Петрова А. Н., Фролов Д. О.
A. N. Petrova, D. O. Frolov

РАЗРАБОТКА МОДЕЛИ ДЛЯ РАНЖИРОВАНИЯ ОБЬЕКТОВ В СИСТЕМАХ БОЛЬШИХ ДАННЫХ

DEVELOPMENT OF A MODEL FOR OBJECT RANKING IN BIG DATA SYSTEMS

Петрова Анна Николаевна – кандидат технических наук, заведующая кафедрой «Проектирование, управление и развитие информационных систем» Комсомольского-на-Амуре государственного университета (Россия, Комсомольск-на-Амуре). E-mail: PetrovaAN2006@yandex.ru.
Anna N. Petrova – PhD in Engineering, Head of Design, Management and Development of Information Systems Department, Komsomolsk-na-Amure State University (Russia, Komsomolsk-on-Amur). E-mail: PetrovaAN2006@yandex.ru.

Фролов Дмитрий Олегович – аспирант Комсомольского-на-Амуре государственного университета (Россия, Комсомольск-на-Амуре). E-mail: optcompanys@mail.ru.
Dmitriy O. Frolov – Graduate Student, Komsomolsk-na-Amure State University (Russia, Komsomolsk-on-Amur). E-mail: optcompanys@mail.ru.

Аннотация. В данной статье предлагается новый подход к обучению векторных представлений текста с использованием многозадачной модели глубоких нейронных сетей. В отличие от существующих методов, которые часто ограничены в использовании контролируемых данных из одной задачи, подход к обучению векторных представлений текста с использованием многозадачной модели глубоких нейронных сетей позволяет использовать контролируемые данные из различных задач. Основной акцент ставится на задачи семантической классификации и поиска информации по смыслу, демонстрируя успешное интегрирование этих задач в единую модель. В статье представлены подробности архитектуры многозадачной модели и обсуждаются её преимущества в контексте улучшения производительности по сравнению с базовыми моделями. Оценка модели проводится на крупномасштабных реальных наборах данных, где для классификации запросов используется площадь под кривой ROC, а для веб-поиска – нормализованный дисконтированный совокупный выигрыш. Полученные результаты подтверждают эффективность предложенного подхода и его превосходство над существующими методами в различных задачах обработки естественного языка.

Summary. This paper proposes a new approach for learning vector representations of text using a multi-task deep neural network model. Unlike existing methods, which are often limited in using supervised data from a single task, our approach allows the use of supervised data from different tasks. The main emphasis is placed on the tasks of semantic classification and information retrieval by meaning, demonstrating the successful integration of these tasks into a single model. The paper presents details of the multitasking model architecture and discusses its benefits in terms of performance improvements over baseline models. The model is evaluated on large-scale real-world datasets where the area under the ROC curve is used for query classification and normalized discounted cumulative gain is used for web search. The results obtained confirm the effectiveness of the proposed approach and its superiority over existing methods in various natural language processing tasks.

Ключевые слова: концепция нейронных сетей для ранжирования документов, использование функции потерь по спискам, оптимизация обучающих данных на основе запросов.

Key words: concept of neural networks for ranking documents, using the list loss function, optimizing training data based on queries.

УДК 517.95

Введение. Последние прорывы в области глубоких нейронных сетей подчеркнули значимость изучения векторных представлений текста, включая слова и предложения, для различных задач обработки естественного языка. Однако существующие методы обучения таким представле-

ниям всё ещё далеки от оптимальных. Большинство предшествующих подходов базируется на неконтролируемых задачах, таких как предсказание слов в процессе обучения. Другие методы прибегают к контролируемым целям обучения для решения одной задачи, что ограничивает их доступ к объёму обучающих данных. В данной статье предлагается многозадачный подход глубоких нейронных сетей для обучения векторным представлениям с помощью контролируемых данных из различных задач. Помимо получения преимуществ от использования большего объёма обучающих данных, многозадачность также способствует регуляризации, что снижает переоснащение модели под конкретную задачу и делает изученные представления универсальными для различных задач.

Предлагается применение многозадачной глубокой нейронной сети для формирования представлений с упором на задачи семантической классификации и поиска информации по смыслу. Созданная модель обучается преобразовывать произвольные текстовые запросы и документы в семантические векторные представления в низкоуровневом скрытом пространстве. Данная модель успешно интегрирует разнообразные задачи, такие как классификация и ранжирование, в едином фреймворке. Кроме того, разработанная модель является компактной и гибкой в возможности внедрения в новые области благодаря способности изученных представлений к адаптации предметной области с использованием гораздо меньшего количества меток.

Обучение многозадачному представлению. Многозадачная модель сочетает в себе задачи классификации и ранжирования. Для уточнения: классификация запросов выполняется в роли задачи классификации, а веб-поиск – в качестве задачи ранжирования.

Классификация запросов представляет собой процесс, при котором модель оценивает, принадлежит ли поисковый запрос Q к определённой категории или домену. Например, если запрос Q содержит фразу «стопорное кольцо установки», классификатор должен определить отнесение к категории «Буровые установки». Точная классификация запросов важна для создания персонализированного пользовательского опыта, т. к. поисковая система может адаптировать интерфейс и результаты под конкретные интересы пользователя. Однако это представляет определённые трудности, поскольку запросы обычно короткие. Поверхностные признаки слов, которые могут быть полезны в традиционных задачах классификации документов, зачастую недостаточно информативны для классификации запросов. В этой статье запросы классифицируются по четырём интересующим областям: «Буровые установки», «Нефтедобывающие установки», «Перерабатывающие установки» и «Промышленные установки». Один запрос может относиться к нескольким категориям, поэтому для выполнения классификации создаётся набор двоичных классификаторов, по одному для каждого домена. Проблема формулируется как 4 задачи двоичной классификации. Таким образом, для области C_t целью является бинарная классификация на основе $P(C_t \mid Q)$ ($C_t = \{0,1\}$). Для каждого домена t предполагаются контролируемые данные ($Q, y_t = \{0,1\}$) с y_t в качестве двоичных меток.

Веб-поиск: учитывая поисковый запрос Q и список документов L, модель ранжирует документы в порядке релевантности. Например, если запрос Q – «прибор для промывки золота», модель возвращает список документов, удовлетворяющих такую информационную потребность. Формально оценивается $P(D_1|Q)$, $P(D_2|Q)$, … для каждого документа D_n и ранжируется в соответствии с этими вероятностями. Предполагается, что контролируемые данные существуют, т. е. для каждого запроса Q существует хотя бы один релевантный документ D_n.

Многозадачная модель глубоких нейронных сетей. Предлагаемая нами модель преобразует любые произвольные запросы Q или документы D в набор фиксированных векторных представлений малой размерности с помощью глубоких нейронных сетей. Эти векторы затем могут быть использованы для классификации запросов или веб-поиска. В отличие от существующих методов обучения представлений, которые либо используют неконтролируемые цели, либо ориентированы на одну задачу, модель изучает эти представления, используя многозадачные цели.

Архитектура многозадачной модели глубоких нейронных сетей показана на рис. 1. Нижние уровни применяются к разным задачам, в то время как верхние уровни генерируют выходные данные для конкретных задач. Важно отметить, что входные данные X (запрос или документ), изна-

чально представленные как набор слов, сопоставляются с вектором (l_2) размером 300. Это общее семантическое представление, которое обучается с помощью многозадачных целей.

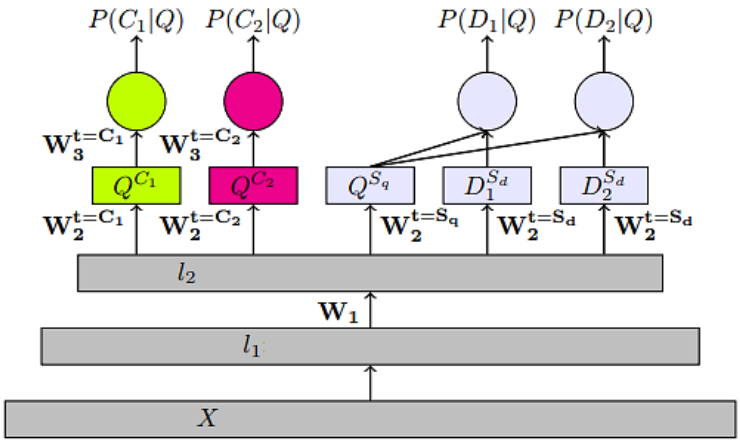

Рис. 1. Архитектура многозадачной
модели глубоких нейронных сетей

Подробный разбор модели. В традиционном подходе каждое слово представляется вектором одного горячего слова, где размерность вектора соответствует размеру словаря. Однако из-за обширного словарного запаса в реальных задачах обучение таких моделей требует значительных временных затрат. Для решения этой проблемы применяется метод хеширования слов, который отображает вектор одного горячего слова с высокой размерностью в ограниченное буквенно-триграммное пространство. Например, слово «*SAP*» хешируется как набор букв триграммы {*#-A-P*, *S-A-P*, *A-P-#*}, где # – граничный символ. Хеширование слов дополняет простое векторное представление в двух аспектах:

1. слова, отсутствующие в словаре, могут быть представлены буквенно-триграммными векторами;

2. различные варианты написания одного и того же слова могут быть сопоставлены с близкими друг к другу точками в буквенно-триграммном пространстве.

Уровень семантического представления (l_2). Это общее представление, полученное при выполнении различных задач. Этот слой отображает входные буквы-триграммы в 300-мерный вектор с помощью

$$l_2 = f(W_1 * l_1),$$

где f (*) – нелинейная активация $f(z) = \frac{1-e^{-2z}}{1+e^{-2z}}$. Эта матрица W_1 размером 50 кБ на 300 отвечает за генерацию семантического представления перекрёстных задач для произвольных текстовых входных данных (например, Q или D).

Представление для конкретной задачи (l_3). Для задачи нелинейное преобразование отображает семантическое представление измерения l_2 в представление измерения для конкретной задачи с помощью

$$l_3 = f(W_2^t * l_2),$$

где t обозначает различные задачи (классификация запросов или веб-поиск).

Наборы данных и метрики оценки. Используются крупномасштабные реальные наборы данных для оценки. Статистические данные представлены в табл. 1. В тестовых данных для веб-поиска содержится 11 000 запросов на русском языке. Каждая пара запрос-документ имеет метку релевантности, которая была вручную аннотирована по пятиуровневой шкале: плохая, удовлетворительная, хорошая, отличная и идеальная. Оценочным показателем для классификации запросов

является площадь под кривой ROC. Для веб-поиска используется нормализованный дисконтированный совокупный выигрыш.

Таблица 1

Статистические данные

Задача	Классификация запросов				Веб-поиск
	Буровые установки	Нефтедобывающие установки	Перерабатывающие установки	Промышленные установки	
Обучение	1.601K	2.201K	2.004K	1.315K	4500 тыс. запросов и переходов по кликам documents
Тест	3.102	6.421	6.240	325	13 051 запрос / 925 600 документов

Результаты точности. Сначала проводится оценка модели на существенное улучшение производительности, которое измеряется точностью при решении нескольких задач.

В табл. 2 приведены совокупные оценки для классификации запросов при сравнении различных классификаторов. Полученные результаты свидетельствуют о том, что предлагаемая многозадачная глубокая нейронная сеть превосходит другие системы.

Таблица 2

Результаты AUC классификации запросов

Система	Классификация запросов			
	Буровые установки	Нефтедобывающие установки	Перерабатывающие установки	Промышленные установки
Модель SVM с функциями слова в форме униграммы, биграммы и триграммы	91.87	81.25	81.15	92.45
Модель SVM с функциями буквенной триграммы	91.54	70.04	85.24	87.53
Однозадачная глубокая нейронная сеть	96.53	76.81	92.24	93.57
Многозадачная глубокая нейронная сеть	**98.25**	**87.53**	**97.14**	**97.15**

Заключение. В заключение можно подчеркнуть значимость предложенного многозадачного подхода глубоких нейронных сетей для обучения векторным представлениям текста. Этот подход демонстрирует превосходство по сравнению с существующими методами, основанными либо на неконтролируемых задачах, либо на однозадачном контролируемом обучении. Основные преимущества многозадачного подхода заключаются в улучшении производительности модели, её способности адаптироваться к различным задачам и регуляризации для снижения переоснащения.

Многозадачная модель, представленная в статье, успешно интегрирует задачи семантической классификации и поиска информации по смыслу, обеспечивая компактность и гибкость в возможности внедрения в новые области. Результаты оценки модели подтверждают её превосходство над другими системами в задаче классификации запросов, а также показывают её высокую производительность при выполнении различных задач.

Таким образом, многозадачный подход глубоких нейронных сетей представляет собой перспективное направление для дальнейших исследований в области обработки естественного языка, обеспечивая эффективное решение сложных задач информационного поиска и классификации текста.

ЛИТЕРАТУРА

1. Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.
2. Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., & Hullender, G. (2005). Learning to rank using gradient descent. In Proceedings of the 22nd International Conference on Machine Learning (ICML) (p. 89-96).
3. Joachims, T. (2002). Optimizing search engines using clickthrough data. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (p. 133-142).
4. Li, H., & Lin, J. (2010). A short introduction to learning to rank. IEICE Transactions on Information and Systems, 94(10), 1854-1862.
5. Zhai, C., & Massung, S. (2016). Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining. Association for Computing Machinery (ACM) Books.
6. Liu, T. Y. (2009). Learning to rank for information retrieval. Foundations and Trends® in Information Retrieval, 3(3), 225-331.
7. Craswell, N., Szummer, M., & Zoeter, O. (2008). An experimental comparison of click position-bias models. In Proceedings of the 17th ACM Conference on Information and Knowledge Management (p. 87-94).
8. Burges, C., Ragno, R., & Le, Q.V. (2007). Learning to rank with nonsmooth cost functions. In Advances in Neural Information Processing Systems (p. 193-200).
9. Zhang, Y., Tay, Y., & Rong, J. (2014). Learning to rank for question retrieval over large-scale Question and Answer archives. In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (p. 1419-1428).
10. Chapelle, O., Metlzer, D., Zhang, Y., & Grinspan, P. (2009). Expected reciprocal rank for graded relevance. In Proceedings of the 18th ACM Conference on Information and Knowledge Management (p. 621-630).