УДК 004.912

А. О. Концевой, О. В. Бісікало

МОДЕЛІ ГЛИБОКОГО НАВЧАННЯ ДЛЯ ВИРІШЕННЯ ЗАДАЧІ КЛАСИФІКАЦІЇ ТЕКСТОВОЇ ІНФОРМАЦІЇ

Вінницький національний технічний університет, Вінниця

Анотація. Аналіз тексту в цілому є новою галуззю вивчення. Такі галузі, як маркетинг, управління продуктами, наукові дослідження та управління, вже використовують процес аналізу та вилучення інформації з текстових даних. У попередньому дописі ми обговорили технологію класифікації тексту, одну з найважливіших частин аналізу тексту. Класифікація тексту або категоризація тексту - це діяльність по позначенню текстів природною мовою відповідними категоріями із заздалегідь визначеного набору. Якщо говорити непросто, класифікація тексту - це процес вилучення загальних тегів із неструктурованого тексту. Ці загальні теги походять із набору заздалегідь визначених категорій. Класифікація вмісту та продуктів за категоріями допомагає користувачам легко шукати веб-сайт чи програму та переходити до них. Класифікація тексту, також відома як категоризація тексту, є класичною проблемою в обробці природної мови (NLP), метою якої є призначення міток або тегів для текстових одиниць, таких як речення, запити, абзаци та документи. Вона має широкий спектр застосувань, включаючи відповіді на запитання, виявлення спаму, аналіз настроїв, категоризацію новин, класифікацію намірів користувача, модерування вмісту тощо. Текстові дані можуть надходити з різних джерел, включаючи веб-дані, електронні листи, чати, соціальні мережі, квитки, страхові виплати, відгуки користувачів, а також запитання та відповіді від служби підтримки клієнтів. Текст є надзвичайно багатим джерелом інформації. Але витягувати корисні дані з тексту зазвичай складно та займає багато часу через неструктурований характер природно-мовної інформації. Моделі, засновані на глибокому навчанні, перевершили класичні підходи на основі машинного навчання в різних завданнях класифікації текстів, включаючи аналіз настроїв, категоризацію новин, відповіді на запитання та умовивід природної мови. У цій статті проводиться огляд найбільш поширених моделей класифікації текстів на основі глибокого навчання, розроблених за останні роки, проаналізовано їхній технічний внесок, схожість та сильні сторони.

Ключові слова: класифікація тексту, аналіз настроїв, відповіді на запитання, категоризація новин, глибоке навчання, висновок з природної мови, класифікація тем.

Abstract. Text analysis as a whole is a new field of study. Fields such as marketing, product management, research, and management already use the process of analysing and extracting information from textual data. In the previous post, we discussed text classification technology, one of the most important parts of text analysis. Text classification or text categorisation is the activity of labelling texts in natural language with appropriate categories from a predetermined set. To put it bluntly, text classification is the process of extracting generic tags from unstructured text. These generic tags come from a set of predefined categories. Categorising content and products helps users easily find and navigate to a website or app. Text classification, also known as text categorisation, is a classic problem in natural language processing (NLP) that aims to assign labels or tags to text units such as sentences, queries, paragraphs, and documents. It has a wide range of applications, including question answering, spam detection, sentiment analysis, news categorisation, user intent classification, content moderation, and more. Text data can come from a variety of sources, including web data, emails, chats, social media, tickets, insurance claims, user feedback, and customer service questions and answers. The text is an extremely rich source of information. But extracting useful data from text is usually difficult and time-consuming due to the unstructured nature of natural language information. Deep learning based models have surpassed classical machine learning based approaches in various text classification tasks, including sentiment analysis, news categorisation, question answering, and natural language inference. In this paper, we provide a comprehensive review of most widespread deep learning based models for text classification developed in recent years, and discuss their technical contributions, similarities, and strengths.

Key words: text classification, sentiment analysis, question answering, news categorisation, deep learning, natural language inference, topic classification.

DOI: https://doi.org/10.31649/1999-9941-2022-55-3-13-20.

Вступ

Класифікація тексту, також відома як категоризація тексту, є класичною проблемою в обробці природної мови (NLP), метою якої є призначення міток або тегів для текстових одиниць, таких як речення, запити, абзаци та документи. Вона має широкий спектр застосувань, включаючи відповіді на запитання, виявлення спаму, аналіз настроїв, категоризацію новин, класифікацію намірів користувача, модерування вмісту тощо. Текстові дані можуть надходити з різних джерел, включаючи веб-дані, електронні листи, чати, соціальні мережі, квитки, страхові виплати, відгуки користувачів, а також запитання та відповіді від служби підтримки клієнтів. Текст є надзвичайно багатим джерелом інформації. Але витягувати корисні дані з тексту зазвичай складно та займає багато часу через неструктурований характер природно-мовної інформації [1].

Актуальність

Соціальні мережі – це феномен сьогоднішнього дня. Переваги використання соціальних мереж полягають у вільному і швидкому зв'язку з друзями зазвичай у вигляді зручних об'єктів, таких як пости, картинки, відео і тексти. Ще одна особливість – широкі можливості у створення власних мереж: друзів, колег, членів сім'ї.

Агрегування інформації із загальнодоступних профілів дуже корисно для актуальних цілей, наприклад, таких як побудова стратегії маркетингу і виявлення груп осіб, пов'язаних із забороненими організаціями. Аналіз соціальних даних стрімко набирає популярність у всьому світі завдяки появі в 1990-х роках онлайнових сервісів соціальних мереж. З цим пов'язаний феномен соціалізації персональних даних, зокрема стали публічно доступними факти біографії, листування, щоденники, фото-, відео-, аудіоматеріали, нотатки про подорожі тощо.

Отже, соціальні мережі є унікальним джерелом даних про особисте життя та інтереси реальних людей. Це новітнє явище відкриває безпрецедентні можливості для вирішення дослідних і бізнес-задач (багато з яких до цього неможливо було вирішувати ефективно через брак даних), а також створення допоміжних сервісів і додатків для користувачів соціальних мереж. Крім того, таким станом речей обумовлюється підвищений інтерес до збору і аналізу соціальних даних з боку компаній і дослідницьких центрів.

Мета

Мета дослідження полягає в обгрунтуванні суттєвих особливостей вибору оптимальної моделі машинного навчання для розв’язання задачі класифікації текстової інформації

Задачі

1. Проаналізувати актуальні моделі машинного навчання для вирішення задач класифікації текстової інформації.
2. Оцінити їх переваги та надоліки.
3. Визначити базові метрики так критерії для вибору моделі для вирішення конкретнох задачі.

Розв’язання задач

Для зручності аналізу розглянемо відомі моделі, що згруповані в декілька типів на основі їх архітектури:

1. Мережі прямої подачі – розглядають текст як мішок слів.
2. Моделі на основі RNN (повторювані нейронні мережі) – розглядають текст як послідовність слів і призначені для захоплення залежностей у текстових структурах.
3. Моделі на основі CNN (згорткові нейронні мережі) – навчаються розпізнавати шаблони в тексті, такі як ключові фрази тощо.
4. Капсульні мережі вирішують проблему втрати інформації, від якої виникають операції об'єднання згорткових нейронних мереж, і нещодавно застосовувалися для текстової класифікації.
5. Мережі з розширеною пам'яттю поєднують нейронні мережі з формою зовнішньої пам'яті, яку моделі можуть читати і записувати.
6. Графові нейронні мережі призначені для захоплення внутрішніх графічних структур природної мови, наприклад дерева синтаксичного та семантичного розбору.
7. Сіамські нейронні мережі призначені для відповідності тексту, окремий випадок текстової класифікації.
8. Гібридні моделі поєднують повторювані нейронні мережі і згорткові нейронні мережі, щоб охопити локальні та глобальні особливості речень та документів [2].

Мережі прямої подачі є одними з найпростіших моделей глибокого навчання для представлення ті обробки тексту, тим не менш, вони досягли високої точності на багатьох тестах. Ці моделі розглядають текст як мішок слів. Для кожного слова вони вивчають векторне представлення, використовуючи модель вбудовування, таку як word2vec або Glove, беруть векторну суму або середнє значення вкладень як представлення тексту, передають його через одне або кілька шарів відомих як багатошарові персептрони (MLP), а потім виконують класифікацію подання кінцевого шару за допомогою класифікатора, такого як логістична регресія, наївний байєс або SVM. Прикладом цих моделей є глибока усереднена мережа (DAN), архітектура якої показана на рисунку 1.

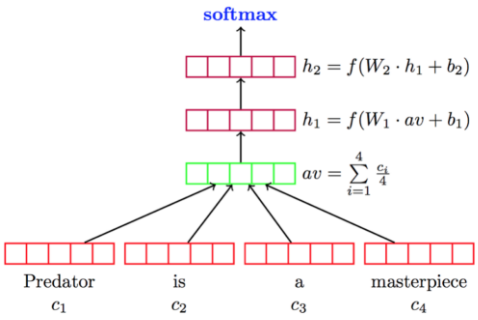

Рисунок 1 − Архітектура глибокої усередненої мережі (DAN)

Незважаючи на свою простоту, глибока усереднена мережа перевершує інші більш складні моделі, які призначені для явного вивчення композиційності текстів. Наприклад, глибока усереднена мережа перевершує синтаксичні моделі на наборах даних з високою синтаксичною дисперсією. Doc2vec викори-

стовує неконтрольований алгоритм для вивчення представлень ознак фіксованої довжини фрагментів текстів змінної довжини, таких як речення, абзаци та документи.

Як показано на рисунку 2, архітектура doc2vec подібна до моделі безперервний мішок слів (CBOW). Єдина відмінність полягає в додатковому маркері абзацу, який зіставляється з вектором абзацу за допомогою матриці D.

В doc2vec конкатенація або середнє значення цього вектора з контекстом із трьох слів використовується для передбачення четвертого слова. Вектор абзацу представляє відсутню інформацію з поточного контексту і може виконувати функцію пам'яті теми абзацу. Після навчання вектор абзацу використовується як ознаки абзацу (наприклад, замість або на додаток до мішку слів) і подається до класифікатора для передбачення. Після публікації Doc2vec досягає нових результатів у виконанні кількох завдань текстової класифікації [3].

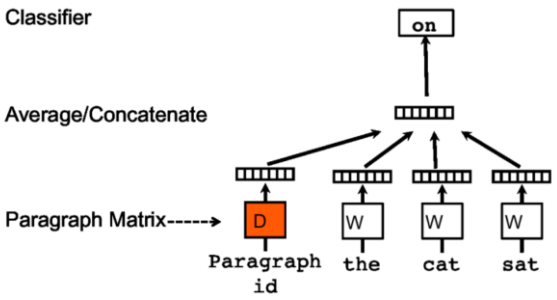

Рисунок 2 − Модель doc2vec

Моделі на основі повторюваних нейронних мереж розглядають текст як послідовність слів і призначені для захоплення залежностей слів і текстових структур для класифікації тексту. Однак звичайні моделі повторюваних нейронних мереж не працюють добре і часто мають низьку продуктивність нейронних мереж із прямим зв'язком. Серед багатьох варіантів повторюваних нейронних мереж найпопулярнішою архітектурою є довготривала пам'ять (LSTM), яка розроблена для кращого захоплення довгострокових залежностей. Довготривала пам'ять вирішує проблеми зникнення або вибуху градієнта, від яких страждають ванільні повторювані нейронні мережі, вводячи комірку пам'яті для запам'ятовування значень через довільні проміжки часу, а також три вентилі (вхідний шлюз, вихідний шлюз, вентиль забуття) для регулювання потоку інформації всередину та з неї. Велись роботи з удосконалення моделей повторюваних нейронних мереж і нейронних мереж з довготривалою пам'яттю для текстової класифікації шляхом захоплення більшої інформації, такої як деревоподібні структури природної мови, відношення між словами в тексті, теми документів тощо.

Варто зазначити, що повторювані нейронні мережі належать до широкої категорії - рекурсивні нейронні мережі. Рекурсивна нейронна мережа застосовує той самий набір ваг рекурсивно до вхідних даних структури для створення структурованого передбачення або векторного представлення на вхідних даних змінного розміру. У той час як повторювані нейронні мережі є рекурсивними нейронними мережами з лінійною ланцюговою структурою вхідних даних, існують рекурсивні нейронні мережі, які працюють на ієрархічних структурах, таких як дерева розбору речень природної мови, об'єднуючи дочірні уявлення в батьківські уявлення. Повторювані нейронні мережі є найпопулярнішими рекурсивними нейронними мережами для текстової класифікації, оскільки вони ефективні та прості у використанні – вони розглядають текст як послідовність маркерів, не вимагаючи додаткових міток структури, таких як дерева аналізу [4].

Моделі на основі повторюваних нейронних мереж навчаються розпізнавати шаблони в часі, тоді як моделі на основі згорткових нейронних мереж вчаться розпізнавати шаблони в просторі. Моделі повторюваних нейронних мереж добре працюють для завдань природомовної обробки тексту, таких як тегування частин мові або QA, де потрібне розуміння дальньої семантики, тоді як моделі на базі згорткових нейронних мереж добре працюють там, де важливо виявляти локальні та інваріантні позиції шаблони. Ці шаблони можуть бути ключовими фразами, які виражають певний настрій, як-от «мені подобається» або тему, як-от «види, що знаходяться під загрозою зникнення». Таким чином, згорткові нейронні мережі моделі стали однією з найпопулярніших архітектур моделей для класифікації текстової інформації.

Одна з перших моделей класифікації текстової інформації на основі згорткових нейронних мереж використовує динамічний k-max-пулінг і називається динамічна модель на основі згорткової нейронної мережі (DCNN).

Як показано на рисунку 3, перший шар динамічної моделі на основі згорткової нейронної мережі створює матрицю речень, використовуючи вбудовування для кожного слова в реченні. Потім згорткова

архітектура, яка чергує широкі згорткові шари з шарами динамічного об’єднання, заданими динамічним k-max-пулінгом, використовується для створення карти ознак над реченням, яка здатна чітко фіксувати короткі та дальні відношення слів і фраз. Параметр об’єднання k можна динамічно вибирати залежно від розміру пропозиції та рівня в ієрархії згортки [5].

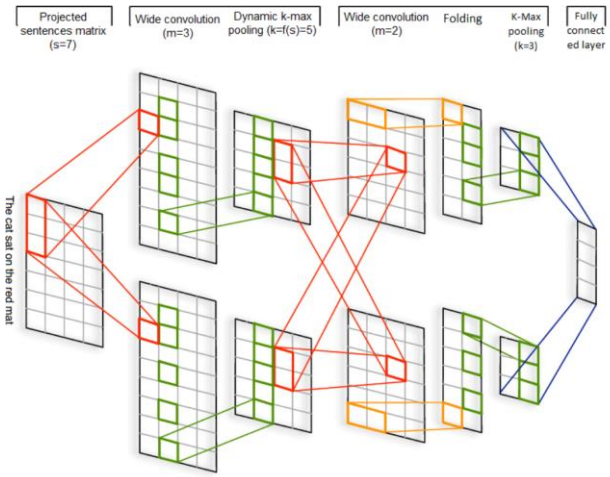

Рисунок 3 − Архітектура динамічної моделі на основі згорткової нейронної мережі

Моделі на основі згорткових нейронних мереж класифікують зображення або тексти, використовуючи послідовні шари згорток і об'єднання. Незважаючи на те, що операції об’єднання визначають основні особливості та зменшують обчислювальну складність операцій згортки, вони втрачають інформацію щодо просторових відносин і, ймовірно, неправильно класифікують об’єкти на основі їх орієнтації або пропорції.

Нещодавно капсульні мережі були застосовані до розв’язання задач класифікації текстів, де капсули адаптовані для представлення речення або документа у вигляді вектора. Модель складається з чотирьох шарів: (1) n-грамовий згортковий шар, (2) шар капсули, (3) згорткий капсульний шар і (4) повністю з’єднаний шар капсули. Автори експериментують із трьома стратегіями для стабілізації процесу динамічної маршрутизації, щоб пом’якшити перешкоди шумових капсул, які містять довідкову інформацію, таку як стоп-слова або слова, які не мають відношення до жодної категорії документів. Вони також досліджують дві архітектури капсул, Capsule-A і Capsule-B, які зображені на рисунку 4. Capsule-A схожа на CapsNet. Capsule-B використовує три паралельні мережі з фільтрами з різними розмірами вікон у згортковому шарі n-грам, щоб дізнатися більш повне представлення тексту [6].

Хоча тексти природною мовою мають послідовний порядок, вони також містять внутрішні структури графів, такі як дерева синтаксичного та семантичного аналізу, які визначають синтаксичні та семантичні відносини між словами в реченнях. Однією з найбільш ранніх моделей побудованих на основі графів що розроблені для природомової обробки тексту, є TextRank. Автори пропонують представити текст природною мовою у вигляді графіка $G\ (V,\ E)$, де V позначає набір вузлів, а E — набір ребер серед вузлів. Залежно від наявних додатків вузли можуть представляти текстові одиниці різних типів, наприклад, слова, словосполучення, цілі речення тощо. Аналогічно, ребра можна використовувати для представлення різних типів відносин між будь-якими вузлами, наприклад, лексичні чи семантичні відносини, контекстне накладання тощо.

Сучасні графічні нейронні мережі (GNN) розробляються шляхом розширення підходів глибокого навчання для даних графів, таких як текстові графіки, які використовує TextRank. Глибокі нейронні мережі, такі як згорткові нейронні мережі, повторюванні нейронні мережі та автокодери були узагальнені протягом останніх кількох років для обробки складних даних графів. Наприклад, двовимірна згортка згорткової нейронної мережі для обробки зображень узагальнюється для виконання згорток графіка, беручи середнє зважене значення інформації про околиці вузла. Серед різних типів нейронних мереж згортки, такі як графові згорткові мережі (GCNs) та їх варіанти, є найпопулярнішими, оскільки вони ефективні та зручні для компонування з іншими нейронними мережами, а також досягли найсучасніших результатів. У багатьох програмах графові згорткові мережі є ефективним варіантом згорткових нейронних мереж на графах.

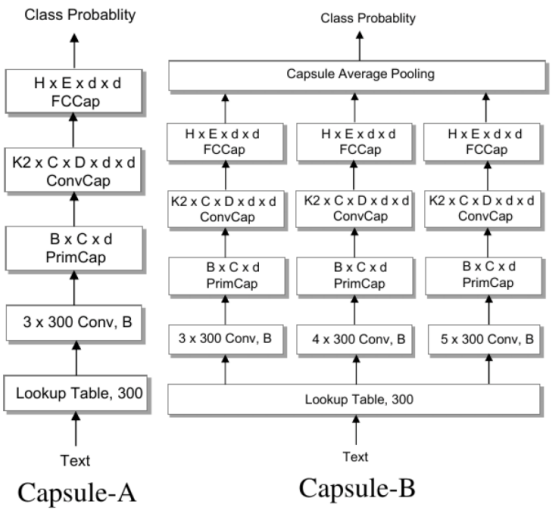

Рисунок 4 − Схема капсул A і B для класифікації текстів

Типовим застосуванням графових згорткових мереж в природомовній обробці тексту є задача класифікації текстової інформації. Графові згорткові мережі використовують взаємозв’язки документів або слів для визначення міток документів.

На рисунку 5 зображена модель, заснована на ієрархічній таксономії та CNN-капсулі з графом уваги. Однією з унікальних особливостей моделі є використання ієрархічних відносин між мітками класів, що у попередніх методах вважаються незалежними.

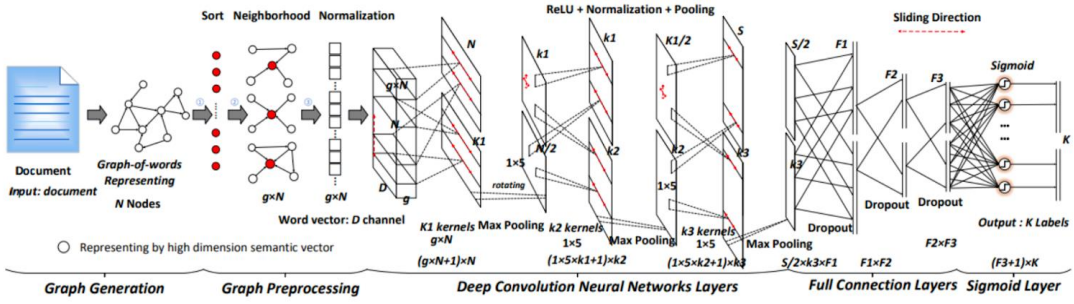

Рисунок 5 − Архітектура моделі на основі ієрархічної таксономії та CNN-капсулі з графом уваги

Розглянемо інший підхід до побудови такої нейронної мережі. Для мережі будується єдиний текстовий граф для корпусу на основі спільного зустрічання слів і взаємовідношень слів у документі, після чого тренується текстова графова нейронна мережа для корпусу, як показано на рисунку 6.

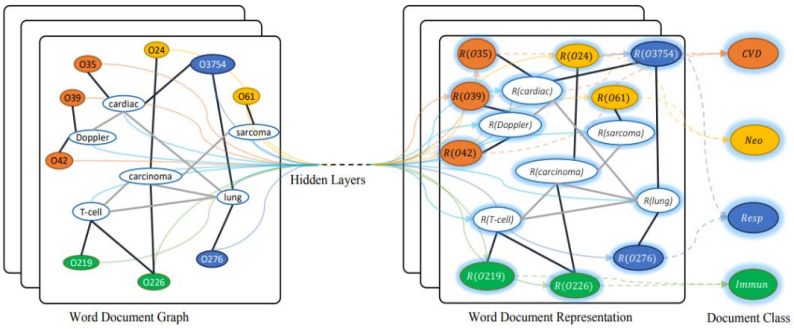

Рисунок 6 − Текстова графова нейронна мережа

Текстова графова нейронна мережа ініціалізується за допомогою однієї гарячої заміни для слова і документа, а потім спільно вивчає вбудовування як для слів, так і для документів, під контролем відомих міток класів для документів. Побудова графової нейронної мережі (GNN) для великомасштабного текстового корпусу коштує дорого. Проводяться дослідження зі зниження вартості моделювання шляхом зменшення складності моделі або зміни стратегії навчання моделі. Прикладом першого є модель проста згортка графів (SGC), де глибока згортка графової нейронної мережі спрощується шляхом багаторазового видалення нелінійності між послідовними шарами та згортання отриманих функцій (матриць ваг) в одну лінійну. трансформація. Прикладом останнього є графова нейронна мережа на текстовому рівні. Замість того, щоб будувати графік для всього текстового корпусу, графова нейронна мережа на рівні тексту створює один графік для кожного фрагмента тексту, визначеного ковзаючим вікном на корпусі тексту, щоб зменшити споживання пам’яті під час навчання. Деякі з інших багатообіцяючих робіт на основі графових нейронних мереж включають GraphSag та контекстуалізовані нелокальні нейронні мережі [7].

Одним із вузьких місць обчислень, яким страждають повторювані нейронні мережі, є послідовна обробка тексту. Незважаючи на те, що згорткові нейронні мережі є менш послідовними, ніж повторювані нейронні мережі, обчислювальні витрати на захоплення зв’язків між словами в реченні також ростуть зі збільшенням довжини речення, подібно до повторюваних нейронних мереж. Трансформаторні моделі долають це обмеження, застосовуючи самоуважність, щоб паралельно обчислювати для кожного слова в реченні або документувати «оцінку уваги», щоб змоделювати вплив кожного слова на інше. Завдяки цій функції, трансформаторні моделі дозволяють набагато більше розпаралелювання, ніж згорткові нейронні мережі і повторювані нейронні мережі, що дає можливість ефективно навчати дуже великі моделі на великих обсягах даних на графічних процесорах.

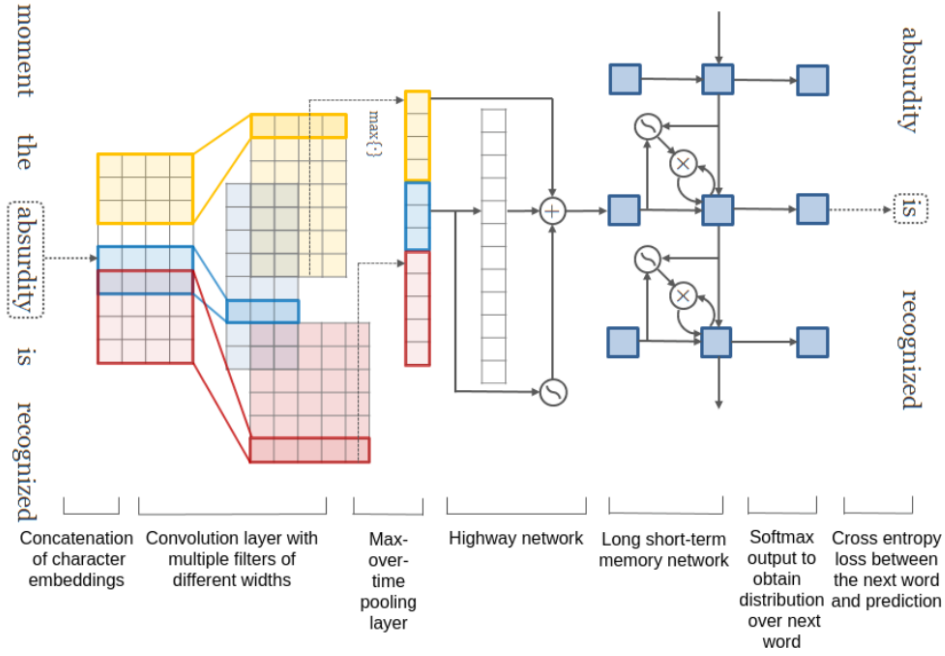

Рисунок 7 − Архітектура нейронної мережі з згортковими нейронними мережами та довготривалою пам’яттю

З 2018 року спостерігається зростання набору великомасштабних попередньо підготовлених мовних моделей (PLM) на основі трансформаторів. У порівнянні з попередніми контекстуалізованими моделями вбудовування, заснованими на згорткових нейроннних мережах, підготовлені мовні моделі на базі трансформацій використовують набагато глибші архітектури мережі (наприклад, 48-шарові трансформатори) і попередньо навчаються на значно більшій кількості тексту. корпусів, щоб вивчити контекстні текстові репрезентації шляхом передбачення слів, обумовлених їх контекстом. Ці підготовлені мовні моделі точно налаштовані за допомогою міток для конкретних завдань і створили новий рівень техніки в багатьох наступних завданнях природомовної обробки текстів, включаючи класифікацію текстової інформації.

Підготовлені мовні моделі можна згрупувати в дві категорії: підготовлені мовні моделі з авторегресією та автокодуванням. Однією з найбільш ранніх підготовлених мовних моделей з авторегресією є OpenGPT, односпрямована модель, яка прогнозує текстову послідовність слово за словом зліва направо (або справа наліво), причому передбачення кожного слова залежить від попередніх передбачень [8].

Для того, щоб обрати модель машинного навчання для вирішення задачі класифікації текстів підхід до вибору сильно змінюється залежно від характеру цільового завдання та домену, доступності міток у домені, затримки та обмежень пропускної здатності програм тощо. Хоча немає жодних сумнівів, що розробка текстового класифікатора є процесом проб і помилок, аналізуючи останні результати загальнодоступних тестів, було вирішено керуватись наступними кроками:

Вибір підготовленої мовної моделі. Використання підготовлених мовних моделей призводить до значних покращень у всіх популярних завданнях класифікації тексту, а підготовлені мовні моделі з автокодуванням часто працюють краще, ніж підготовлені мовні моделі з авторегресією (наприклад, OpenAI GPT).

Адаптація домену. Більшість підготовлених мовних моделей навчаються на текстових корпусах загального домену (наприклад, Web). Якщо цільовий домен різко відрізняється від загального домену, ми можемо розглянути можливість адаптації підготовлених мовних моделей з використанням внутрішньо доменних даних шляхом постійного попереднього навчання вибраної підготовленої мовної моделі загального домену. Для доменів із великою кількістю немаркованого тексту, таких як біомедицина, моделі мови попереднього навчання з нуля також можуть бути хорошим вибором.

Конструкція моделі для конкретного завдання. З урахуванням введеного тексту підготовлена мовна модель створює послідовність векторів у контекстному представленні. Потім один або кілька шарів, що стосуються завдання, додаються зверху, щоб створити кінцевий результат для цільового завдання. Вибір архітектури специфічних для завдання шарів залежить від характеру завдання, наприклад, коли необхідно охопити мовну структуру тексту. Нейронні мережі з прямим зв'язком розглядають текст як мішок слів, повторювані мовні моделі можуть фіксувати порядок слів, згорткові мовні моделі добре розпізнають шаблони, такі як ключові фрази, механізми уваги ефективні для визначення корельованих слів у тексті і графові нейронні мережі можуть бути хорошим вибором, якщо графічні структури природної мови (наприклад, розбір дерев) корисні для цільового завдання.

Точне налаштування для конкретного завдання. Залежно від наявності міток у домені, рівні, що стосуються завдання, можна тренувати окремо з фіксованою підготовленою мовною моделлю або разом із підготовленими мовними моделями. Якщо потрібно створити декілька подібних текстових класифікаторів (наприклад, класифікатори новин для різних доменів), точне налаштування для кількох завдань є хорошим вибором для використання мічених даних подібних доменів.

Стиснення моделі. Підготовлені мовні моделі дорогі в обслуговуванні. Їх часто потрібно стиснути, наприклад, шляхом перегонки знань, щоб задовольнити обмеження затримки та ємності в реальних програмах.

Точність і частота помилок. Це основні показники для оцінки якості моделі класифікації. Нехай TP, FP, TN, FN позначають істинно позитивний, хибнопозитивний, істинно негативний і хибнонегативний відповідно. Точність класифікації та рівень помилок визначені в формулі:

$$\text{Точність} = \frac{TP+TN}{N} \qquad \text{Частота помилок} = \frac{FP+FN}{N}$$

де N − загальна кількість зразків. Очевидно, що Частота помилок = 1 − Точність.

Точність / Відкликання / Оцінка F1. Це також основні показники, і вони використовуються частіше, ніж точність або частота помилок для незбалансованих наборів тестів, наприклад, більшість тестових зразків мають одну мітку класу. Оцінка F1 – це середнє гармонійне значення точності та запам'ятовування і досягає свого найкращого значення при 1 (ідеальна точність і запам'ятовування) і найгіршого при 0.

$$\text{Точність} = \frac{TP}{TP+FP}, \quad \text{Відкликання} = \frac{TP}{TP+FN}, \quad \text{Оцінка } F1 = \frac{2 \cdot \text{Точність} \cdot \text{Відкликання}}{\text{Точність}+\text{Відкликання}}$$

Для задач класифікації з кількома класами ми завжди можемо обчислити точність і відкликання для кожної мітки класу і проаналізувати індивідуальну продуктивність міток класу або усереднювати значення, щоб отримати загальну точність і запам'ятовування.

Точна відповідність (EM). Показник точної відповідності є популярним показником для систем відповідей на запитання, який вимірює відсоток прогнозів, які точно відповідають будь-якій із основних істинних відповідей. EM є однією з основних метрик, що використовуються для SQuAD.

Середній взаємний ранг (MRR). MRR часто використовується для оцінки ефективності алгоритмів ранжирування в задачах НЛП, таких як ранжування запитів-документів і QA. MRR визначається в рівнянні:

$$MRR = \frac{1}{[Q]} \sum_{i=1}^{Q} \frac{1}{rank_i}$$

де Q − набір усіх можливих відповідей, а $ranki$ − рейтингова позиція відповіді, що відповідає дійсності.

Висновки

За останні кілька років класифікація текстової інформації досягла значного прогресу за допомогою моделей глибокого навчання. Було запропоновано кілька нових ідей (таких як нейронне вбудовування, механізм уваги, самоувага, Transformer, BERT і XLNet), які призвели до швидкого прогресу за останнє десятиліття. Незважаючи на наявний прогрес, є ще проблеми, які потрібно вирішити.

Хоча останнім часом було зібрано ряд великомасштабних наборів даних для загальних завдань текстової класифікації, залишається потреба в нових наборах даних для більш складних завдань, таких як класифікація текстів для багатомовних документів і для надзвичайно довгих документів.

Включення загальних знань в моделі глибокого навчання може значно покращити продуктивність моделі, майже так само, як люди використовують загальні знання для виконання різних завдань.

Більшість сучасних нейронних мовних моделей вимагають значного обсягу пам'яті для навчання та висновків. Ці моделі повинні бути стиснуті, щоб відповідати обмеженням обчислень і зберігання граничних додатків. Це можна зробити шляхом побудови моделей за допомогою дистиляції знань, або за допомогою методів стиснення моделей.

Отже, за результатами аналізу найбільш популярних моделей глибокого навчання, які були розроблені за останні шість років, показано, що цей напрям досліджень дозволив суттєво покращити сучасний технологічний рівень та ефективність розв'язання задач класифікації текстової інформації.

References

[1] Bisikalo O. System for definition of indicator characteristics of social networks participants Profiles / Oleg Bisikalo, Anton Kontsevoi // Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2020). – CEUR Workshop Proceedings Volume 2604, 2020. – Lviv, Ukraine, April 23-24, 2020. – Pp. 77-88. – ISSN: 16130073.
[2] I. Goodfellow, Y. Bengio, and A. Courville, Deep learning. MIT press, 2016.
[3] S. Wang and C. D. Manning, "Baselines and bigrams: Simple, good sentiment and topic classification," in Proceedings of the 50th annual meeting of the association for computational linguistics: Short papers-volume 2. Association for Computational Linguistics, 2012.
[4] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in Proceedings of the 2013 conference on empirical methods in natural language processing, 2013.
[5] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in Advances in neural information processing systems, 2015.
[6] W. Zhao, H. Peng, S. Eger, E. Cambria, and M. Yang, "Towards scalable and reliable capsule networks for challenging NLP applications," in ACL, 2019.
[7] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in Advances in neural information processing systems, 2017.
[8] Y. Sun, S. Wang, Y.-K. Li, S. Feng, H. Tian, H. Wu, and H. Wang, "Ernie 2.0: A continual pre-training framework for language understanding." in AAAI, 2020.

Стаття надійшла: 14.06.2022.

Відомості про авторів

Концевой Антон Олександрович − аспірант факультету інтелектуальних інформаційних технологій та автоматизації.

Бісікало Олег Володимирович − доктор технічних наук, професор, факультет інтелектуальних інформаційних технологій та автоматизації.

A. O. Kontsevoi, O. V. Bisikalo

ANALYSIS OF DEEP LEARNING MODELS FOR TEXT INFORMATION CLASSIFICATION TASKS

Vinnytsia National Technical University, Vinnytsia