УДК 004.89

ФОРМАЛИЗОВАННЫЕ МАТЕМАТИЧЕСКИЕ ПАРАДИГМЫ ДЛЯ РЕШЕНИЯ ЗАДАЧ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА

Герман К.О., студент группы ИТб-212, III курс,
Гутова Е.В., старший преподаватель, кафедра математики
Кузбасский государственный технический университет имени Т.Ф. Горбачева, г. Кемерово

В задачах глубокого обучения чаще всего данные стараются представить в виде вектора, так как нейронные сети эффективнее работают с таким видом представления информации [1, 2]. Рассмотрим, каким образом можно представить текст в виде вектора. Текст – это последовательность слов, выстроенная в определенном порядке (в соответствии с лексическими и другими законами).

Если научиться представлять слова в тексте в виде векторов, то сам текст станет аналогом временного ряда (последовательностью, выстроенной по определенным законам и правилам) [3]. Таким образом, появляется возможность векторизации – свертки размерности текста в размерность единичных векторов (рисунок 1).

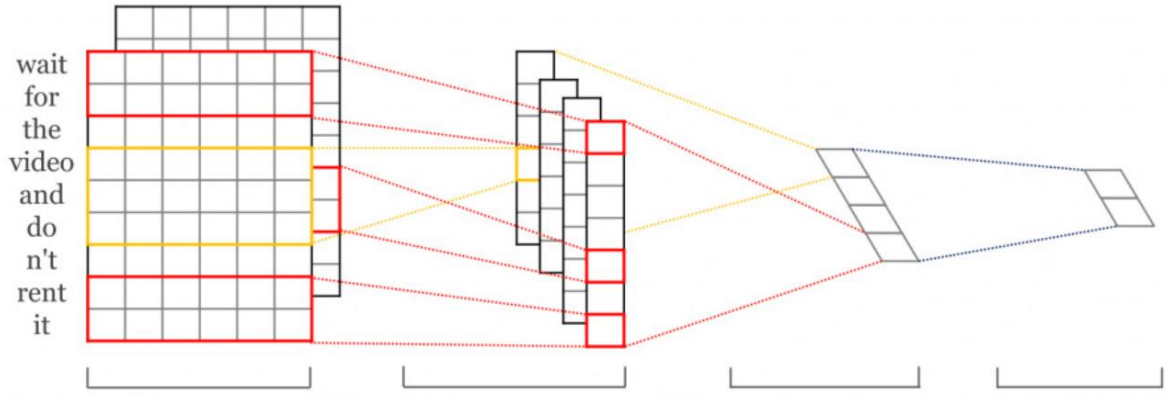

Рисунок 1 – Общая схема векторизации текста

На рисунке 1 представлена общая структура векторизации текстовой информации, подразделенная на составные блоки операций. Первый блок позволяет представлять предложение как совокупность слов, оформленных в виде прямоугольной структуры размерности $n \times k$, где $n -$ это ширина условного прямоугольника, а $k -$ это его длина.

Второй операционный блок выполняет свертку (сверточный слой) по правилам множественных фильтров и карте объектов [4].

Третий операционный слой является объединяющим слоем (от английского «pooling») [5]. Объединяющий слой не имеет обучающихся весовых коэффициентов, его задача сводится к агрегации значений, которые он «видит» в данной окрестности (рисунок 2).

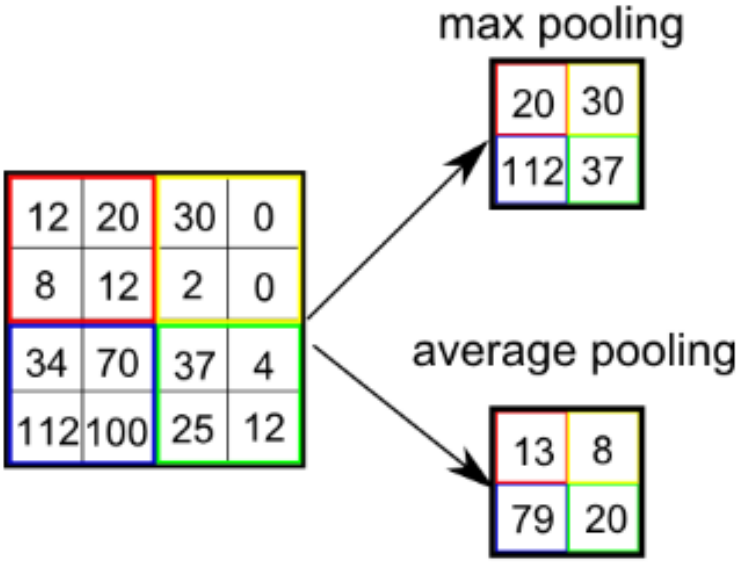

Рисунок 2 – Функционал объединяющего слоя

Например, на рисунке 2, объединяющий слой применяет к каждой ячейке агрегирующую функцию F (она может быть максимизирующей – max pooling, средней – average pooling и любой другой, выполняющей несложные математические операции).

Размер ячейки для свертки задан как 2*2 пикселя, но сам по себе он зависит только от исследователя данных. Свертка позволяет снижать размерность ячейки (агрегирующая функция F «сворачивает» ячейку 2*2 пикселя до размера 1*1 пиксель). В первом случае в качестве функции выбрана максимизирующая, поэтому в качестве единственного значения ячейки 1*1 остается максимальное значение первоначальной ячейки 2*2. Во втором случае функция F является функцией среднего, поэтому единственным значением пикселя остается среднее значение четырех значений.

Наконец, четвертый операционный блок (слой) на рисунке 1 является представлением единичного вектора прецедентов. Единичный вектор характеризует размерность слоя (ширина равна единице), состав же его примерных значений составляет значение $\geq 10^7$ измерений, поскольку данный вектор отражает всю информацию текстового отрывка данных.

Ознакомившись с базовой архитектурой обобщенной нейронной сети для анализа текста (рисунок 1), рассмотрим фундаментальную гипотезу, на которой зиждется обработка естественного языка [6].

Анализ текстов моделями прикладного искусственного интеллекта начал развиваться на основе дистрибутивной гипотезы, которая исследует виды семантической близости слов.

Наиболее логичный вопрос, который сразу возник у истоков машинного анализа текстов – это понимание смысла слова математическими моделями.

В середине прошлого века лингвисты определили, что смысл слова можно определить как совокупность всевозможных контекстов этого слова, в котором оно появляется в языке. Всевозможный контекст – это все возможные предложения, которые могут существовать с этим словом. Таким образом, близкие по смыслу слова появляются в близких контекстах – слово может быть охарактеризовано «компанией слов», в которой оно появляется.

Эти принципы определяют смысл дистрибутивной гипотезы: если слова можно описывать их контекстами, то необходимо формализовать понятие «контекст».

Кроме этого, для дальнейшей формализации, необходимо определить ещё два лингвистических понятия:

- Синтагматическая близость слов (сочетаемость слов в одном контексте, например: здание – строитель; кран – вода; функция – точка);

Рисунок 3 – Схематичное представление синтагматической близости слов

- Парадигматическая близость слов (взаимозаменяемость слов в одном контексте, например: здание – дом; кран – смеситель; функция – отображение).

Рисунок 4 – Схематичное представление парадигматической близости слов

Из рисунков 3 – 4 и их описаний следует, что синтагматическая близость является наблюдаемой, то есть слова можно встретить рядом (в пределах одного предложения), а парадигматическое представление слов – ненаблюдаемая характеристика. Она характеризует взаимозаменяемость слов и очень близка по смыслу к понятию синонимов [7]. Формализуем дистрибутивную гипотезу. Для этого определим начальные условия (1).

$$\left(\begin{array}{c} \text{Дано: текст } (w_1 \dots w_n), \text{который состоит из слов словаря } W \\ \text{Найти: векторные представления слов } v_w \in R^d, \text{так, чтобы} \\ \text{близкие по смыслу слова имели близкие векторы} \\ \text{Модель } CBOW \text{ для оценки вероятности слова } w_i \text{ в заданном контексте} \\ C_i = (w_{i-k} \dots w_{i-1} w_{i+1} \dots w_{i+k}): \\ p(w_i = w|C_i) = SoftMax_{w \in W} \langle u_w, v^{-i} \rangle, \\ \text{где:} \\ v^{-i} = \frac{1}{2k} \sum_{w \in C_i} v_w - \text{средний вектор слов из контекста } C_i, \\ v_w - \text{векторы предсказывающих слов, в общем случае } u_w \neq v_w. \end{array}\right) \quad (1)$$

В качестве исходных данных существует текст на естественном языке. Требуется (1) каждое слово представить таким вектором в пространстве заданной размерности d, чтобы близкие по смыслу слова имели также и близкие векторы.

На данном моменте как раз и необходима рассмотренная ранее дистрибутивная гипотеза, поскольку она позволяет формализовать понятие «близкие по смыслу слова». В системе условий (1) используется модель CBOW (от английского «Continuous bag-of-words» - «непрерывный мешок слов»). В данной модели (и под её определением) имеется в виду то, что порядок слов в тексте не имеет ключевого значения – при этом порядок слов не сохраняется только локально, то есть в окрестности некоторого исследуемого слова.

Далее разрабатывается вероятностная модель $p(w_i = w|C_i)$, которая предсказывает слово в i −ой позиции текста по его контексту. Параметрами данной модели будут векторные представления слов v_w. Разумеется, что само слово исключается из рассмотрения и не принимается в контекст самого себя.

Отметим, что в (1) для каждого слова вводится два альтернативных векторных представления:

- Когда слово находится в контексте, оно предсказывающее, то есть используется для предсказания слова w_i;
- Само слово w_i является предсказываемым и для него принимается другой вектор u_w.

Практически полностью детерминировав задачу системой (1), остается сделать последний шаг – определить критерий максимума правдоподобия (2).

$$\sum_{i=1}^{n} \log p(w_i|C_i) \to max_{U,V} \quad (2)$$

Критерий максимума правдоподобия будет определяться как сумма логарифмов вероятностной модели по всем словопозициям в тексте. Размерность текста определяется матрицей $U \times V$, причем $U, V \in R^{|W| \times d}$, где $W -$ это мощность словаря, а $d -$ это размерность векторных представлений слов.

Список литературы:

1. Методы восстановления непараметрической регрессии в условиях несбалансированных данных / П. А. Пылов, Р. В. Майтак, А. В. Дягилева, А. Д. Салычева. – Вологда : Общество с ограниченной ответственностью "Издательство "Инфра-Инженерия", 2024. – 192 с. – ISBN 978-5-9729-1856-0. – EDN AAJATW.
2. Свидетельство о государственной регистрации программы для ЭВМ № 2023684528 Российская Федерация. Программа интеллектуальной оптимизации процесса добычи тяжелой нефти : № 2023684089 : заявл. 14.11.2023 : опубл. 16.11.2023 / П. А. Пылов. – EDN AJSVLW.
3. Томас Кормен, Чарльз Лейзерсон. Алгоритмы: построение и анализ, 3-е издание – М.: ООО И.Д. Вильямс. 2013. – 1328 с.
4. Дягилева, А. В. Интеллектуальная автоматизация процесса сейсмоакустического мониторинга на основе метода парзеновского окна / А. В. Дягилева, П. А. Пылов, Р. В. Майтак // Вестник научного центра по безопасности работ в угольной промышленности. – 2023. – № 4. – С. 91-94. – EDN JPZLKP.
5. Свидетельство о государственной регистрации программы для ЭВМ № 2023684579 Российская Федерация. Программа интеллектуальной маркировки углеводородных компонентов тяжелой нефти : № 2023684040 : заявл. 14.11.2023 : опубл. 16.11.2023 / П. А. Пылов. – EDN UZMKAX.
6. Дягилева, А. В. Решение задачи превентивного определения токсикоза при выполнении подземных шахтных работ по подтипу клеток крови на основе сверточной нейронной сети / А. В. Дягилева, А. Н. Стародубов, П. А. Пылов // Вестник научного центра по безопасности работ в угольной промышленности. – 2022. – № 2. – С. 28-33. – EDN CSNPKY.
7. Пылов, П. А. Глубокое обучение в задаче ранней диагностики деменции / П. А. Пылов, Р. В. Майтак, А. В. Протодьяконов. – Вологда : Общество с ограниченной ответственностью "Издательство "Инфра-Инженерия", 2024. – 108 с. – ISBN 978-5-9729-2042-6.