

УДК 004.8

Б. Б. Млинко, к.т.н., доц.; І. М. Климко

(Тернопільський національний технічний університет імені Івана Пулюя, Україна)

ЗАСТОСУВАННЯ МЕТОДІВ ШТУЧНОГО ІНТЕЛЕКТУ ДЛЯ ОБРОБКИ ПРИРОДНОЇ МОВИ ЗАСОБАМИ ГЛИБИННОГО НАВЧАННЯ

UDC 004.8

B. B. Mlynko, Ph.D., Assoc. Prof., I. M. Klymko

APPLICATION OF ARTIFICIAL INTELLIGENCE METHODS FOR NATURAL LANGUAGE PROCESSING USING DEEP LEARNING TECHNIQUES

Генерація тексту на основі методів глибинного навчання є одним із ключових напрямів у сучасній обробці природної мови. Враховуючи зростаючі вимоги до автоматизованого створення текстового контенту, актуальним стає завдання розробки моделей, здатних створювати послідовності символів із врахуванням синтаксичних, семантичних і структурних особливостей тексту. LSTM мережі показують високу ефективність у задачах моделювання послідовностей завдяки здатності запам'ятовувати довгострокові залежності. Однак основними викликами при їх застосуванні залишаються оптимізація навчання, подолання ефекту "забування" контексту та забезпечення балансу між узгодженістю і різноманітністю згенерованого тексту.

Метою дослідження було вдосконалити процес генерації тексту за допомогою LSTM-мереж шляхом впровадження методів оптимізації параметрів, поліпшення обробки навчальних даних та адаптації підходів до семплювання, що забезпечує генерування тексту із реалістичною структурою і змістом.

Для досягнення поставленої мети було розроблено експериментальну LSTM-модель, що працює на рівні символів. На етапі підготовки даних виконано токенизацію тексту на символи та побудову словника, що дозволяє перетворювати текст у послідовності числових індексів. Модель включає шар вбудовування, який знижує розмірність входу, а також LSTM-комірки, що обробляють дані послідовно. Навчання здійснювалося за допомогою алгоритму зворотного поширення похибки через час, при цьому дані поділено на батчі фіксованої довжини для ефективного опрацювання.

У процесі оптимізації параметрів моделі використано стохастичний градієнтний спуск із поступовим зменшенням швидкості навчання, що забезпечує стабільне сходження. Для запобігання перенавчанню та зниження ризику переваги локальних мінімумів застосовано методи регуляризації, такі як обмеження норми ваг та масштабування градієнтів. Генерація тексту здійснювалася на основі семплювання із використанням softmax-розподілу. Цей підхід дозволив регулювати баланс між детермінованістю вибору символу та різноманітністю вихідного тексту.

Результати експериментів продемонстрували, що оптимізація параметрів моделі та впровадження регуляризації дозволили покращити якість згенерованого тексту. Модель виявила здатність підтримувати базові патерни англійської мови, відтворювати структуру речень та зберігати контекст на рівні декількох символів. Під час навчання спостерігалось поступове зниження перплексії, що свідчить про ефективне опрацювання довгострокових залежностей.

Література

1. I. Goodfellow. Deep Learning (Adaptive Computation and Machine Learning series) / I. Goodfellow, Y. Bengio, A. Courville., 2016.