DOI 10.37882/2223–2966.2022.10.11

АВТОМАТИЧЕСКОЕ ОПРЕДЕЛЕНИЕ МЕСТА ДЕЙСТВИЯ ТЕКСТА

AUTOMATIC LOCATION DETECTION OF THE TEXT

V. Efimova

Summary. Over the past five years, deep learning models based on neural networks have achieved impressive results in the task of generating images from text. However, the images are generated with artifacts and still in insufficient resolution for printing. To correct this situation, we will divide the problem of image generation into subtasks, including determining the background of the image. From the text, you can try to understand what place is described or implied in it. This article proposes two methods for obtaining information about the place of action from a given text using natural language processing based on a pre-trained BERT transformer. The first method, called Location Extraction Transformer (LET), is designed to extract words from a text that explicitly mentions the place of action. The second method, called Location Inference Transformer (LIT), is designed to determine the location of an action that is implied in the text, but not directly mentioned. The performance of the proposed algorithms is compared by F1-measure with several existing approaches that can be used to extract information about the location of the text. Based on the results obtained during the comparison, it can be concluded that the proposed LET and LIT models turned out to be better than other algorithms.

Keywords: natural language processing, contextual image synthesis, deep learning, neural networks.

Ефимова Валерия Александровна
Аспирант Университета ИТМО
Россия, г. Санкт-Петербург
valeryefimova@gmail.com

Аннотация. За последние пять лет модели глубокого обучения на основе нейронных сетей добились впечатляющих результатов в задаче генерации изображения по тексту. Однако изображения генерируются с артефактами и все еще в недостаточном для печати разрешении. Чтобы исправить эту ситуацию, разобьем задачу генерации изображения на подзадачи, в числе которых определение фона изображения. По тексту можно попытаться понять какое место в нем описано или подразумевается. В данной статье предлагается два метода получения информации о месте действия из заданного текста с помощью обработки естественного языка на основе предобученного трансформера BERT. Первый метод, называемый Location Extraction Transformer (LET), предназначен для извлечения слов из текста, в котором прямо упоминается место действия. Второй метод, называемый Location Inference Transformer (LIT), предназначен для определения места действия, которое подразумевается в тексте, но не упоминается напрямую. Производительность предложенных алгоритмов сравнивается по F_1-мере с несколькими существующими подходами, которые можно использовать для извлечения информации о месте действия текста. На основе результатов, полученных, при сравнении можно сделать вывод, что предложенные модели LET и LIT оказались лучше, чем другие алгоритмы.

Ключевые слова: обработка естественного языка, контекстуальный синтез изображений, глубокое обучение, нейронные сети.

Введение

Визуальное сопровождение важно для успешного восприятия информации человеком. Кроме того, книги с обложкой и иллюстрациями привлекают больше внимания читателей, что важно для их коммерческого успеха. Иллюстрации к тексту обычно рисуют вручную, что требует времени (несколько дней или месяцев) и человеческих усилий. Затраты по времени можно снизить до нескольких минут и не использовать человеческий труд, автоматически генерируя изображения на основе ключевых моментов книги: действующих объектов (людей, предметов), ситуации и места действия (ландшафта, пейзажа).

Автоматическая генерация изображений на основе текста (англ. text-to-image synthesis) широко используется для редактирования фотографий, создания 2D и 3D персонажей, открыток и других графических материалов. Эта задача решается с помощью моделей машинного обучения на основе генеративно-состязательных сетей (GAN) [1] и вариационных автокодировщиков (VAE) [2]. Несмотря на успехи в генерации, создание сложного изображения с несколькими объектами и четким фоном на основе текста остается сложной задачей [3]. Изображения, сгенерированные по тексту с помощью GAN имеют множество артефактов, а с помощью — VAE размыты [4], обе архитектуры позволяют генерировать изображения в разрешении не более 1024x1024, что недостаточно для качественной печати. Даже современные модели не могут генерировать детализированный фон изображения [5]. Предложим способ генерации изображения более высокого качества.

Введем понятие *локации* как места, в котором происходят описываемые в тексте события. Это может быть место на улице или внутри помещения, то есть комната (кухня, кабинет) или место в городе (улица, площадь), географическая достопримечательность или общие особенности ландшафта (море, гора). Добавление до-

полнительных условий позволит уточнить процесс генерации изображения и синтезировать изображение более высокого качества. Разобьем генерацию изображения по тексту на подзадачи [6]: (1) генерация объектов переднего плана, (2) поиск подходящего фона в открытой базе изображений, (3) размещение объектов на фоне и их гармонизация. В открытых базах изображения размечены с помощью ключевых слов (тегов), которые описывают изображенное и позволяют искать нужные изображения по тексту. Однако, если производить поиск по отрывку художественного текста, то список результатов будет пуст или нерелевантен. Поэтому предскажем по тексту слова, которые описывают место действия.

Цель

Цель данной работы — разработка методов автоматического предсказания места действия текста на основе методов обработки естественного языка и глубоких нейронных сетей.

Метод

Проблема определения местоположения по тексту изучается уже много лет для географических координат [7] и определения местоположения пользователей социальных сетей по их сообщениям [8]. Описанные в статьях подходы дают хорошие результаты, однако, локация в этих работах представлена географическим объектом, который ближе к достопримечательностям, а не к месту действия текста. Для автоматического определения места действия текста рассмотрим два случая: локация явно упоминается в тексте («они встретились у горы» — гора) или подразумевается («их ноги утопали в песке» — пляж, пустыня).

Когда **место действия явно упомянуто в тексте**, его достаточно извлечь. Это можно сделать с помощью распознавания именованных сущностей (Named Entity Recognition, NER) [9] в библиотеке Spacy [10] и с помощью извлечения ключевых слов [11]. Однако, эти подходы не очень точны (см. таб. 1), поэтому воспользуемся другими методами обработки естественного языка (NLP), а именно векторным представлением слов с помощью модели-трансформера BERT [12]. Эта модель позволяет уловить контекстуальные отношения между словами, что достигается благодаря множеству уровней само-внимания. Представив слова в виде векторов BERT, применим к каждому логистическую регрессию, чтобы классифицировать, похоже ли исходное слово на локацию или нет. Таким образом, для каждого слова в предложении мы оцениваем вероятность того, что оно является локацией. Обозначим предложенный метод как **Location Extraction Transformer, LET**.

Когда **место действия текста не упоминается явно**, а только подразумевается, локацию следует предсказать. Например, несколько слов-локаций можно сгенерировать с помощью генеративной модели GPT-2 [13], но это также будет не очень успешно (см. таб. 1). Воспользуемся принципом сиамских сетей [14] и для предложения будем определять, какая локация из словаря ему наиболее подходит (словарь построим из 150 наиболее популярных локаций [15]). Для обучения сети используем метод отрицательной выборки [16]: построим набор триплетов, в котором каждый состоит из предложения, слова-кандидата и числа, равного единице, если слово-кандидат является локацией для данного предложения, и равного нулю иначе. Снова используем модель BERT для извлечения информации о контексте, но теперь используем один вектор для представления всего предложения — выходной вектор, соответствующий токену CLS (классификации), специальному токену, описанному в [17]. Затем мы пропускаем вектор предложения через полносвязный слой с функцией активации ReLU и замораживаем веса в модели BERT для стабилизации процесса обучения. Чтобы получить векторное представление для слов-кандидатов, мы пропускаем их через обучаемый слой векторного представления и вычисляем скалярное произведение между вектором предложения и вектором-кандидатом. После этого сигмоидная функция активации применяется для оценки вероятности того, что слово-кандидат будет локацией для входного предложения. Таким образом, для каждого слова в словаре локаций мы можем определить, как оно связано с текстом. Обозначим предлагаемый метод как **Location Inference Transformer, LIT**.

Результаты

Все эксперименты проводились на сервере с видеокартой NVIDIA RTX 3090 и процессором AMD RYZEN9 3950X с 16 ядрами. Обучение проводилось на синтетическом наборе данных на основе набора изображений MS COCO [18] Набор текстов был сгенерирован автоматически с помощью описаний к изображениям и извлечения из них места действия и ручной доработки. Модель LET обучалась на нем восемь часов, LIT — двенадцать часов.

Чтобы оценить эффективность предложенных методов, сравним их с существующими методами обработки текстов, которые способны решить задачу извлечения или вывода локации. Одним из таких решений является алгоритм EmbedRank [11], который используется для извлечения ключевых фраз. EmbedRank использует статические векторные представления sent2vec [19] текста и ключевых фраз-кандидатов. Их замена на контекстно-зависимые векторные представления привела к значительным улучшениям во многих зада-

Таблица 1. F_1 — мера предсказания локации на тестовых наборах данных

Метод	Извлечение локации	Вывод локации	Общее	COCO
Spacy	0,85	0,02	0,74	0,83
EmbedRank × sent2vec	0,52	0,25	0,48	0,18
EmbedRank × RoBERTa Large	0,34	0,15	0,32	0,11
EmbedRank × RoBERTa Base	0,32	0,13	0,03	0,11
EmbedRank × DistilBERT Base	0,39	0,3	0,38	0,11
EmbedRank × DistilBERT msmarco	0,4	0,13	0,37	0,13
EmbedRank × LaBSE	0,38	0,15	0,35	0,14
GPT-2	0,22	0,13	0,2	0,14
GPT-2 x EmbedRank	0,23	0,15	0,22	0,15
LET	0,89	0	0,74	0,85
LIT	0,3	0,28	0,3	0,76
LET+LIT	0,91	0,28	0,81	0,88

чах NLP [20], протестируем подобный подход и в задаче предсказания места действия текста для векторных представлений RoBERTa [21], DistilBERT [22], LaBSE [23]. Все векторные представления использовались со стандартными параметрами, так как настройка не дала заметного эффекта. Кроме того, была протестирована модель GPT-2 в сочетании с методом выделения ключевых слов EmbedRank и без него. Несмотря на то, что локации не относятся к определенным частям речи, было проведено сравнение с библиотекой spaCy, позволяющей частично выделять локации с помощью частеречной разметки и выделения именованных сущностей.

Все решения были протестированы на собранном наборе реальных текстов из блогов и книг, разделенном на две части для извлечения и предсказания локации (всего 300 текстов), и тестовой части синтетического набора данных на основе MS COCO (всего 23898 текстов, в тестовой части 10%), описанного выше. Результаты оценки по стандартной для задач классификации метрике — F_1-мере — представлены в таблице 1, максимально значение в столбце выделено жирным.

Как видно из таблицы 1, существующие решения плохо справляются с обеими задачами предсказания локации. В то время как обе предложенные модели LET и LIT показывают высокие результаты при решении соответствующих задач, а объединение LET и LIT имеет наивысшую точность на всем наборе данных, что доказывает их эффективность.

Выводы

При генерации изображений по тексту нейронные сети генеративно-состязательной архитектуры создают нереалистичные изображения низкого качества. Качество генерируемых изображений можно повысить, разделив задачу на несколько подзадач, включая нахождение ключевых объектов текста и места действия, что упрощает решение задачи.

В статье предложены методы предсказания места действия текста (локации, в которой оно происходит), а именно извлечения локации (если она содержится в тексте) и вывода локации (если она явно не упоминается в тексте) на основе модели-трансформера BERT. Предложенные модели сравнивались с подходом извлечения ключевых слов, извлечением именованных сущностей и генерацией текста по F_1-мере. Сравнение показало, что предложенные модели справляются с предсказанием локации по тексту лучше аналогов. Предложенные методы можно усовершенствовать, расширив обучающий набор данных, продолжив обучение или применив оптимизацию гиперпараметров.

ЛИТЕРАТУРА

1. Creswell A. et al. Generative adversarial networks: An overview //IEEE signal processing magazine. — 2018. — Т. 35. — № . 1. — С. 53–65.
2. Doersch C. Tutorial on variational autoencoders //arXiv preprint arXiv:1606.05908. — 2016.
3. Sylvain T. et al. Object-centric image generation from layouts //Proceedings of the AAAI Conference on Artificial Intelligence. — 2021. — Т. 35. — № . 3. — С. 2647–2655.
4. Hinz T., Heinrich S., Wermter S. Semantic object accuracy for generative text-to-image synthesis //IEEE transactions on pattern analysis and machine intelligence. — 2020.

5. Karras T. et al. Analyzing and improving the image quality of stylegan //Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.— 2020.— С. 8110–8119.
6. Efimova V., Filchenkov A. Text-based sequential image generation //Fourteenth International Conference on Machine Vision (ICMV 2021).— SPIE, 2022.— Т. 12084.— С. 125–132.
7. Huang Y. Conceptually categorizing geographic features from text based on latent semantic analysis and ontologies //Annals of GIS.— 2016.— Т. 22.— № . 2.— С. 113–127.
8. Ajao O., Hong J., Liu W. A survey of location inference techniques on Twitter //Journal of Information Science.— 2015.— Т. 41.— № . 6.— С. 855–864.
9. Schmitt X. et al. A replicable comparison study of NER software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate //2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS).— IEEE, 2019.— С. 338–343.
10. Honnibal M., Montani I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing //To appear.— 2017.— Т. 7.— № . 1.— С. 411–420.
11. Bennani-Smires K. et al. Simple unsupervised keyphrase extraction using sentence embeddings //arXiv preprint arXiv:1801.04470.— 2018.
12. Wang A., Cho K. Bert has a mouth, and it must speak: Bert as a markov random field language model //arXiv preprint arXiv:1902.04094.— 2019.
13. Radford A. et al. Language models are unsupervised multitask learners //OpenAI blog.— 2019.— Т. 1.— № . 8.— С. 9.
14. Bromley J. et al. Signature verification using a" siamese" time delay neural network //Advances in neural information processing systems.— 1993.— Т. 6.
15. Top 1000 phrases that customers use to buy images (market research) [Электронный ресурс].— 2020.— URL: https://www.microstockgroup.com/general-stock-discussion/top-phrases-that-customers-use-to-buy-images-(market-research)
16. Mikolov T. et al. Distributed representations of words and phrases and their compositionality //Advances in neural information processing systems.— 2013.— Т. 2.
17. Kenton J.D.M.W.C., Toutanova L.K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding //Proceedings of NAACL-HLT.— 2019.— С. 4171–4186.
18. Lin T.Y. et al. Microsoft coco: Common objects in context //European conference on computer vision.— Springer, Cham, 2014.— С. 740–755.
19. Mikolov T. et al. Efficient estimation of word representations in vector space //arXiv preprint arXiv:1301.3781.— 2013.
20. Papineni K. et al. Bleu: a method for automatic evaluation of machine translation //Proceedings of the 40th annual meeting of the Association for Computational Linguistics.— 2002.— С. 311–318.
21. Liu Y. et al. Roberta: A robustly optimized bert pretraining approach //arXiv preprint arXiv:1907.11692.— 2019.
22. Sanh V. et al. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter //arXiv preprint arXiv:1910.01108.— 2019.
23. Feng F. et al. Language-agnostic bert sentence embedding //arXiv preprint arXiv:2007.01852.— 2020.

Журнал «Современная наука: актуальные проблемы теории и практики»