УДК 004.912

А. О. Концевой

МОДЕЛІ ГЛИБОКОГО НАВЧАННЯ ДЛЯ ВИРІШЕННЯ ЗАДАЧІ КЛАСИФІКАЦІЇ ТЕКСТОВОЇ ІНФОРМАЦІЇ

Вінницький національний технічний університет

Анотація

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, та обговорюємо їхній технічний внесок, схожість та сильні сторони.

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

Abstract

Deep learning based models have surpassed classical machine learning based approaches in various text classification tasks, including sentiment analysis, news categorization, 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.

Keywords: Text Classification, Sentiment Analysis, Question Answering, News Categorization, Deep Learning, Natural Language Inference, Topic Classification.

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. Він 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 інформацію з тексту може бути складно та займати багато часу через його неструктурований характер.

Огляд моделей глибокого навчання

Розглянемо відомі моделі, що згруповані в декілька типів на основі їх архітектури:

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

• Гібридні моделі поєднують повторювані нейронні мережі і згорткові нейронні мережі, щоб охопити локальні та глобальні особливості речень та документів.

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

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.

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 [1].

Моделі на основі повторюванних 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.

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 [2].

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 [3].

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 за останні роки 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.

У цій статті було досліджено найбільш популярні моделі глибокого навчання, які були розроблені за останні шість років і значно покращили сучасний рівень техніки для різних завдань класифікації текстової інформації.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. I. Goodfellow, Y. Bengio, and A. Courville, Deep learning. MIT press, 2016.
2. 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.
3. 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.

Концевой Антон Олександрович **–** аспірант факультету комп'ютерних систем та автоматики, Вінницький національний технічний університет, м. Вінниця

Науковий керівник: ***Бісікало Олег Володимирович*** **–** д-р техн. наук, професор факультету комп'ютерних систем та автоматики, Вінницький національний технічний університет, м. Вінниця

Kontsevoi Anton O. – post-graduate student, Faculty for Computer Systems and Automatic, Vinnytsia National Technical University, Vinnytsia

Supervisor: ***Bisikalo Oleg V.*** **–** Dr.Sc. (Eng.), Professor, Faculty for Computer Systems and Automatic, Vinnytsia National Technical University, Vinnytsia