УДК 004.08
DOI https://doi.org/10.32838/2663-5941/2021.6/22

Сєнін Ю.І.
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»

ЗГОРТКОВІ НЕЙРОННІ МЕРЕЖІ ДЛЯ ВИЗНАЧЕННЯ АВТОРСТВА ТЕКСТІВ

У статті представляється модель для визначення авторства коротких текстів за допомогою згорткової нейронної мережі (CNN – convolutional neural network) із використанням n-грам символів. Запропонована стратегія, яка покращує інтерпретацію моделі, оцінюючи важливість фрагментів вхідного тексту в прогнозованій класифікації. Штучна нейронна мережа класифікує тексти авторів з використанням набору лексичних дескрипторів і з прямим використанням за допомогою зворотного розповсюдження n-грам.

Використано методи багатофакторного машинного навчання задля визначення авторства, досліджено класифікатор атрибуції з використанням текстів, які мали невеликий обсяг і твітів у соціальних мережах, таких як Twitter, Facebook, Instagram.

Проведено дослідження, яке дає змогу визначити найбільш ймовірного автора статті, новин чи повідомлень. Ідентифікацію авторства можна застосувати до таких завдань, як ідентифікація анонімного автора, виявлення плагіату або пошук письменника-привида. У дослідженні вирішували цю проблему на різних рівнях, із різними моделями глибокого навчання й на різних наборах даних. Серед усіх моделей, які протестовано, GRU на рівні статті досягає найкращого результату: точність 69,1% для набору даних C50 і 89,2% для набору даних Gutenberg.

У результаті проведених досліджень у статті виявлено, що CNN забезпечує кращу продуктивність для тисяч коротких текстів, а використання символьних n-грам, замість простих послідовностей символів, може підвищити продуктивність. Зі спостережень можна визначити, що мережа більше фокусується на деяких розділах тексту, аніж на цілому тексті.

За дослідженням, експериментальна оцінка відображає, що модель згорткової нейронної мережі з використанням n-грам символів є конкурентоспроможною і здатна перевершити попередні методи.

Ключові слова: *класифікація авторів, короткі тексти, нейронна мережа, згорткова нейронна мережа, символи n-грам, аналіз основних компонентів.*

Постановка проблеми. Проблема атрибуції авторства завжди була складнішою для коротких текстів порівняно з довгими текстами. Проте в сучасному світі, де більшість людських взаємодій відбувається в Інтернеті й у короткі терміни, використання коротких текстів стає все більш актуальним, особливо в таких галузях, як фішингові листи, спам і спільні проекти з масовими джерелами, такі як Вікіпедія. З появою соціальних мереж можна стверджувати, що створення систем, які працюють із короткими текстами є важливішим, аніж систем, які працюють з довгими текстами, такі як книги, наприклад. Ця потреба також відображається в інтересі, що зростає, до визначення авторства невеликих текстів, таких як твіти й ті самі наукові публікації.

Аналіз останніх досліджень і публікацій. На момент написання статті не знайдено жодної попередньої роботи, у якій успішно застосовувалися символьні n-грами з використанням CNN, ні яких-небудь методів CNN, що мали справу з визначенням авторства коротких текстів. Тим не менше знайдено дослідження в галузі атрибуції авторства коротких текстів, що використовують як традиційні, так і суміжні підходи. Символи й n-грами слів використовувалися як ядра багатьох систем атрибуції авторства.

Символи та слова n-грами допомагають визначити автора документа, фіксуючи синтаксис і стиль автора. Розглядаючи підходи до глибокого навчання, знайшли ще одну роботу, у якій CNN використовується для визначення авторства. Однак вони використовують словесні подання текстів, а не символьне подання для коротких текстів. Крім того, у праці [1] використовується багатоголова рекурентна нейронна мережа (RNN), модель мови символів, яка дає набір імовірностей наступного символу для кожного автора на кожному кроці моделі. Це була найбільш ефективна система для завдання ідентифікації автора PAN 2015 з макроусередненої площі під кривою (AUC) 0,628 [1]. Незважаючи на багатообіцяючі результати, які

показують CNN і RNN, результати не піддаються інтерпретації, лише деякі із робіт намагаються проаналізувати, чого насправді навчаються мережі.

Постановка завдання. Метою статті є огляд вирішення проблеми визначення авторства коротких текстів, пропонування архітектури нейронної мережі, яка здатна вивчати подання тексту з послідовності символів. Запропонована архітектура – це CNN, який використовує послідовність n-грамів символів як вхідні дані. Це контрастує з традиційним підходом до CNN, згідно з яким використовується послідовність слів або послідовність символів.

Виклад основного матеріалу дослідження. N-грам – згорткові нейронні мережі. Запропонована архітектура отримує послідовність n-грам символів як вхідних даних. Ці n-грами потім обробляються трьома модулями: модулем вставки символів, згортковим модулем і підключеним модулем softmax. Модуль вбудовування символів оснований на успіху інших розподілених векторних представлень, таких як вбудовування слів. Цей модуль вивчає безперервне, нерозбірливе d-виміряння векторного представлення n-грам символів. Максимальна довжина l навчальних послідовностей визначає розмір вхідного сигналу, а вхідні дані, коротші за l, доповнюються. Цей модуль дає матрицю $C \in Rd \times l$, де стовпці знаходяться в позиції j. вкладення n-грами c_j

Наступний компонент – згортковий модуль. До частини першої застосовується фільтр згортки, $C H \in Rd \times w$, де w – ширина фільтра. Отримана матриця O використовується як вхідні дані для сигмоподібної функції g разом із терміном зміщення b для створення представлень *функцій* f для тексту [2]

$$O=H \cdot C[i:i+w-1]$$

$$f=q(H \cdot C[i:i+w-1]+b), f \in Rl-w+1.$$

Як видно з рисунка 1, використовується згортковий шар з різною шириною w, що дає змогу фіксувати шаблони, включаючи все, від морфем до слів. Потім буде об'єднано отримані карти об'єктів f шляхом максимального об'єднання в часі, щоб отримати y_k, максимальне значення кожної карти об'єктів:

$$fk : yk = maxfk[i], k = 1...m,$$

де m – кількість карт об'єктів.

Це дає змогу представляти текст з його найбільш важливими ознаками, незалежно від їх позиції. Після об'єднання об'єктів y_k отримуємо компактне представлення тексту. Нарешті, це подання передається через повністю підключений модуль, що містить softmax шар. Моделі навчання основані на нейронних мережах, які намагаються знайти функції, корисні для автоматичного вирішення проблеми навчання. У разі визначення автора тексту стилістичні особливості можуть бути виявлені на морфологічному, лексичному та синтаксичному рівнях. Припускається, що запропонована модель здатна автоматично фіксувати

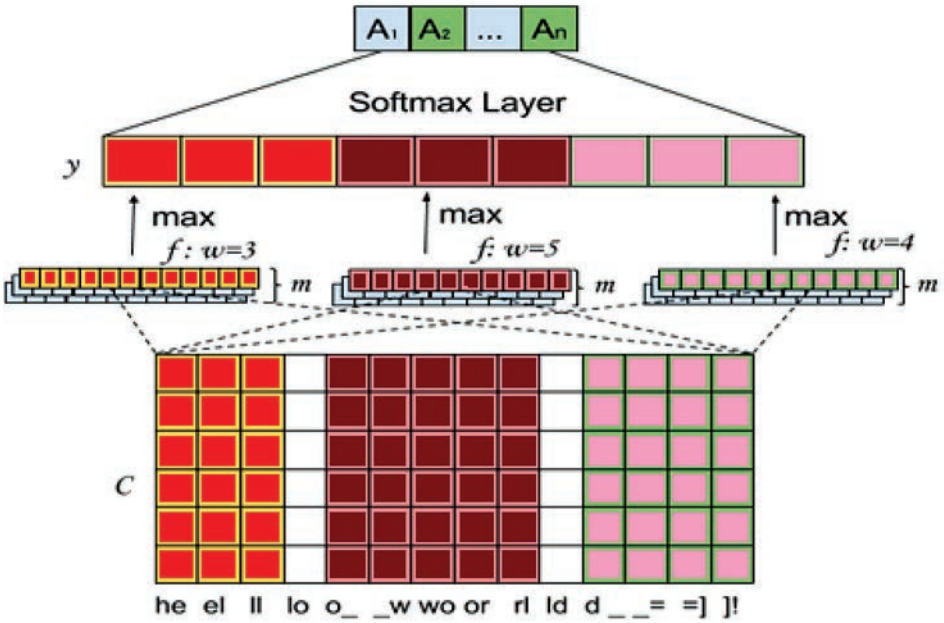

Рис. 1. N-грам CNN. Вкладення N-грам передаються в згорткові й максимальні шари об'єднання, а остаточна класифікація виконується з допомогою softmax шару, застосовуваного до остаточного текстового поданням

шаблони на всіх цих рівнях, починаючи з коротких послідовностей символів, а потім використовуючи згортку для створення уявлень для більш довгих послідовностей.

Деталі реалізації.

Таблиця 1

Гіперпараметри архітектури нейронної мережі

Шар	Шари	Гіперпараметри	
Шари вбудовування	2	*l* *d*	170 330
Згорток	4	*m w* Об'єднання в пул	[600, 600, 600, 600,] [2,3,4,3,] Максимальне об'єднання
Повністю підключених	2	Одиниці	Залежить від автора

Таблиця 1 містить комбінацію гіперпараметрів для трьох модулів, які генерують кращий результат перевірки. Крім того, додано шар відсіву з 25%, що фільтрується після першого шару вбудовування для регуляризації. Потім змішуємо та групуємо зразки в міні-партії розміром 32 байти для пришвидшеного навчання, використовуючи адаптивну оцінку моменту зі швидкістю навчання $1e-4$ для навчання нейронної мережі. Здійснюємо тренування моделі на 200 точках і вибираємо з найменшою помилкою перевірки.

Експериментальна оцінка.

Таблиця 2

Точність для 50 авторів із 1000 коротких текстів кожен

CNN-2	CNN-1	SCH	SCH	LSTM-2	CNN-W
0,761	0,757	0,712	0,703	0,645	0,548

На основі набору даних [2] оцінено запропонований підхід, який містить 10000 користувачів Facebook, кожен із яких має до 500 коротких текстів, використовуючи однакові розбиття. Навчено окремі моделі CNN з n-грамами символів (n = 1,2,3,2,3) на невеликому наборі перевірки. Тут оцінюємо дві наші найбільш ефективні моделі, одну на юніграмах (CNN-1) та іншу на біграмах (CNN-2). Порівняння з трьома іншими системами наведено нижче.

Усі системи використовують перехресну перевірку в процесі навчання для налаштування гіперпараметрів. Спочатку було проведено експеримент з відносно невеликим набором авторів – із 25 авторів, і їх 100 короткими текстами кожен. Результати наведено в таблиці 2. Результати показують, що наша модель CNN [3] bigram (CNN-2) дуже добре працює із цим набором даних і перевершує систему SCH майже на 5%. CNN-1 також перевершує метод SCH, але трохи гірше, ніж CNN-2, показуючи, що є сенс навчити модель CNN n-грам, а не тільки на окремі символи.

Таблиця 3

Порівняння точності для збільшення числа авторів зі 100 коротких текстів на одного автора

Автори	CNN-1	CNN-2	CHAR	SCH	LSTM-2	CNN-W
100	0,504	0,506	0,435	0,415	0,328	0,246
200	0,487	0,476	0,413	0,419	0,337	0,204
500	0,425	0,447	0,357	0,345	0,299	0,163
1000	0,361	0,356	0,313	0,292	0,247	0,126

Також проведено дослідження, як запропонований метод працює порівняно з іншими методами, коли проблема стає більш складною, тобто коли число авторів збільшується або коли кількість коротких текстів на одного автора зменшується, як це зроблено в Schwartz et al. Результати для збільшення числа авторів наведено в таблиці 3. Обидві представлені моделі CNN працюють досить добре порівняно з іншими методами для всіх експериментів, хоча точність знижується зі збільшенням числа авторів, навіть за наявності 1000 авторів наша модель досягає точності значно вище за 36%, є поліпшення на 6% порівняно із сучасним (SCH).

Таблиця 4

Порівняння точності для зменшення кількості твітів одного автора для 50 різних авторів

Кількість твітів	CNN-2	CNN-1	SCH	CHAR	LSTM-2	CNN-W
500	0,734	0,727	0,671	0,633	0,546	0,509
200	0,655	0,645	0,623	0,547	0,512	0,460
100	0,623	0,627	0,554	0,537	0,465	0,427
50	0,532	0,565	0,521	0,421	0,325	0,364

Можемо зробити аналогічні висновки з результатів, у яких зменшуємо кількість коротких текстів на одного автора, як показано в таблиці 4. Після роботи в SCH ці результати є середніми значеннями точності, отриманими з 10 непересічних наборів даних. Продуктивність системи досить стабільна, навіть якщо кількість коротких текстів на одного автора невелика. Різниця в поліпшенні трохи збільшується по мірі того, як ми наближаємося до меншої кількості твітів. Статистичний t-тест результатів по 20 непересічних наборах даних показує, що відмінності між CNN-2 і CHAR, LSTM-2 і CNN-W статистично

значущі при $p < 0{,}001.001$. Проте з результатами SCH виконання тестування стало неможливе, так як результати окремих зв'язаних наборів даних не повідомляються. В обох таблицях можемо побачити, що модель CNN-2 перевершує за продуктивністю модель CNN-1 для експериментів з великою кількістю точок даних (авторів і/або твітів вище немає), це може бути пов'язано з тим, що CNN-2 має більшу кількість параметрів для навчання. CNN-W працює гірше, ніж інші системи. Вхідні дані на основі символів спеціалізуються на стилістичних моделях, тоді як словесні моделі фокусуються на пов'язанні з утриманням шаблонів, які менш важливі для визначення авторства. Цей висновок узгоджується з попередніми дослідженнями в галузі визначення авторства [4].

Що фіксує CNN? Незважаючи на конкурентоспроможну ефективність методів нейронного вистави в кілька NLP, існує недостатнє розуміння того, що саме вивчають ці моделі або як параметри співвідносяться з вхідними даними. Деякі емпіричні дослідження намагалися зрозуміти роль компонентів RNN. Щоб проаналізувати, що робить навчання нейронних уявлень, які підходять для визначення авторства, розглянемо найбільш помітні розділи одного вхідного короткого тексту.

Обрано двох авторів та одну роботу, щоб проаналізувати, які шаблони вивчаються для конкретних авторів. Представлено два короткі тексти від автора бота. Чим темніший відтінок, тим помітніший цей розділ твіту в рішенні про атрибуції. Цей автоматичний бот, схоже, вартий назви шаблону: *URL*, звичайно ж, він виявлений моделлю CNN-2, про що свідчить темне затінення в кінці обох твітів. Аналогічно, на рисунку 3 показано два твіти від людини-автора. Відразу можемо помітити, що в цього автора є бажання використовувати *uhm* і що цей розділ виділений на рисунку. Автор також схильний використовувати послідовні точки, це теж виділено [5]. На рисунку 4 показано значення значущості для твіти з моделей CNN-2 (вгорі) і CNN-1 (усередині).

Нижня фігура затінена з використанням ознак із логістичної регресії для CHAR.

Для моделі CNN-1, хоча виділяються *uhm*, значення значущості більш розподілені по всьому твіту, виділяючи навіть *«є»* і *«боляче»*. Водночас ми бачимо, що модель CNN-2 фокусується саме на *uhm*, що є дуже характерним стилем цього автора. На рисунку 4 наведена s-образна діаграма для моделі CHAR внизу, яку створено за допомогою вагомих ознак із класифікатора логістичної регресії [6]. Хоча більше уваги приділяється *uhm* частини *uhm*, знову ж таки, розподіл більш поширений і для цієї моделі порівняно з моделлю CNN-2.

N-грами з найбільшим внеском. Деякі n-gram програми активують кілька фільтрів, але генерують низькі значення активації, водночас інші n-грами генерують більш високі значення активації, але тільки для декількох фільтрів. Обидва типи містять важливі підказки в розумінні запропонованої моделі. Використаємо проміжне представлення фільтрів CNN, що складається в матриці $O \in R^{n*m}$, де n – кількість n-грамів. Спочатку визначте n-грами, які генерують найбільші значення активації, агреговані по всіх фільтрах. Можемо відзначити, що багато найвищих біграм є незвичайними варіантами символів, таких як *(:*, *: p* і *;D*), які, імовірно, пов'язані з конкретними авторами. Для авторів роботів *[U]* має найвищу продуктивність, так як більшість автоматичних твітів мають URL-адреси в кінці якісної характеристики.

Наступним кроком буде збір n-грамів, у яких найбільша кількість фільтрів, де їх продуктивність входить у топ-3. Можемо побачити, що біграми стають біграмами з афіксами. Можемо пояснити цей факт важливістю морфологічних особливостей для характеристики людських твітів.

Висновки. У статті представлено стратегію використання CNN із символьними n-грамами для визначення авторства коротких текстів і порівняння зі стандартними підходами. Виявлено, що CNN забезпечує кращу продуктивність для визначення авторства твору, а використання символьних n-грам може підвищити продуктивність порівняно з використанням простої послідовностей символів. Протягом усього процесу дослідження проаналізовано, що мережа більше фоку-

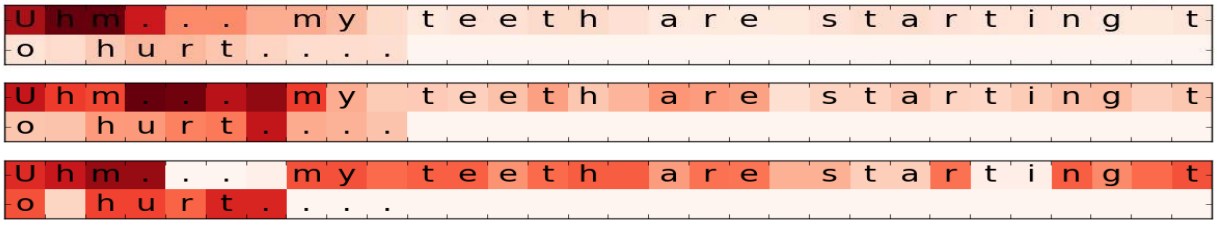

Рис. 2. Порівняння опуклих перерізів CNN-2 (вгорі) і CNN-1 (середина)

сується на деяких розділах тексту, аніж на цілому тексті. Це створює передумови для застосування моделей уваги. Дослідження, які проводитимуться в майбутньому, будуть спрямовуватися на покращення алгоритмів і реалізацію запропонованого програмного забезпечення.

Список літератури:

1. Бэгвелл Д. Ідентифікація автора з використанням багатоголових рекурентних нейронних мереж. *Робочі примітки до документів Лабораторії оцінки CLEF.* 2015. Том 1391.
2. Stamatatos, 2009; Koppel and Winter Koppel and Schler, 2004.
3. Lstm: Пошукова космічна одіссея / К. Грефф, Рупеш К. Шривастава, Я. Кутник, Бас Р. Штюнебринкі Ю. Шмидхубер. *Транзакції IEEE по нейронних мереж і систем навчання*. 2016. С. 99.
4. Обробка природної мови (майже) з нуля / Р. Колберт, Д. Уинстон, Л. Ботту, М. Карлен, К. Кавукчуоглуі П. Кукса. 2011. *Журнал досліджень машинного навчання*. 2011. № 12 (серпень). С. 2493–2537.
5. Юн Ким. 2014 рік. Згорткові нейронні мережі для класифікації речень. *Матеріали Конференції 2014 року з емпіричних методів обробки природної мови (EMNLP)*. 2014. С. 1746–1751.
6. Кингма Д., Ба Дж. Адам: метод стохастичної оптимізації. На *Міжнародна конференція з вивчення уявлень*. 2015.

Sienin Yu.I. CONVOLUTIONAL NEURAL NETWORKS TO DETERMINE THE AUTHORSHIP OF TEXTS

The article presents a model for determining the authorship of short texts using convolutional neural networks (CNN) with n-grams of characters. A strategy that improves the interpretation of the model by assessing the importance of input text fragments in the predicted classification will also be presented. The artificial neural network proposes to classify the authors' texts Heusing a set of lexical descriptors and the neural network with direct use by inversely distributing n-grams.

The effect of stylometry was analyzed. This document used the methods of multifactor machine learning of authorship, the attribution classifier was investigated using texts from novels as a data set. An artificial neural network is proposed for the classification of authors' texts using a set of lexical descriptors and a forward neural network using inverse propagation.

A study was conducted to determine the most likely author of articles, news or reports. Authorship identification can be applied to tasks such as identifying an anonymous author, detecting plagiarism, or finding a ghost writer. This project addressed this issue at different levels, with different models of deep learning and on different data sets. Of all the models tested, the article-level GRU achieves the best result: an accuracy of 69.1% for the C50 dataset and 89.2% for the Gutenberg dataset.

Research has shown in a written article that CNN provides better performance for thousands of short texts, and using symbolic n-grams instead of simple sequences of symbols can also increase productivity. From the observations it could be determined that the network focuses more on some sections of the text than on the whole text.

Experimental evaluation shows that the CNN text works competitively and is able to outperform previous methods.

Key words: *classification of authors, short texts, neural network, convolutional neural network, n-gram symbols, analysis of main components.*