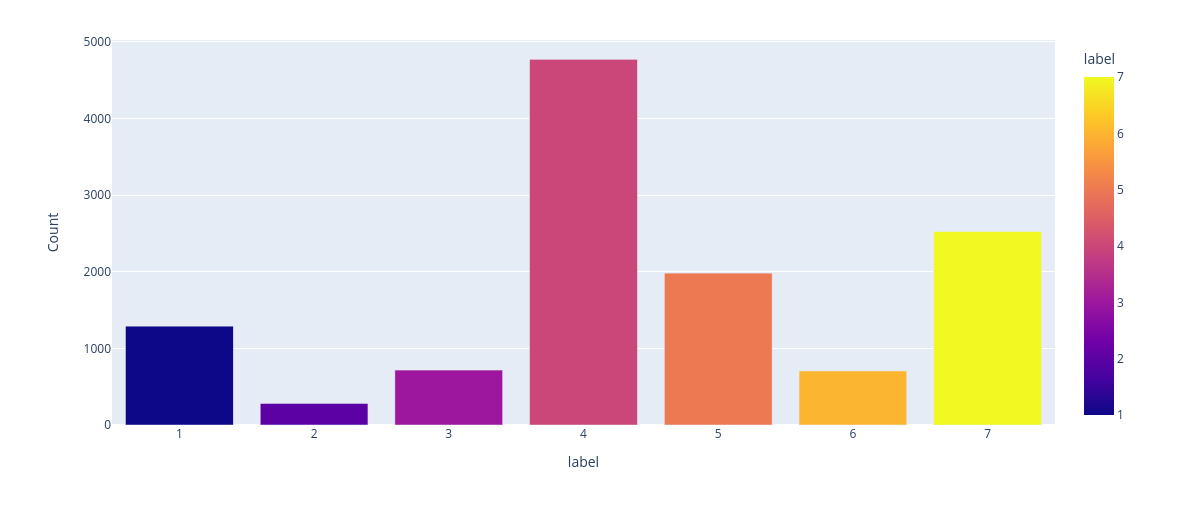


Рис. 1. Тренировочный датасет

Классы в тестовой выборке были уравновешены по полноте, что представлено на следующем графике:

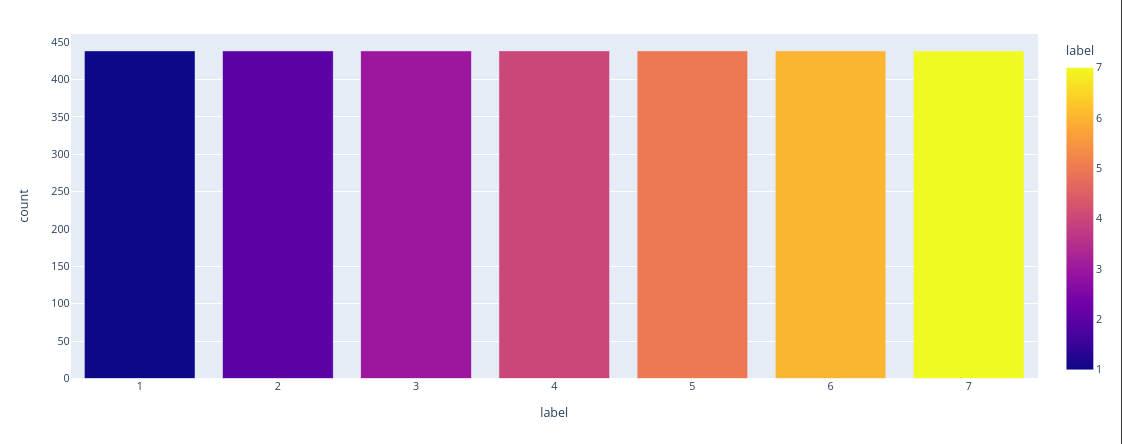


Рис. 2. Тестовый датасет

Целью исследования классических методов машинного обучения является применение метода многоклассовой классификации и байесовской оптимизации для настройки гиперпараметров модели распознавания эмоций на лицах. Был использован датасет, содержащий изображения с лицами, размеченные по эмоциональной окраске. Изображения были преобразованы, чтобы обеспечить одинаковый размер и серый цветовой формат (Gray resize).

Далее проводится предобработка данных, включающая нормализацию и преобразование изображений в числовой формат, пригодный для обучения модели. Для каждой комбинации гиперпараметров проводилось обучение и оценка модели с использованием кросс-валидации. Модели для которых использовался метод: PCA+RF, RF, PCA+SVC, SVC, PCA+GDM(catboost), GDM(catboost). В конце модели оцениваются на предмет точности на тестовом наборе данных.

Затем производится предобработка данных, включая загрузку и изменение размера изображений, преобразование в градиентные признаки с помощью метода HOG, а также стандартизацию значений для обеспечения единообразия данных и повышения эффективности модели. Далее определяются архитектура модели и параметры оптимизации Optuna.

На основе градиентных признаков HOG, создается и обучается модели классификации PCA+RF, RF, PCA+SVC, SVC, PCA+GDM(catboost), GDM(catboost). В процессе обучения модель многократно перебирает комбинации параметров с использованием метода оптимизации Optuna с целью нахождения оптимальных значений. После обучения модели оценивают производительность на отложенном тестовом наборе данных.

Далее описывается методология для использования CNN Feature Extractor и CatBoost, включает в себя две основные части: извлечение признаков с помощью сверточной нейронной сети (CNN Feature Extractor) и классификацию с использованием алгоритма градиентного бустинга CatBoost.

CNN Feature Extractor - это метод, использующий сверточные нейронные сети для извлечения важных признаков. В процессе извлечения признаков CNN применяет операцию свертки, которая позволяет выделить пространственные зависимости в данных. Затем полученные признаки сжимаются с помощью операции пулинга, чтобы уменьшить размерность данных. Наконец, выходы CNN используются в качестве признаков для дальнейшей классификации.

CatBoost - это алгоритм градиентного бустинга, который используется для классификации эмоций. Градиентный бустинг - это метод машинного обучения, который комбинирует несколько слабых моделей (в данном случае деревьев решений), чтобы получить более сильную модель классификации. CatBoost обладает высокой эффективностью и способностью обрабатывать большие наборы данных, его преимущество заключается в том, что он автоматически обрабатывает категориальные признаки и может работать с разными типами данных. Оценивается accuracy, как базовая метрика модели на тестовой выборке.

Этот раздел описывает эксперименты, проведенные для классификации эмоций на основе CNN. В данной методологии применяется стандартная архитектура CNN, включающая сверточные слои, слои пулинга и полносвязные слои. Количество и размерности слоев определяются на основе анализа входных данных и требований к задаче классификации эмоций. Модель CNN обучается на обучающих данных с использованием оптимизатора и функции потерь, выбранных на основе экспериментов. Также проводится аугментация данных для увеличения разнообразия обучающей выборки и улучшения обобщающей способности модели. Результаты классификации эмоций с использованием CNN оцениваются на основе метрики точность.

Визуальное преобразование (Visual Transform) представляет собой метод аугментации изображений, который способен генерировать разнообразные варианты исходных изображений с целью улучшения качества обучения нейронных сетей. Целью визуального преобразования является создание новых вариаций исходных изображений, которые сохраняют информацию о классе исходных изображений, но имеют разные атрибуты, такие как размер, положение, цветовые характеристики и масштаб. Для использования метода визуального преобразования необходимо загрузить изображения из исходного набора данных, создать экземпляр класса VisualTransform, применить различные функции преобразования к загруженным изображениям, генерируя новые вариации, сохранить новые варианты изображений в качестве аугментированного набора данных, использовать аугментированный набор данных для обучения нейронной сети. Эффективность метода визуального преобразования оценена на основе метрики точности.

**4. Результаты**

Код моделей: <https://github.com/lm-cyber/nero/tree/master>

В данном разделе мы представляем результаты экспериментов, проведенных для сравнения алгоритмов машинного и глубокого обучения в задаче классификации эмоций по изображениям. Были использованы все вышеупомянутые модели, а также была рассчитана метрика accuracy для каждой из них.

В первую очередь, была использована модель CNN с различными параметрами. Были произведены несколько запусков с различными значениями архитектуры сети, такими как количество слоев, количество нейронов и функции активации. Детали конфигурации сети приведены в приложении. Для модели CNN было проведено обучение на 100 эпохах с коэффициентом обучения 0.0001 и пакетным размером 128. Был также применен метод ранней остановки для предотвращения переобучения модели. В качестве функции потерь использовалась перекрестная энтропия.

Результаты экспериментов показали следующие значения метрики accuracy для модели CNN:

Densenet-121: 0.85

Inception v3: 0.89

ResNet-101: 0.83

EfficientNetV2\_s: 0.93

Для модели VIT мы задали следующие параметры: интервал масштабирования [0.9, 1.1], случайное изменение размера изображения от -10% до +10% от исходного размера, количество эпох обучения — 40, learning rate — 0.0001, batch size — 128. Мы также использовали раннюю остановку для предотвращения переобучения модели и функцию потерь cross entropy.

Результаты эксперимента с моделью VIT с архитектурой vit\_b16 показали точность (accuracy) равную 0.69. Следует отметить, что данная модель имеет некоторые ограничения в предсказательной способности и требует дальнейшей оптимизации. На наш взгляд это плохой результат из-за сложности модели и отсутствия карт внимания.

Мы применили классические методы машинного обучения, включая использование гистограмм градиентов (HOG) с параметрами orientations=4, pixels\_per\_cell=(16,16), cells\_per\_block=(2,2), и channel\_axis=-1. Провели эксперименты с несколькими моделями, включая Histogram of Oriented Gradients (HOG) и Canny, для которых использовали несколько классификаторов, таких как [Principal component analysis](https://en.wikipedia.org/wiki/Principal_component_analysis)(PCA) Random Forest (RF), Support Vector Machine (SVC) и Gradient Descent Machine (GDM) с использованием генерализационного градиентного спуска (catboost). Метрика точности была измерена для каждого из этих подходов и комбинаций из них:

HOG, модели:

PCA+RF: 0.46

RF: 0.43

PCA+SVC: 0.694

SVC: 0.688

PCA+GDM(catboost): 0.61

GDM(catboost): 0.62

Resize(32,32)+gray\_scale, модели:  
 PCA+RF: 0.45

RF: 0.44

PCA+SVC: 0.67

SVC: 0.68

PCA+GDM(catboost): 0.57

GDM(catboost): 0.6

В финальном подходе мы использовали гибридную модель, объединяющую признаки, извлеченные с помощью нейронной сети, с классическим методом машинного обучения. Результаты эксперимента:

InceptionV3+GDM: 0.5  
 autoencoder+GDM: 0.67

Наилучшим результатом является значение accuracy, полученное в CNN EfficientNetV2\_s: 0.93.

**5. Обсуждение**

В разделе обсуждение рассмотрим применение вышеперечисленных моделей к датасету RAF-DB, которое позволило лучше изучить сложность их архитектуры, количество времени для тренировки и точность, как сравнительную метрику классификации. Обсудим подходы в написании моделей и эксперименты.

Модели Feature Extractor показали интересное преобразование черт лица. Так на рисунке 3 показано исходное изображение лица грустной девочки, а выделение фич для Canny и HOG показано на рисунках 4 и 5 соответственно. Мы можем видеть конкретное представление грустных черт.

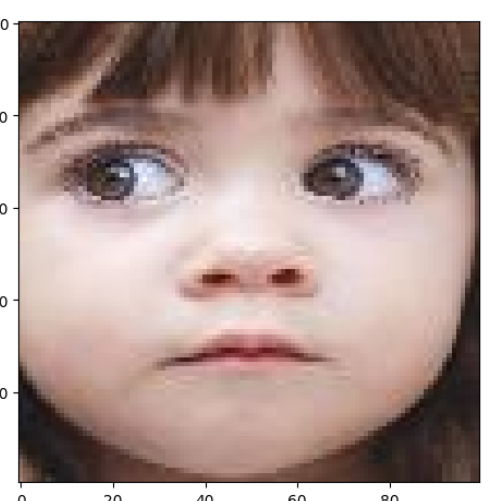


Рис. 3. Исходное изображение

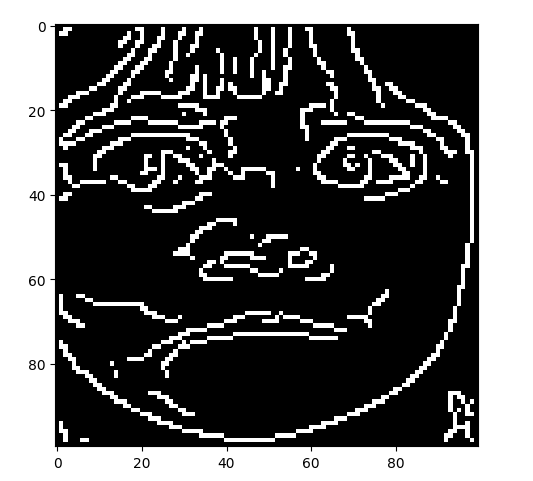


Рис 4. Преобразованное Feature extractor Canny

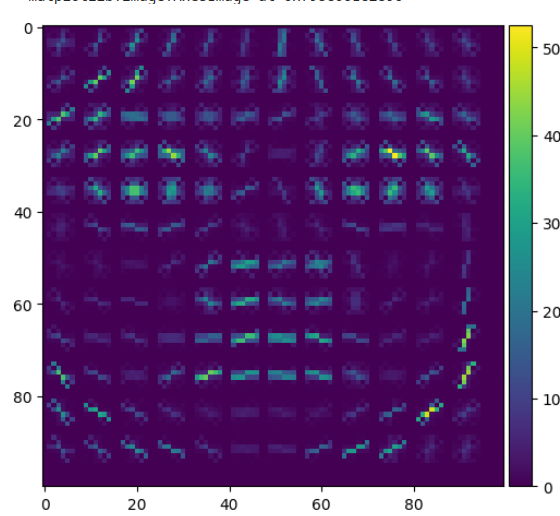


Рис. 5 Преобразованное Feature extractor HOG

Представление изображений моделью с autoencoder показало не самый однозначный результат, поэтому данная область исследования может быть улучшена в дальнейшем (см. рис. 6,7). Плохая архитектура фича экстрактора, поэтому имеем такой результат.

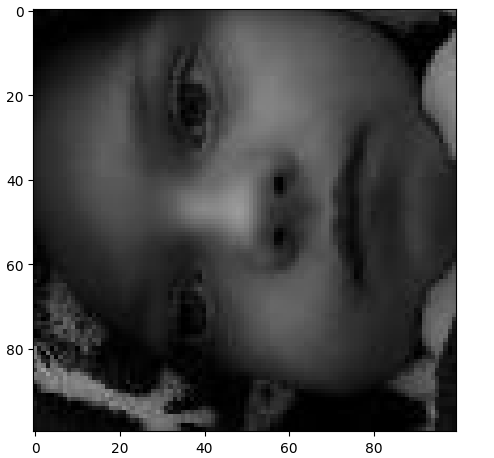


Рис. 6. Изображение Autoencoder

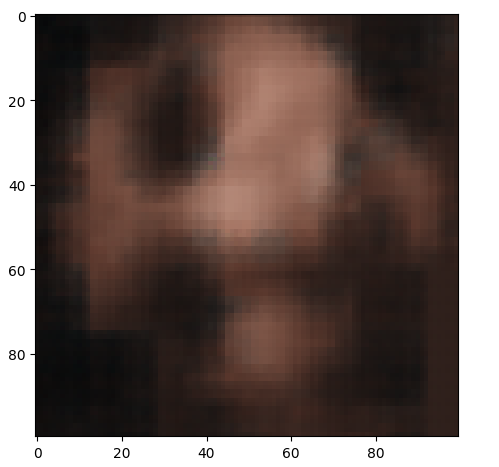


Рис. 7 Результат Autoencoder

Имеем лучшие результаты точности для этих моделей равными 0.68, 0.694 и 0.67 соответственно. Анализируя данные показатели можно сделать вывод о неравномерности в выделении фич между классами эмоций, что привело к неточности в их классификации. Автоэнкодер имеет довольно высокий показатель, несмотря на изображения, которые посылал в модель. Из этого напрашивается вывод о его доработке, ведь метод является одним из наиболее актуальных в задачах нейротехнологий. Ниже представлена архитектура автоэнкодера (см. рис. 8).

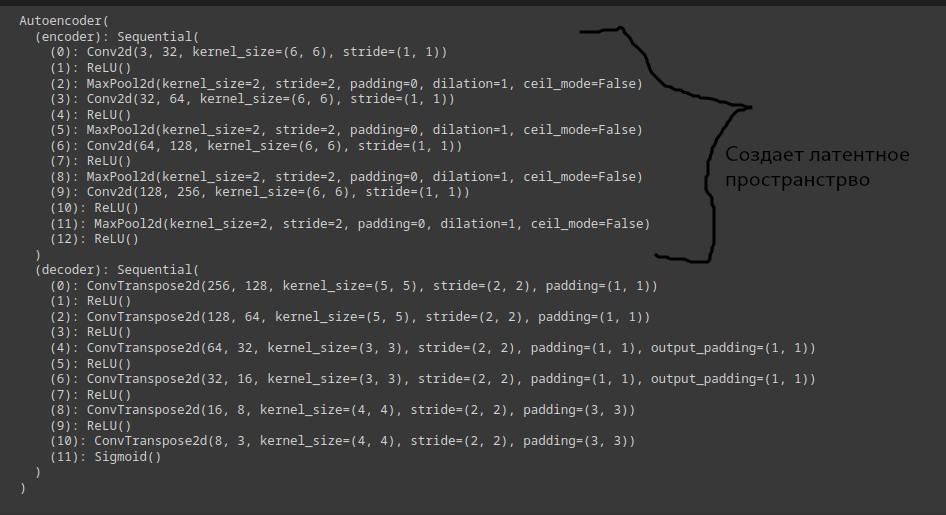
****

Рис. 8. Архитектура autoencoder

Далее рассмотрим лучшие результаты классического пайпа HOG+PCA+SVC, список параметров настройки:

{'C': 2.1344516909404465, 'decision\_function\_shape': 'ovo', 'kernel': 'rbf, ‘class\_weight’:class\_weight'}

{'n\_components': 0.9732684111337544, 'power\_iteration\_normalizer': 'LU', 'svd\_solver': 'full'}

Оценивая матрицу ошибок по тестовой выборке можно сделать вывод о небольшом разбросе в точности классификации эмоций, модель примерно с одинаковой точностью определяла каждый класс (см. рис. 9)

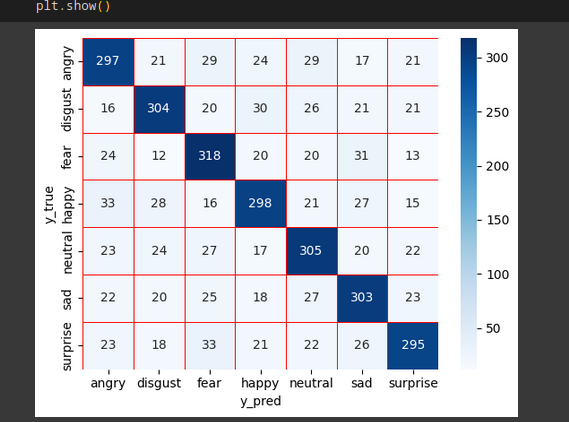
****

Рис.9. Матрица ошибок модели HOG+PCA+SVC

Ниже представлены основные метрики и их изменение для этой модели (рис. 10). Имеем точность, равную 0.69, что является довольно хорошим показателем для классического машинного обучения в задаче классификации эмоций.

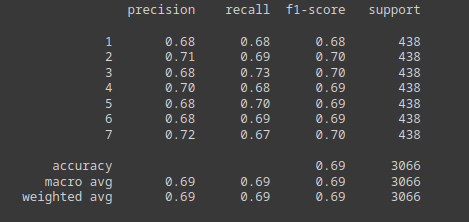
****

Рис. 10. Основные метрики модели HOG+PCA+SVC

Наилучшие результаты в исследовании показал класс CNN, а именно модель EfficientNetV2\_s, имея точность 0.93. Отметим, что это наиболее долгая по обучению модель из представленных, время обучения составило около 50 минут. Рассмотрим матрицу ошибок и основные метрики для этой нейронной сети (см. рис. 11,12). Здесь разброс между классами еще меньше, чем в описанной матрице выше. С задачей CNN справилась отлично.

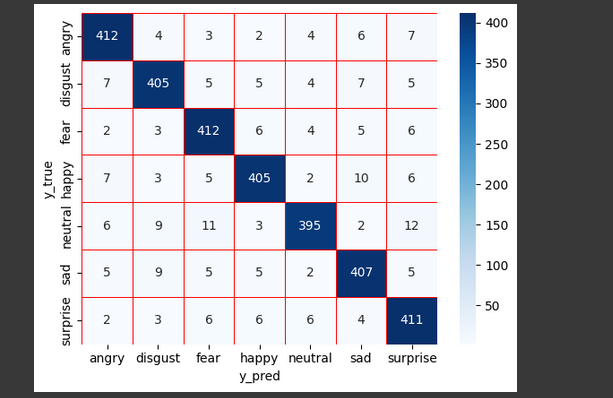


Рис.11. Матрица ошибок EfficientNetV2\_s

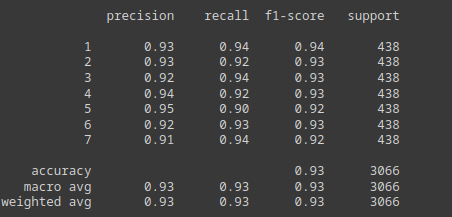


Рис.12. Основные метрики EfficientNetV2\_s

Сравнивая результаты экспериментов, CNN показали себя лучше всего, но если потоково проверять эмоции, то скорость классических методов будет лучше, что лишь подчеркивает их важность в задаче классификации. Необходимо заранее подбирать метод исходя из того, для какой конечной цели он будет использовать и понимать, чем разумнее пожертвовать.

Дальнейшая разработка может предполагать улучшения для методов feature extractor исходя из наших ошибок и доработка autoencoder.  
Применением может быть анализ реакции пользователей на рекламу, покупки, различного рода контент в связке с детектором.

**6. Заключение**

Многие задачи ML - это трейдофф, и не стоит отметать возможность обучения классического ML

**7. Библиография**

https://scholarworks.sjsu.edu/cgi/viewcontent.cgi?article=1643&context=etd\_projects

<https://habr.com/ru/companies/speechpro/articles/418151/>

<https://ieeexplore.ieee.org/document/9763165>

<https://www.diva-portal.org/smash/get/diva2:1677831/FULLTEXT02>

https://www.sciencedirect.com/science/article/pii/S1877050920318019

<https://drive.google.com/file/d/1-mVVbabm8ePTMJKwO0itdMXB3j5vEw7h/view>

<https://towardsdatascience.com/emotion-detection-a-machine-learning-project-f7431f652b1f>

<https://arxiv.org/pdf/1705.07871.pdf>

<http://www.diva-portal.org/smash/get/diva2:1174434/FULLTEXT01.pdf>

<https://arxiv.org/abs/1711.04598>