УДК 681.12

РОЗПІЗНАВАННЯ ЕМОЦІЙ ЛЮДИНИ ЗА ДОПОМОГОЮ ЗГОРТКОВОЇ НЕЙРОННОЇ МЕРЕЖІ

Зінов'єв Євгеній[1], *Арсенюк Ігор*[1]

[1]Вінницький національний технічний університет

Анотація

Проаналізовано основні варіанти архітектур згорткових нейронних мереж у контексті розв'язання задачі розпізнання емоцій. Запропоновано додатковий проміжний згортковий шар, який дозволяє знизити вимоги до потужності обчислювальних ресурсів без вагомих втрат точності. Розроблено відповідний програмний продукт.

Abstract

Basic variants of convolutional neural network architectures are analyzed in the context of emotion recognition resolution. An additional intermediate convolutional layer is proposed to reduce the power requirements of computing resources without significant loss of accuracy. A suitable software product has been developed.

Вступ

Задача розпізнавання емоцій людини є досить актуальною на сучасному етапі розвитку людства. Її розв'язання є важливим для багатьох сфер життєдіяльності людей. Так, наприклад, спостереження за емоціями може допомогти помітити людину, яка потенційно може створити проблеми на публічному заході. Аналіз емоцій людей на платформах станцій метро, у аеропортах, автовокзалах, залізничних вокзалах і т. п. допоможе визначати підозрілу поведінку та сповіщати про це відповідні органи (наприклад, про потенційну терористичну загрозу). Такі емоції, як сумнів і злість, можуть ховатися під маскою і контрастувати з тим, що говорить людина. Супермаркети та великі точки продажу також можуть отримати з цього вигоду, аналізуючи емоції людей, та використовуючи це у цільовому маркетингу та під час розв'язання задачі раціонального розміщення товарів. Розпізнавання емоцій обіцяє бути корисним і під час різноманітних опитувань, оскільки це дає можливість побачити, які саме питання (чи яка частина рекламного роліку) добре працюють і викликає емоційний відгук людей. Також, розпізнавання емоцій можна ефективно використати у медичній сфері та у сфері освіти, в тому числі і під час дистанційного навчання, яке отримало небувалої актуальності в умовах карантину внаслідок поширення коронавірусної інфекції Covid-19.

Метою роботи є аналіз та обґрунтування вибору архітектури згорткової нейронної мережі для розпізнавання емоцій людини з максимальною точністю та мінімальним використанням ресурсів, а також створення відповідного програмного продукту.

Стислий аналітичний огляд основних напрацювань у сфері розпізнавання емоцій людини

Можливість якісного аналізу та інтерпретації зображень з'явилась завдяки швидкому росту обчислювальної потужності комп'ютерів на початку теперішнього сторіччя. Загалом, в області розпізнавання зображень вченими окреслено основний підхід – виділення заздалегідь відомих ознак шуканого об'єкту за допомогою розбиття його зображення. Даний метод реалізовано за допомогою нейронних мереж із можливістю навчання та самонавчання. Уже при навчанні на тестовому наборі даних (зображеннях) система виділяє певні ознаки, розробляючи на їх основі власні правила класифікації об'єктів. Натепер існує велика кількість наборів даних та достатня обчислювальна потужність для реалізації якісних моделей, що базуються на нейронних мережах.

Розглянемо найвагоміші розробки в області розпізнавання емоцій. Для цього потрібно спочатку розпізнати обличчя людини, а вже потім розпізнавати контури його складових, щоб визначити емоцію. Про методи розпізнавання облич більш докладно наведено у роботі [1].

Розпізнавання емоцій – процес, що виконується за допомогою нейронних мереж із особливою архітектурою. Найпомітнішою особливістю мережі є концепція ієрархічного парсингу облич-

чя [2]. Зображення передається через мережу декілька разів, щоб вперше виявити загальний контур обличчя, а після цього його основні ознаки: очі, ніс і рот, і, нарешті, належну емоцію.

Інша робота у напрямку розпізнання емоцій [3] використовує фільтрацію Gabor для обробки зображень та підтримку векторної машини (SVM) для класифікації. Фільтр Габора особливо підходить для розпізнавання емоцій у зображеннях і, як стверджується, імітує функцію зорової системи людини. Точність розпізнавання образів досить висока, коливаючись від 88% у випадку гніву до майже 100% у випадку здивованості людини. Недоліком такого підходу є досить жорсткі вимоги на вхідне зображення (обов'язкова відповідність його строгому формату), що вимагає додаткових затрат на попередню обробку даного зображення.

Одне з останніх досліджень з розпізнавання емоцій описує нейронну мережу, здатну розпізнавати расу, вік, стать та емоції. Набір даних, що використаний для останньої категорії, походить з виклику розпізнавання виразів обличчя (FERC-2013) [4]. Особливістю дослідження є використання чітко організованої глибокої нейронної мережі, що складається з трьох згорткових шарів (одного повністю зв'язного шару та декількох невеликих шарів між ними) дозволила досягти середньої точності 67% за класифікацією емоцій, що еквівалентно результатам, отриманим у попередніх сучасних публікаціях на тому ж наборі даних [2, 3].

Як вказано у роботі [5], найперспективнішою концепцією аналізу виразів обличчя є використання глибоких згорткових нейронних мереж.

Отже, враховуючи вищенаведене, для вирішення проблеми розпізнавання емоцій у дослідженні детальніше зупинимось на глибоких архітектурах.

Аналіз та обґрунтування вибору нейронної мережі для розпізнавання емоцій

Своєю назвою згорткові нейронні мережі зобов'язані використанню математичної операції згортки (особливий вид лінійної операції). Згорткові мережі – нейронні мережі, в яких замість спільної операції множення на матрицю, хоча б в одному шарі, використовується згортка. У загальному вигляді згортка – операція над двома функціями дійсного аргументу. Функцію згортки можна подати у вигляді [6]:

$$S(t) = (x \cdot w)(t). \tag{1}$$

У термінології згорткових мереж перший аргумент (у нашому прикладі функція x) називається входом, а другий (функція w) – ядром. Вихід $S(t)$ називається картою ознак.

Оскільки під час розпізнавання емоцій на вхід буде подано зображення, представлене у вигляді двовимірної матриці, формула (1) набуде такого вигляду:

$$S(i,j) = (I \cdot K)(i,j) = \sum_m \sum_n I(i+m, j+n) K(m,n), \tag{2}$$

де I – вхідне зображення, K – ядро, i, j – координати елементів вхідного зображення, m, n – координати елементів ядра.

Далі розглянемо три основні архітектури згорткових нейронних мереж, які найчастіше використовують для розпізнання емоцій людини.

(A) Перша мережа для тестування базується на наведених дослідженнях Крижевського та Хінтона [7]. Це найпростіша мережа (з трьох розглянутих), і вона потребує найменших обчислювальних вимог. Варто зауважити, що оскільки розроблювальний додаток може бути представлений у вигляді розпізнавання емоцій в реальному часі у вбудованих системах, алгоритми швидкої роботи є дуже бажаними. Дана мережа складається із трьох згорткових шарів та двох повнозв'язних шарів, для зменшення розміру зображення, та шаром проріджування для зменшення ймовірності перенавчання. Гіперпараметри вибираються такими, що кількість обчислень у кожному згортковому шарі залишається приблизно однаковою (забезпечує збереження інформації по всій мережі). Навчання здійснюється за допомогою різного числа згорткових фільтрів, щоб оцінити їх вплив на результативність.

(B) У 2012 році була розроблена згорткова мережа AlexNet для класифікації зображень у більш ніж 1000 різних класах, використовуючи 1,2 мільйона зразкових зображень із набору даних ImageNet [8]. Внаслідок того, що у цьому дослідженні модель має розрізняти сім основних [9] емоцій, а завдяки обмеженим обчислювальним ресурсам розмір оригінальної мережі вважається занадто великим, що може негативно відобразитись на продуктивності обробки зображення, особ-

ливо в умовах відеопотоку, тому автори спробували зменшити кількість згорткових шарів з п’яти до трьох. В отриманих трьох повнозв’язних шарах кількість вузлів кожного повнозв’язного шару також було зменшено з 4096 до 1024. Хоча початкова мережа була оптимізована для паралельних тренувань, було відзначено, що це не було необхідно для простішої її версії. Мережа також використовує локальну нормалізацію для прискорення шарів тренувань та відсіву, щоб зменшити ступінь перенавчання.

(C) Ще ряд експериментів було проведено на основі мережі, наведеної у роботі Гуді [4]. Оскільки це дослідження також мало на меті розпізнати сім емоцій за допомогою набору даних FERC-2013, архітектура повинна стати гарною відправною точкою для наших досліджень. Початкова мережа починається з шару введення розмірністю 48x48, що відповідає розміру вхідних даних. За цим шаром роташований один згортковий шар. Взагалі дана мережа містить два згорткових шари та одним повнозв’язний шар, з’єднаний з вихідним шаром. Метод проріджування було застосовано до повністю пов'язаного шару, і весь шар містить блоки ReLu.

Для нашого дослідження вирішено застосовати другий проміжний згортковий шар для зменшення кількості параметрів. Це дозволить знизити вимоги до потужності обчислювальних ресурсів, здатних реалізувати дану мережу [4]. Крім того, покращується швидкість навчання без вагомих втрат точності розпізнавання емоцій.

Усі мережі (A) – (C) пройшли навчання протягом 60 епох. На рисунку 1 наведено різні деталі процесу навчання [4]. Для мережі (A), точність даних під час валідації становить близько 63%. Вже через 10 епох точність піднялася вище 60%, що свідчить про можливості швидкого навчання. Крім того, варто зазначити, що коригування розміру фільтра не мало великого впливу на точність, хоча це і не вплинуло на час обробки. Це означає, що швидкі моделі можна реалізувати з досить високою продуктивністю. Важливо зазначити, що значно складніша мережа (B), також швидко вчиться, але з точністю до 54%. Таким чином видно, що зменшення розміру мережі сутєво знизили показники точності оригінальної мережі (B). Разом із значно вищими вимогами до потужності обчислювальних ресурсів, а отже, і повільнішими показниками роботи, ця модель не є достатнім викликом двох інших архітектур. Мережа (C) показує дещо повільнішу криву навчання, але показники точності розпізнавання емоцій, під час тестування на наборі валідацій, аналогічні відповідним показникам мережі (A). Вимоги до обробки вхідних зображень не такі жорсткі як у мережі (B), але жорсткіші за мережу (A). Тому виходячи з цього факту, мережа (A) здається найбільш перспективною для нашого завдання розпізнавання емоцій. Однак продуктивність мережі (C) на додатковому тестовому наборі RaFD суттєво краща (60%), ніж у мережі (A) (50%). Це свідчить про кращі узагальнюючі можливості, що дуже важливо для розробки на її основі нашого програмного додатку.

Отже, враховуючи вище проаналізоване, остання мережа (C) виявилася для нас найперспективнішою для практичних застосувань.

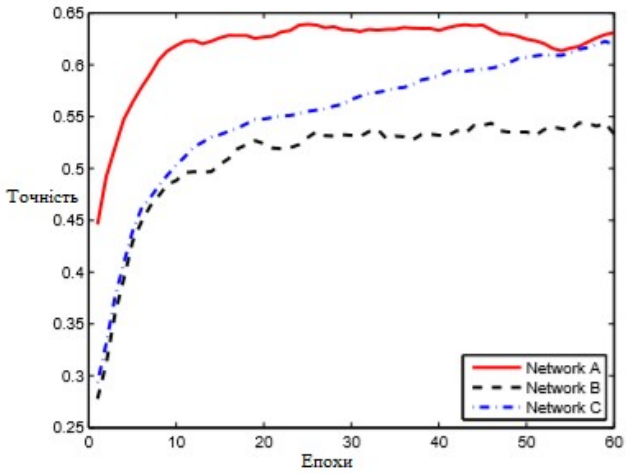

Рисунок 1 – Порівняння точності нейронних мереж (A) – (C)

Практичні дослідження

На основі наведеної архітектури згорткової нейронної мережі (С) розроблено програмний продукт на базі мови програмування Python, бібліотек NumPy, OpenCV, TensorFlow (Keras).

Основний модуль цієї прогами передбачає виконання таких етапів.

1. У кожному кадрі відеоптоку веб-камери методом каскаду Хаару [10] виділяється зображення людини.
2. Область зображення, що містить обличчя, зменшується до 48x48 і передається на вхід згорткової нейронної мережі.
3. На виході нейронної мережі формується список балів для семи емоцій.
4. Визначається назва емоції, що отримала максимальний бал.

Приклад результату роботи розробленого програмного продукту наведено на рис. 2

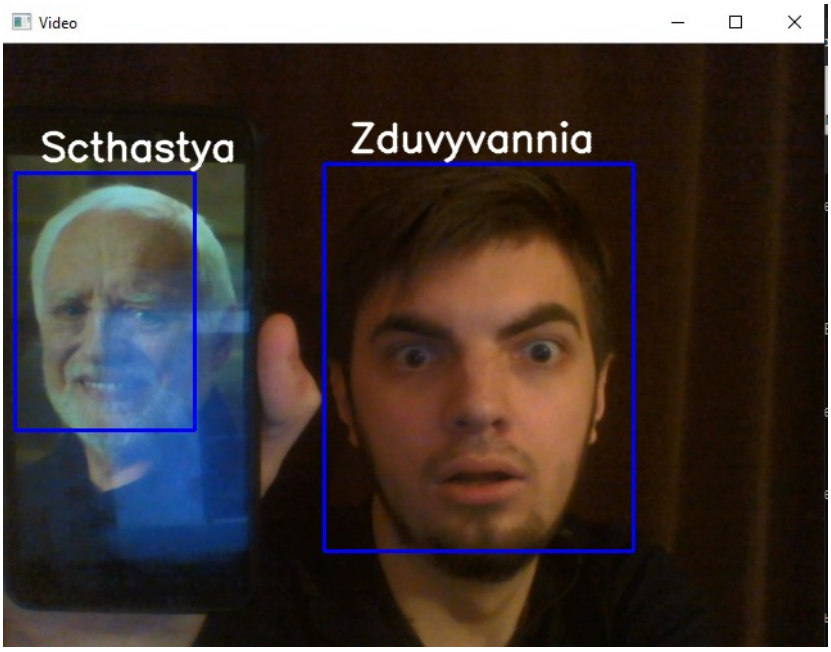

Рисунок 2 – Результат роботи програмного продукту

Для дослідження отриманих результатів, було розроблено модуль візуалізації історії тестування нейронної мережі у процесі навчання. У процесі навчання і автоматичного тестування після кожної епохи зберігається результат тестування. Таким чином, аби відобразити залежність кількості правильних результатів розпізнавання емоцій (точності, Accuracy) відносно до кількості епох (Epoch), які зменшують швидкодію побудовано графік Model Accuracy (рис.3). Тестування здійснювалося на двох наборах даних: відформатований для навчання тестовий набір із приблизно 29000 зображень певного формату (train) і набір із більше 7000 фотографій довільного формату (val). На графіку Model Loss (рис. 3) відображена залежність помилкового розпізнання від кількості епох на тих самих наборах даних.

За допомогою загорткової нейронної мережі (С) точність тестування в процесі навчання досягала близько 90%, а в процесі валідації на невідформатованих зображеннях різної роздільної здатності досягала приблизно 64% за 50 епох. Отже результат розробки довів доцільність використання згорткової нейронної мережі (С) із запропонованим застосуванням додаткового проміжного згорткового шару в процесі розпізнання емоцій людини.

Висновки

Проаналізовано основні варіанти архітектур згорткових нейронних мереж та доведено доцільність використання глибокої згорткової нейронної мережі для розв’язання задачі розпізнання емоцій.

На відміну від існуючих архітектурних рішень згорткових нейронних мереж для розпізнавання емоцій запропоновано застосувати додатковий проміжний згортковий шар, який дозволяє знизити вимоги до потужності обчислювальних ресурсів без вагомих втрат точності розпізнавання емоцій.

На основі запропонованого рішення розроблено програмний продукт для розпізнавання емоцій людини, що продемонстрував точність 90% на відформатованому наборі даних, та близько 64 % точності у складних випадках (невідформатовані зображення довільної роздільної здатності).

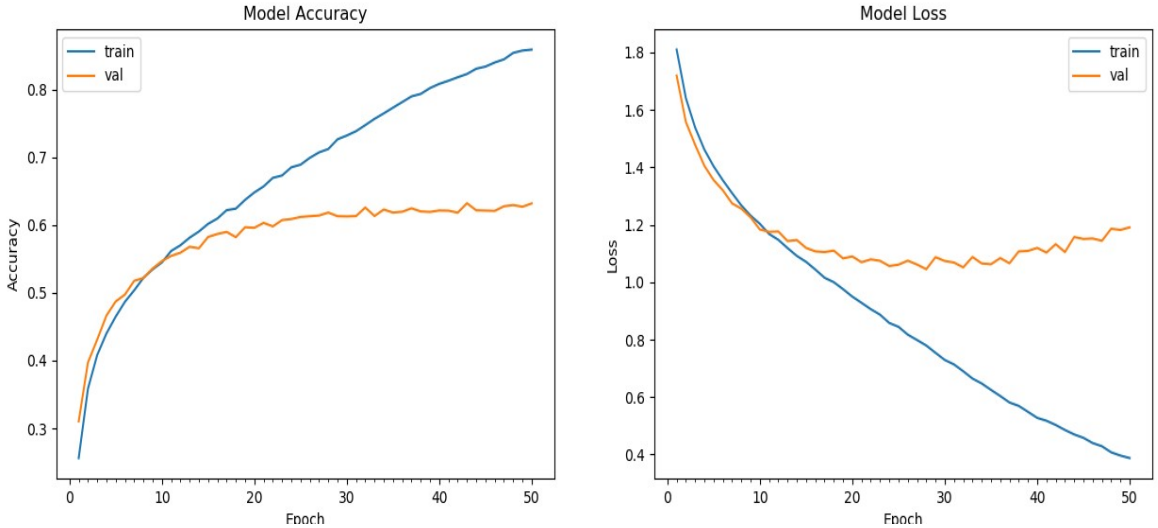

Рисунок 3 – Результат тестування та валідації запропонованої нейронної мережі

Список використаних джерел:

1. Зінов'єв Є., Арсенюк І. Дослідження методів розпізнавання емоцій за допомогою нейронних мереж // Матеріали XLIX науково-технічної конференції підрозділів ВНТУ. Вінниця, 2020. URL: https://ir.lib.vntu.edu.ua/bitstream/handle/123456789/29505/8981.pdf?sequence=3

2. Y. Lv, Z. Feng, and C. Xu. Facial expression recognition via deep learning. In Smart Computing (SMARTCOMP), 2014 International Conference on, pages 303–308. IEEE, 2014.

3. T. Ahsan, T. Jabid, and U.-P. Chong. Facial expression recognition using local transitional pattern on gabor filtered facial images. IETE Technical Review, 30(1):47–52, 2013.

4. A. Gudi. Recognizing semantic features in faces using deep learning. arXiv preprint arXiv:1512.00743, 2015.

5. Яровий А. А. Розпізнавання мімічних мікровиразів обличчя людини на основі Time Delay Neural Network / Яровий А. А., Кашубін С. Г., Кулик О. О., Липкань І. М. // Вісник Хмельницького національного університету. Технічні науки. – 2015. – № 1. – С. 122 – 126.

6. Гудфеллоу Я., Бенджио И., Курвилль А. Г93 Глубокое обучение / пер. с анг. А. А. Слинкина. – 2-е изд., испр. – М.: ДМК Пресс, 2018. – 652 с.: цв. ил.

7. A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images, 2009.

8. A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.

9. P. Ekman and W. V. Friesen. Constants across cultures in the face and emotion. Journal of personality and social psychology, 17(2):124, 1971

10. Viola P., Jones M. Rapid Object Detection using a Boosted Cascade of Simple Features // 2013 IEEE Conference on Computer Vision and Pattern Recognition. 2001. Vol. 01. 511 p

11. Інтелектуальна система нейромережевого розпізнавання мімічних мікровиразів обличчя людини / Кашубін С., Яровий А.: Збірник праць IX Міжнародної науково-практичної конференції [Інтернет — Освіта — Наука (ІОН-2014)], (Вінниця, 14 –17 жовтня 2014 р.) – Вінниця, ВНТУ, 2014. – с. 60 –62.