УДК 000.000:000

РОЗРОБКА КРОССПЛАТФОРМНОЇ СИСТЕМИ ДЛЯ ВИРІШЕННЯ ЗАДАЧІ МУЛЬТИМОДАЛЬНОГО НЕЙРОМАШИННОГО ПЕРЕКЛАДУ

Майоров Микита

Науковий керівник: доктор ф.-м. наук, професор Авраменко О.В.

Центральноукраїнський державний педагогічний університет

імені Володимира Винниченка, м. Кропивницький, Україна

Штучні нейронні мережі в останній час є одним з найбільших напрямів досліджень у галузі штучного інтелекту. При наявності великої кількості даних для навчання, нейронні мережі можуть проявляти себе якісніше ніж традиційні алгоритми машинного навчання, особливо у задачах обробки природної мови. Однією с основних проблем побудови систем нейромашинного перекладу є втрата контексту із збільшенням довжини речень. Для її вирішення, у даній статті запропоновано модифікований метод навчання, що залучає до основного набору даних зворотне представлення їх векторизованної форми.

Ключові слова: машинний переклад, штучні нейронні мережі, обробка природної мови, розробка кроссплатформних систем.

DEVELOPMENT OF A CROSS-PLATFORM SYSTEM FOR SOLVING THE PROBLEM OF MULTIMODAL NEURO-MACHINE TRANSLATION

Maiorov Mykyta

Supervisor: *Ph.D. of Physical and Mathematical Sciences*, Prof. Avramenko O.V.

Volodymyr Vynnychenko Central Ukrainian State

Pedagogical University, Kropyvnytskyi, Ukraine

Artificial neural networks have recently been one of the largest areas of research in artificial intelligence. In the presence of a large amount of training data, neural networks can perform better than traditional machine learning algorithms, especially in human language processing tasks. One of the major problems with the construction of neural machine translation systems is the loss of context with increasing sentence length. To resolve this problem, a modified training method is proposed in this article, which involves a backward representation of their vectorized form in the main data set.

Keywords: machine translation, artificial neural networks, natural language processing, cross-platform systems development.

Постановка проблеми

Нейромашинний переклад (НМП) був представлений як багатообіцяюча технологія із потенціалом до виправлення багатьох недоліків традиційних систем машинного перекладу. Основна ідея НМП полягає у тому, щоб навчити модель штучної нейронної мережі безпосередньому перетворенню вхідного тексту на цільовий вихідний текст, з однієї мови на іншу. Перевага нейромашинного перекладу полягає в тому, що даний метод використовує у своїй моделі багато рішень, схожих на традиційний машинний переклад на основі фраз. Однак, на практиці, система НМП досить часто показує більш гірші результати, аніж система перекладу на основі фраз, особливо при навчанні на дуже великих наборах даних. Для підвищення точності перекладу речень, структура яких суттєво відрізняється в залежності від мови, було запропоновано модифікований метод навчання моделі, що полягає у додаванні до навчального потоку даних зворотного векторного представлення речень вхідної мови.

Аналіз досліджень

Під час дослідження було проаналізовано архітектури існуючих систем нейромашинного перекладу, включаючи Google NMT [5], що використовує рекурентні блоки (LSTM), розташовані на восьми шарах, та покращений режим паралельної обробки завдяки з'єднанню механізму уваги (attention) від нижнього рівня мережі до верхнього. Також було розглянуто саму технологію "Self-Attention" [1; 2], яка окрім покращення якості роботи моделі, суттєво підвищує можливості інтерпретації її роботи.

Мета статті

Метою написання даної статті було дати загальну постановку задачі мультимодального нейромашинного перекладу, описати процес реалізації кроссплатформної комп'ютерної системи для вирішення даної задачі а також викласти основні результати дослідження модифікованого методу навчання моделі, деталі його імплементації та вплив на якість перекладу.

Виклад основного матеріалу

Загальна постановка задачі: Нехай *x* позначає мову джерела, а *y* позначає мову перекладу, з урахуванням набору параметрів моделі θ, метою алгоритму машинного перекладу є пошук перекладу з максимальною ймовірністю $\hat{y}$, де:

$$\hat{y} = \arg\max_y P(y|x;\theta) \quad (1)$$

Навчання в НМП здійснюється шляхом максимізації логарифмічної достовірності у якості цільової функції:

$$\hat{\theta} = \arg\max_\theta \sum_{ij=1}^{l} \log P(y_i|x_i;\hat{\theta}) \quad (2)$$

End-to-End моделі машинного перекладу, які також називаються моделями нейромашинного перекладу (НМП), спрямовані на пошук відповідностей між вхідною та цільовою мовами, використовуючи багатошарові нейронні мережі. Головна відмінність між підходами, заснованими на методах НМП і звичайного статистичного машинного перекладу (СМП), полягає в тому, що нейронні моделі здатні знаходити та вивчати складні взаємовідношення між природними мовами безпосередньо з самих даних, не вдаючись до ручної інженерії характеристик. Але основна проблема залишається такою ж, з огляду на послідовність слів у реченнях вхідної мови $X = x_1 \dots x_j, \dots x_J$ та реченнях цільової мови $Y = y_1 \dots y_i, \dots y_I$, модель НМП намагається зв'язати ймовірність перекладу на рівні речення із ймовірністю перекладу, що залежить від контексту, де $y < i$ є частковим перекладом:

$$P(x|y;\theta) = \prod_{i=1}^{I} P(y_i|x, y < i;\theta) \quad (3)$$

В контексті вхідного та цільового речення може існувати розрідженість, якщо речення стають занадто довгими. Для вирішення цієї проблеми, K. Cho та ін. у 2014 році запропонував нову архітектуру мережі кодер-декодер (Encoder-Decoder) [3]. Роль кодера у даному типі мереж полягає у представленні речень довільної довжини у вигляді вектору фіксованої довжини, який називається вектором контексту. Цей контекстний вектор містить всі необхідні ознаки, які можна вивести з початкового речення. Декодер приймає цей вектор як вхідний з

метою виведення цільового речення слово за словом. Очікується, що ідеальний декодер виведе речення, що містить повний контекст речення вхідної мови.

За допомогою емпіричного тестування, спостерігалося, що якість перекладу залежить від розміру вхідного речення і значно зменшується із його збільшенням. Щоб вирішити цю проблему D. Bahdanau та ін. у 2014 році запропонували інтегрувати механізм уваги (attention) [1] всередину шару кодера і показав, що це допомагає динамічно вибирати відповідні частини контексту вхідного речення для побудови цільового речення. Для цього використовувалася двонаправлена рекурсивна мережа (BRNN) для захоплення глобальних контекстів:

$$\overrightarrow{h_s} = f(x_s, \overrightarrow{h_{s-1}}, \theta) \tag{4}$$

$$\overleftarrow{h_s} = f(x_s, \overleftarrow{h_{s-1}}, \theta) \tag{5}$$

Прямий прихований стан $\overrightarrow{h_s}$ і зворотній прихований стан $\overleftarrow{h_s}$ об'єднуються для захоплення контексту на рівні всього речення.

$$h_s = [\overrightarrow{h_{s-1}}, \overleftarrow{h_{s-1}}] \tag{6}$$

Основною ідеєю, що стоїть за обчисленням уваги, є пошук відповідних частин у вхідному реченні, для допомоги при генерації цільових слів речення. Це досягається шляхом обчислення ваг уваги.

$$\alpha_{i,j} = \frac{\exp(a(t_{j-1}, h_i, \theta))}{\sum \exp(a(t_{j-1}, h_i, \theta))} \tag{7}$$

Де $a(t_{j-1}, h_i, \theta)$ є функцією вирівнювання, яка оцінює, наскільки добре входи та виходи вирівнюються по відношенню до i. Контекстний вектор c_j обчислюється як зважена сума прихованих станів джерела:

$$c_j = \sum_{i=1}^{l=+1} \alpha_{i,j} h_i \tag{8}$$

Цільовий прихований стан обчислюється наступним чином:

$$t_j = f(y_{j-1}, s_{j-1}, c_j, \theta) \tag{9}$$

Різниця між механізмом уваги та оригінальною архітектурою кодер-декодер полягає в тому, як обчислюється контекст джерела. У випадку із кодер-декодером,

прихований стан джерела використовується для ініціалізації вхідного прихованого стану цільового речення, тоді як при механізмі уваги використовується зважена сума прихованого стану. Це гарантує, що релевантність кожного вхідного слова у реченні буде добре зберігатися в контексті, що значно покращує продуктивність перекладу. Також, за допомогою візуалізації ваг уваги можна інтерпретувати результат роботи мережі даного типу у вигляді графіку типу heatmap (рис.1).

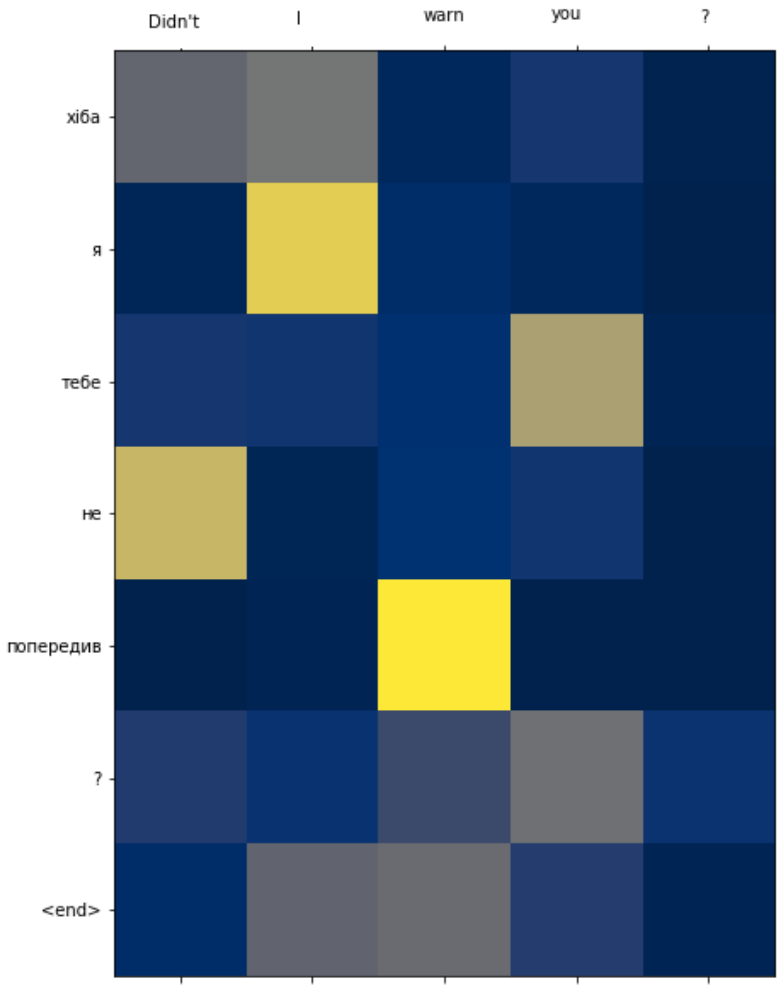

Рис.1 Візуалізація ваг уваги в результаті перекладу речення з англійської мови на українську

Під час навчання моделі було помічено низьку якість перекладу на реченнях із сильною різницею їх структур на вхідній та цільовій мовах. Наприклад «Where are you from?» – «Звідки ти?», «What are you talking about?» – «Про що ти говориш?». Проаналізувавши існуючі дослідження у даному напрямку, зокрема

роботу команди компанії Baidu [4], було вирішено додавати зворотне векторне представлення речень до основного потоку навчальних даних з метою отримання знань про структуру більш складних словосполучень між двома мовами.

Окрім покращення якості перекладу, даний метод додавав і деякі недоліки: процес навчання вимагав від операційної системи більшої кількості оперативної пам'яті та займав більше часу, аніж оригінальна версія. Але враховуючи цілі та технічні можливості розробника, вплив однієї з цих двох проблем можна зменшувати шляхом формування зворотного векторного представлення окремих речень саме під час навчання (але збільшуючи час на дану процедуру), або виконання даної операції для повного набору перед початком першої ітерації навчання (що збільшує потребу в більшій кількості оперативної пам'яті).

Як результат, при використанні даного методу значення цільової loss-функції зменшується швидше, ніж при класичній реалізації (рис.2). Також було помічено покращення значення спеціалізованої функції оцінки якості для задач обробки природної мови – BLEU приблизно на 1,6062.

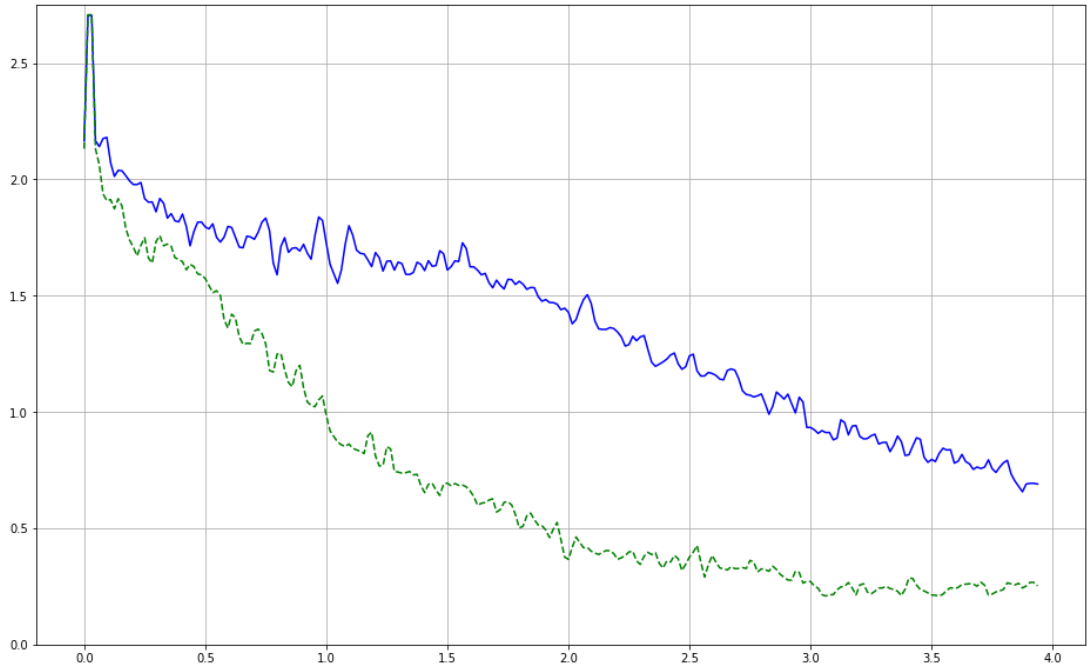

Рис.2 Значення loss-функції в процесі навчання мережі. Синім кольором зображено оригінальний метод навчання, зеленим пунктиром – модифікований

Програмна реалізація системи

Для реалізації системи мультимодального нейромашинного перекладу було використано клієнт-серверну архітектуру. Для імплементації моделі НМП та написання серверної частини системи було використано програмну мову Python, та бібліотеки її екосистеми для спрощення процесу побудови та навчання штучної нейронної мережі (numpy, tensorflow, keras), обробки природної мови (nltk), обробки запитів від програм-клієнтів (flask, nginx), а також для реалізації мультимодальних функцій image-to-text (pyocr) та speech-to-text (speech_recognition).

Клієнтську частину написано із використанням фреймворків для побудови кросплатформних додатків, таких як Qt (десктоп платформа), Flutter (мобільна платформа) та VueJS (web платформа). При проектуванні користувацького інтерфейсу додатків використовувався Material Design (рис.3).

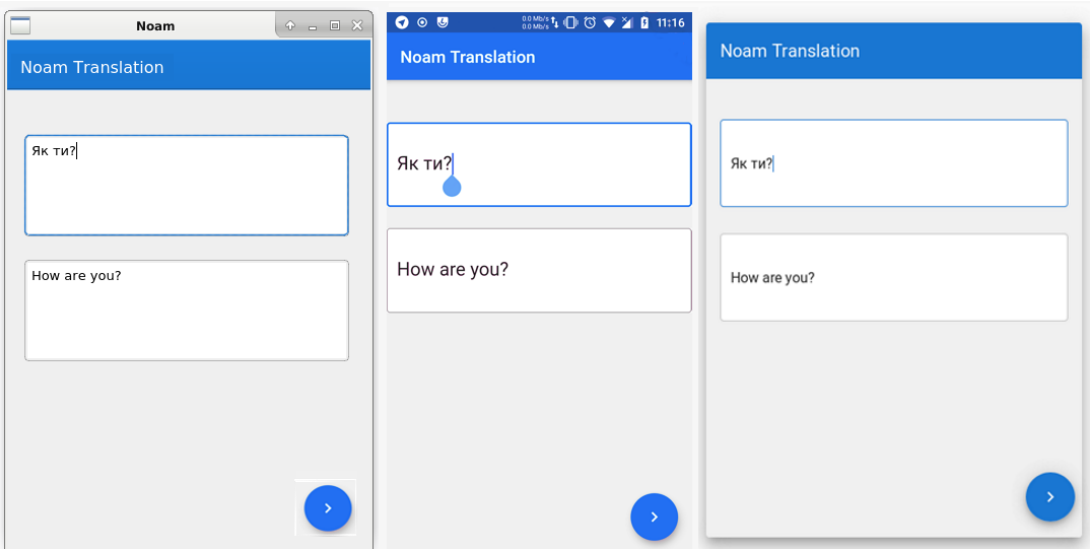

Рис.3 Зовнішній вигляд клієнтських додатків на різних платформах. Зліва-направо: десктоп, мобільна, web

Висновки та перспективи подальших пошуків у напрямі дослідження

У цій статі було описано основну задачу нейромашинного перекладу, пов’язані з нею характерні труднощі, а також представлено реалізацію

кросплатформнної системи НМП, включаючи опис її архітектури, та оцінку якості. Окрім цього, викладені результати дослідження впливу модифікованого методу навчання на якість вирішення поставлених задач.

Подальші напрями дослідження включають в себе аналіз впливу довжини речень навчальних даних на якість виконання поставленої задачі, з метою створення підсистеми генерації більш якісного навчального набору даних.

Список використаної літератури

1. D. Bahdanau Neural Machine Translation by Jointly Learning to Align and Translate / D. Bahdanau, K. Cho, Y. Bengio. arXiv preprint: 1409.0473, 2014. URL: https://arxiv.org/pdf/1409.0473.pdf (дата звернення: 16.10.2019).
2. A. Vaswani Attention Is All You Need / A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin. arXiv preprint: 1706.03762, 2017. URL: https://arxiv.org/pdf/1706.03762.pdf (дата звернення: 16.10.2019).
3. K. Cho Learning phrase representations using rnn encoder–decoder for statistical machine translation / K. Cho, B. Van Merriënboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio // In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), P.1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics. URL: http://www.aclweb.org/anthology/D14-1179 (дата звернення: 20.10.2019).
4. M. Sun Baidu Neural Machine Translation Systems for WMT19 / M. Sun, B. Jiang, H. Xiong, Z. He, H. Wu, H. Wang // Proceedings of the Fourth Conference on Machine Translation. P.374–381. URL: https://www.aclweb.org/anthology/W19-5341.pdf (дата звернення: 25.10.2019).
5. Y. Wu Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation / Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi. arXiv preprint: 1609.08144, 2018. URL: https://arxiv.org/pdf/1609.08144.pdf (дата звернення: 17.10.2019).