УДК 004.91

ДОСЛІДЖЕННЯ СУЧАСНИХ МЕТОДІВ АУГМЕНТАЦІЇ ТЕКСТОВИХ ДАНИХ

Абросімов Є. О.
Науковий керівник – к.т.н., доц. Дейнеко А. О.
Харківський національний університет радіоелектроніки, каф. ШІ
м. Харків, Україна
e-mail: yehor.abrosimov@nure.ua

The goal of this work is to explore methods of text data augmentation, which involves creating new synthetic data similar to real ones, for machine learning tasks where available data is limited. Generative data augmentation is used to combat overfitting, but it has found limited application in Natural Language Processing. Simple augmentation methods like random insertions, replacements, and shuffling are too limited in their effectiveness. Substituting n-grams with synonyms is another method that can be used for data augmentation, as well as the application of intelligent models like Back translation and Style augmentation. The use of generative models such as C-BERT is a popular solution for the augmentation task. Prompt engineering is also becoming increasingly popular for creating queries that prompt the model to provide optimal responses.

Ціллю даної роботи є дослідження методів аугментації текстових даних. Аугментація це процес генерації нових синтетичних даних, схожих на реальні, та таких, які є репрезентативними для процесу що породжує реальні дані. Техніки аугментації використовуються в задачах, в яких наявних даних недостатньо.

Для більшості реальних задач машинного навчання (МН) недостатня кількість даних є значною проблемою адже алгоритми МН схильні до перенавчання. Основною ознакою перенавчання є зниження якості передбачення моделі на нових даних при збільшенні складності моделі. Аугментація є одним з основних засобів боротьби із перенавчанням в Computer Vision задачах. Однак, в задачах обробки природної мови застосування відповідних підходів є значно більш обмеженим.

Найпростішою групою підходів є Simple data augmentations що включають в себе випадкові видалення, вставку, заміну та обмін позиціями між словами або реченнями. Вставка та заміна зазвичай відбувається випадковим словом зі словника. Дана група підходів використовується найчастіше, але має дуже обмежену ефективність.

Наступна група евристичних підходів це Mix-up. Для отримання нових прикладів ми випадковим чином перемішуємо токени з різних спостережень, отримуючи таким чином абсолютно нові спостереження. Хоча цей підхід має обмежену ефективність, однак, є доволі популярним.

Заміна синонімами. Ця група підходів передбачає заміну певних слів або словосполучень синонімічними. Синоніми можна отримати двома способами: словники синонімів та семантичні графи знань. Для застосування семантичного графу знань нам необхідно щоб в ньому містилися відношення типу «є еквівалентним», яке ми можемо використовувати як відношення синонімічності.

Досі ми розглядали підходи, що не вимагають моделей МН. Першим підходом що використовує інтелектуальні моделі обробки даних є Back translation. Він є доволі ефективним засобом аугментації. Суть підходу полягає в тому щоб перекласти текст з мови оригіналу на довільну (зазвичай схожої мовної групи) мову, а отриманий результат знову перекласти мовою оригіналу. Цей підхід є доволі доступним і його ефективність основною мірою залежить від специфіки задачі та моделі для перекладу тексту.

Схожим і доволі перспективним підходом є Style augmentation, що дозволяє переносити стиль та емоційне забарвлення на наявні текстові дані. Наприклад, ми можемо перетворити науковий текст в суворо діловому стилі в текст, що своєю формою нагадує відомих літераторів.

Генеративна аугментація, безумовно, є найбільш перспективним фронтиром в контексті задачі що розглядається в даній роботі. Найпопулярніше рішення задачі аугментації генеративними моделями це використання pre-trained моделей, з опціональним донавчанням. Зазвичай донавчання відбувається на supervised learning задачах. В подальшому приховані шари нейронної мережі (НМ) та їх ваги використовуються для створення нової моделі, що виконує задачу трансформації тексту. Доволі популярною, в рамках цього підходу є модель C-BERT (Conditional BERT).

Також великої популярності наразі набирає застосування великих мовних моделей (LLM) для даної задачі. Цей підхід передбачає використання promt-engineering для створення запитів, що змушують модель надавати найбільш оптимальну відповідь.

Список використаних джерел:

1. Shorten C., Khoshgoftaar T. M., Furht B. Text Data Augmentation for Deep Learning. Journal of Big Data. 2021. Т. 8, № 1. URL: https://doi.org/10.1186/s40537-021-00492-0 (дата звернення: 25.02.2024).

2. Wei J., Zou K. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. Protago Labs Research, 2019. 9 p. (Препринт. Dartmouth College). URL: https://arxiv.org/pdf/1901.11196.pdf.