УДК 004.8 **Информационные технологии**

Галичий Д. А., *магистрант, Московский государственный технический университет им. Н.Э. Баумана*

Афанасьев Г. И., *кандидат технических наук, доцент, Московский государственный технический университет им. Н.Э. Баумана*

Нестеров Ю. Г., *кандидат технических наук, доцент, Московский государственный технический университет им. Н.Э. Баумана*

РАСПОЗНАВАНИЕ ЭМОЦИЙ ЧЕЛОВЕКА ПРИ ПОМОЩИ СОВРЕМЕННЫХ МЕТОДОВ ГЛУБОКОГО ОБУЧЕНИЯ

Аннотация: Рассмотрено несколько современных подходов к построению нейронных сетей для распознавания эмоций человека. Продемонстрированы результаты работы модели свёрточной нейронной сети и моделей, спроектированных по правилам переносимого обучения. Сделаны выводы об эффективности изученных нейросетей и о существующих проблемах в распознавании эмоций рассмотренными моделями. Предложены пути решения возникающих трудностей в определении эмоций.

Ключевые слова: распознавание эмоций, свёрточные нейронные сети, переносимое обучение, ResNet50, SeNet50, VGG16.

Annotation: The emotions reflected on the face of a person serve as one of the main tools of nonverbal communication, which in most cases allows you to accurately assess the person's attitude to what is happening. Currently, thanks to the capabilities of neural networks, this indicator is finding new applications in various fields of science. Automated assessment of user emotions can be useful in analyzing the convenience and quality of software products, in setting up the means of human interaction with a computer system, during testing the effectiveness of educational services, in monitoring the condition of vehicle drivers, in psychology, in medicine and in many other areas. This article discusses several modern approaches to the

construction of emotion recognition systems belonging to seven classes: expressions of anger, disgust, fear, joy, sadness, surprise, and a neutral state. In order to see the objective results of the neural networks the training was performed on a single data set – «FER2013». The results of the four-layer convolutional neural network model and models ResNet50, SeNet50, VGG16 designed according to the rules of transfer learning are demonstrated in the form of graphs of accuracy and loss metrics. Conclusions about the effectiveness of the studied neural networks and about the existing problems in the recognition of emotions by the considered models are given. Among the disadvantages, the unbalanced classes in the data set and the frequent problem of retraining the neural network were noted. The ways of solving the emerging difficulties in recognizing emotions are proposed, namely, increasing the volume of the data set and using «Dropout» and «Early stopping» techniques. The constructed neural networks can be further improved and applied as a module for emotion recognition in software products.

Keywords: face expression recognition, convolution neural networks, transfer learning, ResNet50, SeNet50, VGG16.

Введение

Эмоции человека, отражающиеся на лице, служат одним из главных инструментов невербального общения, который в большинстве случаев позволяет безошибочно оценить отношение человека к происходящему. В настоящее время благодаря возможностям нейронных сетей этот показатель находит всё новые применения в различных отраслях науки. Автоматизированная оценка эмоций пользователя может принести пользу при анализе удобства и качества программных продуктов, при настройке средств взаимодействия человека с компьютерной системой, во время тестирования эффективности образовательных сервисов, в наблюдении за состоянием водителей транспортных средств, в психологии, медицине и во многих других сферах.

Проблематика в задачах распознавания эмоций на лице человека определяется характером условий, в которых производится оценка: они могут быть контролируемыми или естественными. В первом случае достигается высокая (порядка 98%) точность работы моделей с обрабатываемыми изображениями за счёт заранее определённых параметров, таких как ракурс, поза, освещение. Во втором случае, когда фотографии, подаваемые на вход модели, варьируются и не отбираются по продуманному шаблону, достичь такой точности всё ещё очень сложно [1]. В связи с потребностью в результативных алгоритмах по распознаванию эмоций человека поиск оптимальной нейросетевой архитектуры продолжает оставаться актуальной задачей.

Основная часть

В этой статье рассматриваются несколько современных подходов к построению систем распознавания эмоций, принадлежащих семи классам: выражению гнева, отвращения, страха, радости, грусти, удивления и нейтрального состояния. В рамках проведённых экспериментальных исследований делаются выводы о существующих сложностях в повышении точности работы моделей, предпринимаются попытки улучшения качества распознавания эмоций за счёт методик, которые предлагаются в последних научных публикациях, посвящённых проблематике компьютерного зрения.

Набор данных «FER2013»

Большинство экспериментальных исследований на тему распознавания эмоций человека проводятся с использованием набора данных «FER2013» (face emotion recognition). Этот набор был создан в 2013 году с целью проведения соревнований на платформе «Kaggle» учёными, работающими в области машинного обучения, - Пьером Люком Каррье и Аароном Курвиллом [2]. В результате соревнования лучшими стали три команды, которые использовали свёрточные нейронные сети (CNN – convolutional neural networks) и методы преобразования изображений. Победитель, Ичуань Тан, смог достичь точности 71,162%, использовав в качестве функции потерь машину опорных векторов - SVM (support vector machine) и L2-SVM [3]. После окончания соревнования

набор фотографий «FER2013» продолжает активно использоваться во многих исследованиях как наиболее объёмная имеющаяся в открытом доступе коллекция изображений с эмоциями [4]. В данной работе обучение нейронных сетей также производилось на основе «FER2013», который содержит 35887 фотографий людей, выражающих семь разных эмоций (рис. 1).

Рис. 1. Примеры изображений каждого из классов, содержащихся в «FER2013»

Следует отметить, что классы в наборе данных несбалансированы: в классе «радость» хранятся 8989 фотографий, в то время как класс «отвращение» состоит лишь из 547 фотографий. На категории «гнев», «страх», «грусть», «удивление» и «нейтральное выражение лица» приходится 4953, 5121, 6077, 4002 и 6198 фотографий соответственно. Размер изображений в наборе данных – 48*48 пикселей, все они являются чёрно-белыми.

Построение свёрточной нейронной сети

Свёрточные нейронные сети, возникшие в результате исследований зрительной коры головного мозга, на сегодняшний день являются самым популярным способом анализа изображений [5]. Согласно последним исследованиям, свёрточные нейронные сети с несложной архитектурой способны показать приемлемую точность распознавания эмоций (точность на тестовых данных может превышать 60%), а модели, состоящие из ансамбля свёрточных нейросетей, - точность, сравнимую с результатами победителя соревнования «Kaggle» (точность на тестовых данных достигает 75%), при этом не требуя дополнительного сбора данных для обучения [6]. В качестве первой модели распознавания эмоций в данной работе была выбрана архитектура нейросети, содержащая четыре свёрточных слоя (рис. 2).

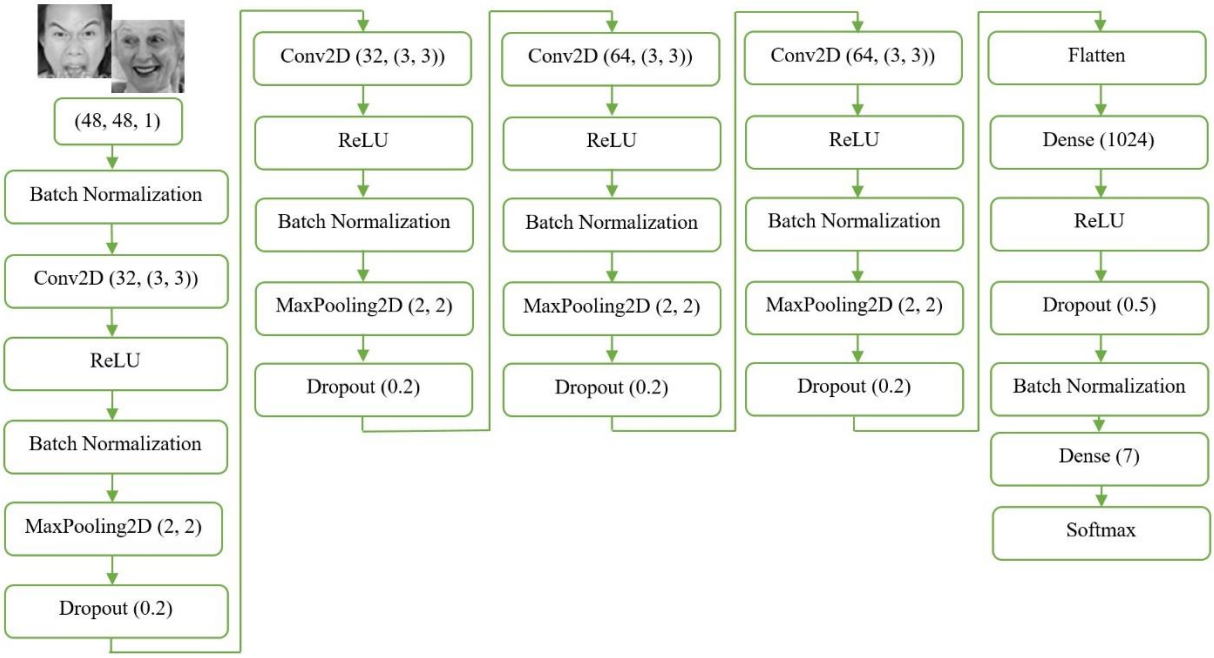

Рис. 2. Архитектура свёрточной нейронной сети

В первых двух свёрточных слоях использовались 32 фильтра с ядрами 3*3 пикселя, в последних двух свёрточных слоях – 64 фильтра с ядрами 3*3 пикселя. Выходные карты признаков на каждом канале были той же размерности, что и входные изображения (48*48), а функцией активации послужила кусочно-линейная ReLU (rectified linear units):

$$\mathrm{ReLU}(x) = \begin{cases} 0, \text{если } x < 0 \\ x, \text{если } x \geq 0. \end{cases}$$

В соответствии с правилами построения свёрточных нейронных сетей, для укрупнения масштаба получаемых признаков после каждого свёрточного слоя применялась операция MaxPooling с размером окна (2*2). Чтобы ускорить процесс обучения, был задействован алгоритм batch normalization, а в качестве метода регуляризации, предотвращающего переобучение сети, - техника Dropout. Перед полносвязным слоем, состоящим из 1024 нейронов и имеющим функцию активации ReLU, проводилась линеаризация при помощи слоя «Flatten». Финальный полносвязный слой, в соответствии с количеством классов в наборе данных «FER2013», содержал семь нейронов и завершался функцией активации «Softmax» (сглаженный максимум), которая лучше всего подходит для задач классификации, когда количество возможных классов больше двух:

$$\mathrm{softmax}(x)_j = \frac{\exp(x_j)}{\sum_i \exp(x_i)}.$$

В качестве функции потерь была выбрана стандартная категориальная кросс-энтропия:

$$H_t(y) = -\sum_i t_i \log y_i,$$

где y – предсказанное значение, а t – правильный ответ, а в качестве оптимизатора – «Adam» со скоростью обучения 0,0001.

Дополнительно для предотвращения переобучения нейронной сети был использован механизм «EarlyStopping», который каждую эпоху оценивал метрику качества «accuracy» на обучающем и валидационном наборах. В случае, если метрика качества «val_acc» (доля правильных ответов) на обучающем наборе увеличивалась, а на валидационном – уменьшалась в течение пяти эпох, то производилась ранняя остановка процесса обучения.

Графики, отражающие изменение доли правильных ответов и значения функции потерь на обучающем и тестовом наборах данных, в течение процесса обучения свёрточной нейронной сети, приведены на рисунке 3.

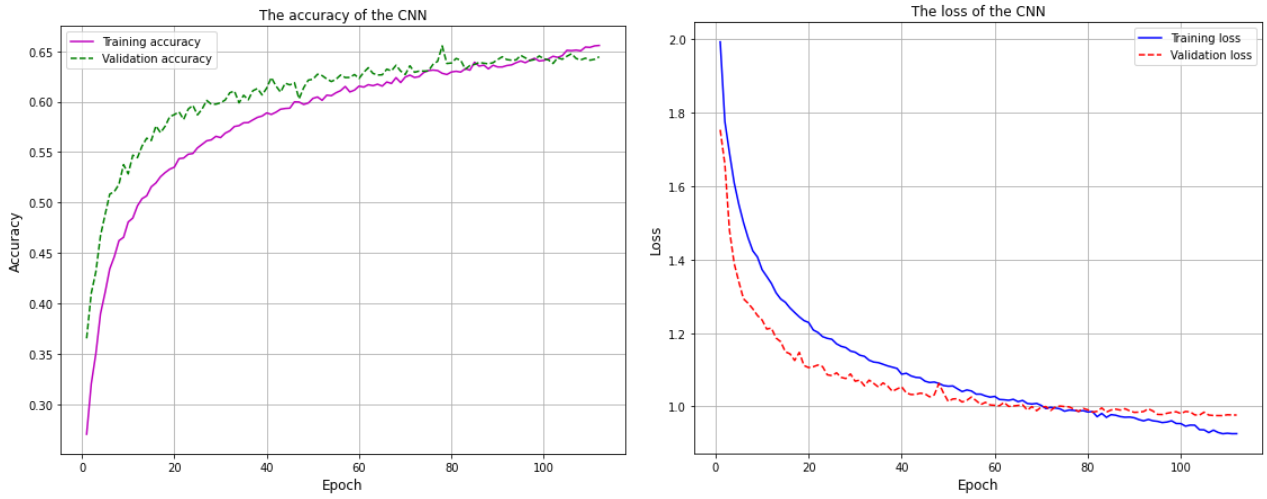

Рис. 3. Изменение метрик «accuracy» и «loss» на обучающем и валидационном наборах данных в процессе обучения свёрточной нейронной сети

При помощи библиотеки «wandb» были визуализированы предсказания нейросети на тестовом наборе данных. По рисунку 4 видно, что большинство ответов нейронной сети действительно являются правильными.

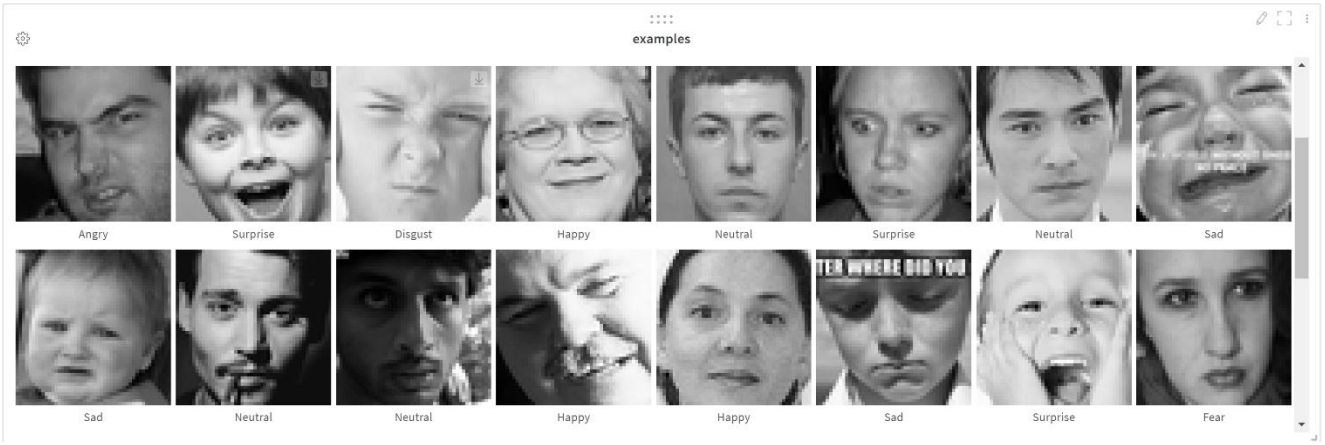

Рис. 4. Визуализация предсказаний модели CNN на тестовом наборе данных

Таким образом, нейронная сеть, содержащая лишь четыре свёрточных слоя, продемонстрировала хороший результат обучаемости и за 112 эпох приблизилась к отметке 65% точности определения эмоции человека на проверочном наборе данных.

Дообучение нейронной сети ResNet50

Высокой эффективностью обладают нейронные сети, построенные при помощи модификации и дообучения заранее обученных моделей свёрточных нейронных сетей с большим количеством слоёв (Transfer Learning), поэтому в данной работе также был использован этот приём.

Следующая модель для распознавания эмоций, которая была рассмотрена в рамках исследования, основана на предобученной нейронной сети ResNet50. Реализованная в библиотеке «Keras VGG-Face» ResNet50 – это свёрточная нейронная сеть, состоящая из 175 слоёв [7]. Перед тем, как использовать ResNet50, было необходимо преобразовать входные изображения к размеру 197*197 и цветовому формату RGB [8]. Далее была произведена настройка модели, полученная архитектура представлена на рисунке 5.

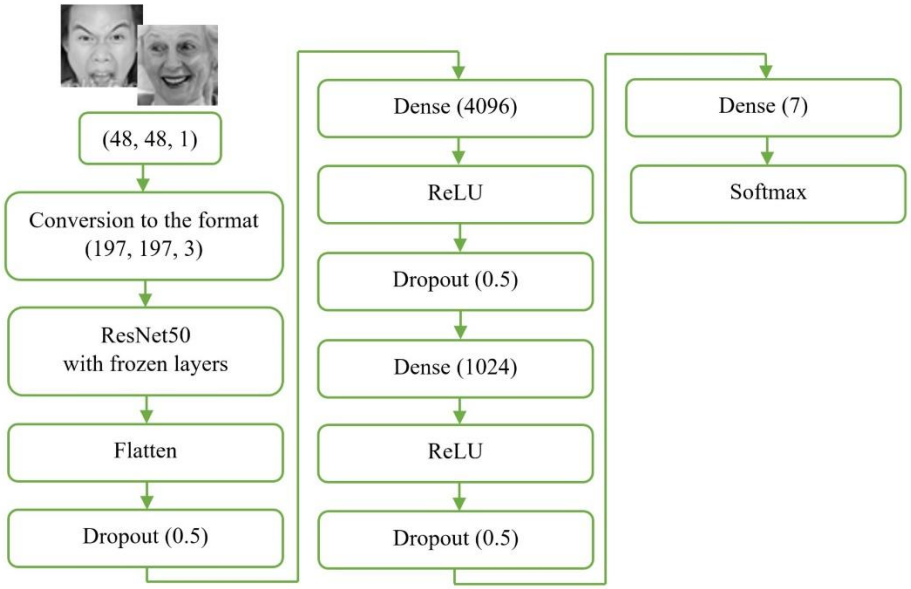

Рис. 5. Архитектура нейросети, построенной на основе ResNet50

170 слоёв ResNet50 были заморожены, чтобы избежать потери предварительно обученных весов. После выходного слоя нейронной сети была произведена линеаризация при помощи Flatten и «прореживание» нейронов с помощью Dropout с вероятностью 0,5. Первоначальный выходной слой ResNet50 был заменён на два последовательных полносвязных слоя, содержащих 4096 и 1024 нейрона соответственно, с функцией активации ReLU. После каждого из этих слоёв применялся Dropout с вероятностью 0,5. За классификацию снова отвечал полносвязный слой из семи нейронов с функцией активации Softmax. В качестве оптимизатора модели был выбран механизм SGD (стохастический оптимизатор градиентного спуска), скорость которого динамически уменьшалась во время обучения модели.

На следующих графиках отражено изменение метрик «accuracy» и «loss» в процесс обучения сконструированной нейронной сети (рис. 6).

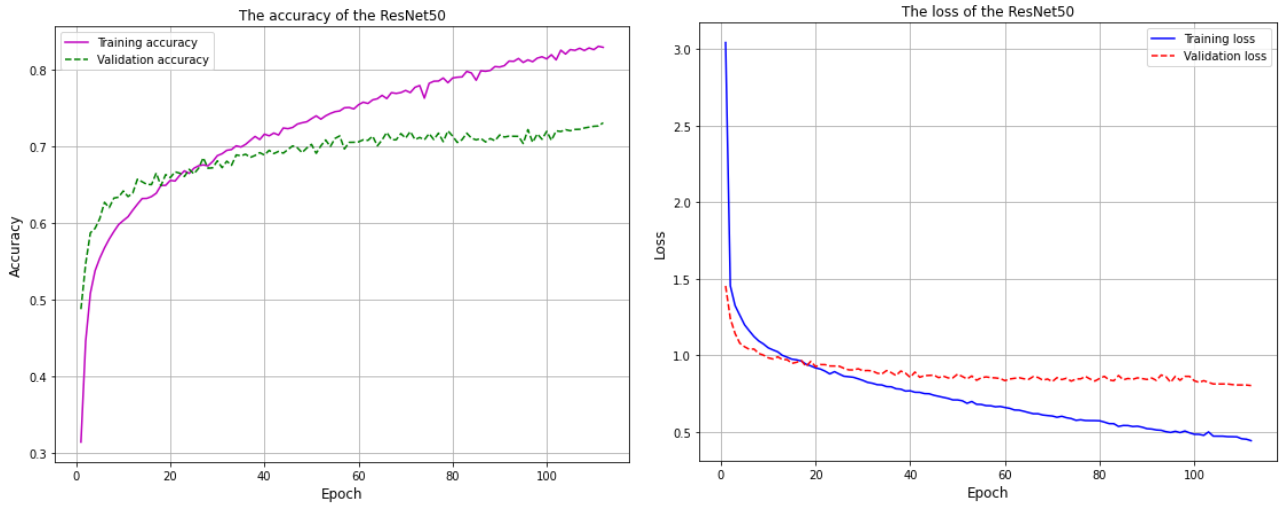

Рис. 6. Изменение метрик «accuracy» и «loss» на обучающем и валидационном наборах данных в процессе обучения нейронной сети, построенной на основе ResNet50

После прохождения 112 эпох точность работы модели на тестовых данных достигала 73%, что подтвердило эффективность методов Transfer Learning в решении задачи распознавания эмоций.

Дообучение нейронной сети SeNet50

Ещё одна предварительно обученная нейронная сеть, содержащаяся в библиотеке «Keras VGG-Face» - это SeNet50. Она так же, как и ResNet50, активно используется в качестве основы при создании многих нейросетевых алгоритмов классификации [9]. Её архитектура напоминает структуру модели ResNet50, поэтому в данной работе дообучение производилось способом, аналогичным вышеописанному. Схема основанной на SeNet50 нейросети для распознавания эмоций представлена на рисунке 7.

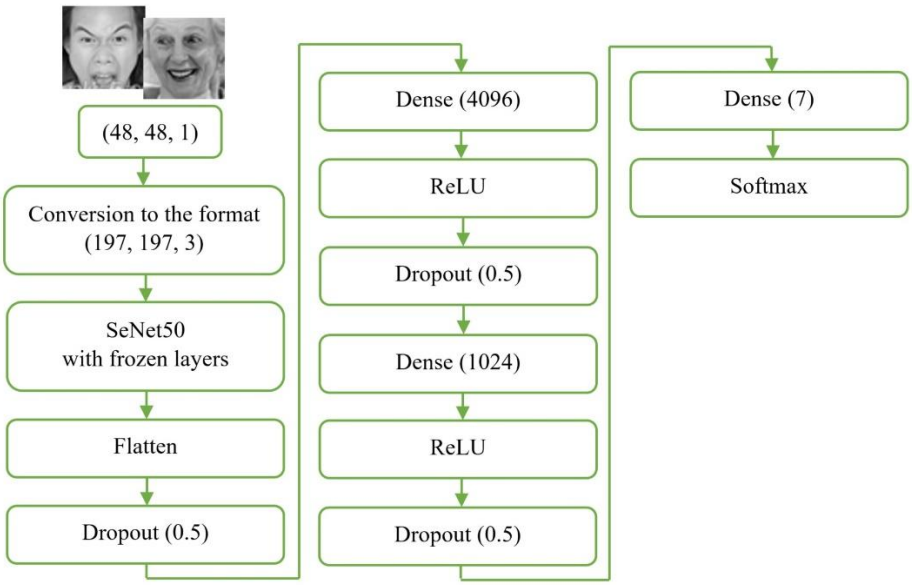

Рис. 7. Архитектура нейросети, построенной на основе SeNet50

При формировании этой модели применялись уже описанные методики замораживания слоёв нейросети, использования слоёв Dropout, добавления в архитектуру полносвязных слоёв из 4096 и 1024 нейронов с функцией активации ReLU и классификационного слоя из 7 нейронов с Softmax-активацией. Настройки оптимизатора SGD остались прежними, автоматически снижающими скорость обучения при необходимости.

В результате обучения полученной нейронной сети были построены графики метрик, которые показаны на рисунке 8.

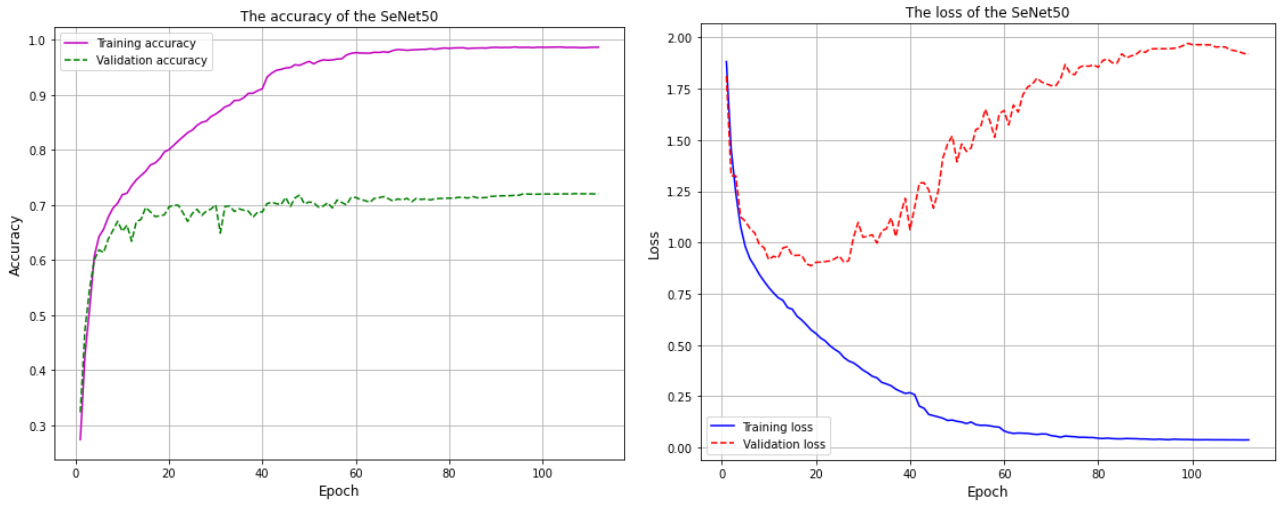

Рис. 8. Изменение метрик «accuracy» и «loss» на обучающем и валидационном наборах данных в процессе обучения нейронной сети, построенной на основе SeNet50

По прошествии 110 эпох обучения нейронная сеть, основанная на SeNet50, показывала практически такую же точность предсказаний на тестовом наборе данных, как и дообученная модель ResNet50, - 72%.

Дообучение нейронной сети VGG16

В завершение исследования была построена ещё одна модель с предобученной нейросетью в качестве основы. VGG16 – 16-слойная нейронная сеть, имеющая глубокую, но довольно несложную архитектуру, которая является очень популярной в задачах, решаемых переносимым обучением [10]. Добавленные к первоначальной структуре VGG16 слои представлены на рисунке 9.

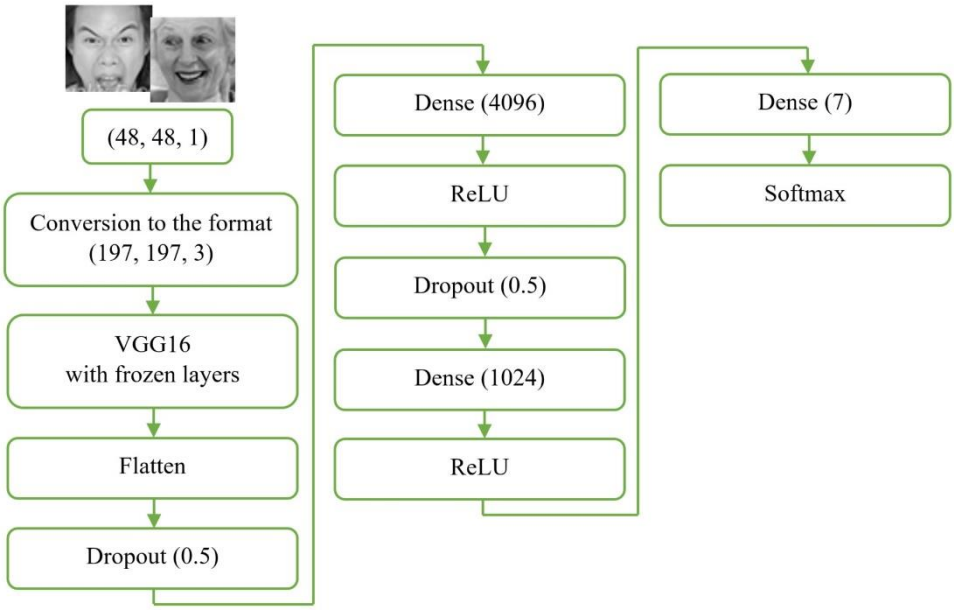

Рис. 9. Архитектура нейросети, построенной на основе VGG16

Ключевые операции в построении сети остались неизменными: заморозка предобученных слоёв модели, добавление полносвязных слоёв из 4096 и 1024 нейронов с функцией активации ReLU и последующим применением Dropout, в конце – классификация полносвязного слоя из 7 нейронов с функцией активации Softmax, оптимизатор – Adam.

Графики изменения метрик на протяжении обучения нейронной сети приведены на рисунке 10.

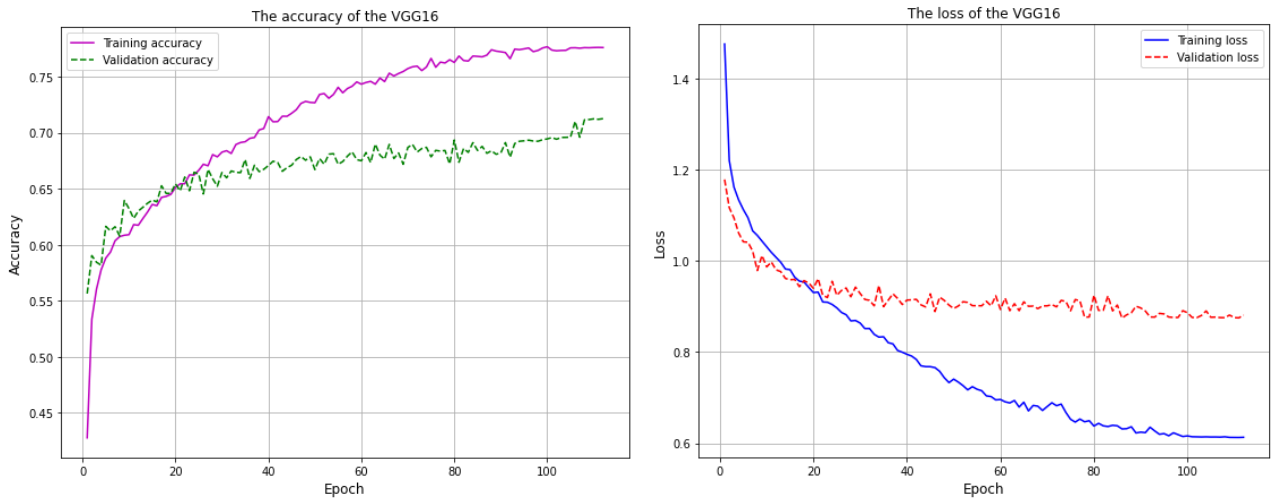

Рис. 10. Изменение метрик «accuracy» и «loss» на обучающем и валидационном наборах данных в процессе обучения нейронной сети, построенной на основе VGG16

Как следует из результатов, представленных на графиках, за счёт повторного использования предварительно обученной модели VGG16 и тренировки лишь весов связей, ведущих к новому выходному слою, удалось достигнуть точности 71% в задаче классификации эмоций человека.

Заключение

Таким образом, построив несколько часто встречающихся в задачах классификации моделей, можно сделать вывод об их эффективности в рамках работы по распознаванию эмоций человека при помощи компьютерного зрения. Точность работы этих нейронных сетей определяется грамотным выбором гиперпараметров и тонкой настройкой модели.

Проведённое исследование продемонстрировало, что задачу распознавания эмоций человека можно решить несколькими современными методами, каждый из которых способен показать достойный внимания результат. Высокой эффективностью может обладать как простая свёрточная нейронная сеть, состоящая лишь из нескольких слоёв, так и сложные дообученные модели, например, ResNet50, SeNet50, VGG16, построенные по правилам переносимого обучения.

Основные сложности, возникающие при обучении моделей – несбалансированность классов в исходном наборе данных, а также проблема

переобучения нейронных сетей. Путём увеличения количества изображений в наборе данных можно избавиться от имеющегося дисбаланса в наборе «FER2013» и, как результат, устранить сложность в распознавании эмоций, образцов которых недостаточно для качественного обучения нейронной сети. С переобучением моделей можно бороться, применяя технику «прореживания» нейронов – Dropout, а также механизм «Early stopping», прерывающий процесс обучения, если метрика точности не показывает положительной динамики развития в течение нескольких эпох подряд.

Рассмотренные модели нейронных сетей могут быть дополнительно доработаны и использованы в качестве модуля распознавания эмоций в программных продуктах.

Библиографический список:

1. Pramerdorfer C., Kampel M. Facial expression recognition using convolutional neural networks: state of the art // URL: https://www.researchgate.net/publication/311573401_Facial_Expression_Recognition_using_Convolutional_Neural_Networks_State_of_the_Art (дата обращения 12.01.2021).

2. Challenges in Representation Learning: Facial Expression Recognition Challenge // URL: https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data (дата обращения 12.01.2021).

3. Goodfellow I. J., Erhan D., Carrier P.L., Courville A. Challenges in representation learning: A report on three machine learning contests // URL: https://arxiv.org/pdf/1307.0414.pdf (дата обращения 12.01.2021).

4. Mollahosseini A., Chan D., Mahoor M.H. Going deeper in facial expression recognition using deep neural networks // URL: https://arxiv.org/pdf/1511.04110.pdf (дата 12.01.2021).

5. Mehendale N. Facial emotion recognition using convolutional neural networks (FERC) // URL: https://link.springer.com/article/10.1007/s42452-020-2234-1#citeas (дата обращения 12.01.2021).

6. Talegaonkar I., Joshi K., Valunj S., Kohok R., Kulkarni A. Real time facial expression recognition using deep learning // URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3421486 (дата обращения 12.01.2021).

7. Raghu V. N, Bharathi R. S. Facial expression recognition using deep learning // URL: https://arxiv.org/ftp/arxiv/papers/2006/2006.04057.pdf (дата обращения 12.01.2021).

8. Selvaraju R.R., Cogswell M., Das A., Vedantam R., Parikh D., Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization // URL: https://openaccess.thecvf.com/content_ICCV_2017/papers/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.pdf (дата обращения 12.01.2021).

9. Kim H.-R., Kim Y.-S., Kim S. J., Lee I.-K. Building emotional machines: recognizing image emotions through deep neural networks // URL: https://arxiv.org/pdf/1705.07543.pdf (дата обращения 12.01.2021).

10. Minaee S., Abdolrashidi A. Deep-emotion: facial expression recognition using attentional convolutional network // URL: https://arxiv.org/pdf/1902.01019v1.pdf (дата обращения 12.01.2021).