Генерация текста на основе нейронной сети LSTM

Н.А. Кривошеев[1], К.В. Вик[1], Ю.А. Иванова[1], В.Г. Спицын[1,2]

[1]*Национальный исследовательский Томский политехнический университет, Ленина, 30, Томск, Россия, 634050*
[2]*Национальный исследовательский Томский государственный университет, Ленина 36, Томск, Россия, 634050*

Аннотация

Целью данной работы является исследование качества генерации коротких текстов с помощью нейронной сети LSTM на русском и английском языках. Для обучения нейронной сети используются следующие подход оценка максимального правдоподобия (Maximum Likelihood Estimation, MLE). Предложен и реализован подход на основе возведения значений вектора вероятностей в степень больше 1, данная операция позволяет увеличить качество генерируемого текста, но снижает его разнообразие. Обучение и тестирование проводятся на основе следующих выборок данных: сборника русских стихов с сайта Stihi.ru и подписей к изображениям на английском языке из выборки COCO Image Captions. Проведена оценка качества генерации текстов на основе метрики. Приведены примеры сгенерированных текстов. Проведен анализ аналогичных решений. На основе полученных результатов сделан вывод, что алгоритм MLE способен генерировать короткие тексты.

Ключевые слова

MLE, генерация текста, возведения значений вектора вероятностей в степень

1. Введение

Целью данной работы является оценка качества автоматической генерации коротких текстов на основе сети с долгой краткосрочной памятью (Long Short-Term Memory, LSTM) [1]. Для обучения нейронной сети LSTM используется контролируемое обучение на основе метода максимального правдоподобия (maximum likelihood estimation, MLE [2]).

Длина генерируемых текстов составляет 10 и 20 слов. Для обучения и тестирования нейронных сетей используются следующие выборки данных: сборник русских стихов с сайта Stihi.ru [3] и подписи к изображениям на английском языке из выборки COCO Image Captions [4]. Применяется пословная генерация текста.

Качество генерации текста оценивается с использованием метрики BLEU [5]. Было проведено обучение и тестирование подходов на основе MLE. На основе представленных результатов можно сделать вывод, что нейронная сеть, обученная на основе MLE, позволяет генерировать тексты близкие по качеству с примерами из обучающей выборки по метрике BLEU.

2. Программная реализация

В рамках данной работы была создана программная реализация исследуемых моделей на языке Python и проведены тестовые эксперименты. В реализации используется библиотека PyTorch. Программная реализация представлена на сайте [6].

2.1. Оценка качества генерации текстов по метрике BLEU

Для оценки качества генерации текстов в данной работе используется метрика BLEU, реализованная в библиотеке nltk.

При оценке по метрике BLEU используется сглаживание на уровне предложений, что позволяет получить более корректные оценки качества текста. Но данная модификация не позволяет сравнить результаты с аналогичными работами, представленными в данной статье, в которых она не использовалась.

2.2. Результаты тестирования нейронной сети LSTM обученной на основе MLE

Было проведено тестирование нейронной сети LSTM, обученной на основе MLE на выборках данных: стихи с сайта Stihi.ru [3] и подписи к изображениям из выборки COCO Image Captions [4]. Тестирование проводилось на тестовых выборках. Для тестирования использовалось 500 сгенерированных примеров. Оценка качества проводилась на основе метрики BLEU. Результаты тестирования представлены в табл. 1:

Таблица 1

Результаты тестирования LSTM обученной на основе MLE

Выборка	BLEU-2	BLEU-3	BLEU-4	BLEU-5
Stihi.ru	0,719	0,376	0,186	0,115
COCO Image Captions	0,523	0,384	0,248	0,149

3. Заключение

В данной работе рассмотрен и реализован подход MLE [2] для генерации коротких текстов на основе нейронной сети LSTM [1]. Рассматриваемый подход протестирован на выборках данных: сборник русских стихов с сайта Stihi.ru [3] и подписи к изображениям из выборки COCO Image Captions [4]. Нейронная сеть LSTM, обученная на основе MLE, генерирует примеры близкие по качеству к реальным примерам из обучающей выборки, по метрике BLEU.

4. Литература

[1] Hochreiter, S. LONG SHORT-TERM MEMORY / S. Hochreiter, J. Schmidhuber [Электронный ресурс] – Режим доступа: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.676.4320&rep=rep1&type=pdf (дата обращения 20.10.2020).

[2] Cramer, J.S. Econometric Applications of Maximum Likelihood Methods / J.S. Cramer. – Cambridge University Press, 1986.

[3] StihiData [Электронный ресурс]. – Режим доступа: https://github.com/DenisVorotyntsev/StihiData/tree/194107ff98249fd11e8da5c3ee2d (дата обращения 20.10.2020).

[4] Chen, X. Microsoft COCO Captions: Data Collection and Evaluation Server / X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, C.L. Zitnick [Электронный ресурс]. – Режим доступа: https://arxiv.org/pdf/1504.00325.pdf (дата обращения 20.10.2020).

[5] Papineni, K. BLEU: a Method for Automatic Evaluation of Machine Translation / K. Papineni, S. Roukos, T. Ward, W.-J. Zhu [Электронный ресурс]. – Режим доступа: https://www.aclweb.org/anthology/P02-1040.pdf (дата обращения 20.10.2020).

[6] Программная реализация SeqGAN [Электронный ресурс]. – Режим доступа: https://github.com/NikolayKrivosheev/Generation-of-short-texts-SeqGAN (дата обращения 10.12.2020).