УДК 004.93'1 **doi: 10.20998/2522-9052.2019.4.11**

О. В. Гороховатський, О. О. Передрій

Харківський національний економічний університет імені Семена Кузнеця, Харків, Україна

АНСАМБЛЬ ДРІБНИХ ЗГОРТКОВИХ НЕЙРОННИХ МЕРЕЖ ДЛЯ КЛАСИФІКАЦІЇ СТАТІ ЛЮДИНИ У ВІДЕОПОТОЦІ

Анотація. Предметом досліджень є нейромережеві моделі класифікації статі особи на зображенні обличчя при обробці відеопотоку. **Метою** є дослідження ефективності окремих дрібних згорткових мереж та ансамблів, що створені з них, для вирішення задачі класифікації статі людини у відеопотоці, що обробляється як послідовність окремих фреймів. **Завданнями є** розробка математичних моделей обробки послідовностей фреймів із накопичуванням за різними стратегіями, дослідження їх ефективності при вирішення задачі класифікації, компіляція ансамблів дрібних згорткових нейронних мереж. Застосовуваними **методами** є: моделі нейронних мереж, інтелектуальний аналіз даних, математична статистика, функціональний аналіз, комп'ютерне моделювання. Отримані **результати**: показано, що точність класифікації може бути підвищена як за рахунок використання різних моделей голосування результатів окремих фреймів, так і за рахунок використання ансамблів неглибоких загорткових нейронних мереж. Незначні апаратні та програмні ресурси, необхідні для їх навчання та використання, дають можливість підвищити швидкість класифікації в декілька разів порівняно із результатами класифікації нейронними мережами, що мають складнішу архітектуру. **Висновки. Наукова новизна** полягає у створенні ансамблів неглибоких нейронних мереж, загальне рішення в яких приймається після узагальнення різними методами голосування з довірою як результатів класифікації окремих фреймів, так і результатів класифікації одного і того ж фрейму різними мережами, що дає можливість підвищити надійність та швидкість класифікації. **Практична значущість** роботи полягає у створенні метода, що дає можливість зберегти прийнятний рівень точності класифікації та значно пришвидшити процес класифікації за рахунок використання неглибоких архітектур нейронних мереж.

Ключові слова: ансамбль; неглибокі нейронні мережі; детектування облич; класифікація статі; розпізнавання зображень; згорткові нейронні мережі; голосування із довірою; агрегація; фрейм; відеопотік.

Вступ

Задачі, пов’язані із автоматичним аналізом відеопотоків, набули надзвичайної популярності в останнє десятиліття завдяки синхронному швидкому розвитку різних напрямків та технологій штучного інтелекту та відповідної апаратної складової. Класифікація статі людей у відеопотоці може бути корисною в різних комерційних та рекламних застосунках, збиранням статистики покупців, адаптації об’єктів доповненої реальності тощо. Кількість різних наукових досліджень щодо класифікації статі за зображенням чи відео залишається величезною в останні роки. Більшість з підходів, що існують, використовують умовні два етапи – пошук характерних ознак та їх порівняння. Обидва з них можуть бути реалізовані як із застосуванням штучних нейронних мереж (ШНМ), так і без них.

Згорткові нейронні мережі (Convolutional Neural Networks, CNN) є сучасним напрямком у вирішенні різних проблем аналізу зображень, таких як аналіз, класифікація, розпізнавання тощо. Було запропоновано багато різних архітектур CNN [1, 2, 3], але більш розповсюдженим є використання попередньо відомих та навчених реалізацій мереж типу VGG [4, 5]. Глибокі архітектури нейронних мереж і CNN з десятків шарів вимагають налаштувань багатьох параметрів, а також значну кількість часу для навчання, тому кількість робіт про компактні (дрібні) архітектури нейронних мереж невпинно зростає [6, 7-9] останнім часом. Доцільною видається побудова компактних та водночас якісних ШНМ, які можуть використовуватися на звичайному домашньому персональному комп’ютері. Приклади реалізацій подібних дрібних мереж можна знайти в [10, 11, 12]. Наприклад, у [12] запропоновано і оцінено CNN (названу GenderCNN), що складається лише з трьох згорткових та двох повнозв'язних шарів. Використання ансамблю дрібних CNN з усередненням їхніх виходів для прийняття спільного рішення розглянуто в [10]. Інші результати досліджень про залежність якості класифікації від глибини CNN, а також залежність ефективності класифікації від розміру вікна, що обмежує область із обличчям, представлені в [11]. Варто зазначити, що більшість робіт описують експериментальне моделювання на відносно малих наборах даних, що не дозволяє зрозуміти стабільність наведених результатів.

Багато досліджень описують в основному параметри архітектури CNN та їх налаштування, але не враховують можливі переваги від аналізу послідовностей кадрів і можливих зв'язків між кадрами. Наприклад, у випадку, коли кожний кадр класифікується окремо, можуть виникати помилки класифікації [13, 14]. Крім того, результати, наведені в роботі [14], показують, що якість класифікації може бути підвищена за рахунок накопичування результатів класифікації окремих кадрів.

Основні ідеї цієї роботи стосуються тренування дрібних CNN і застосування їх окремо та в ансамблях для класифікації статі людини за зображенням обличчя, захопленого з безперервного відеопотоку в реальному часі.

Наукова новизна полягає у створенні ансамблів неглибоких нейронних мереж, загальне рішення в яких приймається після узагальнення різними методами голосування з довірою як результатів класифікації окремих фреймів, так і результатів класифікації одного і того ж фрейму, що дає можливість підвищити надійність та швидкість класифікації.

Дрібні згорткові нейронні мережі

Архітектуру CNN будемо вважати дрібною. якщо її можна вдало навчити за декілька годин і швидко використовувати, обмежуючись лише не більш як десятьма шарами [7]. Структура традиційної CNN зазвичай являє собою послідовність різних типів шарів, що дозволяє застосовувати конкретні операції на кожному етапі обробки початкового зображення. CNN зазвичай включає в себе згортку, максимізаційне агрегування, відкидувальні та повноз'єднані шари, і наші мережі включають всі ці шари.

Оператор згортки в комп'ютерному зорі і в програмах обробки зображень в основному використовується в якості фільтра, що дозволяє отримати характерні особливості зображення. Згортання означає сканування зображення піксель на пікселем за допомогою накладання на окіл пікселя вікна ядра K розміру $k \times k$ і обчислення нових значень згорнутого зображення. Різні значення ядра K дозволяють застосовувати різноманітність специфічних фільтрів, таких як різкість, розмиття, виявлення країв і т.д. Агрегування (максимізаційне або усереднювальне) є популярним методом зменшення розмірності, який залишає лише більш цінні значення ознак.

Повнозв'язні (щільні) шари зазвичай використовуються для узагальнення особливостей після попередніх шарів обробки. Навчання щільних шарів відбувається повільніше порівняно з шарами інших типів, тому видається доцільним використовувати тільки один чи два таких шари в дрібних архітектурах ШНМ. Відкидувальні шари використовуються для запобігання перенавчанню і прискорення процесу навчання, їх ідея полягає в тому, щоб встановити нульові вагові значення для деякої кількості випадкових вхідних нейронів. В нашому випадку ми відкидали половину таких нейронів. Існує багато різних функцій активації, але ми використовували лише два типи. Нейрони у внутрішніх шарах використовують активацію випрямляча згідно з

$$f(x) = \max(0, x),$$

де x – вхід з попереднього нейрона. Таку активацію можна використовувати, щоб зробити навчання швидшим завдяки простому градієнту і дещо ефективнішим через нульову реакцію на негативні входи. Останні повнозв'язні шари використовують сигмовидну активацію

$$f(x) = 1/(1+e^{-x})$$

для отримання значення виходу в діапазоні від 0 до 1.

Розглянемо використання різних архітектур дрібних загорткових мереж (рис. 1), які містять два шари згортки, два агрегувальних і три щільних шари (включно із останнім). Крім того, між ними було використано три відкидувальних шари для зменшення можливості перенавчання. На рис. 1 наведено три різні архітектури нейронних мереж, які відрізняються кількістю фільтрів на першому етапі (для першої мережі було використано 32 фільтри, другої – 40, третьої – 48). Розмір вхідного зображення для двох перших архітектур склав 32x32 пікселя, для останньої – 64x64.

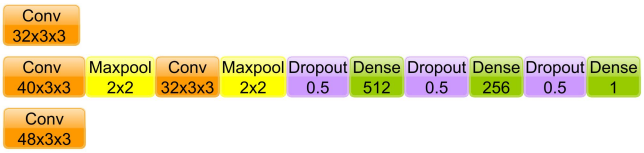

Рис. 1. Архітектури дрібних мереж

Тренування

Набір даних Labeled Face in the Wild [15, 16] (LFW) з глибоким вирівнюванням був використаний для навчання запропонованих дрібних нейронних мереж. Цей набір даних містить 13233 зображення (2966 жіночих і 10267 чоловічих). Цей набір є незбалансованим, тож ми створили його збалансовану модифікацію, яка містить всі 2966 жіночих зображення і таку саму кількість випадкових зображень з обличчями чоловіків, 75% цих зображень використовуються під час навчання і перевірки, а інші 25% використовуються для тестування.

Під час навчання кожної з наведених раніше трьох неглибоких нейронних мереж було виконано 30 ітерацій (кожна з яких складалася з 5 епох), розмір валідаційного набору склав 30% від загального. Лише одне обличчя на кожному зображенні було оброблено, для кожного такого фрагменту зображення з обличчям навколо нього було створено додаткову рамку розміром у 20% від величини фрагменту. Знаходження обличчя було реалізовано з використанням каскадного класифікатора на основі Хаара [17] з коефіцієнтом масштабування 1,3 та 5 мінімальними номерами кожного кандидата з прямокутним обличчям, реалізованого в OpenCV 2.4.13. [18]. Якщо детектор не знайшов обличчя, ціле зображення було використано замість регіону з обличчям.

Точність навчених дрібних мереж було протестовано на наборах даних CelebA [19] і Adience [20]. Важливо відзначити, що запропоновані мережі не були натреновані на цих наборах, а просто перевірені. Під час обробки обох цих наборів даних ми залишили тільки зображення з обличчями (знайдені і обрізані з використанням описаної вище процедури), зокрема, обличчя були знайдені на 182298 зображеннях з 202599 (близько 90%) для набору даних CelebA. 9615 з 19370 (близько 49%) зображень набору даних Adience були протестовані після виключення зображень з дітьми (до 13 років) і невизначеною статтю. Результати класифікації, отримані при використанні цих мереж, наведено в таблиці 1. Не дивлячись на те, що перша мережа обробляє менший початковий розмір зображення та має меншу кількість ядер у першому згортковому шарі, порівняно із третьою, вона дає кращі результати.

Таблиця 1 – **Точність класифікації для дрібних мереж 1-3**

Архітектура	Успішність класифікації		
	LFW (тестовий набір)	CelebA	Adience
1	93,93%	90,91%	79,38%
2	95,08%	91,63%	80,38%
3	91,1%	84,59%	80,58%

Агрегування окремих результатів

Описані вище дрібні моделі CNN для класифікації статі використовуються в режимі онлайн, що означає, що зображення надходять один за одним і обробляються негайно. Передбачається наявність тільки однієї людини в кадрі, а кадри без облич автоматично ігноруються.

Метод виявлення очей на зображенні використовується для запобігання помилкового виявлення облич, оскільки кількість таких помилок може бути значною залежно від відео та значень параметрів детектора облич.

Найпоширенішим підходом у багатьох роботах є класифікація незалежних кадрів відеопотоку, тому цей метод будемо використовувати в якості основного для порівняння.

Надійним способом підвищення точності класифікації є використання голосування або деяке усереднення результатів класифікації, отриманих з різних джерел або при різних умовах. Під голосуванням ми маємо на увазі прийняття рішення, яке базується не тільки на результаті класифікації поточного кадру, але і на результатах останніх N кадрів, де N – константа, яка встановлюється до початку класифікації. Значення з останнього шару CNN використовується як рішення для розбиття між чоловічим і жіночим класами.

Визначимо порогову функцію $f_1(\mathrm{y})$, результати якої агрегуються для створення функції прийняття рішень $D_1(f(\mathrm{y}))$:

$$f_1(\mathrm{y}) = \begin{cases} 1, \mathrm{y} \ge 0.5 \text{ (чоловік)}; \\ 0, \mathrm{y} < 0.5 \text{ (жінка)}; \end{cases}$$

$$D_1(f(\mathrm{y})) = \begin{cases} 1, & \sum_{\mathrm{l}=1}^{N} f_1(\mathrm{y}_i) > N/2 \text{ (чоловік)}; \\ 0 & \text{(жінка)}, \end{cases} \quad (1)$$

де y_i – вихідне значення для кадру i (тут і пізніше ми називаємо i останнім відомим кадром, $i+1$ попереднім і так далі).

Такий підхід є традиційним способом збору інформації з різних джерел і об'єднання в єдине вихідне значення шляхом голосування.

Модель голосування (1) не включає жодної інформації про рівень впевненості результату класифікації, тому розглянемо іншу модель голосування достовірностей для формування значення рішення

$$f_2(\mathrm{y}) = \begin{cases} \mathrm{y} - 0.5, \mathrm{y} \ge 0.5 \text{ (чоловік)}; \\ 0.5 - \mathrm{y}, \mathrm{y} < 0.5 \text{ (жінка)}; \end{cases}$$

$$D(f_2(\mathrm{y})) = \begin{cases} 1, \sum_{\mathrm{l}=1}^{N} f_2^{mal}(\mathrm{y}_i) > \sum_{\mathrm{l}=1}^{N} f_2^{fem}(\mathrm{y}_i) \text{ (чоловік)}; \\ 0 \text{ (жінка)}, \end{cases} \quad (2)$$

де $f_2^{mal}(y)$, $f_2^{fem}(y)$ – окремі функції для чоловічого і жіночого класів відповідно.

Наприклад, якщо значення результату класифікації неглибокої CNN становить 0,49, то правильний клас обличчя на зображенні є жіночим (якщо виходити з того, що мережу натреновано для класифікації реальних жіночих зображень як 0 і чоловічих зображень як 1, природній поріг прийняття рішення становить 0,5), але рівень довіри такої класифікації є низьким і дорівнює 0,5 – 0,49 = 0,01.

Експериментальне моделювання всіх вищевикладених підходів виконувалося на відеопотоці в режимі онлайн, коли рішення відразу ж приймається, коли це можливо. Накопичення значень функціоналів $D_1(f(\mathrm{y}))$ та $D_2(f(\mathrm{y}))$ припиняються і починаються з нуля знову, якщо обличчя втрачено під час обробки, наприклад, зустрівся кадр без обличчя. Після того, як нові значення $f_1(\mathrm{y})$, $f_2^{mal}(y)$, та $f_2^{fem}(y)$ накопичено протягом обробки останніх N кадрів, нове значення додається до накопичувача тільки після видалення найбільш застарілого, тому для прийняття рішення завжди використовуються тільки результати класифікації останніх N кадрів. Якщо поточна кількість накопичених кадрів менше, ніж N, рішення не приймається, що призводить до ігнорування деяких кадрів без класифікації.

Результати експериментальних досліджень

Якість класифікації з використанням (1) і (2) потрібно перевірити на такому наборі відеоданих, який має коректні позначки статі для окремих кадрів або цілих відеофрагментів і є достатньо великим, але відповідний набір знайти важко.

Таким чином, експерименти виконано на зібраному нами наборі 92 фрагментів відео, кожен з яких містить лише одну людину (або більше осіб з однаковою статтю). Ці відео взято з різних онлайн-курсів і інтерв'ю, майже всі вони містять стабільний фон. 41 з цих фрагментів містять відео жінок, 51 – чоловіків.

Етап виявлення очей під час обробки відео з динамічним фоном і різними локаціями здається дуже важливим, тому прямокутну область ми вважатимемо за обличчя, тільки якщо знайдено положення очей у ній. З іншої точки зору, обробка може бути набагато швидше без цієї стадії, якщо фон є сприятливим для обраних параметрів детектора обличчя.

Загальна кількість кадрів у нашому наборі даних склала 1 793 529, 991 312 з яких – у відео із чоловіками, 802 217 – із жінками. 564 238 чоловічих та 450 403 жіночих зображень було вдало ідентифіковано як області з обличчями та очима одночасно.

Кількість відкинутих через агрегування кадрів для цього набору даних становить приблизно 10% для випадків з N = 10 і приблизно 15%, якщо N = 20. Ці значення є трохи вищими для відео із чоловіками, що означає, що відео з жіночими обличчями є більш стабільними і кадри з обличчям втрачаються частіше при детектуванні чоловічих облич.

Результати класифікації зображень осіб з використанням першої з побудованих нейронних мереж наведені в табл. 3.

Таблиця 2 – **Класифікація статі у відеопотоці дрібною мережею із застосуванням першої архітектури**

Метод	Всі зображення	Чоловіки	Жінки
Незалежні кадри	84.94%	92.06%	76.10%
Голосування (N=10)	86.87%	92.78%	79.75%
Голосування (N=20)	87.68%	92.84%	81.55%
Голосування із довірою (N=10)	86.91%	92.83%	79.81%
Голосування із довірою (N=20)	87.72%	92.86%	81.60%

Як можна побачити, процедура голосування дозволяє отримати кращі результати порівняно з класифікацією кожного кадру незалежно у всіх випадках, а випадки з більшою кількістю останніх кадрів N = 20 перевершують відповідні з N = 10. Варто зауважити, що голосування з агрегацією із рівнем довіри згідно з (2) було кращим, ніж просто голосування (1) у всіх випадках.

Час класифікації, необхідний для запропонованих дрібних CNN, можна порівняти із існуючою CNN [21], що відноситься до архітектури wide residual networks [22].

Варто зауважити, що CNN з [21] містить 24 456 656 навчальних параметрів, тоді як третя з наших дрібних архітектур (найпотужніша) – 3 358 561. Обидві ці мережі отримують вхідне зображення розміром 64x64 пікселів, середній час класифікації зображення для WideResNet склав 0,37 секунди, нашої мережі – 0,004 секунди.

Ансамблі мереж

Швидкість класифікації, що забезпечується використанням дрібних CNN, дозволяє будувати мережеві ансамблі, що не тільки робить обробку відеопотоку більш плавною, отримуючи не єдиний результат класифікації кадру, а й допомагає отримати більш впевнений загальний результат класифікації, об'єднаний з різних мереж. Звичайні підходи до агрегації виходів з різних мереж в єдине значення включають голосування, усереднення (чи зважене усереднення), стекування та інші [23-26].

Наша модель ансамблю працює наступним чином. Виходи з кожної з описаних раніше трьох CNN агрегуються згідно з простою процедурою голосування (1) або голосування з довірою (2), так що накопичуються не тільки результати класифікацій незалежних кадрів, але й множинні результати класифікацій одного кадру, отримані з різних дрібних мереж.

Ми додатково вводимо модифікацію голосування (2), яка включає відкидання максимальних та мінімальних результатів класифікації ансамблю, так само, як, наприклад, під час оцінювання спортивних змагань зі стрибків у воду.

Результати класифікації ансамблем дрібних загорткових нейронних мереж наведено в таблиці 3, і вони, безумовно, є кращими за результати класифікації тільки за допомогою однієї мережі, навченої на збалансованому наборі (табл. 2).

Таблиця 3 – **Результати класифікації статі ансамблем з трьох мереж та різними моделями голосування**

Метод	Всі зображення	Чоловіки	Жінки
Незалежні кадри	86.71%	93.13%	78.66%
Голосування (N=10)	89.18%	93.83%	83.49%
Голосування (N=20)	89.61%	93.42%	85.00%
Голосування із довірою (N=10)	89.25%	93.15%	84.48%
Голосування із довірою (N=20)	89.63%	93.27%	85.21%
Голосування із довірою без макс. та мін. (N=10)	89.95%	93.69%	85.46%
Голосування із довірою без макс. та мін. (N=20)	90.59%	93.87%	86.70%

Кількість прийнятих рішень є на 3% і 5% (4% і 7% для голосування без максимальних і мінімальних значень) меншою за початкову кількість кадрів для N = 10 і N = 20 відповідно.

Необхідно відзначити, що голосування без максимального та мінімального значень дозволили отримати найкращі результати.

Висновки

У статті запропоновано підходи до вирішення задачі класифікації статі людини на базі зображення обличчя, захопленого з відеопотоку. Для класифікації використано дрібні моделі згорткових нейронних мереж, ансамбль таких мереж та узагальнення результатів класифікації різних кадрів та навіть одного кадру із використанням різних підходів.

Перевірено точність підготовлених дрібних CNN і отримано близько 87-90% правильних рівнів класифікації з різними архітектурами мереж.

Виконано дослідження голосування та голосування із довірою як способу підвищення точності, просте голосування дозволяє підвищити точність на 4-5%. Класифікація, заснована на голосуванні із довірою, у більшості випадків є дещо кращою (найкраще поліпшення становить близько 1,5%), але є аналогічною до голосування після встановлення порогу.

Використання дрібних CNN може зменшити час класифікації в десятки разів порівняно з більш сучасними глибокими CNN, що робить можливим створення ансамблів нейронних мереж для прийняття остаточного рішення. Використання ансамблю з трьох різних дрібних мереж на базі голосування із довірою дозволило підвищити точність класифікації на 2-5%.

Експерименти, проведені із більшою кількістю послідовних кадрів (N = 20) виявилися кращими за аналогічні, проведені із N = 10. В той самий час, агрегація призводить до пропуску деяких кадрів з обличчями в оригінальному відеопотоці. Кількість таких пропущених кадрів залежить від відеопотоку та методу класифікації (єдина CNN або ансамбль) і знаходилася в діапазоні від 5% до 15% для нашого набору даних.

Переваги дрібних CNN пов'язані з спрощенням їх архітектури, швидкою реалізацією навчання та класифікації, низькими апаратними або програмними вимогами до обладнання чи програмного середовища, зручністю використання для цілей дослідження та / або навчання.

Основним недоліком є обмеження класу задач, які можливо вирішити із їх використанням, як було показано раніше [7].

References

1. Dehghan, A., Ortiz, E.G., Shu, G. and Masood, S.Z. (2017), *DAGER: Deep Age, Gender and Emotion Recognition Using Convolutional Neural Networks*, available at: https://arxiv.org/pdf/1702. 04280.pdf.
2. El Khiyari, H., Wechsler, H. (2016), *Face Verification Subject to Varying (Age, Ethnicity, and Gender) Demographics Using Deep Learning*, DOI: https://doi.org/10.4172/2155-6180.1000323
3. Levi, G. and Hassner, T. (2015), "Age and Gender Classification Using Convolutional Neural Networks", *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, DOI: https://doi.org/10.1109/cvprw.2015.7301352.
4. Rothe, R., Timofte, R. and Gool, L.V. (2015), "Dex: Deep expectation of apparent age from a single image", *Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCVW)*, DOI: https://doi.org/10.1109/iccvw.2015.41.
5. Simonyan, K. and Zisserman, A. (2015), *Very deep convolutional networks for large-scale image recognition*, available at: https://arxiv.org /pdf/1409.1556.pdf.
6. Ekmekji, A. (2016), *Convolutional Neural Networks for Age and Gender Classification*, available at: http://cs231n.stanford.edu/reports/2016/ pdfs/003_Report.pdf.
7. Gorokhovatskyi, O. (2018), "Shallow Convolutional Neural Networks for Pattern Recognition Problems", *Proceedings of the IEEE International Conference on DataStream Mining & Processing*, 23-27 August 2018, Lviv, Ukraine, pp. 459-463, DOI: https://doi.org/10.1109/dsmp.2018.8478540.
8. Hebda, B. and Kryjak, T. (2016), "A compact deep convolutional neural network architecture for video based age and gender estimation", *Proceedings of the Federated Conference on Computer Science and Information Systems*, pp. 787–790.
9. Hogervorst, J., Okafor, E. and Wiering, M. (2017), *Deep Colorization for Facial Gender Recognition*, available at: http://www.ai.rug.nl/ ~mwiering/GROUP/ARTICLES/Facial_Gender_Classification.pdf.
10. Antipov, G., Berrani, S. and Dugelay, J. (2016), "Minimalistic CNN-based ensemble model for gender prediction from face image", *Pattern Recognition Letters*, Vol. 70, Issue C, pp. 59-65, DOI: 10.1016/j.patrec.2015.11.011.
11. Jia, S., Lansdall-Welfare, T. and Cristianin, N. (2016), *Gender Classification by Deep Learning on Millions of Weakly Labelled Images*, available at: http://www.lansdall-welfare.com/wp-content/uploads/2016/11/deep_gender.pdf.
12. Selim, M., Sundararajan, S., Pagani, A. and Stricker, D. (2018), *Image Quality-Aware Deep Networks Ensemble for Efficient Gender Recognition in the Wild*, available at: http://av.dfki.de/ ~pagani/papers/Selim2018_VISAPP.pdf.
13. Bekios-Calfa, J., Buenaposada, J. M. and Baumela, L. (2011), "Revisiting Linear Discriminant Techniques in Gender Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 4, pp. 858-864, DOI: https://doi.org/10.1109/tpami.2010.208.
14. Demirkus, M., Toews, M., Clark, J. J. and Arbel, T. (2010), "Gender classification from unconstrained video sequences", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Workshops*. DOI: https://doi.org/10.1109/cvprw.2010.5543829.
15. Huang, G. B., Mattar, M., Lee, H. and Learned-Miller, E. (2012), "Learning to Align from Scratch", *Advances in Neural Information Processing Systems*, pp. 764-772.
16. Huang, G. B., Ramesh, M., Berg, T. and Learned-Miller, E. (2007), *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*, University of Massachusetts, Amherst, Technical Report 07-49.
17. Viola, P. and Jones, M. (2001), "Rapid object detection using a boosted cascade of simple features", *Proceeding of the International Conference on Computer Vision and Pattern Recognition* , vol. 1, pp. 511-518.
18. OpenCV Open Source Computer Vision, available at: https://docs.opencv.org/master/index.html.
19. Liu, Z., Luo, P., Wang, X. and Tang, X. (2015), "Deep Learning Face Attributes in the Wild", *Proceedings of International Conference on Computer Vision (ICCV)*, DOI: https://doi.org/10.1109/iccv.2015.425.
20. Eidinger, E., Enbar, R. and Hassner, T. (2014), "Age and gender estimation of unfiltered faces", *IEEE Transactions on information forensics and security*, Vol. 9, Issue 12, DOI:.1109/tifs.2014.2359646.
21. *Easy Real time gender age prediction from webcam video with Keras* (2017), available at: https://github.com/Tony607/Keras_age_gender.
22. Zagoruyko, S. and Komodakis, N. (2017), *Wide Residual Networks*, available at: https://arxiv.org/pdf/1605.07146.pdf.
23. Shu, C. and Burn, D. H. (2004), "Artificial neural network ensembles and their application in pooled flood frequency analysis", *Water Resources Research*, Vol. 40, W09301, DOI: https://doi.org/10.1029/2003WR002816.
24. Frazao, X., Alexandre, L. A. (2014), Weighted Convolutional Neural Network Ensemble, available at: https://www.di.ubi.pt/~lfbaa/ pubs/ciarp2014.pdf.
25. Jiménez, D. (1998), "Dynamically Weighted Ensemble Neural Networks for Classification", *Proceedings of the IEEE International Joint Conference on Neural Networks*, DOI: https://doi.org/10.1109/ijcnn.1998.682375.

26. Ju, C., Bibaut, A. and Van der Laan, M.J. (2017), "The Relative Performance of Ensemble Methods with Deep Convolutional Neural Networks for Image Classification", *Journal of Applied Statistics*, 45(15), DOI: https://doi.org/10.1080/02664763.2018.1441383.

Надійшла (received) 29.09.2019
Прийнята до друку (accepted for publication) 20.11.2019

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Гороховатський Олексій Володимирович – кандидат технічних наук, доцент, доцент кафедри інформатики та комп'ютерної техніки, Харківський національний університет радіоелектроніки, Харків, Україна;
Oleksii Gorokhovatskyi – Candidate of Technical Sciences, Associate Professor, Associate Professor of Computer Science and Computer Engineering Department, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine;
e-mail: oleksii.gorokhovatskyi@gmail.com; ORCID ID: http://orcid.org/0000-0003-3477-2132

Передрій Олена Олегівна – кандидат технічних наук, старший викладач кафедри інформатики та комп'ютерної техніки, Харківський національний університет радіоелектроніки, Харків, Україна;
Olena Peredrii – Candidate of Technical Sciences, Senior Lecturer of Computer Science and Computer Engineering Department, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine;
e-mail: elena_peredriy@ukr.net; ORCID ID: http://orcid.org/0000-0003-0390-1931

Ансамбль мелких сверточных нейронных сетей для классификации пола человека в видеопотоке

А. В. Гороховатский, Е. О. Передрий

Аннотация. Предметом исследования являются нейросетевые модели классификации пола человека по изображению лица при обработке видеопотока. **Целью** является исследование эффективности отдельных неглубоких нейронных сетей и созданных из них ансамблей для решения задачи классификации пола человека в видеопотоке, который обрабатывается как последовательность отдельных фреймов. **Задачи:** разработка математических моделей обработки последовательности фреймов с накоплением с использованием разных стратегий, исследование их эффективности для решения задачи классификации, компиляция ансамблей мелких свёрточных нейронных сетей. Применяемые **методы**: модели нейронных сетей, интеллектуальный анализ данных, математическая статистика, функциональный анализ, компьютерное моделирование. Полученные **результаты**: показано, что точность классификации может быть повышена как за счет использования разных моделей голосования результатов отдельных фреймов, так и за счет использования ансамблей неглубоких свёрточных нейронных сетей. Незначительные аппаратные и программные ресурсы, которые требуются для их обучения и использования, дают возможность повысить скорость классификации в несколько раз в сравнении с результатами классификации нейронными сетями более сложной архитектуры. **Выводы. Научная новизна** заключается в создании ансамблей неглубоких нейронных сетей, общее решение в которых принимается после обобщения различными методами голосования с доверием как результатов классификации отдельных фреймов, так и результатов классификации одного и того же фрейма разными сетями, что дает возможность повысить точность и быстродействие классификации. **Практическая значимость** работы состоит в создании метода, который дает возможность обеспечить допустимый уровень точности классификации и значительно повысить быстродействие за счет использования неглубоких архитектур нейронных сетей.

Ключевые слова: ансамбль; неглубокие нейронные сети; определение лиц; классификация пола; распознавание изображений; свёрточные нейронные сети; голосований с доверием; агрегация; фрейм; видеопоток.

Ensemble of shallow convolutional neural networks for classification of gender in video stream

O. Gorokhovatskyi, O. Peredrii

Abstract. Subjects of the research are neural network models for a person's gender classification by the image of a person when processing a video stream. The **goal** is to investigate the effectiveness of individual shallow neural networks and ensembles created from them to solve the problem of classifying a person's gender in a video stream, which is processed as a sequence of individual frames. **Tasks** include the development of mathematical models to process a sequence of frames with accumulation using different strategies, investigation of their effectiveness for solving the classification problem, compiling ensembles of shallow convolutional neural networks. Following **methods** are used: neural networks modeling, data mining, mathematical statistics, functional analysis, computer modeling. **Results** follows: it is shown that the classification accuracy can be improved both through the use of different voting models of the individual frames classification results, and through the use of ensembles of shallow convolutional neural networks. The insignificant hardware and software resources that are required for their training and use make it possible to increase the classification speed by several times in comparison with the results of classification by neural networks, that have more complex architecture. **Conclusions. The contribution is in** the creation of ensembles of shallow neural networks, the general decision in which is made after the generalization by various voting methods with confidence both the classification results of individual frames and the classification results of the same frame by different networks, which makes it possible to increase the accuracy and speed of classification. The **practical significance** of the work is in the creation of a method that makes it possible to provide an acceptable classification accuracy and significantly improve performance by using shallow neural network architectures.

Keywords: ensemble; shallow neural networks; face detection; gender classification; image recognition; convolutional neural networks; trusted voting; aggregation; frame; video stream.