УДК 004.7

Фильтрация текстов сгенерированных нейронной сетью

Е. Ю. Котляров*

* *Кафедра прикладной информатики и теории вероятностей,*
Российский университет дружбы народов,
ул. Миклухо-Маклая, д.6, Москва, Россия, 117198

Email: 1032161509@pfur.ru

В настоящее время тексты, в том числе новости, могут быть сгенерированны нейронными сетями, что может привести к распространению недостоверной информации. Усложняет ситуацию то, что сгенерированная и опубликованная статья становится мгновенно доступной тысячам людей в сети интернет, поэтому очень важно проверять подобную информацию на достоверность, что особенно актуально для новостных источников, так как их репутация напрямую зависит от публикуемой ими информации. В данной работе будет рассмотрен вариант решения проблемы классификации текстов на натуральные и сгенерированные. Основной целью работы будет выявление возможности отличать тексты которые были сгенерированны нейронной сетью, от текстов которые написали люди. В качестве генератора будет использоваться нейронная сеть архитектуры GPT-2, обученная на массиве текстов полученных с различных веб-сайтов. Для обучения классификатора, помимо данных полученных из генератора, будут использоваться тексты схожие с данными на которых обучался генератор. То есть тексты взятые с различных сайтов. По результатам работы будет сделан вывод об эффективности используемых методов и их недостатках. Данное исследование может быть использовано в новостных источниках для выявление достоверности предоставленной новости.

Ключевые слова: нейронные сети, классификация, GPT-2.

1. Введение

За последние десятки лет алгоритмы генерации текста сильно продвинулись в развитии. На данный момент лучший результат показывают нейронные сети. Современное развитие нейронных сетей для генерации текста позволяют генерировать текст, по которому сложно понять написал его человек, или сгенерировал алгоритм. Это может привести к тому, что люди смогут генерировать рассказы или новости которые будут мало отличимы от своих аналогов написанных человеком. Что позволяет, например, создавать в автоматическом режиме новости которые будут похожи на настоящие, но по факту не будут такими являться. Становится актуальна задача автоматического разделения текстов которые написала нейронная сеть и текстов которые написал человек. Дальше всех в генерации текстов продвинулись разработчики компании OpenAI с нейронной сетью GPT-2 [1], которая является второй версией сети GPT [2]. Данная сеть доступна в нескольких версиях. Они отличаются между собой количеством признаков внутри нейронной сети.

В данной работе будет рассмотрена возможность классификации текстов на написанные человеком и сгенерированные сетью GPT-2 с 117 миллионами параметров внутри сети.

2. Основная часть

Данные

Сеть GPT-2 обучалась на текстах написанных человеком, взятых с различных сайтов и разной тематики. Поэтому сгенерированные ей тексты отличаются по тематике и содержанию. Данные написанные человеком взяты с соревнования Telegram Data Clustering Contest(*https : //contest.com/docs/data_clustering*). В

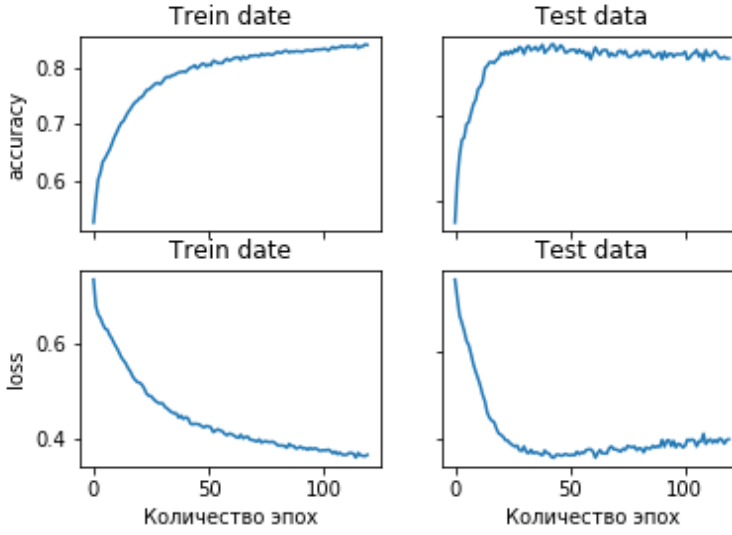

Рис. 1. Качество при обучении

данном соревновании были предоставлены тексты с различных сайтов разной направленности. То есть примерно те же данные что использовались для обучения сети GPT-2.

После предварительной очистки данных от метаданных были получены тексты содержащие только слова. Очистка проводилась с помощью регулярного выражения вида "$[a - zA - Z]$", все пробельные символы и символы переноса строки были заменены на одиночный пробел. Это нужно что бы из текста удалились все символы не являющимися словами, например знаки препинания, знаки переноса строки, числа и эмодзи. Очистка данных позволяет провести токенизацию, то есть разбить весь текст на отдельные слова которые затем можно преобразовывать в вектор. Очистка данных позволяет провести токенизацию, то есть разбить весь текст на отдельные слова, которые затем можно преобразовывать в вектор.

Существуют различные алгоритмы векторизации текстовых данных. Два основных направления это кодирование слов[3] и embeding[4]. В данной работе используется hash кодирование, то есть каждому слову в тексте присваивается индивидуальное случайное число, и получается вектор чисел. Так как большинству алгоритмов машинного обучения, в том числе который используется в этой работе, на вход необходимо подавать вектор фиксированной длины, то если количество слов в тексте меньше какой то фиксированной длины, то вектор дополняется нулями. Данный метод позволяет получить матрицу, количество столбцов которой является фиксированным. Такое представление данных позволяет алгоритмам обрабатывать текстовые данные.

После очистки и обработки данных датасет делится на тестовый и обучающий. В тестовом датасете используется 20% данных от общей выборки. Такое количество данных позволяет получить различные примеры выборки в тестовом наборе, что позволит более объективно оценивать качество модели. Всего в выборке 41 тысяча текстов, поделённых поровну между написанными человеком и сгенерированные

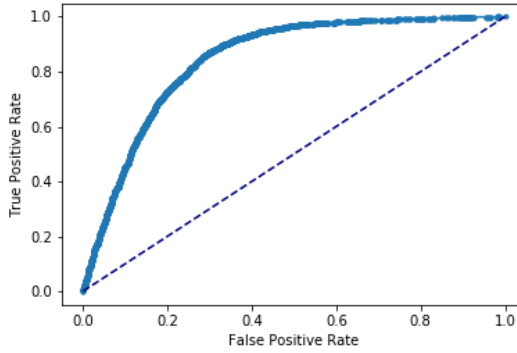

Рис. 2. Кривая roc-auc

нейронной сетью. Равенство количества классов позволяет более эффективно обучать алгоритмы, так как не будет смещения в сторону большего класса.

Модель

В качестве модели используется нейронная сеть, обучающаяся методом обратного распространения ошибки [5]. В качестве функции ошибки используется loglos [6] который выглядит как

$$J = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \log\left(\hat{y}_i\right) + (1 - y_i)\log\left(1 - \hat{y}_i\right)\right)$$

Где y_i это настоящая метка класса, 0 - текст написанный человеком, 1 - текст сгенерированный нейронной сетью. А $\hat{y}_i$ метка класса которую поставила классифицирующая нейронная сеть. Модель содержит 4 скрытых слоя. В первом скрытом слое нейронной сети в качестве функции активации используется гиперболический тангенс[7].

$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}.$$

Во втором и третьем скрытом слое использовалась функция активации relu[7].

$$relu(x) = max(0, x)$$

Выходным слоем является сигмоида[7]. Она подходит для задачи бинарной классификации, так как выдаёт вероятность принадлежности к одному или другому классу.

$$sigmoid(x) = \frac{1}{1 + e^{-x}}$$

В итоге на выходе получается число, которое выражает принадлежность входного вектора к классу. Так же в нейронной сети слой dropout. Он удаляет часть связей в нейронной сети. Это позволяет бороться с переобучением, так как нейронной сети становится сложнее запоминать обучающий датасет, и она менее подстраивается под оценочную выборку, находя более общие зависимости.

Как видно на Рис. 1, метрика достигает максимального значения в тестовом датасете примерно на 40-ка эпохах. Функция ошибки ожидаемо достигает минимума так же примерно на 40-й эпохе. На обоих графиках видно, что после 40-й эпохи модель начинает переобучаться. Признаком этого служит понижение качества на тестовых данных, при росте качества на обучающих данных.

Для проверки качества использовалась roc кривая. Как видно по Рис. 2, качество классификации достаточно высокое, площадь под кривой равняется 0.76. Это говорит о высокой способности классификатора различать классы.

Итоги

Нейронная сеть справляется с задачей различия текстов написанных человеком и текстов написанных нейронной сетью. Получившееся качество позволяет утверждать что эта задача решаема. По крайней мере для текстов сгенерированных нейронной сетью такого уровня. Эффективным методом оказалось использование различных функций активации на разных слоях внутри нейронной сети.

3. Заключение

В данной работе была рассмотрена задача отличия текстов написанных человеком и сгенерированных нейронной сетью. Разобрана обработка текстовых данных для этой задачи и приведён пример архитектуры классификатора. При рассмотрении рисунков обучения модели был выявлен этап переобучения. Проведена оценка качества классификации с помощью roc кривой. Получившуюся нейронную сеть можно использовать для фильтрации текстов, например в новостных изданиях, которые публикуют новости из различных источников.

Литература

1. Alec Radford, Alec and Wu, Jeff and Child, Rewon and Luan, David and Amodei, Dario and Sutskever, Ilya, Language Models are Unsupervised Multitask Learners, 2019
2. Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, Improving Language Understanding by Generative Pre-Training, 2018
3. Benjamin Bengfort, Rebecca Bilbro, Tony Ojeda, Applied Text Analysis with Python: Enabling Language-Aware Data Products with Machine Learning Bilbro, O'Reilly Media; 1 edition (July 1, 2018), 332 pages
4. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean, Distributed Representations of Words and Phrases and their Compositionality, 2013
5. Goodfellow, Ian; Bengio, Yoshua; Courville, Aaron (2016). 6.5 Back-Propagation and Other Differentiation Algorithms. Deep Learning. MIT Press. pp. 200–220. ISBN 9780262035613.
6. Loss Functions For Binary Classification and Class Probability Estimation Shen, Yi 2005

UDC 004.7

Filtration of texts generated by a neural network

E. Y. Kotliarov*

* *Department of Applied Probability and Informatics*
Peoples' Friendship University of Russia(RUDN University)
Miklukho-Maklaya str. 6, Moscow, 117198, Russia

Email: 1032161509@pfur.ru

Nowadays news articles can be generated by neural networks which might result in the spreading of fake news. What makes matters more complicated is that once an article is published it is immediately available to thousands of people on the internet which is why it is crucial to monitor the authenticity of the news. The aim of this paper is to analyze the solutions for the aforementioned problem. The main objective is to make it possible to

distinguish between the news articles generated by a neural network and the articles written by people. Para 2(still in the making) A neural network of GPT-2 architecture was used as a generator which was trained by a dataset of articles acquired from various websites. A dataset produced by the generator among with the similar datasets utilized for the training of the generator were used to train a classifier . At the end of the research the conclusion is made concerning the effectiveness of the used methods and their disadvantages. The results of this research might be further used to establish the authenticity of the news articles.

Key words and phrases: neural networks, classification, GPT-2.