Технічні науки

УДК 004.8.032.26

Карпович Артем Валерійович
аспірант факультету кібернетики
Київського національного університету імені Тараса Шевченка
Карпович Артем Валерьевич
аспирант факультета кибернетики
Киевского национального университета имени Тараса Шевченко
Karpovych Artem
Post-Graduate Student of the Department of Cybernetics of
Taras Shevchenko National University of Kyiv

ВИКОРИСТАННЯ ЗГОРТКОВИХ НЕЙРОННИХ МЕРЕЖ ДЛЯ ЗАДАЧІ КЛАСИФІКАЦІЇ ТЕКСТІВ
ИСПОЛЬЗОВАНИЕ СВЕРТОЧНЫХ НЕЙРОННЫХ СЕТЕЙ ДЛЯ ЗАДАЧИ КЛАССИФИКАЦИИ ТЕКСТОВ
USE OF CONVOLUTIONAL NEURAL NETWORKS FOR THE TASK OF CLASSIFYING TEXTS

Анотація. *Згорткові нейронні мережі - потужний інструмент машинного навчання, який націлений на ефективне розпізнавання і класифікацію зображень. Успіх застосування згорткових нейронних мереж для зображень породив безліч спроб використання цього інструменту в інших завданнях. У даній роботі досліджено основні методи використання згортальних нейронних мереж для задачі класифікації текстів. Виконано експерименти на текстових даних великого обсягу, що показали, що згорткові нейронні мережі для задачі класифікації текстів дозволяють досягти якості,*

аналогічної або кращого в порівнянні з традиційними методами.

Ключові слова: *нейронна мережа, згортка, класифікація.*

Аннотация. *Сверточные нейронные сети - мощный инструмент машинного обучения, который нацелен на эффективное распознавание и классификацию изображений. Успех применения сверточных нейронных сетей для изображений породил множество попыток использования этого инструмента в других заданиях. В данной работе исследованы основные методы использования сверточных нейронных сетей для задачи классификации текстов. Выполненные эксперименты на текстовых данных большого объема, показали, что сверточные нейронные сети для задачи классификации текстов позволяют достичь качества, аналогично или лучше по сравнению с традиционными методами.*

Ключевые слова: *нейронная сеть, свертка, классификация.*

Summary. *Convolutional neural networks are a powerful tool of machine learning, which is aimed at efficient recognition and classification of images. The success of using convolutional neural networks for images has given rise to many attempts to use this tool in other problems. In this paper, we study the basic methods of using convolutional neural networks for the task of classifying texts. Experiments were performed on large-scale text data, which showed that convolutional neural networks for a word classification problem can achieve a quality similar to or better than traditional methods.*

Key words: *neural network, convolution, classification.*

Завдання класифікації текстів стає все більш актуальною в зв'язку з постійно зростаючим обсягом інформації в інтернеті і потребою в ній

орієнтуватися. Наприклад, класифікація текстів необхідна для вирішення наступних завдань:

- Боротьба зі спамом.

  Спам - це небажані розсилки, які можуть приходити на адресу електронної пошти. Вони можуть містити рекламні пропозиції або комп'ютерні віруси. Завдання боротьби зі спамом полягає в тому, щоб класифікувати всі листи на два класи: спам і не спам.

- Розпізнавання емоційного забарвлення текстів.

  Завдання полягає в тому, щоб оцінити думку автора по відношенню до об'єктів, наприклад на основі відгуків про ці об'єкти. Часто таке завдання необхідно вирішувати для видачі релевантних рекомендацій.

- Поділ сайтів по тематичним каталогам.

  Дане завдання вирішується пошуковими системами і передбачає обробку документів і віднесення їх до однієї з декількох категорій, перелік яких заздалегідь заданий.

- Персоніфікація реклами.

  Контекстна реклама є основним джерелом доходу IT компаній. Вона відображається відвідувачам інтернет-сторінки, сфера інтересів яких потенційно збігається або перетинається з тематикою рекламованого товару або послуги, цільової аудиторії, що підвищує ймовірність їх відгуку на рекламу. Сфера інтересів визначається за текстом інтернет-сторінок переглянутих користувачем.

У зв'язку з важливістю даного завдання, по її вирішенню проводяться безліч змагань по машинному навчанню з цінними призами, досліджуються нові методи для досягнення кращої якості класифікації. У даній роботі розглянемо

основні методи класифікації тексту, а так само посимвольний підхід з використанням згортальних нейронних мереж.

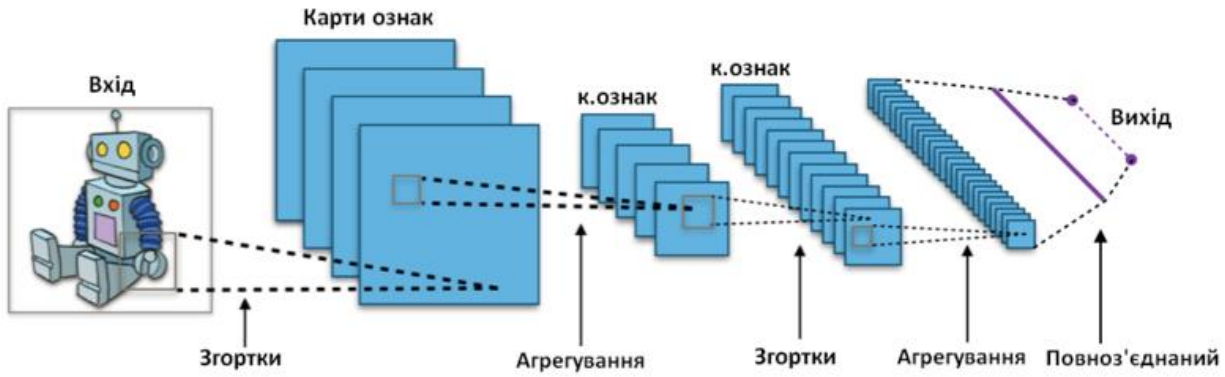
Згорткові нейронні мережі

З появою великих обсягів даних і великих обчислювальних можливостей стали активно використовуватися нейронні мережі. Особливу популярність отримали згорткові нейронні мережі, архітектура яких була запропонована Яном Лекуном [12] і націлена на ефективне розпізнавання зображень. Свою назву архітектура мережі отримала через наявність операції згортки, суть якої в тому, що кожен фрагмент зображення множиться на матрицю (ядро) згортки поелементно, а результат підсумовується і записується в аналогічну позицію вихідного зображення. В архітектуру мережі закладені апріорні знання з предметної області комп'ютерного зору: піксель зображення сильніше пов'язаний з сусіднім (локальна кореляція) і об'єкт на зображенні може зустрітися в будь-якій частині зображення.

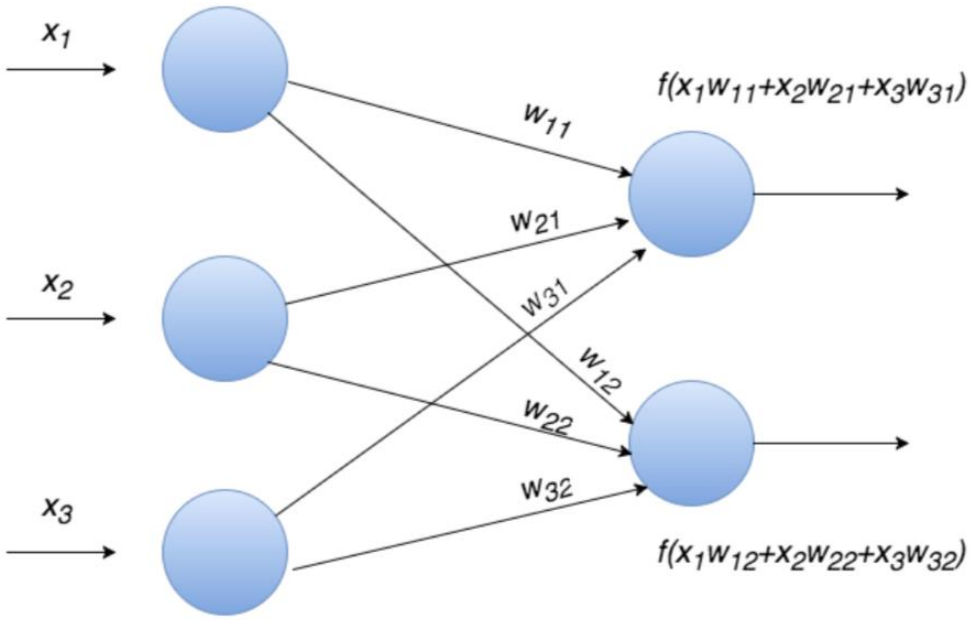
Особливу увагу згорткові нейронні мережі отримали після конкуса ImageNet, який відбувся в жовтні 2012 року і був присвячений класифікації об'єктів на фотографіях. У конкурсі було потрібно розпізнавання образів в 1000 категорій. Переможець цього конкурсу - Алекс Крижевський, використовуючи згорткову нейронну мережу, значно перевершив інших учасників [6]. Успіх застосування згорткових нейронних мереж до класифікації зображень привів до безлічі спроб використовувати даний метод до інших місій. Останнім часом їх стали активно використовуватися для завдання класифікації текстів.

Згорткова нейронна мережа зазвичай являє собою чергування згортальних шарів (convolution layers), агрегувальних шарів (subsampling layers) і при наявності повнозв'язних шарів (fully-connected layer) на виході. Всі три види шарів можуть чергуватися в довільному порядку [12].

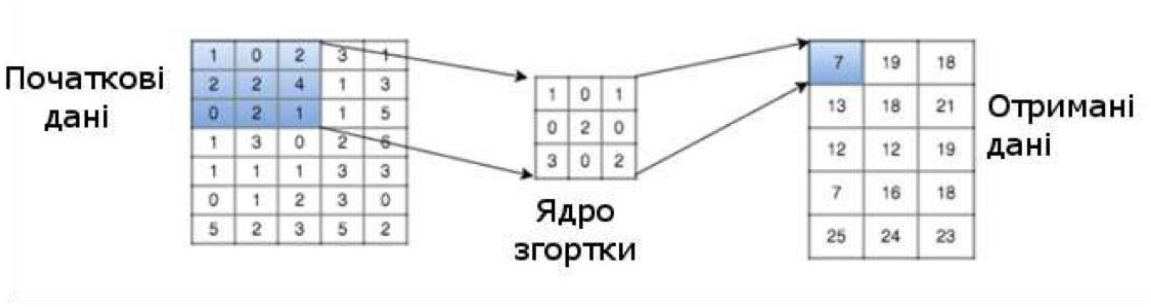
У згорткові шарі нейрони, які використовують одні і ті ж ваги, об'єднуються в карти ознак (feature maps), а кожен нейрон карти ознак пов'язаний з частиною нейронів попереднього шару. При обчисленні мережі виходить, що кожен нейрон виконує згортку деякої області попереднього шару (яка визначається безліччю нейронів, пов'язаних з даними нейроном).



Шар в якому кожен нейрон з'єднаний з усіма нейронами на попередньому рівні, причому кожна зв'язок має свій ваговий коефіцієнт



На відміну від повнозв'язну, в згортковому шарі нейрон з'єднаний лише з обмеженою кількістю нейронів попереднього рівня, згортковий шар аналогічний застосуванню операції згортки, де використовується лише матриця ваг невеликого розміру (ядро згортки), яку «рухають» по всьому оброблюваного шару. Ще одна особливість згорткового шару в тому, що він трохи зменшує зображення за рахунок крайових ефектів.



Агрегувальний шар

Шари цього типу виконують зменшення розмірності (зазвичай в кілька разів). Це можна робити різними способами, але найчастіше використовується метод вибору максимальному елементу (max-pooling) - вся карта ознак поділяється на осередки, з яких вибираються максимальні за значенням.

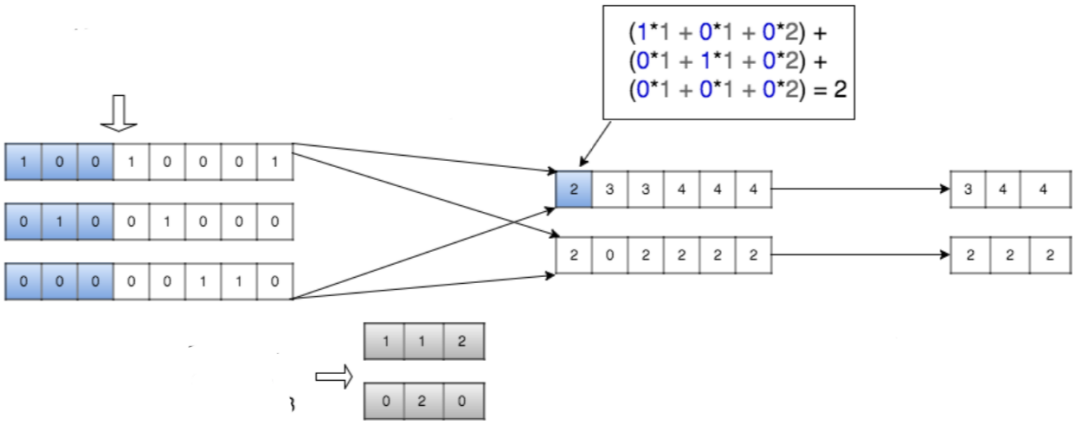
Dropout шар (dropout регуляризація) [14] - спосіб боротьби з перенавчанням в нейронних мережах, навчання яких зазвичай виробляють стохастичним градієнтним спуском, випадково вибираючи деякі об'єкти з вибірки. Dropout регуляризація полягає в зміна структури мережі: кожен

нейрон викидається з певною ймовірністю p. За такою прорідженості мережі проводиться навчання, для для решти ваг робиться градієнтний крок, після чого всі викинуті нейрони повертаються в нейромережу. Таким чином, на кожному кроці стохастичного градієнтного спуску ми налаштовуємо одну з можливих 2N архітектур мережі, де під архітектурою ми розуміємо структуру зв'язків між нейронами, а через N позначаємо сумарне число нейронів. При тестуванні нейромережі нейрони вже не викидаються, але вихід кожного нейрона множиться на (1 - p) - завдяки цьому на виході нейрона ми будемо отримувати маточікування його відповіді по всіх 2N архітектурах. Таким чином, навчену за допомогою dropout-регуляризації нейромережу можна розглядати як результат усереднення 2N мереж.

Посимвольний підхід

Назвемо алфавітом упорядкований набір символів. Нехай обраний алфавіт складається з m символів. Кожен символ алфавіту в тексті закодований з допомогою на гою 1 - m - кодування (кожному символу буде підтверджено вектор довжини m елемент якого дорівнює одиниці, в позиції рівної порядковому номеру символу в алфавіті, а нулю у всіх інших позиціях). Якщо в тексті зустрілися символ, який не увійшов до алфавіту, то необхідно закодувати його вектором довжини m якій складається з одних нулів. З тексту вибираються перші L символів. Параметр L повинен бути великим, щоб в перших L символах містилося достатньо інформації для визначення класу всього тексту. Далі отримані вектори складаються в матрицю розміру m × l, в якій в кожен стовпець матиме не більше однієї одиниці. Кожен рядок отриманої матриці використовується як окрема карта ознак. На вхід згорткової нейронної мережі подається m карт ознак розміру L × L аналогічно зображенню. Архітектуру мережі необхідно вибирати виходячи з завдання. На

Приклад посимвольного підходу для L = 6, m = 3.



Підхід с використанням кодування слів

Підхід був описаний в статті [5]. В даному підході кожному слову в тексті відповідає вектор фіксованої довжини, потім з отриманих векторів для кожного об'єкта вибірки складається матриця, яка аналогічно зображень подається на вхід згорткової нейронної мережі. Для експериментів в статті [5] була реалізована нейронна мережа з одним згортковим, одним агрегувальним і одним повнозв'язним шаром. Дана нейронна мережа використовувалася для класифікації текстів невеликого розміру, що складаються з одного речення.

Методи перекладу слова в вектор фіксованої довжини

One-hot кодування

В даному методі кожне слово кодується за допомогою вектора фіксованої довжини, що дорівнює кількості використовуваних слів в вибірці. Кожен вектор складається з нулів і однієї одиниці.

Word2vec

Робота цієї технології здійснюється наступним чином: word2vec приймає великий текстовий корпус в якості вхідних даних і зіставляє кожному слову вектор, видаючи координати слів на виході. Спочатку він створює

словник, «навчаючись» на вхідних текстових даних, а потім обчислює векторне подання слів. Векторне подання ґрунтується на контекстної близькості: слова, що зустрічаються в тексті поруч з однаковими словами (а отже, мають схожий зміст), у векторному поданні матимуть близькі координати векторів-слів. Отримані вектори-слова можуть бути використані для обробки природної мови та машинного навчання.

У word2vec існують два основних алгоритму навчання: CBOW (Continuous Bag of Words) і Skip-gram. CBOW - «безперервний мішок зі словами" модельна архітектура, яка передбачає поточне слово, виходячи з навколишнього його контексту. Архітектура типу Skip-gram діє інакше: вона використовує поточний слово, щоб передбачати навколишні його слова. Користувач word2vec має можливість перемикатися і вибирати між алгоритмами. Порядок слів контексту не впливає на результат ні в одному з цих алгоритмів.

Експерименти проводилися на даних з Ag news. Обсяг навчальної вибірки 120000 об'єктів, обсяг тестової вибірки 7600 об'єктів. Статті необхідно класифікувати на 4 класу - світові, спортивні, бізнес і наукові новини та Amazon Review Full - коментарі з сайту Amazon.com. Обсяг навчальної вибірки 3000000 об'єктів, обсяг тестової вибірки 600000 об'єктів. Тексти необхідно класифікувати на 5 класів - відгуки користувачів від негативного до позитивного за п'ятибальною шкалою

Реалізована згорткова нейронна мережа з посимвольним підходом для класифікації текстів. В даному підході використовувався алфавіт з символу перекладу рядка і наступних 69 символів:

abcdefghijklmnopqrstuvwxyz0123456789 -,;.!? : ' "/ \ | _ @ # $% & *' + - = <> () [] {}

З кожного об'єкта обрані перші 1014 символів і далі тільки вони враховували ються при класифікації. Цим символом переводяться в матрицю розміру 70 × 1014, а потім подаються на вхід згорткової нейронної мережі. Всі букви англійського алфавіту в тексті наводяться до нижнього регістру. Ваги нейронної мережі започатковано з нормального розподілу N (0, 0.05).

Експериментальні результати		
Data	Iter	Accuracy
Ag news	5000	0.829
Amazon Rev.	30000	0.563

Література

1. Damashek, M. Gauging similarity with n-grams: Language-independent categorization of text / Marc Damashek // Science, New Series. — 1995.
2. Efficient estimation of word representations in vector space / Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean // ICLR. — 2013.
3. Harris, Z. Distributional structure / Zellig Harris // Word. — 1954.
4. John Duchi Elad Hazan, Y. S. Adaptive subgradient methods for online learning and stochastic optimization / Yoram Singer John Duchi, Elad Hazan // JMLR. — 2011.
5. Kim, Y. Convolutional neural networks for sentence classification / Yoon Kim // IEMNLP. — 2014. — Sep. — 1746-1751 p.

6. Krizhevsky, A. Imagenet classification with deep convolutional neural networks / Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton // NIPS. — 2012. — 1106-1114 p.
7. Cun, X. Z. Y. Text understanding from scratch / Xiang Zhang Yann LeCun // Computer Science Department. — 2016.
8. McCulloch, W. S. A logical calculus of the ideas immanent in nervous activity / Warren S. McCulloch, Walter Pitts // Springer New York. — 1943.
9. Pennington, J. Glove: Global vectors for word representation / Jeffrey Pennington, Richard Socher, Christopher D // EMNLP. — 2014. — 1532 - 1543 p.
10. S, J. K. A statistical interpretation of term specificity and its application in retrieval / Jones K. S // Journal of Documentation. — 1972.
11. X, R. word2vec parameter learning explained / Rong X. // arXiv:1411.2738. — 2014.
12. Yann LeCun Leon Bottou, Y. B. Gradient-based learning applied to document recognition / Yoshua Bengio Yann LeCun, Leon Bottou, Patrick Haffner // IEEE. — 1998.
13. Zhang, X. Character-level convolutional networks for text classification / Xiang Zhang, Junbo Zhao, Yann LeCun // In Advances in Neural Information Processing Systems. — 2015. — Feb. — 649-657 p.
14. Воронцов, К. В. Курс лекций по машиному обучению / К. В. Воронцов. — 2015.