УДК 004.912: 004.93: 004.853

ЗАСТОСУВАННЯ ГЛИБОКОЇ РЕКУРЕНТНОЇ НЕЙРОННОЇ МЕРЕЖІ ІЗ ВИКОРИСТАННЯМ АЛГОРИТМУ LSTM У СИСТЕМАХ ІНТЕЛЕКТУАЛЬНОЇ ВЗАЄМОДІЇ

Яровий Андрій, Кудрявцев Дмитро, Кулик Олександр

Вінницький національний технічний університет

Анотація

В ході проведеного дослідження проаналізовано прикладні аспекти використання різних типів нейронних мереж у галузі інформаційних відносин. Відзначено актуальність використання технології глибокого навчання у сфері обробки текстових масивів змінної довжини. Розроблено прототип системи із використанням нейронної мережі, на прикладі чат-боту.

Abstract

In the given research applies aspects of using different types of neural networks in information relations was analyzed. The actuality of using deep neural network in sphere of processing text arrays with variable length was noted. The prototype of system with using neural network on example of chat-bot was made.

Вступ

Сучасна сфера інформаційних відносин є однією з найбільш розвинених в останній період [1]. Базуючись на даному твердженні, більшість інформації, що оброблюється натепер складається із ланцюгів та послідовностей у вигляді діалогів. Спостерігаючи за популярністю та активністю соціальних мереж та корпоративних мереж [2], можна зробити висновок, що більшість сучасної інформації можливо розбити на кластери і класифікувати за певними ознаками. Наступним етапом, систематизувати у вигляді відповідних тематик та предметних областей, що значно підвищить рівень обробки вхідної інформації. Особливо, це стосується сфери інформаційної підтримки та систем підтримки прийняття рішень. Однією з ключових проблем є вибір засобу для обробки вхідної інформації, у даному випадку типу, архітектури та методу навчання нейронної мережі для задачі класифікації та розпізнання.

Порівняльний аналіз основних типів нейронних мереж

Натепер, значну швидкодію виявляють нейронні мережі глибокого навчання, які дозволяють обробляти інформацію з великою пропускною та результативною здатністю [3]. Для обробки вхідної інформації, на вхід нейронної мережі подається набір сигналів, що характеризуються функціонально повноцінним змістовим значенням для розподілу між усіма нейронами першого шару. При використанні змінної розмірності вхідного набору даних, виникає необхідність застосування нейронної мережі з можливістю повторного використання. Дана вимога формується у разі використання чат-ботів та систем підтримки користувачів у веб-орієнтованих додатках.

У вказаних сферах, інформація являє собою діалоги між користувачем та інтелектуальною інформаційною системою, що представлена для користувача чат-ботом, системою підтримки прийняття рішень чи набором бізнес-правил [4]. При обробленні текстових повідомлень основна увага звертається на швидкодію та точність розпізнання повідомлення. Для цього застосовується алгоритм LSTM (long short-term memory) [5], що використовує два вектори для кодування та декодування вхідних та вихідних наборів даних відповідно. Сутність даного алгоритму полягає у розбитті вхідної послідовності на менші частини, що значно спрощує задачу класифікації [4]. Основний принцип роботи даного алгоритму базується на використанні рекурентної нейронної мережі.

Передумовами створення даного алгоритму стала недостатня ефективність використання технології машинного навчання для типу послідовність-послідовність [4]. Застосовуючи даний алгоритм було досягнуто значного підвищення якості обробки вхідних наборів даних, а також підвищення точності матриці кореляцій [6]. В літературних джерелах описано спробу застосування згорткової нейронної мережі при тестових вибірках із, результатом якої було досягнуто точності розпізнання – 93%. Проте, при цьому загальне навчання нейронної мережі із використанням 8 GPU-систем зайняло 10 діб [6].

Глибокі нейронні мережі (Deep Neural Networks) – достатньо потужний засіб машинного навчання [4, 5], що надає високу продуктивність при розв'язанні складних задач, таких як розпізнання текстів та графічних зображень, що складають значну частину сучасних інформаційних об'єктів [3]. Однією з головних переваг мереж даного типу є їх спроможність до паралелізму [7], що значно пришвидшує оброблення вхідного набору даних. Загальна структура глибокої нейронної мережі, що застосовується при паралельній обробці зображена на рисунку 1.

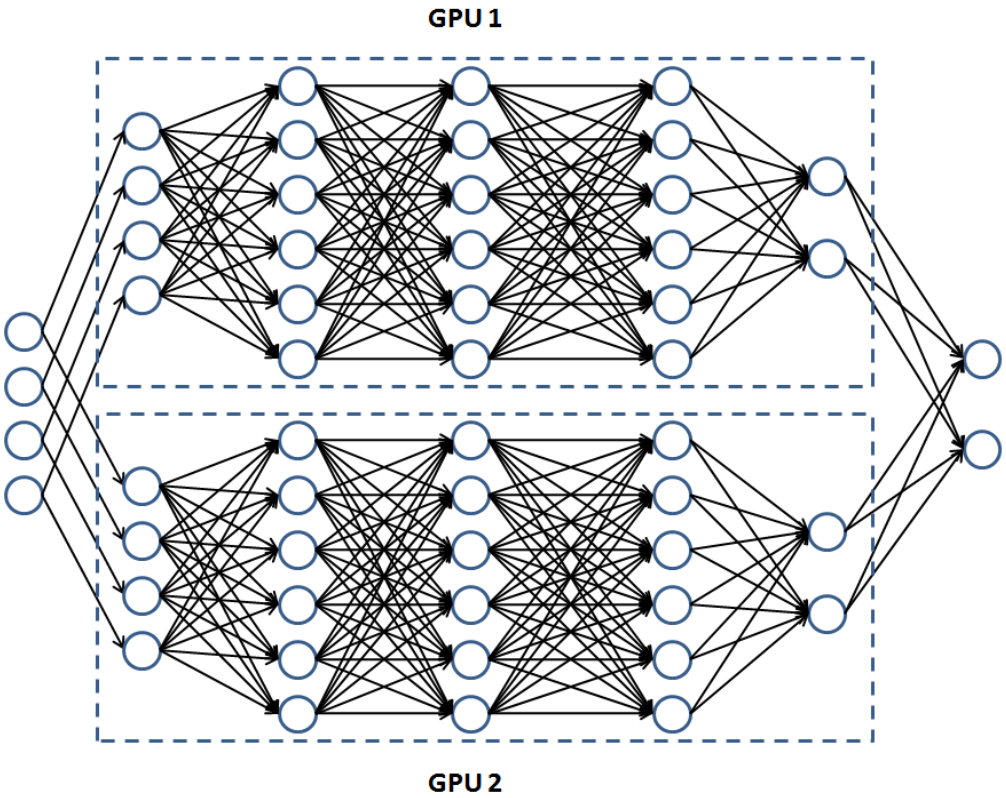

Рисунок 1 – Загальна схема багатошарової глибокої нейронної мережі із використанням двох GPU систем

При застосуванні глибокої рекурентної нейронної мережі виникає складність в задачах розпізнання тексту через змінну довжину вхідної та вихідної послідовностей зі складними немонотонними семантичними зв'язками [8]. Звичайна рекурентна нейронна мережа формує послідовність вхідних сигналів та формує вихідну послідовність за виразами [4]:

$$h_t = \text{sigm}(W^{hx}x_t + W^{hh}h_{t-1}) \quad (1)$$

$$y_t = W^{yh}h_t \quad (2)$$

Даний тип нейронної мережі із використанням алгоритму LSTM є досить перспективним рішенням для задачі розпізнання текстових повідомлень у сфері інформаційних відносин. Різниця у якості розпізнання тексту між двома рекурентними нейронними мережами з використанням алгоритму LSTM та без, склала понад 10% на користь алгоритму LSTM [9, 10].

Приклад використання

Для прикладу було створено чат-бот із використанням глибокої рекурентної нейронної мережі на основі пакетів прикладних бібліотек від Microsoft Cognitive Toolkit [11] та когнітивних сервісів Microsoft Cognitive Services. Також, застосовано фреймворк Microsoft Bot Framework та хмарний сервіс Azure Bot Service. Після тренування нейронної мережі (10^3 речень), на апробаційній вибірці (100 речень) було досягнуто точність розпізнання 94%.

Висновки

В ході проведених досліджень, здійснено порівняльний аналіз різних типів нейронних мереж для задачі розпізнання тексту змінної довжини, зокрема, згорткову, рекурентну та глибоку рекурентну нейронні мережі. Виокремлено проблемні аспекти використання звичайних рекурентних нейронних мереж для вказаної задачі. Відзначено ефективність використання алгоритму LSTM для розпізнання тексту змінної довжини та запропоновано його використання в роботі глибокої рекурентної нейронної мережі. Розроблено чат-бот із використанням когнітивного сервісу Microsoft, експериментальні результати якого підтверджують доцільність та перспективність застосування глибокої рекурентної нейронної мережі із використанням алгоритму LSTM у системах інтелектуальної взаємодії.

Список використаних джерел:

1. Joy Goodman-Deane, Anna Mieczakowski, Daniel Johnson and more. The impact of communication technologies on life and relationship satisfaction / Joy Goodman-Deane, Anna Mieczakowski, Daniel Johnson and more. – Engineering Design Centre, University of Cambridge, vol 57, pp. 219-229, April 2016.
2. C. Canali, M. Colajanni, R. Lancelotti. Data Acquisition in Social Networks: Issues and Proposals / C. Canali, M. Colajanni, R. Lancelotti – Department of Information Engineering, University of Modena and Reggio Emilia, pp 1-12, June 2011.
3. Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic and more. A Network-based End-to-End Trainable Task-oriented Dialogue System / Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic and more – Engineering Department, University of Cambridge, pp 1-12, 2016.
4. Ilya Sutskever, Oriol Vinyals and Quoc V. Le. Sequence to Sequence Learning with Neural Networks / Ilya Sutskever, Oriol Vinyals and Quoc V. Le, Advances in neural information processing systems, pp 3104-3112, 2014.
5. K. Yao and G. Zweig. Sequence-to-Sequence Neural Net Models for Grapheme-to-Phoneme Conversion, INTERSPEECH, pp 1-5, 2015.
6. Ross Girshick. Fast R-CNN, International Conference on Computer Vision, 2015.
7. А. Яровий, С. Кашубін і О. Кулик, Розпізнавання мімічних мікровиразів обличчя людини на основі комбінування Time Delay Neural Network та Deep Belief Network, Оптико-електронні інформаційно-енергетичні технології, vol 29, no 1, pp 76-83, Лип 2015.
8. Ramesh, Nallapati, Zhou, Bowen, Gulcehre, Caglar, Xiang, Bing et al. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In Proc. of EMNLP, 2016.
9. Dahl, G. E., Yu, D., Deng, L & Acero, A. Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition. IEEE Trans. Audio Speech Lang. Process, vol 20, pp 33-42, 2012.
10. Graves, Alex. Generating Sequences With Recurrent Neural Networks. Technical report, arXiv preprint arXiv:1308.0850, p 43, 2013.
11. Cho, K. Et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In Proc. Conference on Empirical Methods in Natural Language Processing, pp 1724-1734, 2014.