STUD NET

ОБЗОР МОДЕЛЕЙ НЕЙРОННЫХ СЕТЕЙ ДЛЯ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА

OVERVIEW OF NEURAL NETWORK MODELS FOR NATURAL LANGUAGE PROCESSING

УДК-004

Богомолов Юрий Алексеевич, студент магистратуры «Национального исследовательского университета ИТМО», Россия, г. Санкт-Петербург
Bogomolov Yuri Alekseevich, bogomolov.yuriy@gmail.com

Аннотация: В данной обзорной статье рассматриваются проблемы, связанные с машинной обработкой естественных языков, а затем приводится обзор существующих на данный момент моделей нейронных сетей, которые каким-либо образом решают задачи обработки естественного языка. В заключении приводятся сильные стороны и ограничения перечисленных моделей. Данная статья будет полезна тем, кто впервые сталкивается с необходимостью решать задачи обработки естественных языков с помощью нейронных сетей.

Abstract: this review article discusses the problems associated with machine processing of natural languages, and then provides an overview of the currently existing models of neural networks that somehow solve the problems of natural language processing. In conclusion, the strengths and limitations of these models are given. This article will be useful for those who are faced with the need to solve problems of natural language processing using neural networks for the first time.

Ключевые слова: обработка естественных языков, машинное обучение, глубокое обучение, нейронные сети, искусственный интеллект.

Keywords: natural language processing, machine learning, deep learning, neural networks, artificial intelligence.

Введение. Обработка естественного языка (англ. *Natural Language Processing, NLP*) – развивающееся направление искусственого интеллекта, задачами которого являются компьютерный анализ и синтез естественных языков.

Хотя сначала может показаться, что обрабатывать естественный язык – задача простая, на деле это оказывается не так. Для этого необходимо рассмотреть несколько примеров.

Словообразование. Проблемы здесь возникают для каждого языка свои. Например, в английском языке, несмотря на довольно простые правила образования новых форм слов, существуют исключения в виде неправильных форм прошедшего времени глаголов. В русском языке у одного и того же слова могут быть десятки (а в некоторых случаях и сотни) форм – за счёт

разнообразия падежей, склонений и прочих морфологических приёмов. В немецком языке довольно частое явление – конкатенация нескольких слов с образованием нового длинного слова. В японском языке при записи слов используется сразу два слоговых алфавита, заимствованные из китайского языка иероглифы и латиница; а один и тот же иероглиф может иметь сразу несколько вариантов чтения, или наоборот – одно и то же слово может записываться разными иероглифами в зависимости от значения. А также независимо от языка можно складывать сколь угодно длинные слова, характеризующие конкретное химическое вещество. Таким образом, можно сделать вывод, что максимальная длина слова в любом естественном языке не зафиксирована.

Немного отступимся от темы и рассмотрим другую задачу искусственного интеллекта – компьютерное зрение. На данный момент в этой сфере получены значительные успехи: компьютер может классифицировать объекты лучше человека [1], генерировать лица несуществующих людей и другие объекты [2] и многое другое. И подход, там использующийся, не зависит от размера входных данных: можно масштабировать входную картинку к удобному для нейросети размеру и она сделает своё дело. Если же мы вернёмся к теме обработки естественного языка, мы увидим, что входные данные чувствительны к размеру: мы не можем отмасштабировать слово так, чтобы не потерять его смысл, соответственно, подход из задачи компьютерного зрения неприменим (по крайней мере, в том же самом виде). Вследствие богатых возможностей словообразования в языках также неэффективна идея создать словарь фиксированной длины и кодировать слово одним битом в векторе битов размера словаря (*one-hot representation*). Это не спасёт от проблемы редких и уникальных слов – слов, использующихся во всём тексте (или даже во всём языке) единожды. Примером может служить слово «пятидесятирублируй», смысл которого понятен человеку из контекста, но для компьютера совершенно неведом.

Так мы плавно подошли к теме контекста. Это ещё одна проблема, создающая гораздо большие сложности в обработке естественного языка, чем проблема словообразования. Для примера можно взглянуть на современные автоматические переводчики: хоть за последние годы качество таких переводов и возросло, «живой» переводчик всё равно лучше справится с текстом больше пары предложений – поскольку машине сложно учитывать контекст, как предшествовавший переводимому предложению, так и отсутствующий в тексте, но выводимый из человеческого опыта.

Целью данного исследования является обзор существующих на данный момент технологий обработки естественного языка с помощью глубокого обучения (англ. *Deep Learning*, *DL*).

Word2Vec. Одним из самых известных и популярных подходов по обработке естественного языка является Word2Vec, представленный компанией Google в 2013 году в [3] и [4]. Суть этого метода заключается в преобразовании слов в вектор, отражающий семантические свойства слова. Так, слова, близкие по значению, будут находиться рядом в N-мерном пространстве результирующего вектора. Пример продемонстрирован на рисунке 1. На нём многомерное пространство спроецировано на двумерное по методу главных компонент – таким образом, семантическая близость слов показана с минимальными потерями.

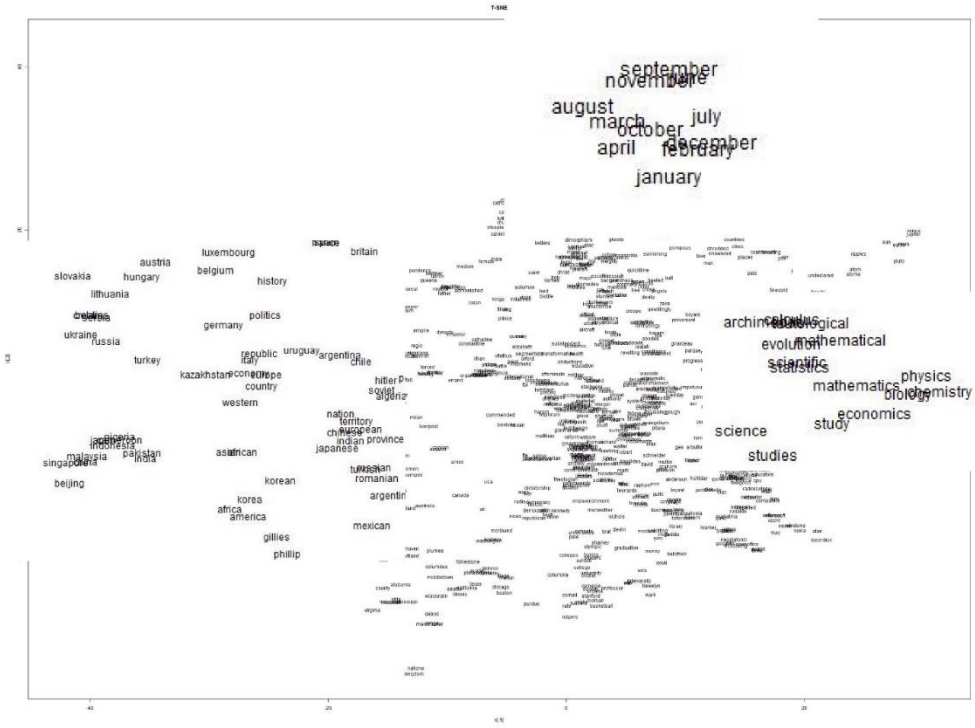

Рисунок 1 – Близость семантически близких слов в конечном векторе

Кроме того, у пар слов со схожим семантическим отношением будет схожее смещение в конечном векторе – пример показан на рисунке 2.

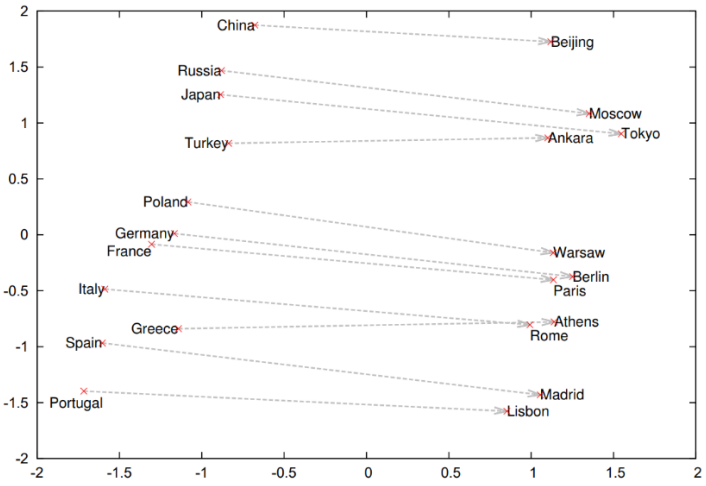

Рисунок 2 – Схожее смещение у пар со схожим семантическим отношением

Также у таких векторов есть и другие свойства: например, одни и те же слова в разных языках при несложных трансформациях дают схожие вектора [5].

Несмотря на такое количество плюсов данный подход обладает существенным минусом – на вход такой модели подаётся тот самый *one-hot representation*, неэффективный при изменяющемся размере словаря.

RNN. Ещё одним популярным и очень эффективным методом обработки естественного языка являются рекуррентные нейронные сети (англ. *Recurrent Neural Network, RNN*). Однако базовый вариант (см. рисунок 3) подвержен проблеме так называемых исчезающих (англ. *vanishing gradients*) и взрывающихся (англ. *exploding gradients*) градиентов.

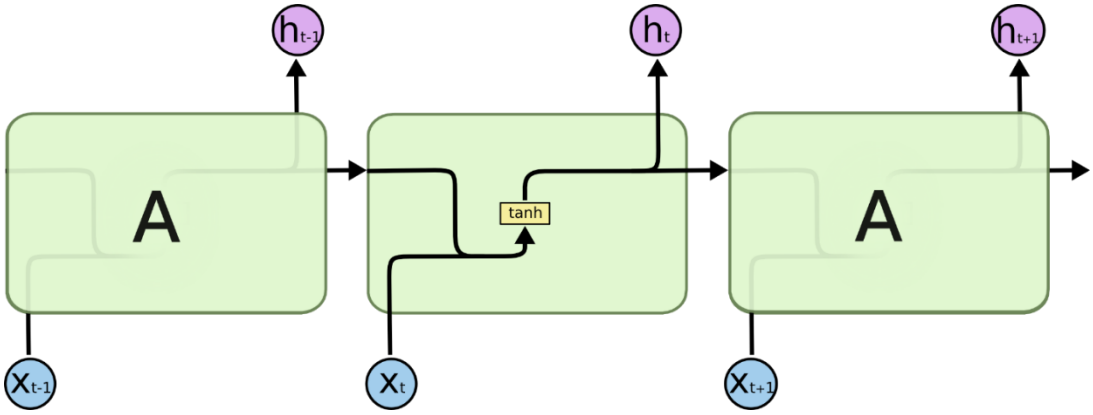

Рисунок 3 – Архитектура стандартной модели RNN

В связи с этим были разработаны две более усложнённые архитектуры: долгосрочная кратковременная память (англ. *Long Short-Term Memory, LSTM*), представленная в декабре 1997 года в [6], а также управляемые рекуррентные блоки (англ. *Gated Recurrent Units, GRU*), представленная в 2014 году в [7]. Современная архитектура стандартной модели LSTM представлена на рисунке

4, а GRU – на рисунке 5. Обе архитектуры устойчивы к проблеме исчезающих и взрывающихся градиентов.

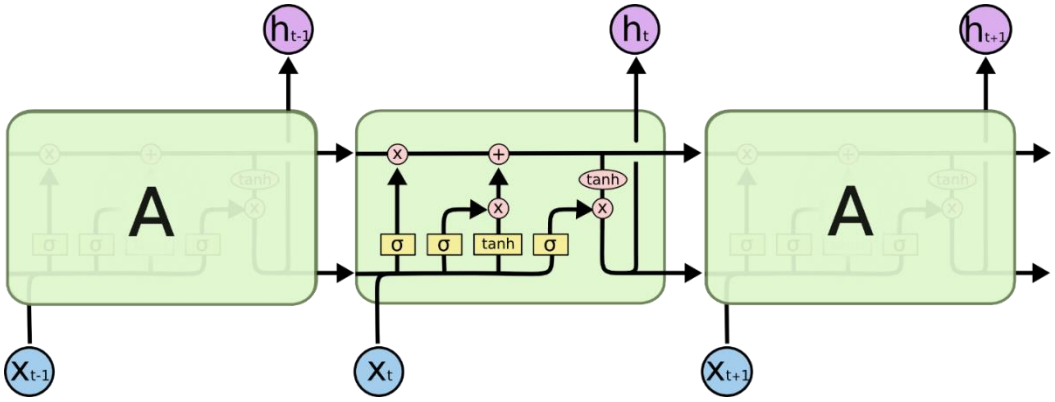

Рисунок 4 – Архитектура стандартной модели LSTM

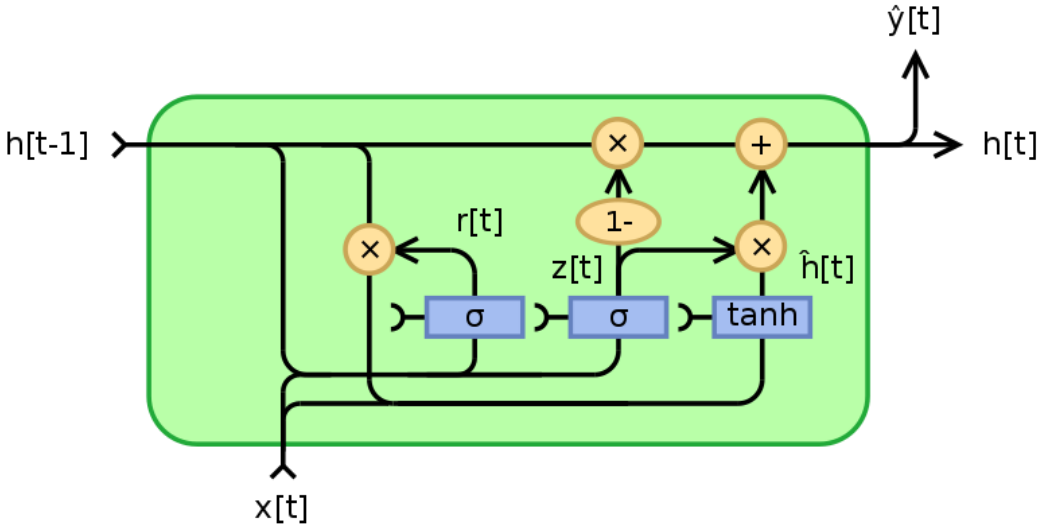

Рисунок 5 – Архитектура стандартной модели GRU

Также предпринимались попытки реализовать более простую архитектуру сети LSTM, но в [8] было показано, что модификации либо уменьшают вычислительную сложность с незначительным ухудшением результатов, либо значительно ухудшают результаты. Пример продемонстрирован на рисунке 6. На нём показано, что восемь различных модификаций не дают улучшения результатов по сравнению со стандартной (V – vanilla) моделью на трёх различных наборах данных.

Однако существуют и «хорошие» модификации, которые хоть и усложняют архитектуру, позволяют показывать превосходные результаты в задачах определения частей речи (англ. *Part of Speech Tagging, POS*) [9] и определения именованных сущностей (англ. *Named Entity Recognition*) [10].

В [11] было показано, что GRU с аналогичными задачами может справляться не хуже, чем LSTM, хотя последнее всё равно остаётся более популярным подходом.

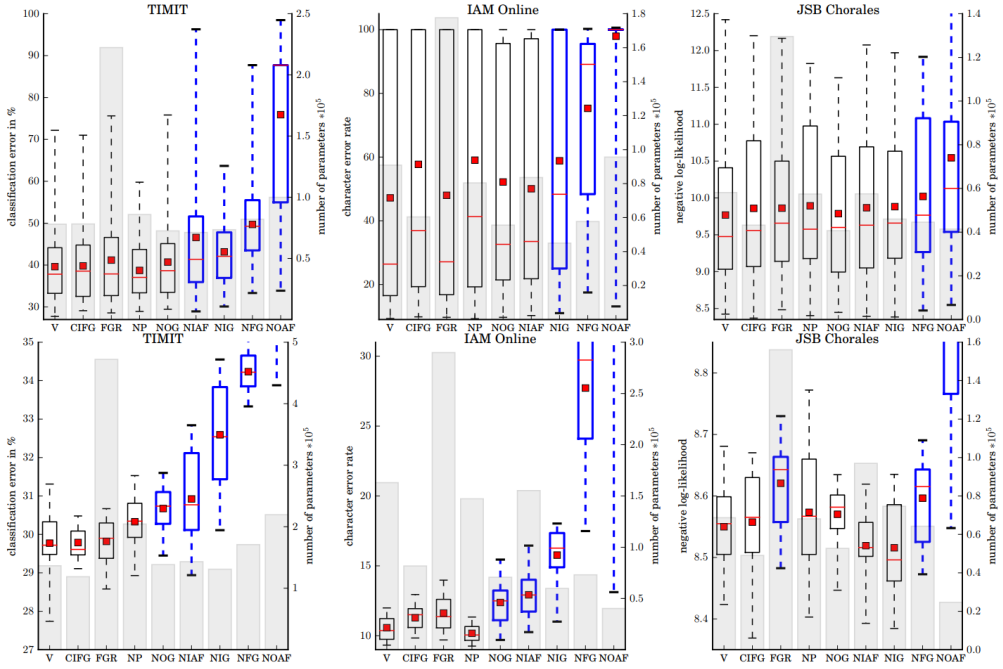

Рисунок 6 – Увеличение ошибки при упрощении стандартной модели (V – Vanilla) LSTM

Если говорить о величине эффективно используемого контекста, то в статье [12] авторы выяснили, что для LSTM это значение порядка 200 токенов. Кроме того, она чувствительна к порядку слов только в ближайшем контексте.

CNN. Ещё одной разновидностью нейронных сетей являются свёрточные нейронные сети (англ. *Convolutional Neural Networks, CNN*). Эти нейросети хорошо зарекомендовали себя в компьютерном зрении (англ. *Computer Vision, CV*), и в данный момент показывают результаты в задаче распознавания образов лучше, чем человек [1], что продемонстрировано на графике на рисунке 7.

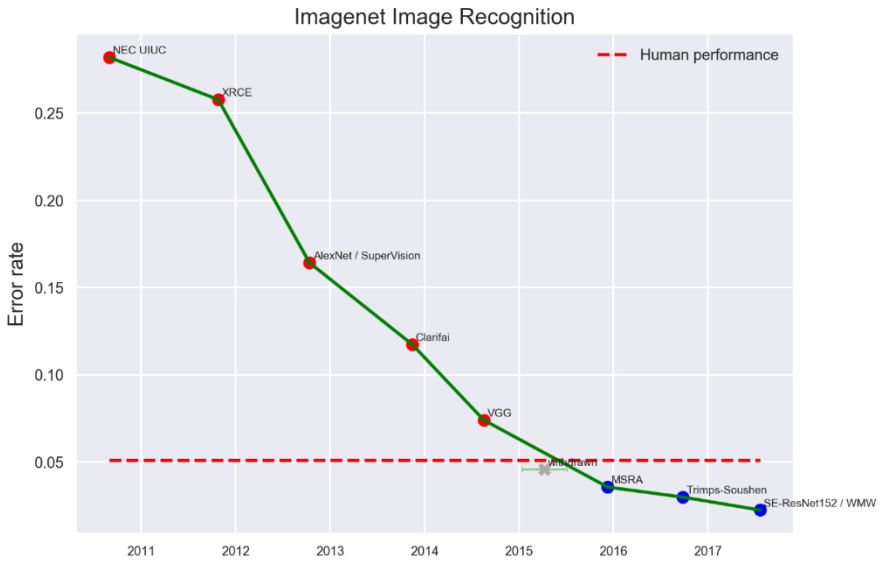

Рисунок 7 – Улучшение результатов нейросетей в задаче распознавания образов

И хотя свёрточные нейронные сети в основном используются в компьютерном зрении, им также можно найти применение в обработке естественного языка, например, в задачах генерации [13] (англ. *speech generation*) и распознавания речи (англ. *speech recognition*) [14]. Структура последней показана на рисунке 8.

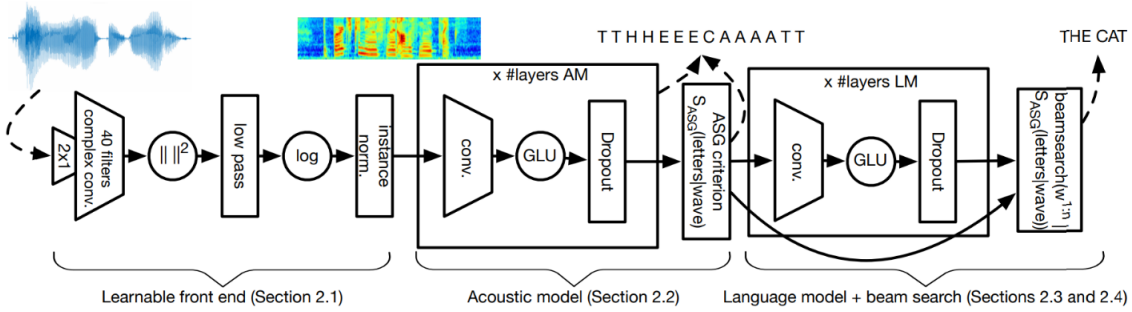

Рисунок 8 – Архитектура полностью свёрточной нейросети для распознавания речи

Attention. Заслуживающей внимания архитектурой нейросетей является *Attention*, впервые представленная в [15]. В работе авторы вместе с этой архитектурой использовали двусторонние рекуррентные нейронные сети, в результате чего добились превосходных (англ. *state-of-the-art*) результатов в переводе предложений с английского на французский. На рисунке 9 показано, что обученная модель показывает хорошие результаты даже на более длинных предложениях, чем были в тренировочных данных.

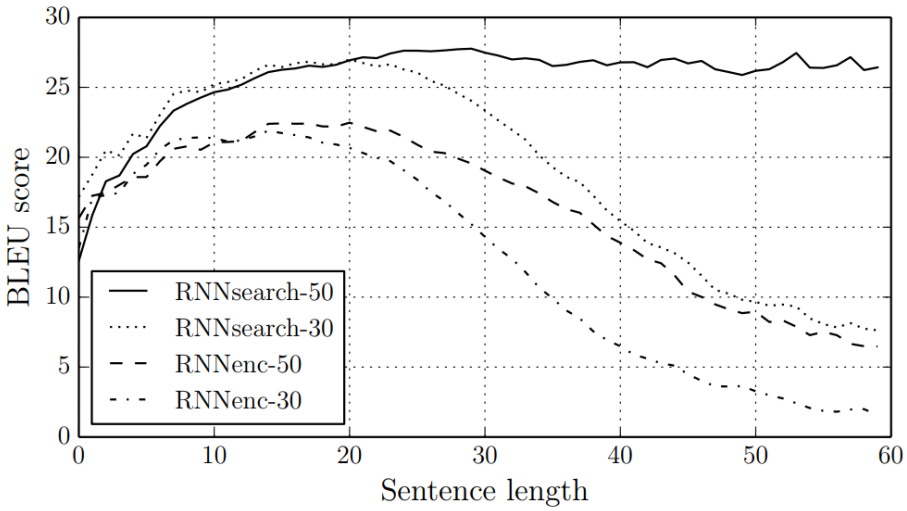

Рисунок 9 – Модель показывает хорошие результаты даже на очень длинных предложениях

Transformer. Развитием архитектуры Attention стала разработка компании Google, которая называется Transformer и представленная в [16]. Разработчики исключили из архитектуры все рекуррентные (RNN) и свёрточные (CNN) слои, и внедрили механизм *self-attention*.

На данный момент эта архитектура показывает одни из лучших результатов в переводе текстов по сравнению с другими архитектурами, что продемонстрировано на рисунках 10 и 11, опубликованных в блоге Google [17].

Также эта архитектура на данный момент используется в Яндекс-переводчике, о чём сообщил Антон Фролов – заместитель руководителя отдела машинного перевода Яндекса в интервью [18, тайм-код: 32:25].

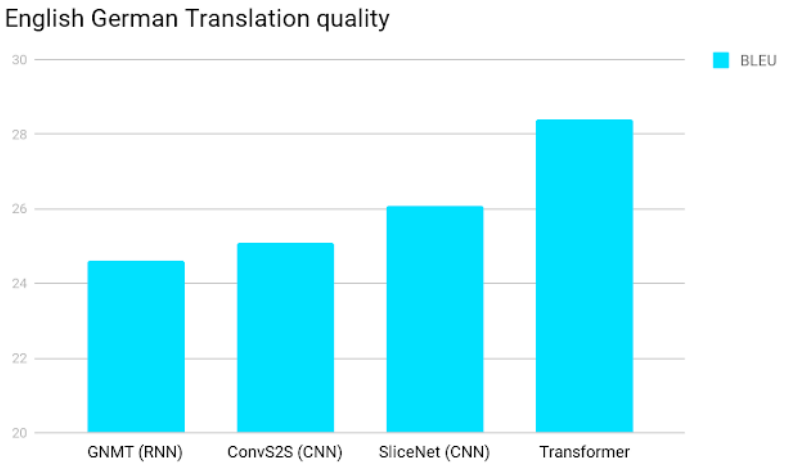

Рисунок 10 – Лучшие результаты в переводе с английского на немецкий у Transformer по сравнению с другими архитектурами (чем выше показатель BLEU, тем лучше)

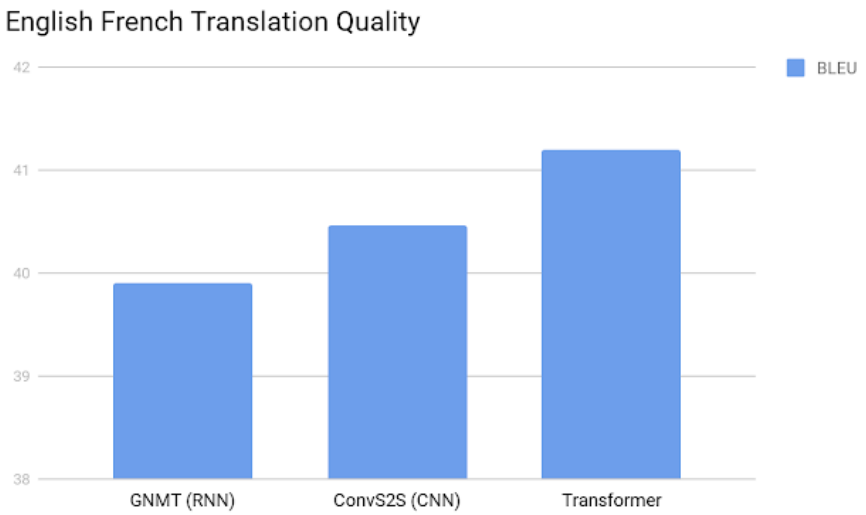

Рисунок 11 – Лучшие результаты в переводе с английского на французский
у Transformer по сравнению с другими архитектурами (чем выше показатель BLEU, тем лучше)

HTM. Отдельно стоящей категорией нейронных сетей является иерархическая временна́я память (англ. *Hierarchical Temporal Memory, HTM*). Фундаментальное отличие от методов глубокого обучения заключается в том, что в HTM нейросеть не сводится к упрощённой математической модели, а моделирует структуру, гораздо более похожую на неокортекс человеческого мозга. В связи с этим обучение здесь происходит не с помощью градиентного спуска и обратного распространения ошибки, а по специальному алгоритму. Этот вид нейросетей разрабатывается компанией Numenta с Джеффом Хокинсом во главе. Первое упоминание HTM было в книге [19], а подробный алгоритм обучения опубликован в [20]. В статье [21] было показано, что при использовании графических процессоров (GPU) для распараллеливания алгоритм может использоваться в режиме реального времени.

Как можно выяснить из названия, иерархическая временная память позволяет выделять распределённые во времени паттерны, т.е. позволяет работать с последовательностями. В качестве входных данных HTM получает разреженные распределённые представления (англ. *Sparse Distributed Representations, SDR*), которые имеют реальный биологический аналог – кортиев орган человеческого уха. Он состоит из множества волосковых клеток. Каждый из этих волосков реагирует на свой диапазон частот. Таким образом, примерно одинаковые звуковые колебания всё время активируют примерно одинаковый набор волосков. В какой-то мере это похоже на свёрточные нейронные сети. Благодаря этому принципу можно кодировать

самые различные типы данных [22], такие как числа, категории, геопространственные данные (англ. *geospatial data*), данные естественного языка (слова, предложения и документы) и даже множества значений. Пример показан на рисунке 12. Как было продемонстрировано в [23], такой вид входных данных позволяет кодировать достаточно большое количество паттернов с достаточно малой вероятностью ложных точных совпадений.

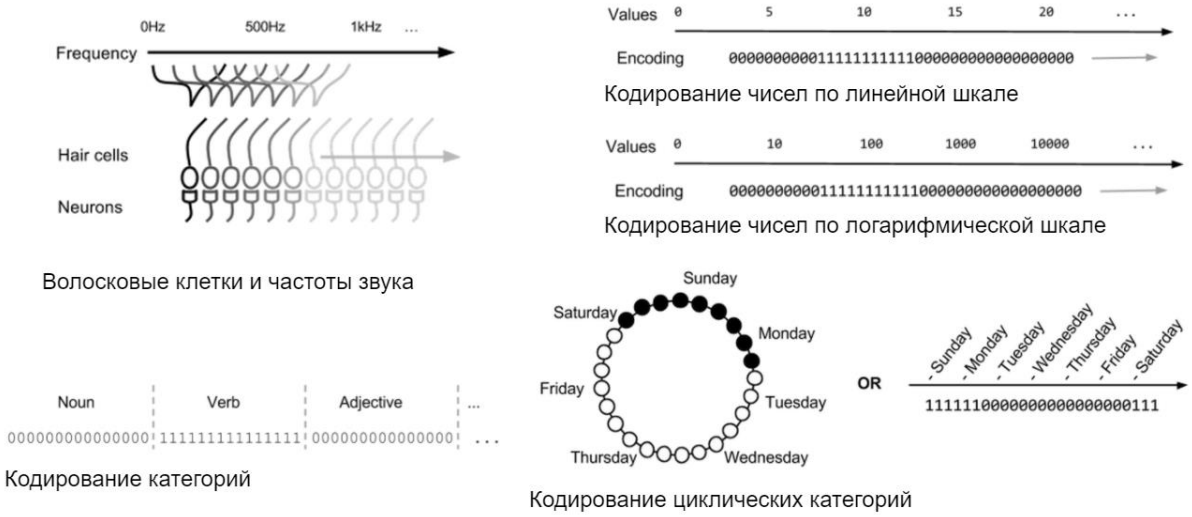

Рисунок 12 – Способы кодирования некоторых типов данных с помощью Sparse Distributed Representations (SDR)

В документе [24] подробно рассказывается о том, как применять иерархическую временную память в задачах обработки естественного языка – методы семантического сворачивания (англ. *Semantic Folding*) и семантических отпечатков (англ. *Semantic Fingerprinting*).

Семантические отпечатки слов имеют интересное свойство: их можно складывать, чтобы получить семантический отпечаток целого предложения. Пример продемонстрирован на рисунке 13.

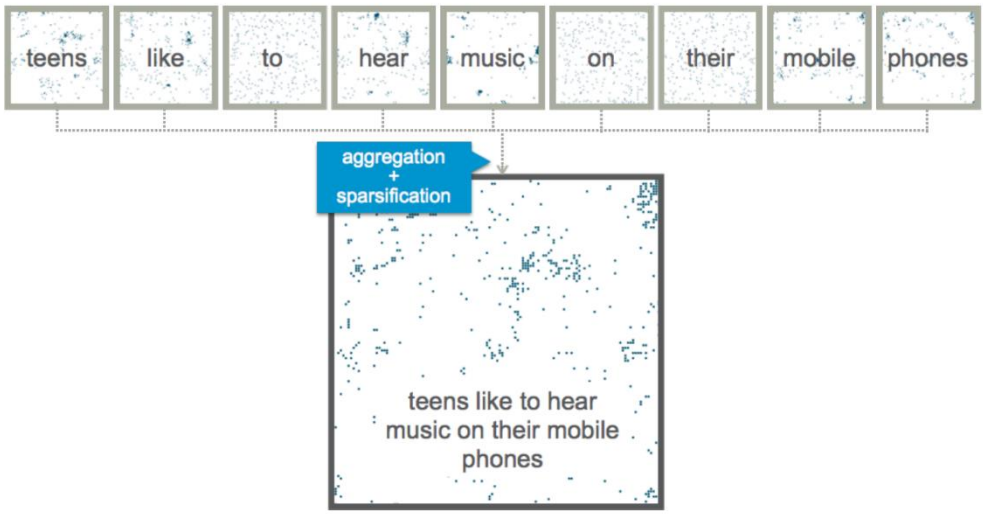

Рисунок 13 – Семантический отпечаток предложения получается путём сложения (aggregation) и разрежения (sparsification) отпечатков слов

Кроме того, семантические отпечатки обладают свойствами Word2Vec: с отпечатками можно производить различные операции (рис. 14); у семантически близких слов похожие отпечатки (рис. 15); в разных языках отпечатки у одного и того же слова могут быть схожими (рис. 16).

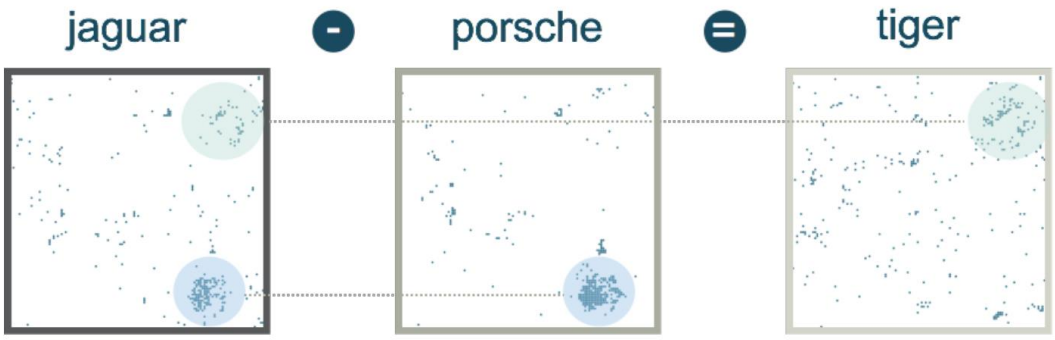

Рисунок 14 – Над семантическими отпечатками можно производить различные операции

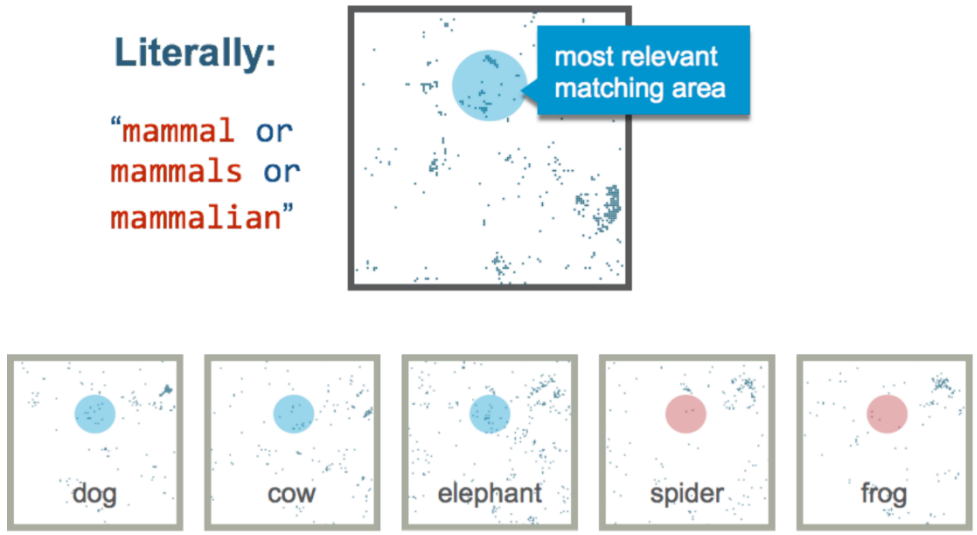

Рисунок 15 – Классификация с помощью семантических отпечатков

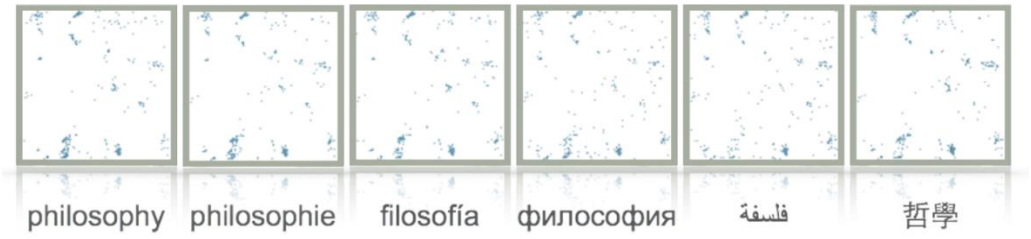

Рисунок 16 – Схожие семантические отпечатки у одного и того же слова в разных языках

Заключение. В данной статье были рассмотрены многие из ныне существующих моделей нейронных сетей, способных решать задачи обработки естественного языка. Модели Word2Vec и HTM хорошо отражают семантические свойства слов, однако, первая ограничена фиксированным размером словаря, а вторая на фундаментальном уровне отличается от других широко используемых подходов. Рекуррентные нейронные сети могут решать задачи обработки последовательностей, однако, из-за своих особенностей «забывают» слишком далёкий контекст. Свёрточные нейронные сети способны решать задачи обработки речи, причём на отличном уровне. Архитектуры Attention и Transformer хорошо справляются с задачей перевода текста.

Литература/ Reference

1. Measuring the Progress of AI Research [Электронный ресурс]. – Режим доступа: URL: https://www.eff.org/ai/metrics
2. Progressive Growing of GANs for Improved Quality, Stability, and Variation [Электронный ресурс] / T. Karras, T. Aila, S. Laine, J. Lehtinen. – arXiv.org: 26.02.2018. – Режим доступа: arXiv: 1710.10196

3. Efficient Estimation of Word Representations in Vector Space [Электронный ресурс]: Архив эл. публ. науч. статей / T. Mikolov, K. Chen, G. Corrado, J. Dean. – arXiv.org: 07.09.2013. – Режим доступа: arXiv: 1301.3781
4. Distributed Representations of Words and Phrases and their Compositionality [Электронный ресурс]: Архив эл. публ. науч. статей / T. Mikolov, I. Sutskever, K. Chen и др. – arXiv.org: 16.10.2013. – Режим доступа: arXiv: 1310.4546
5. Mikolov T. Exploiting Similarities among Languages for Machine Translation [Электронный ресурс]: Архив эл. публ. науч. статей / T. Mikolov, Q. V. Le, I. Sutskever. – arXiv.org: 2018. – Режим доступа: arXiv: 1309.4168
6. Hochreiter S. Long Short-Term Memory [Текст] / S. Hochreiter, J. Schmidhuber. – Neural Computation, 15.11.1997. – Том 9, выпуск 8, стр. 1735-1780. – Режим доступа: DOI: 10.1162/neco.1997.9.8.1735
7. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation [Электронный ресурс]: Архив эл. публ. науч. статей / K. Cho, B. van Merrienboer, C. Gulcehre и др. – arXiv.org: 03.09.2014. – Режим доступа: arXiv: 1406.1078
8. LSTM: A Search Space Odyssey [Электронный ресурс]: Архив эл. публ. науч. статей / K. Greff, R.K. Srivastava, J. Koutník и др. – arXiv.org: 04.10.2017. – Режим доступа: arXiv: 1503.04069
9. Huang Zh. Bidirectional LSTM-CRF Models for Sequence Tagging [Электронный ресурс]: Архив эл. публ. науч. статей / Zh. Huang, W. Xu, K. Yu. – arXiv.org: 09.08.2015. – Режим доступа: arXiv: 1508.01991
10. Chiu J.P.C. Named Entity Recognition with Bidirectional LSTM-CNNs [Электронный ресурс]: Архив эл. публ. науч. статей / J. P.C. Chiu, E. Nichols. – arXiv.org: 19.07.2016. – Режим доступа: arXiv: 1511.08308
11. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling [Электронный ресурс]: Архив эл. публ. науч. статей / J. Chung, C. Gulcechre, K. Cho, Y. Benjio. – arXiv.org: 11.12.2014. – Режим доступа: arXiv: 1412.3555
12. Sharp Nearby, Fuzzy Far Away: How Neural Language Models Use Context [Электронный ресурс]: Архив эл. публ. науч. статей / U. Khandelwal, H. He, P. Qi, D. Jurafsky – arXiv.org: 12.05.2018. – Режим доступа: arXiv: 1805.04623
13. WaveNet: A Generative Model for Raw Audio [Электронный ресурс]: Архив эл. публ. науч. статей / A.v.d. Oord, S. Dieleman, H. Zen и др. – arXiv.org: 19.09.2016. – Режим доступа: arXiv: 1609.03499

14. Fully Convolutional Speech Recognition [Электронный ресурс]: Архив эл. публ. науч. статей / N. Zeghidour, Q. Xu, V. Liptchinsky и др. – arXiv.org: 17.12.2018. – Режим доступа: arXiv: 1812.06864
15. Bahdanau D. Neural Machine Translation by Jointly Learning to Align and Translate [Электронный ресурс]: Архив эл. публ. науч. статей / D. Bahdanau, K. Cho, Y. Bengio. – arXiv.org: 19.05.2016. – Режим доступа: arXiv: 1409.0473
16. Attention Is All You Need [Электронный ресурс]: Архив эл. публ. науч. статей / A. Vaswani, N. Shazeer, N. Parmar и др. – arXiv.org: 06.12.2017. – Режим доступа: arXiv: 1706.03762
17. Transformer: A Novel Neural Network Architecture for Language Understanding [Электронный ресурс]. – Режим доступа: URL: https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html
18. «Прямой эфир» с Антоном Фроловым [Видеозапись]: Интервью с заместителем руководителя отдела машинного перевода Яндекса / компания Яндекс. – 25.09.2017. – Режим доступа: URL: https://vk.com/video-11283947_456239228
19. Hawkins J. On Intelligence: How a New Understanding of the Brain will Lead to the Creation of Truly Intelligent Machines [Текст] / J. Hawkins, S. Blakeslee. – United States: Times Books, 2004. – 272 с.
20. Hawkins J. Hierarchical Temporal Memory including HTM Cortical Learning Algorithms [Электронный ресурс] / J. Hawkins, S. Ahmad, D. Dubinsky. – Redwood City, California, US: Numenta, Inc, 2011. – Режим доступа: URL: https://numenta.com/assets/pdf/whitepapers/hierarchical-temporal-memory-cortical-learning-algorithm-0.2.1-en.pdf
21. Pietron M. Parallel Implementation of Spatial Pooler in Hierarchical Temporal Memory [Текст] / M. Pietron, M. Wielgosz, K. Wiatr. – Proceedings of the 8th International Conference on Agents and Artificial Intelligence, 2016. – Том 2, стр. 346-353. – Режим доступа: DOI: 10.5220/0005706603460353
22. Purdy S. Encoding Data for HTM Systems [Электронный ресурс]: Архив эл. публ. науч. статей / S. Purdy. – arXiv.org: 18.02.2016. – Режим доступа: arXiv: 1602.05925
23. Ahmad S. Properties of Sparse Distributed Representations and their Application to Hierarchical Temporal Memory [Электронный ресурс]: Архив эл. публ. науч. статей / S. Ahmad, J. Hawkins. – arXiv.org: 25.03.2015. – Режим доступа: arXiv: 1503.07469

24. Webber F.D.S. Semantic Folding Theory And its Application in Semantic Fingerprinting [Электронный ресурс]: Архив эл. публ. науч. статей / F.D.S. Webber. – arXiv.org: 2018. – Режим доступа: arXiv: 1511.08855