УДК 004.8:004.9

АНАЛІЗ МОДЕЛЕЙ РЕКУРЕНТНИХ НЕЙРОННИХ МЕРЕЖ ДЛЯ РОЗРОБКИ ІНТЕЛЕКТУАЛЬНИХ ЧАТ-БОТІВ

Яровий Андрій, Прозор Олена, Щипський Юрій

Вінницький національний технічний університет

Анотація

В ході проведених досліджень здійснено аналіз моделей рекурентних нейронних мереж та обґрунтовано їх вибір для розробки інтелектуальних чат-ботів. Наведено аналіз методів навчання нейронних мереж. Здійснено порівняння засобів програмного забезпечення для реалізації рекурентної нейронної мережі

Abstract

In the course of the researches the models of recurrent neural networks were analyzed and their choice for the development of intelligent chatbots was substantiated. The methods of learning neural networks are analyzed. Comparison of software for implementation of recurrent neural network is made.

Вступ

Сучасне життя тісно пов’язане з інформаційними технологіями, і користувачам важко уявити своє життя без гаджетів, Інтернету, соціальних мереж, тощо. Всі ці технології помітно допомагають у вирішенні різного роду задач, і в тому числі нейронні мережі є каталізатором цього процесу, зокрема рекурентні нейронні мережі, які натепер активно використовуються.

В зв’язку із цим актуальним є їх дослідження, аналіз та широке використання у практичній діяльності.

Аналіз моделей і методів навчання рекурентних нейронних мереж та засобів їх програмної реалізації

Штучна нейронна мережа – програмне забезпечення або система апаратних (програмно-апаратних) засобів, що функціонують з певним ступенем подібності до нейронів нервової системи людини (рис. 1) [1-2]. Проаналізувавши різні архітектури нейронних мереж, такі як пряма нейрона мережа, радіально-базисні мережі, багатошаровий перцептрон, згорткові нейронні мережі та рекурентні нейронні мережі, встановлено, що рекурентні нейронні мережі найкраще підходять для реалізації інтелектуальних чат-ботів [3].

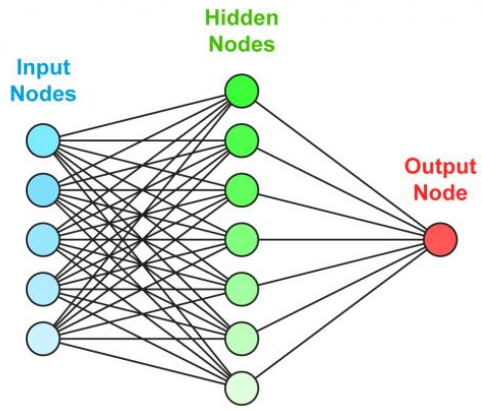

Рисунок 1 – Загальна структура рекурентної нейронної мережі

Рекурентні нейронні мережі (Recurrent Neural Networks, RNN) – мережі, які містять зворотні зв’язки та дають змогу зберігати інформацію (рис. 2) [4].

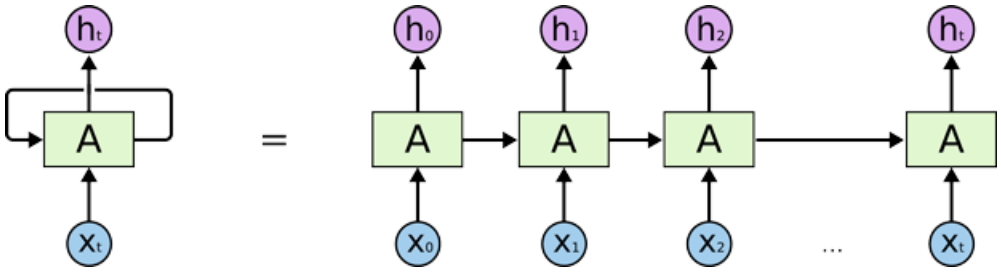

Рисунок 2 – Загальна структурно-функціональна схема рекурентної нейронної мережі

У даній схемі фрагмент нейронної мережі A приймає значення x_t та повертає значення h_t. Існування зворотного зв'язку дозволяє передавати дані з попередньої ітерації навчання нейронної мережі до поточної.

Учені виділяють декілька типів/варіацій рекурентної нейронної мережі [5-8]:

- звичайна рекурентна нейронна мережа (*conventional RNN*);
- нейронна мережа Хопфілда (*Hopfi1eld network*);
- мережа Елмана (*Elman network*);
- мережа з довгою короткостроковою пам’яттю (*Long short-term memory, LSTM*);
- двонаправлені рекурентні нейронні мережі (*Bidirectional RNNs*)

В більшості рекурентних нейронних мереж є проблема, що важлива інформація, яка була доступна 20-30 ітерацій тому, з часом втрачає свій вплив на мережу. Мережа Елмана досить ефективно працює з одношаровим або багатошаровим перцептроном, але не більше. І тут більш ефективною є архітектура мережі з довгою короткостроковою пам’яттю (*LSTM*).

Основна ідея *LSTM* полягає в тому, що оскільки різні приклади послідовності по-різному впливають на результат поточної ітерації, то необхідно розробити метод, який деяким елементам в контексті в минулих ітерацій буде надавати більшу вагу, тобто більший вплив, а іншим елементам – менший.

Структура *LSTM* також нагадує ланцюг, але модулі виглядають по-іншому. Замість одного шару нейронної мережі, вони містять чотири шари, які взаємодіють особливим чином (рис. 3) [4,9,10].

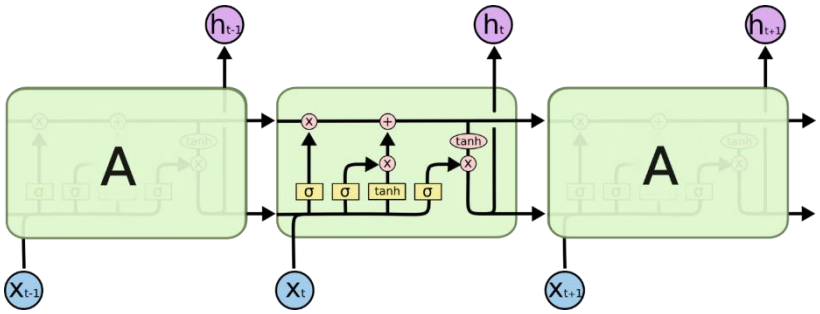

Рисунок 3 – Модель LSTM, що складається з чотирьох шарів із взаємодією

Ключовий компонент *LSTM* – це стан комірки (*cell state*) – верхня горизонтальна лінія (рис.3). Дана модель має певні структури – фільтри (gates), за допомогою яких можна керувати інформацією: пропускати / не пропускати, видаляти за необхідності. Також, дані фільтри дозволяють захищати та контролювати стан комірки.

Варто зазначити, що на основі проведених досліджень, саме рекурентна нейронна мережа з довгою короткостроковою пам'яттю найкраще підходить для реалізації інтелектуального чат-бота [11-13].

Для того, щоб нейронна мережа, як складова інтелектуального чат-бота, змогла розпізнати текст користувача та дати правильну відповідь, її потрібно навчити. Процес навчання досить складний та потребує багато часу та зусиль. Виділяють декілька методів навчання нейронних мереж [5]:

- метод градієнтного спуску;
- метод зворотного поширення (*Backpropagation*);
- навчання з вчителем;
- навчання без вчителя.

Для навчання рекурентних нейронних мереж найчастіше використовують метод зворотного поширення з певними модифікаціями, що має назву метод зворотного поширення в часі (*Backpropagation Through Time, BPTT*).

Концептуально, метод зворотного поширення в часі працює розгорнувши всі вхідні часові кроки. Кожен часовий крок має один вхідний крок, одну копію мережі та один вихідний крок. Потім помилки розраховуються та накопичуються для кожного часового кроку. Мережа повертається до певного стану, ваги оновлюються [10].

Для реалізації поставленої у даному дослідженні задачі було порівняно такі програмні засоби:

- Python/TensorFlow;
- Python/Keras;
- Matlab.

Проаналізувавши дані програмні засоби, можна зробити висновок, що найдоцільніше використовувати мову програмування Python з його бібліотеками TensorFlow/Keras. Оскільки Python досить проста та гнучка мова програмування з потужними інструментами для аналізу та обробки даних. Програмні бібліотеки TensorFlow/Keras досить легко встановити та в подальшому з їх допомогою розробляти нейронну мережу, модифікувати її та використовувати. Додатково TensorFlow має можливість завантажити датасети MNIST для навчання нейронної мережі, що підвищує ефективність комп'ютерного моделювання та експериментальних досліджень.

Висновки

В ході проведення даного дослідження здійснено аналіз моделей рекурентних нейронних мереж та методів їх навчання в контексті задачі розробки інтелектуального чат-бота. Серед розглянутих моделей рекурентних нейронних мереж та методів їх навчання обрано LSTM-мережу із Backpropagation Through Time, що найдоцільніше використовувати для реалізації інтелектуального чат-бота засобами мови програмування Python із використанням його бібліотеки TensorFlow.

Список використаних джерел

1. Нейронные сети: полный курс, 2-е издание. / Хайкин С. – Москва, 2006. – 1104с – ISBN 5-8459-0890-6.

2. A. Yarovyi, D. Kudriavtsev, S. Baraban, V. Ozeranskyi, L. Krylyk, A. Smolarz, G. Karnakova Information technology in creating intelligent chatbots. – Proc. SPIE 11176, Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments 2019, 1117627 (6 November 2019); doi: 10.1117/12.2537415; https://doi.org/10.1117/12.2537415

3. Аналіз технологій нейронних мереж для розробки інтелектуальних чат-ботів у соціальних мережах / Щипський Ю.О., Прозор О.П.: Збірник матеріалів XLVIIX Науково-технічної конференції Вінницького національного технічного університету, (Вінниця, 27-28 квітня 2020 р.). – В.: ВНТУ, 2020. – С. 1-3. [Електронний ресурс] – Режим доступу:https://conferences.vntu.edu.ua/index.php/all-fitki/all-fitki-2020/paper/view/8759/7556

4. Азбука ИИ: «Рекуррентные нейросети» [Електронний ресурс] – Режим доступу: https://nplus1.ru/material/2016/11/04/recurrent-networks

5. Нейронные сети для начинающих. Часть 2 [Електронний ресурс] – Режим доступу: https://habr.com/ru/post/313216/

6. Застосування глибокої рекурентної нейронної мережі із використанням алгоритму LSTM у системах інтелектуальної взаємодії / Яровий А., Кудрявцев Д., Кулик О. : Збірник праць XI Міжнародної науково-практичної конференції [Інтернет-Освіта-Наука (ІОН-2018)], (Вінниця, 22-25 травня 2018 р.) – Вінниця, ВНТУ, 2018. – с. 30-32

7. Finding structure in time / Elman, J.L.: Cognitive Science. — 1990. — С. 179-211. [Електронний ресурс] – Режим доступу: http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=87F767D3C21027CEF557110E2867E556?doi=10.1.1.117.1928&rep=rep1&type=pdf

8. The capacity of the Hopfield associative memory / McEliece R.J., Posner E.C., Rodemich E.R., Venkatesh S.S.: IEEE Transactions on Information Theory, Volume 33, Issue 4 (July 1987), 461 – 482с. [Електронний ресурс] – Режим доступу: https://authors.library.caltech.edu/6929/1/MCEieeetit87.pdf

9. LSTM – сети долгой краткосрочной памяти [Електронний ресурс] – Режим доступу: https://habr.com/ru/company/wunderfund/blog/331310/

10. A Gentle Introduction to Backpropagation Through Time [Електронний ресурс] – Режим доступу: https://machinelearningmastery.com/gentle-introduction-backpropagation-time/

11. Інтелектуальний чат-бот для задачі розпізнавання природної мови / Яровий А.А., Кудрявцев Д.С. : Збірник матеріалів XLVIII Науково-технічної конференції Вінницького національного технічного університету, (Вінниця, 13–15 березня 2019 р.). – В.: ВНТУ, 2019. – С. 1-3. [Електронний ресурс] – Режим доступу: https://conferences.vntu.edu.ua/index.php/all-fitki/all-fitki-2019/paper/view/7290/6071

12. Прикладна реалізація моделі інтелектуального чат-бота у сфері інформаційних відносин. / А.А. Яровий, Д.С. Кудрявцев, Л.В. Крилик : Збірник тез доповідей VI Міжнародної науково-технічної конференції "Оптоелектронні інформаційні технології "Фотоніка ОДС-2018", м. Вінниця, 2-4 жовтня 2018 року. – Вінниця: Видавництво "ТД Едельвейс і К", 2018. – С. 75-76.

13. Чат-бот як система інтелектуальної взаємодії / Яровий А.А., Кудрявцев Д.С. : Збірник матеріалів XLVII Науково-технічної конференції Вінницького національного технічного університету, (Вінниця, 21–23 березня 2018 р.). – В.: ВНТУ, 2018. – С. 1-3. [Електронний ресурс] – Режим доступу: https://conferences.vntu.edu.ua/index.php/all-fitki/all-fitki-2018/paper/view/4847/4269