DOI: 10.36910/6775-2524-0560-2019-37-6
УДК: 004.89
Лавренчук С. В., Ілюшик Р. С.
Луцький національний технічний університет

ДОСЛІДЖЕННЯ ТЕХНОЛОГІЇ ОБРОБКИ ПРИРОДНОЇ МОВИ ТА МАШИННОГО НАВЧАННЯ ПРИ СТВОРЕННІ CHAT-BOT ЗАСОБАМИ PYTHON

Лавренчук С. В., Ілюшик Р. С. Дослідження технології обробки природної мови та машинного навчання при створенні chat-bot засобами Python. У статті розглянуто сучасні технології побудови додатка chat-bot, зокрема досліджено ефективність використання алгоритмів природної обробки мови, машинного навчання, а також застосування нової, більш продуктивної архітектури нейронної мережі.
Ключові слова: Bi Encoder, GloVe, python, chat-bot, Word2Vec, штучний інтелект, нейронні мережі, Tensorflow

Лавренчук С. В., Илюшик Р. С. Исследование технологии обработки естественного языка и машинного обучения при создании chat-bot средствами Python. В статье рассмотрены современные технологии построения приложения chat-bot, в частности исследована эффективность использования алгоритмов естественной обработки речи, машинного обучения, а также применение новой, более производительной архитектуры нейронной сети.
Ключевые слова: Bi Encoder, GloVe, python, chat-bot, Word2Vec, искусственный интеллект, нейронные сети, Tensorflow

Lavrenchuk S. V., Iliushyk R. S. Research of technology of natural language processing and machine learning through chat-bot creation by Python means. The article discusses modern technologies for building a chat-bot application, in particular, researched the effectiveness of using natural language processing algorithms, machine learning, and the use of a new, more productive neural network architecture.
Keywords: Bi Encoder, GloVe, python, chat-bot, Word2Vec, artificial intelligence, neural networks, Tensorflow

Постановка наукової проблеми та аналіз досліджень.

Із розвитком сучасних технічних можливостей, комунікаційні функції в Інтернеті набувають нових форм. Особливе місце посіли технології, пов'язанні з використанням месенджерів та чат-ботів. Глобальна мережа стає середовищем спілкування, яке займає важливе місце в усіх сферах життєдіяльності суспільства.

Внаслідок отримання підтримки швидкісного доступу до інформації, асинхронної та одночасної роботи в мережі багатьох користувачів, можливістю збору та обміну різноманітною інформацією, навчання і підвищення кваліфікації, комунікацій для вирішення особистих і бізнес-питань з клієнтами і партнерами, проведення всіляких аналітичних досліджень, на сьогоднішній день проблема віртуального спілкування в мережі стала актуальною. Тому дослідження ефективної побудови і навчання чат-ботів за допомогою машиного навчання та з використаням технології природної обробки мови – перспективне завдання.

Метою роботи є вивчення основних принципів побудови сервісів для платформ Телеграм та Фейсбук, застосування та дослідження нейронної мережі та методів природньої обробки мови з використаням машиного навчання для покращення унівесальності та ефективності чат-бота; провести експерименти із навчанням даних помічників. Об'єктом дослідження є чат-бот, заснований на рекурентній нейронній мережі з використаням машиного навчання та технології обробки природньої мови.

Завдання, які треба вирішити:

– дослідити способи та методи застосування технологій машинного навчання нейронних мереж для покращення ефективності чат-бота та його «природного» функціонування;

– виконати вибір засобів та технологій розробки;

– проаналізувати отриманні результати застосування нейронних мереж і вибрати найефективнішу;

– визначити ефективний спосіб для використання технології NLP;

– розробити chat-bot засобами мови Python з використанням машинного навчання та технології обробки природної мови.

Наукова новизна одержаних результатів полягає у впровадженні нової ефективної архітектури Bi Encoder Recurrent Neural Network з LSTM та її подальшого використання, за допомогою машинного навчання і NLP.

Практична цінність проведеної роботи полягає у можливості застосування отриманих результатів для ефективного навчання нейромережі, що використовується в задачі обробки

природної мови чат-ботом, також покращення виведення інформації ботом і можливості його застосування в різних сферах. При вирішенні поставленого завдання використовувалися наукові досягнення в областях штучного інтелекту, а саме нейронних мереж і машинного навчання.

Виклад основного матеріалу й обґрунтування отриманих результатів.

Залежно від архітектури на якій базується бот, можна поділити їх на дві великі групи: скриптові (працюють на основі заздалегідь заготовлених фраз) та розумні (здатні навчатися).

Перший тип ботів, як правило, використовується для вузько направлених цілей і не розрахований на масового споживача. Дані боти, працюють через команди, опираючись на заздалегідь написані ключові слова, які вони розуміють. Кожна з таких команд повинна бути запрограмована розробником окремо, із впровадженням регулярних виразів або інших форм аналізу термінів. Якщо користувач задав питання, не використавши жодного ключового слова, робот не може зрозуміти його і, як правило, відповідає повідомленнями типу «вибачте, я не зрозумів».

До другої категорії відносяться більш складні програми. Ці боти, опираються на штучний інтелект (ШІ) [1], у вигляді нейронних мереж, щоб спілкуватися з користувачами. Замість заздалегідь підготовлених відповідей, робот генерує адекватні відповіді та пропозиції що стосуються теми. До того ж, всі слова, сказані чи написані в процесі діалогу, зберігаються для подальшої обробки та самонавчання. В поняття нейронних мереж вкладаються моделі, описані математичним способом, а також їх реалізації, що побудовані за правилами функціонування біологічних мереж та їх організації.

Машинне навчання - це підрозділ штучного інтелекту, що вивчає методи побудови алгоритмів, здатних навчатися. Першопрохідцем в цій області вважається фахівець з обчислювальної техніки з компанії IBM Артур Самуель, який написав в 1959 році комп'ютерну програму Checkers-playing для гри в шашки, яка вважається однією з перших самонавчальних програм в світі і є демонстрацією базових понять штучного інтелекту [2].

В даний час, машинне навчання застосовується в областях зазначених на рисунку 1.

Рисунок 1 – Сфери застосування машинного навчання

Алгоритм навчання мережі прямого поширення називається зворотним поширенням помилки («back propagation»). Алгоритм був представлений в 1970, але основний розвиток отримав в 1986 р. [3].

Алгоритм складається з двох фаз:

1. Пряма фаза. На цій фазі ваги і зміщення нейронів залишаються невідомими. На вхід мережі подаються сигнали і зчитується результат. Потім обчислюється сумарна помилка за різницею очікуваного результату на виході. Як приклад для сумарної помилки часто застосовується середньоквадратичне відхилення [4].

2. Зворотна фаза. Сумарна помилка представляє функцію багатьох змінних, де змінні – це параметри мережі. Обчислення градієнта цієї функції покаже напрямок зростання функції помилки, отже, завдання навчання мережі зводиться до мінімізації функції помилки.

Техніка «back propagation» полягає в обчисленні градієнта і помилки на кожному шарі нейронної мережі. Для навчання мережі необхідно повторювати процес зворотного поширення помилки кілька разів. Існує кілька режимів обчислення помилки [5]:

- online mode: обчислення помилки відбувається для кожного зразка;

- batch mode: кілька зразків проходять пряму фазу, потім обчислюються нові параметри;
- full-batch: використовуються всі навчальні зразки;
- mini-batch: всі зразки розбиваються на групи.

У даній роботі застосовується метод навчання з учителем.

В останні кілька років ми спостерігаємо вибух інтересу до нейронних мереж, які успішно застосовуються в різних областях: бізнесі, медицині, техніці, геології, фізиці. Нейронні мережі увійшли в практику всюди, де потрібно вирішувати завдання прогнозування, класифікації або управління. Однак особливе місце вони посідають у процесі створення віртуальних помічників, з подальшим машинним навчанням та удосконаленням результатів. Таку зацікавленість розробників до даної задачі можна пояснити проблемою заміни дороговартісного утримання працівників більш дешевими програмами, продуктивність яких набагато вища ніж у людей. Було доведено [6], що нейронна мережа може апроксимувати будь-яку функцію. Цей процес називається лінійної регресією.

Одна з найскладніших речей в NLP – це підготовка даних. Для того, щоб нейронна мережа зрозуміла введений текст, слова повинні бути представлені як вектори.

У ролі вчителя нейронної мережі виступає набір даних (дата сет), по якому проводилось навчання і подальша генерація відповідей чат-ботом, був використаний корпус запитань та відповідей. З цього дата сету сформований словник для нейронної мережі. Словник буде оброблений через Word2Vec і сформована матриця ваг.

Нижче представлений результат формування словника (рисунок 2).

```
(venv) roma@roma-HP-ProBook-4535s:~/DIPLOMA/chatbot$ python run_train.py

Size of Vocabulary : 15897
Word embeddings: 20105
Number of words missing from CN: 567
Percent of words that are missing from vocabulary: 2.820 %
Total number of unique words: 15897
```

Рисунок 2 – Формування словника

З результатів видно, що:

– кількість слів, що важливі для контексту: 15897 (це і буде нашим словником);
– всього слів в Word2Vec: 20105;
– кількість слів, які не співпали з сетом Word2Vec: 567;
– відсоток від параметра, зазначеного вище від загального числа слів: 2.820%;
– загальна кількість унікальних слів: 15897.

На наступному етапі програма формує матрицю векторів слів (рисунок 3). Вектор кожного слова формується у 250 вимірів.

```
(venv) roma@roma-HP-ProBook-4535s:~/DIPLOMA/chatbot$ python run_train.py
[

[0.4578445 0.4511225 -0.4564548 …, 0.1122354 -0.1547874 0.0554472]

[-0.0574175 -0.2336815 0.4427541 …, 0.0172237 -0.4178941 0.0454545

[0.4353755 0.0454545 -0.1222121 …, 0.0100014 0.1455564 0.2404647]

…,

[0.2236891 -0.0684532 -0.1610332 …, 0.1719013 0.1483761 0.1545497]

[0.0071415 0.1112325 -0.2547811 …, 0.1124524 -0.0001254 0.0234999]

]
```

Рисунок 3 – Матриця векторів слів розмірністю 250

Від кількості вимірів буде залежати й якість відповідей чат-бота.

Було розглянуто випадкове представлення слів, а потім застосовано векторне представлення Word2Vec, що тестувалось на навчальному діалоговому корпусі. Також використано попередньо підготовлене вбудовування векторів GloVe для порівняння. При тестуванні попередньо навченого, навченого вбудовування в UDC та випадкового вбудовування, найкращим результатом є вбудоване загальне повзання.

Діаграма векторного представлення слів Word2Vec показана на рисунку 4. Як видно на діаграмі, подібні слова (наприклад, "дякую", "спасибі" та "ласка") вбудовуються близько один до одного.

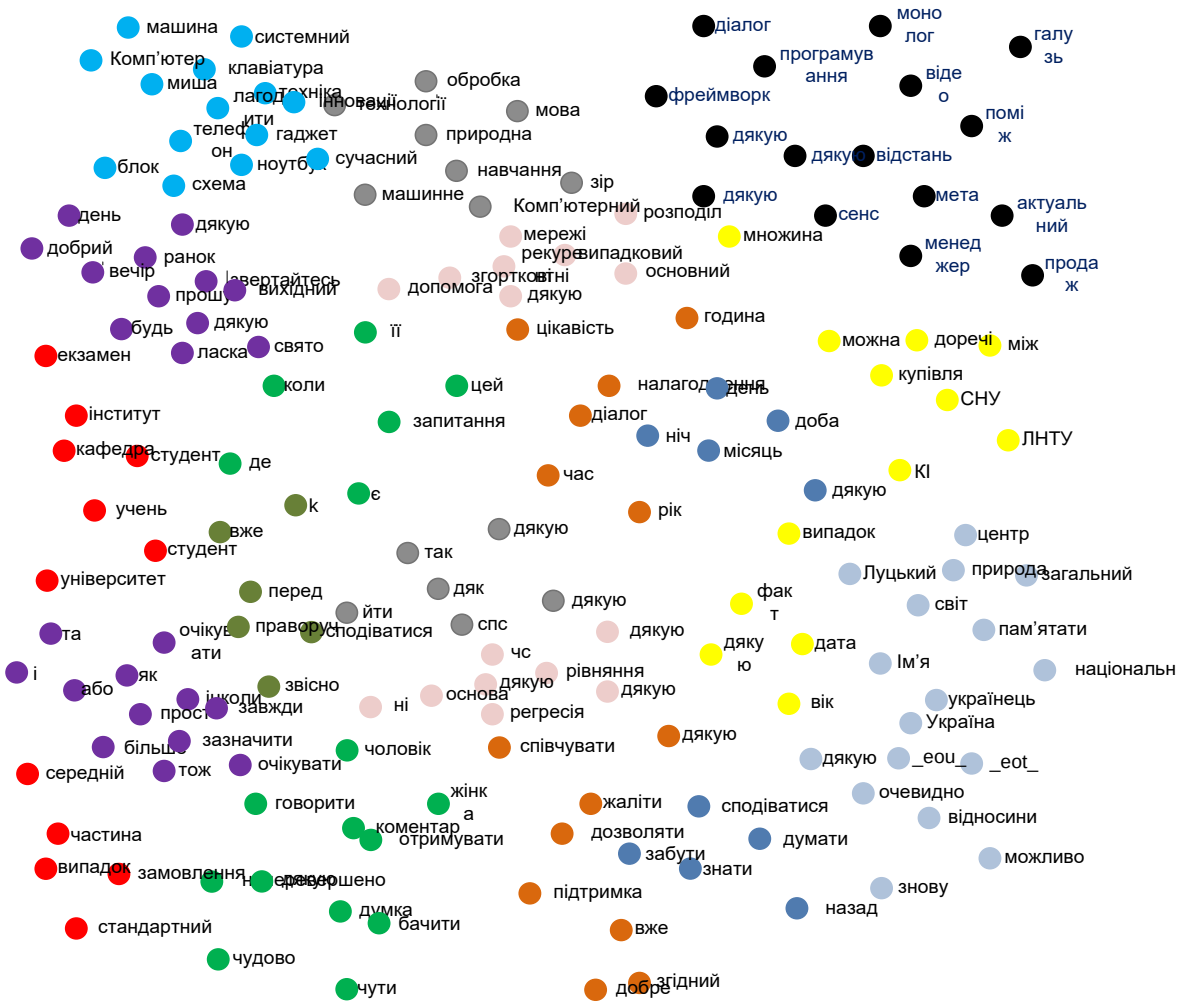

Рисунок 4 – Діаграма векторного представлення деяких слів, що часто зустрічаються

Графіки (рисунки 5,6) показують, що Word2Vec з методом CWOB та Skip-Gram точніше та швидше формують вектори слів, ніж схожі до нього методи. Це стає зрозуміло, тому що GloVe показує набагато кращі результати у своєму підході до представлення слів, коли середня довжина всіх вхідних речень більша, ніж та, яка використовується для даного чат-бота.

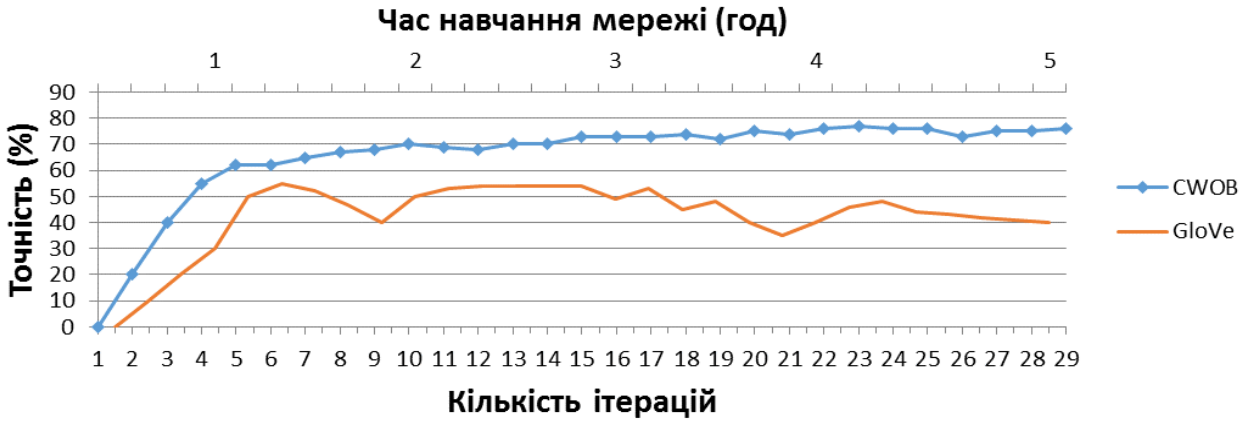

Рисунок 5 – Графік ефективності застосування Word2Vec та GloVe

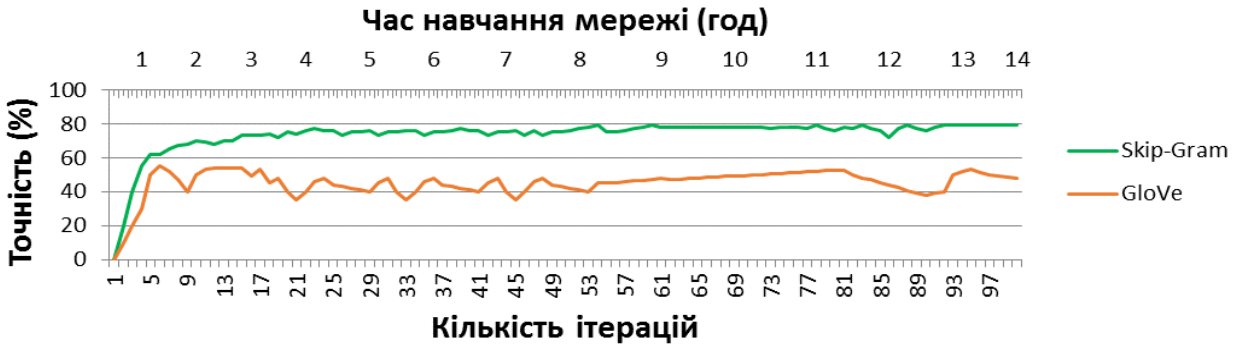

Рисунок 6 – Графік ефективності застосування Skip-Gram та GloVe

В таблиці 1 наведено узагальнене порівняння алгоритмів, які були протестовані на наборі даних для перевірки.

Таблиця 1 – Порівняння характеристик моделі BE на різних типах векторного представлення слів.

Назва	«Виклик1»	«Виклик2»	«Виклик3»
Random	41.7	61.1	87.8
Word2Vec (CVOB)	56.55	73.61	92.7
GloVe	52.3	71.1	90.3
Word2Vec (Skip-Gram)	56.55	73.61	92.7
Common Crawl	51.5	72.1	88.9

Для визначення ефективної архітектури нейронної мережі було порівняно модель Dual Encoder (DE) та Bi Encoder (BE). На відміну від BE (рисунок 7), модель DE має одну комірку LSTM, що кодує як питання, так і відповідь, тобто нейромережа нечутлива до типу даних (запитання, відповідь) та номеру входу, на який вхід ці дані подаються.

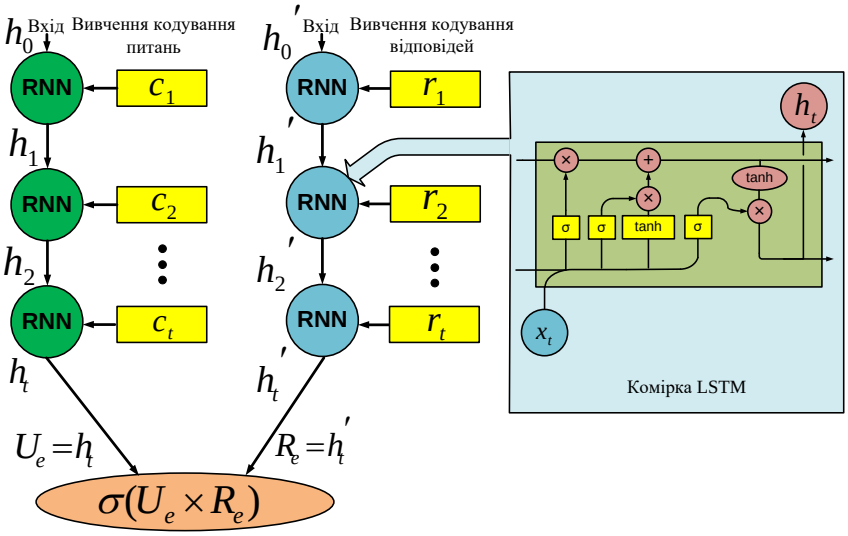

Рисунок 7 – Bi Encoder LSTM

Моделі навчалися на тисячі пар висловлювань та відповідей навчального набору та оцінювались на основі тестового набору. У таблиці 2 порівнюється ефективність різних моделей, заснованих на пошуку. Порівняно з базовою моделлю DE, запропонована модель BE досягає 1,8%, 2,6% та 1,27% вищої точності для «Виклик1», «Виклик2» та «Виклик3» відповідно. Зауважимо, що порівняно з відтвореною моделлю DE модель BE працює краще, ніж коли вона порівнюється з базовою моделлю. Здійснено порівняння між запропонованою моделлю BE і базовою моделлю DE, зокрема порівняно модель BE з регулярною RNN (BE-RNN) з моделлю DE з регулярною RNN (DE-RNN). Це порівняння показано в таблиці 2 із однаковими гіперпараметрами.

Таблиця 2 – Порівняння моделі BE з архітектурою RNN та моделі DE з RNN. Результати з точність % на наборі перевірки

Модель	Опис	«Виклик1» %	«Виклик2» %	«Виклик3» %
DE	Dual Encoder LSTM	54.2	71.09	91.43
BE	Bi Encoder LSTM	56.0	73.15	92.7
DE-RNN	Dual Encoder RNN	37.9	56.00	83.60
BE-RNN	Bi Encoder RNN	34.6	53.10	82.3

Проведені експерименти показали, що кращі результати отримуються з використанням Bi-Encoder LSTM.

Висновки та перспективи подальшого дослідження.

Показано потенціал використання chat-bot на основі нейронних мереж для спілкування, консультації або інших видів надання допомоги користувачам.

Був згенерований власний дата-сет діалогів, що задовольняв поставлені вимоги, і на якому проводились всі експерименти та навчання, що налічує близько 4500 тисяч реплік. А також розглянуто і проаналізовано алгоритми та методи векторного представлення слів для машинного навчання нейронної мережі. Визначено найефективніший спосіб подачі корпусу слів у нейрону мережу.

Представлено і досліджено навчання нової RNN-архітектури на основі LSTM та Bi Encoder за допомогою технології машинного навчання. Дана мережа може оцінити набір заздалегідь визначених відповідей.

Досліджено і показано, що нейрона мережа генерує правильну відповідь в середньому у 92,7%, 73,15% та 56,0% у «Виклику 1», «Виклику 2» та «Виклику 3» відповідно, що перевищує точність моделі, яка була використана в якості стандартної. Оскільки створений набір даних відповідає всім вимогам, можна продовжувати подальшу роботу з вивчення моделі BE зі втратою

рангу, поступово розширюючи дата-сет українських діалогів. Досліджувана архітектура може бути розширена на більш ієрархічні шари RNN, захоплюючи довший контекст.

В результаті проведення навчання моделі, було визначено, що для отримання якісних результатів при використанні рекурентних мереж LSTM Bi Encoder, потрібно затратити кілька десятків годин, на графічному процесорі Nvidia GTX GeForce 1070, також можна продовжувати цей процес і на звичайному CPU, проте це займе досить багато часу.

Також, проведено огляд та аналіз бібліотек машинного навчання, визначено найоптимальнішу для проведення даних експериментів. Підходи досліджені ефективні, і володіють гнучкістю, що дозволяє їх масштабувати для різних застосувань.

В процесі розробки інтелектуального помічника, було протестовано можливість його правильного функціонування в невеликих групах. Таким чином, можна зробити висновок, що всі поставленні задачі виконуються, а мета дослідження досягнута.

Список бібліографічного опису

1. Паскану. Р. Про труднощі навчання періодичних нейронних мереж / Р. Паскану, Т. Міколов, Ю. Бенджо. / arXiv препринт, 1211.5063, 2012.
2. Дж. Куццола, Й. Йованович, Е. Багері та Д. Гасевич, "Еволюційна тонка настройка автоматизованих семантичних анотаційних систем", Експертні системи з додатками, 42, 2015, с. 6864-6877.
3. Рассел, Стюарт Дж .; Норвіг, Пітер (2003), Штучний інтелект: сучасний підхід (2-е видання), Верхнє сідло, річка Нью-Джерсі: Прентіс Холл, ISBN 0-13-790395-2, стор. 939.
4. Принципи нейродинаміки: рецептори та теорія механізмів мозку. - М .:, 1965. - 480 с.
5. Сінцзянь Ши; Чжуронг Чень; Хао Ван; Діт-Ян Єунг; Вай-Кін Вонг; Ван-чун Ву. Конволюційна мережа LSTM: підхід до машинного навчання для випадання атмосферних опадів. Праці 28-ї міжнародної конференції з нейронних систем обробки інформації: журнал. - 2015. - С. 802—810.
6. Барцев С.І., Гілев С.Є., Охонін В.А., Принцип подвійності в організації адаптивних мереж обробки інформації, В: Динаміка хімічної та біологічної систем. - Новосибірськ: Наука, 1989. - С. 6-55.

References

1. Pascanu. R. On the difficulty of training recurrent neural networks / R. Pascanu, T. Mikolov, Y. Bengio. / arXiv preprint, 1211.5063, 2012.
2. J. Cuzzola, J. Jovanovic, E. Bagheri and D. Gasevic, "Evolutionary fine-tuning of automated semantic annotation systems", Expert Systems with Applications, 42, 2015, pp. 6864-6877.
3. Russell, Stuart J.; Norvig, Peter (2003), Artificial Intelligence: A Modern Approach (2nd ed.), Upper Saddle River, New Jersey: Prentice Hall, ISBN 0-13-790395-2, page 939.
4. Principles of Neurodynamic: Perceptrons and the Theory of Brain Mechanisms. — M.: , 1965. — 480 c.
5. Xingjian Shi; Zhourong Chen; Hao Wang; Dit-Yan Yeung; Wai-kin Wong; Wang-chun Woo. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. Proceedings of the 28th International Conference on Neural Information Processing Systems : journal. — 2015. — P. 802—810.
6. Bartsev SI, Gilev SE, Okhonin VA, The principle of duality in the organization of adaptive information processing networks, In: Dynamics of chemical and biological systems. - Novosibirsk: Science, 1989. - P. 6-55.