УДК 004.89
https://doi.org/10.36906/AP-2020/33

ОПРЕДЕЛЕНИЕ ТОНАЛЬНОСТИ ТЕКСТОВ ПРИ ПОМОЩИ НЕЙРОННЫХ СЕТЕЙ

Тагиров К. М.
Нижневартовский государственный университет
г. Нижневартовск, Россия
Катермина Т. С.
канд. техн. наук
Нижневартовский государственный университет
г. Нижневартовск, Россия

Аннотация. В данной статье рассматриваются рекуррентные нейронные сети, и примеры решений задач, связанных с обработкой естественного языка

Ключевые слова: искусственные нейронные сети; машинное обучение; анализ естественного языка.

С развитием компьютерных технологий широко распространилась тема виртуального общения. Стали актуальны проблемы, связанные с обработкой естественного языка. А область машинного обучения искусственных нейронных сетей, возникла из вопроса: может ли компьютер выйти за рамки того, «что мы и сами знаем, как выполнять», и самостоятельно научиться решать некоторую определенную задачу [1, 2]? Задачи обработки естественного языка, решаемые искусственными нейронными сетями:

1. классификация текста;
2. извлечение сущностей из текста;
3. реферирование (аннотирование) текста;
4. генерация текста (чат-боты);
5. автоматический перевод.

Именно нейронные сети, на этих задачах, показывают самый лучший результат. Для анализа текста применяются рекуррентные нейронные сети (2); одномерные нейронные сети, сети с вниманием (attention) и другие.

Для анализа нейронной сетью текст необходимо векторизовать. Текст можно разбить на токены: символы, слова и предложения. Далее используется векторное представление токенов, которое основывается на контекстной близости: слова, встречающиеся в тексте рядом с одинаковыми словами (а, следовательно, имеющие схожий смысл), будут иметь близкие (по косинусному расстоянию) векторы. Существуют также несколько методов извлечения признаков текста, которые используются, когда данных не слишком много:

1. N-граммы (последовательности слов от 1 слова до N), например, биграммы — словосочетания;
2. Мешок слов (множество всех слов, встречаемых в тексте);

Плотные векторные представления слов (world embedding) [3] в нейронных сетях определяются в процессе обучения сети. На первом этапе элементы векторов инициализируются случайными числами, далее значения изменяются с помощью метода обратного распространения ошибки. В отличие от векторов, полученных прямым кодированием, — бинарных, разреженных (почти полностью состоящих из нулей) и с

большой размерностью (их размерность совпадает с количеством слов в словаре) — векторные представления слов являются малоразмерными векторами вещественных чисел. Для этого нужно очень много данных и большие вычислительные ресурсы, поэтому используются предварительно обученные векторные представления слов для русского языка это RUSSE (Russian Semantic Evaluation) (https://russe.nlpub.org/downloads/). В векторном представлении можно выполнять векторные операции над представлениями слов и получать осмысленные вектора, где закономерности сохраняются. Например, если рассмотреть вектор «мужчина женщина», и к слову «король» прибавить этот вектор получится — «королева», то есть король относится к мужчине также, как королева к женщине.

Далее для того чтобы не терять смысл таких высказываний в тексте как: «не хороший» — то есть плохой и другие — текст, нужно анализировать как последовательность токенов, так как порядок слов или предложений в тексте имеет большой смысл. Для этого используются специальные архитектуры нейронных сетей — рекуррентные нейронные сети (RNN). Зададим функцию:

$$f_w(h_{t-1}, x_t), \tag{1}$$

где t — время (шаг), x — векторное представление текущего слова, h_{t-1} — вектор, полученный на предыдущем шаге вычислений (для начального подбирается эффективно). Применим данную функцию рекурсивно к скрытым состояниям h_{t-1} (то есть последовательно будем обрабатывать слова предложения, получая новую информацию) как представлено на рис. 1. И так до конца предложения получим 10 состояний — h_i.

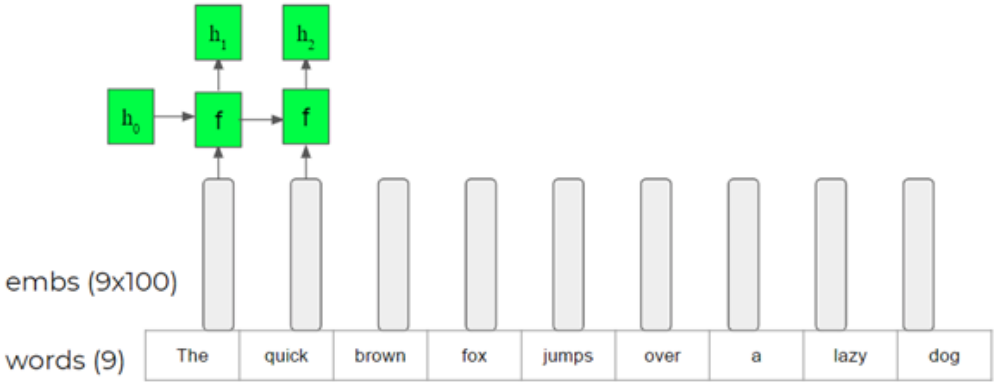

Рис. 1. Рекуррентный метод векторного представления слов

Вместо функции f в формуле (1) используют гиперболический тангенс:

$$f_h(h,x) = \tanh(A_h h + A_x x),$$
$$tanh(x) = \frac{\sinh(x)}{\cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}},$$

где A_h и A_x — это матрицы коэффициентов весов, которые можно домножить на вектора h и x. Сложение происходит поэлементно, поэтому на выходе функции мы получим вектор. Это воспринимают как однослойную нейронную сеть, у которой в качестве функции активации $A(x)$ формула (1) стоит гиперболический тангенс (рис. 2).

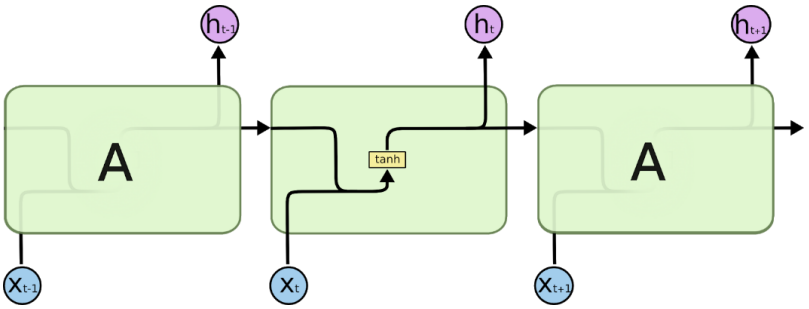

Рис. 2. Структура RNN

Так получается много выходов внутри искусственной нейронной сети RNN, и теперь подставим это в формулы, где выходы представляются векторами:

$$h_t = Wf(h_{t-1}) + W^{(hx)}x,$$

$$y_t = W^S f(h_t),$$

а *W*-различные матрицы весов. Тогда, с учетом функции потерь (ошибок — *E*), производную по матрице *W* можно представить, как сумму производных от суммы всех полученных векторов, тогда из формул (1) получим:

$$\frac{\partial E}{\partial W} = \sum_{t=1}^{T} \frac{\partial E_t}{\partial W},$$

где:

$$\frac{\partial E_t}{\partial W} = \sum_{t=1}^{T} \frac{\partial E_t}{\partial y_t} \frac{\partial y_t}{\partial h_t} \frac{\partial h_t}{\partial h_k} \frac{\partial h_k}{\partial W}.$$

Это слагаемое можно представить, как произведение слагаемых:

$$\frac{\partial h_t}{\partial h_k} = \prod_{j=k+1}^{t} \frac{\partial h_j}{\partial h_{j-1}},$$

следовательно, получим:

$$\frac{\partial E_t}{\partial W} = \sum_{t=1}^{T} \frac{\partial E_t}{\partial y_t} \frac{\partial y_t}{\partial h_t} \frac{\partial h_k}{\partial W} \prod_{j=k+1}^{t} \frac{\partial h_j}{\partial h_{j-1}}.$$

Метод рекуррентных нейронных сетей может обрабатывать текст произвольной длины, учитывая информацию с любого момента времени, в отличии от полносвязной нейронной сети.

Литература

1. Разинкин В. Б., Катермина Т. С. Распознавание лица по фотографии // International journal of advanced studies. 2018. №1-2. С. 171-180.

2. Шолле Ф. Глубокое обучение на Python. СПб. Питер, 2018. 400 с.

3. Mikolov T., Yih W., Zweig G. Linguistic regularities in continuous space word representations // Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies. 2013. P. 746-751.