Фильтрация текстов сгенерированных нейронной сетью

Котляров Е.Ю.

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

Вступление

В настоящее время тексты, в том числе новости, могут быть сгенерированны нейронными сетями. Качество этих текстов не уступает человеческому написанию. Автоматизация процесса генерации и выкладывания в сеть таких текстов может заполнить социальные сети несуществующими новостями или спамом, и используемые фильтры не будут справляться со своей задачей.

В данной работе рассмотрен вариант решения проблемы классификации текстов на натуральные и сгенерированные. В качестве генератора используется нейронная сеть GPT-2 с 117M параметров, обученная на массиве текстов полученных с различных веб-сайтов. Для обучения классификатора, помимо данных полученных из генератора, используются тексты схожие с данными на которых обучался генератор.

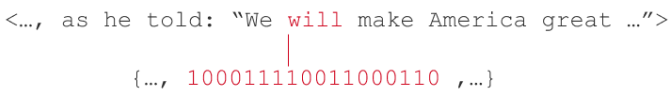
В качестве классификатора используется нейронная сеть прямого распространения. Для её построения применялся язык программирования python3, а так же популярные библиотеки для написания нейронных сетей keras и tensorflow. Данные библиотеки обладают всеми необходимыми инструсентами для проектирования и разработки нейронных сетей. По результатам работы будет сделан вывод об эффективности используемых методов и их недостатках.

Работа с данными

Данные написанные человеком взяты из соревнования Telegram Data Clustering Contest. Датасет состоит из веб страниц разной направленности. В них включены такие категории как новости, культура, спорт, политика, социальная сфера и др.

После предварительной очистки страниц от метаданных были получены исходные тексты, не содержащие html код. Следующим этапом с помощью регулярного выражения вида "[A – zA – Z]+", все пробельные символы и символы переноса строки были заменены на одиночный пробел, а символы не являющимися буквами, например знаки препинания, знаки переноса строки, числа и эмодзи были удалены. Очистка данных позволила провести токенизацию, то есть разбить весь текст на отдельные слова которые затем можно преобразовывать в вектор.

В данной работе используется hash кодирование, то есть каждому слову в тексте присваивается индивидуальное случайное число. Затем каждое слово заменяется на это число, и получается вектор, количество элементов которого равно количеству слов в тексте. 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 этой фиксированной длины, то вектор дополняется нулями. Данный метод позволяет получить матрицу, количество столбцов которой является фиксированным. Такое представление данных позволяет алгоритмам обрабатывать текстовые данные.

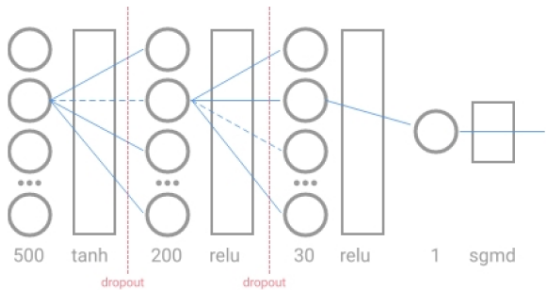


Модель

Для данной задачи средствами keras построена нейронная сеть прямого распространения. Такая архитектура подходит для задачи классификации. На вход сети подаются вектора с закодированными в числа словами, на выходе получается вероятностная оценка принадлежности вектора к тому или иному классу.

В архитектуре были использованы слои активации tanh, relu. Для получения вероятности на выходе сети был использован слой sigmoid с одним нейроном. Так же между слоями был встроен слой dropout который удаляет часть взаимосвязей между слоями, что позволяет бороться с переобучением, так как модели становится сложнее запоминать обучающий датасет, и она менее подстраивается под оценочную выборку, находя более общие зависимости.

В качестве функции ошибки используется logloss. Данная функция ошибки хорошо подходит для задачи бинарной классификации, она позволяет обучать модель на основе вероятностных предсказаний, минимизируя вероятностную ошибку модели. Для минимизации logloss используется adam. Это алгоритм оптимизации с адаптивной скоростью обучения.

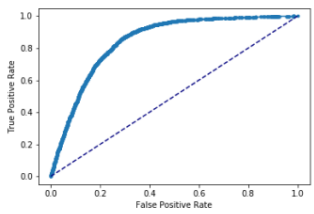


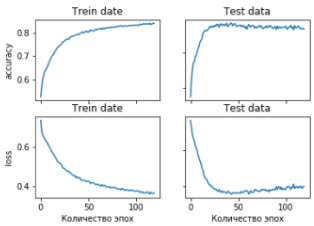
Анализ работы модели

Для тестирования используется 20% данных от общей выборки. Всего в выборке 41 тысяча текстов, поделённых поровну между написанными человеком и сгенерированными нейронной сетью. Равенство количества классов позволяет более эффективно обучать и оценивать алгоритм, так как не будет смещения в сторону большего класса.

Лучшего качества алгоритм достиг примерно на 40-й эпохе. Затем алгоритм начал переобучаться. Признаком этого служит понижение качества на тестовых данных, при росте качества на обучающих данных. Функция ошибки уменьшается достаточно быстро, это означает что алгоритм достаточно хорошо справляется со своей задачей.

Для итоговой проверки качества модели использовалась roc-auc кривая. Качество классификации достаточно высокое, площадь под кривой равняется 0.76. Это говорит о высокой способности классификатора разделять классы.





Итоги

В данной работе была рассмотрена задача классификации текстов написанных человеком и сгенерированных нейронной сетью. Так же была разобрана обработка текстовых данных для этой задачи и показан пример архитектуры классификатора. Проведена оценка качества классификации с помощью roc кривой. Получившееся качество позволяет утверждать что эта задача решаема. 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

SCAN ME