ОЛИМПИАДА ШКОЛЬНИКОВ «ШАГ В БУДУЩЕЕ»

НАУЧНО-ОБРАЗОВАТЕЛЬНОЕ СОРЕВНОВАНИЕ «ШАГ В БУДУЩЕЕ, МОСКВА»

30

регистрационный номер

ИУ — Информатика и системы управления
название факультета

ИУ-6 — Компьютерные системы и сети
название кафедры

Написание нейронной сети, определяющей принадлежность текста к заданному естественному языку

название работы

Автор: ***Пешков Дмитрий Витальевич***
фамилия, имя, отчество
ГБОУ школа № 1208, 11 класс
наименование учебного заведения, класс

Научный руководитель: **Минитаева Алина Мажитовна**
фамилия, имя, отчество
МГТУ имени Н.Э. Баумана
место работы
доцент
звание, должность

_____
подпись научного руководителя

Москва - 2019

Аннотация

XXI-й век открывает перед нами множество возможностей по использованию компьютерных систем в тех областях, в которых ранее использовался лишь человеческий разум. К примеру, распознавание естественного языка всегда являлось одной из важнейших проблем лингвистики. Тем не менее, её можно решить, используя программную реализацию математический модели, построенной по принципу организации и функционирования биологических нейронных сетей. Целью проекта является написание искусственной нейронной сети для определения принадлежности предложений к тому или иному естественному языку. Реализация такой сети позволяет упростить работу систем машинного перевода и, таким образом, облегчит работу переводчиков. Кроме того, данная работа является еще одним шагом к полноценному анализу естественных языков и может стать опорой для дальнейших исследований механизмов их формирования и изменения. Обучение сети происходит методом обратного распространения ошибки. В дальнейшем данная нейронная сеть может быть приспособлена для определения принадлежности предложений к практически любому естественному языку. В качестве языка программирования выбранPython из-за своей простоты, лаконичности и выразительности. В результате данной работы написана, обучена и проанализирована искусственная нейронная сеть, а также проведен анализ полученных результатов, на основании которых сделан вывод о созданной нейронной сети. Наиболее важным результатом является возможность использования нейронной сети в различных отраслях лингвистики.

Содержание

ВВЕДЕНИЕ

Цель и задачи работы

Целью данной работы является написание искусственной нейронной сети для определения принадлежности предложений к тому или иному естественному языку. Реализация такой сети позволит утверждать, что создание таких систем существенно упрощает работу ученых и переводчиков, что и является целью данной работы.

Задачи могут быть четко сформулированы в следующих пунктах:

1) изучить и проанализировать теорию создания искусственных нейронных сетей;

2) написать рабочую нейронную сеть на языке Python;

3) обучить нейронную сеть на распознавание предложений определенного языка;

4) подвести итоги эффективности нейронной сети.

Актуальность работы

Нейронные сети можно использовать для решения трудоемких задач, которые требуют аналитических вычислений подобных тем, что делает человеческий мозг, поэтому их использование наиболее оптимально для решения задач по распознаванию, классификации и предсказыванию. В данной работе искусственная нейронная сеть будет использоваться для распознавания принадлежности предложения к естественному языку — проблемы, которая, несмотря на практически повсеместное проникновение английского языка в каждую страну мира, до сих пор стоит остро.

Новизна

Лингвистика, как наука, тесно связанная с человеческой психологией и напрямую отражающаяработу человеческого мозга, мало подчиняется алгоритмизации, а, следовательно, и автоматизации. Тем не менее, с помощью

искусственной нейронной сети можно использовать аналитические вычисления для решения этой проблемы. Впервые ученые задумались о таком использовании нейросетей практически сразу же после формализации правил и теории их создания. Однако активное проникновение нейронных сетей в лингвистику произошло только на рубеже XX-ого и XXI-ого веков; это связано с прогрессом в сфере компьютерных технологий и, как следствие, с повышением мощности компьютеров.

Гипотеза

С помощью искусственной нейронной сети можно создать удобный инструмент для определения принадлежности предложений к естественному языку.

Методы исследования

—теоретические (моделирование);

— эмпирические (изучение источников, анализ полученных сведений, эксперимент).

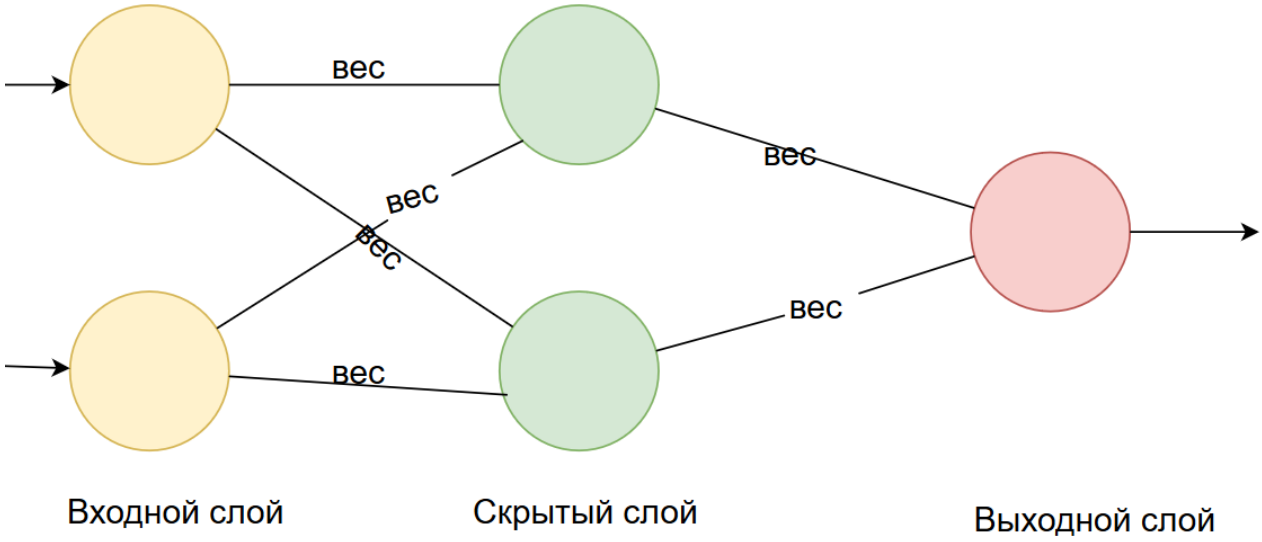
Предполагаемый результат

Предполагается, что данный проект установит наиболее подходящее решение поставленной проблемы использования искусственных нейронных сетей в лингвистике. Данную работу можно использовать как в образовательных, так и в практических целях — для определения принадлежности предложений к естественному языку.

ОСНОВНАЯ ЧАСТЬ РАБОТЫ

Общие сведения о нейронных сетях

Нейронная сеть — математическая модель, созданная по образу биологических нейронных сетей. Сеть представляет собой граф, где вершины являются нейронами, представленный на рисунке 1. Данная нейронная сеть является перцептроном, то есть информация передаётся прямолинейно от входных нейронов к выходным.



(Рис.1)

Так, представленный на данной схеме перцептрон получает входные данные в два входных нейрона (жёлтый), расположенных на входном слое. Значения затем умножаются на соответствующие весы и переходят на следующий слой. В данном примере существует лишь один скрытый слой.

Нейронные сети отличаются:
— по типу обучения:

1. Обучение с учителем — нейронная сеть настраивает весы, зная во время обучения вход и выход (обучающую выборку);

2. Обучение без учителя — нейронная сеть обучается, выискивая закономерности, взаимосвязи и зависимости, существующие между объектами.

— по типу настройки весов (синапсов):

1. Сети с динамическими связями — в процессе обучения сети весы изменяются;
2. Сети с фиксированными связями — весовые коэффициенты выбираются сразу и в дальнейшем не изменяются.

— по архитектуре:

1. Нейронные сети прямого распространения и перцептроны, в которых информация прямолинейно передается от входа к выходу. Каждый слой, который может состоять из входных, скрытых или выходных нейронов, связан с каждыми из нейронов предыдущего и последующего слоёв; нейроны одного слоя между собой не связаны. В нейронных сетях этой архитектуры предпочтительно использование обучения с учителем (по методу обратного распространения ошибки);
2. Нейронная сеть Хопфилда —полносвязная нейронная сеть с симметричной матрицей связей. Во время получения входных данных каждый нейрон является входным, в процессе обучения он становится скрытым, а затем становится выходным. Обучение сети происходит с помощью шаблонизации ситуаций: установленные входные и выходные значения организуются в шаблон, для которого настраиваются весы и к которому позднее сеть приводит остальные ситуации;
3. Машина Больцмана тоже является полносвязной нейронной сетью, однако в ней некоторые нейроны помечены как скрытые, а другие — как входные, которые в дальнейшем становятся выходными. Алгоритм обучения в целом схож с алгоритмом обучения сети Хопфилда. Машина Больцмана имеет архитектурный подвид — ограниченную машину Больцмана, в которой отсутствует симметрия в матрице связей, что

позволяет использовать метод обратного распространения ошибки при обучении сети;

4. Автокодировщики и разреженные автокодировщики представляют собой нейронные сети, в которых производится автоматическое кодирование информации. В случае автокодировщиков скрытых слов меньше, чем входных и выходных, благодаря чему информация сжимается; в разреженных автокодировщиках же наоборот, скрытых слоёв больше. Обе сети можно обучать с помощью метода обратного распространения ошибки;
5. Вариационные автокодировщики — эти сети занимаются приближением вероятностного распределения входных образцов, учитывая «влияние» каждого нейрона в общей сети, отличаясь в этом от всех остальных видов нейронных сетей, представленных здесь;
6. Сетьтипа «deepbelief» — нейронная сеть, которая состоит из нескольких соединенных блоков-сетей. Очевидно, что и обучение происходит поблочно, причем каждый блок кодирует предыдущий;
7. Свёрточные нейронные сети и глубинные свёрточные нейронные сети — в данных типах архитектуры используются сверточные слои, число которых сжимается с определенной глубиной, зачастую являющейся степенью двойки. В глубинных свёрточных нейронных сетях на конце находится связанная с выходным слоем свёрточной сети сеть прямого распространения, используемая для дальнейших вычислений;
8. Развёртывающие нейронные сети — сети, обратные по своей структуре к свёрточным сетям. Их также можно соединять с нейронными сетями прямого распространения.

Используемые методы

1. Использование искусственных нейронных сетей.Искусственные нейронные сети могут быть смоделированы как простыми, так и

сложными — для повышенной точности. В данной работе представленперцептрон с одним скрытым слоем, обновляющий свои весы методом обратного распространения ошибки. Данный выбор обусловлен тем, что перцептроны с одним скрытым слоем достаточно наглядны, что позволяет изучить и исследовать механизм их работы. Кроме того, сети данного типа прекрасно справляются с задачами распознавания. Немаловажным является и тот факт, что данный тип сетей является одним из самых частых и хорошо изученных;

2. Использование языка Python 3.6.4. Язык Pythonбыл создан Гвидо ванРоссумом в 1991 году; в этом языке делается акцент на производительности и удобочитаемости кода. Благодаря гибкости и простоте языка он подходит для статистической сферы и для создания искусственных нейронных сетей.Кроме того, были использованы дополнительные модулиrandom и Math. Первый — расширение языка, позволяющее генерировать псевдослучайные числа, второй — модуль, позволяющий работать с математическими функциями и иррациональными бесконечными числами (так, в данной нейросети использовалось число Эйлера);
3. Представление слов в числовом виде. Для представления слов в числовом виде была написанаотдельная функция, которая разбивает слово на символы, присваивая им индексы по собственной алфавитной таблице;
4. Метод обратного распространения ошибки.Этот метод обучения многослойной нейронной сети также называется обобщенным дельта-правилом. Метод был предложен в 1986 г. Румельхартом, Макклеландом и Вильямсом. Это ознаменовало возрождение интереса к нейронным сетям, который стал угасать в начале 70-х годов. Данный алгоритм является первым и основным практически применимым для обучения многослойных нейронных сетей.Само дельта-правило заключается в следующем — изменение величины весового коэффициента должно быть равно:

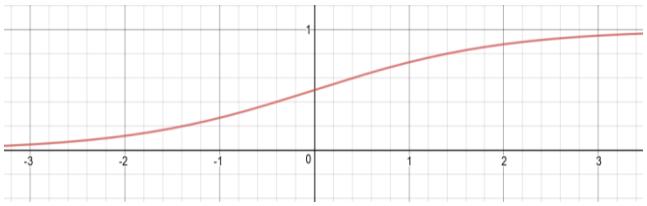
$$\Delta\omega_{jk} = \eta * \delta_k * x_j, \quad (1)$$

Где η — норма обучения, которая задается до начала тренировки; δ_k— ошибка элемента k; x_j — сигнал, приходящий к элементу k от элемента j.

Для выходного слоя корректировка весов понятна — известен конечный результат выходных нейронов, но для скрытых слоев долгое время не было известно алгоритма. Веса скрытого нейрона должны изменяться прямо пропорционально ошибке тех нейронов, с которыми данный нейрон связан. Именно поэтому этот метод позволяет корректно настраивать веса связей между всеми слоями. В этом случае величина ошибок уменьшается и сеть обучается.

СОЗДАНИЕ ИСКУССТВЕННОЙ НЕЙРОННОЙ СЕТИ

Логистическая функция



(Рис.2)

В данной искусственной нейронной сети используется логистическая функция, имеющая формусигмоиды, представленная на рисунке 2:

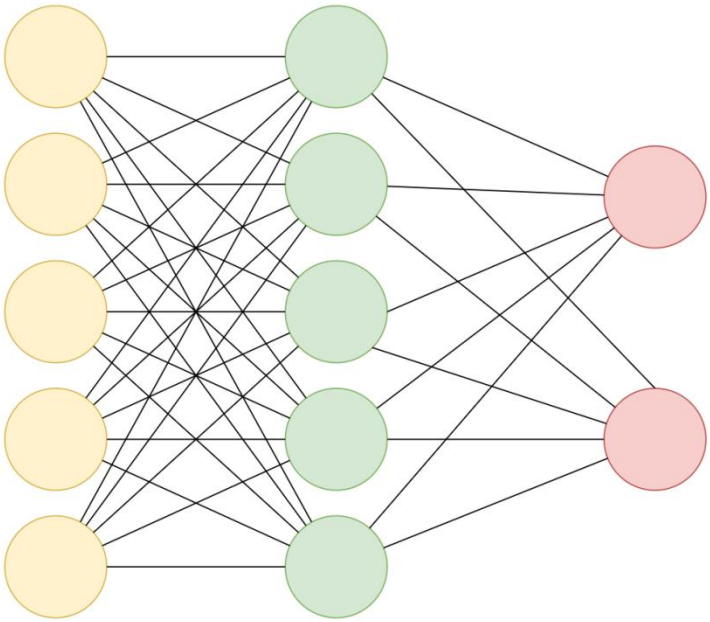
$$y = \frac{1}{1+e^{-x}} \quad (2)$$

Здесь y, x— координаты в двумерном пространстве, e — число Эйлера. Благодаря тому, что эта функция дифференцируема на всей оси абсцисс, она подходит для использования сетью, обучаемой с помощью метода обратного распространения ошибки; а из-за легкости выражения производной функции вычислительная сложность этого метода сокращается.

$$y' = \frac{1}{1+e^{-x}} * \left(1 - \frac{1}{1+e^{-x}}\right) \quad (3)$$

(x,y, аналогично, координаты в двумерном пространстве; e — число Эйлера)

Алгоритм работы нейронной сети



(Рис. 3) Схема ограниченного участка нейронной сети. Желтый цвет – входные нейроны, зеленый цвет – скрытые нейроны, красный цвет – выходные нейроны.

Как уже было упомянуто выше, данная нейронная сеть – трёхслойный перцептрон (рис. 3). На входной слой подается представленное в числовом виде слово, разбитое на символы. Далее сигналы первого слоя умножаются на соответствующие синапсы; эти значения суммируются в нейронах скрытого слоя, имеющих связь с нейронами первого. Сумма подается как аргумент в функцию для нормализации, и значением нейрона второго слоя является уже нормализованное значение. Аналогично происходит и на следующем шаге, и сигналы скрытого слоя, умножаемые на синапсы, суммируются и нормализуются в нейронах второго слоя, который и является выходным слоем.

Алгоритм обучения сети

Нейронная сеть обучается с помощью случайно генерируемых предложений как на английском, так и на русском языках. Предложения генерируются с помощью словарейанглийского и русского языков, содержащих по 10000 слов каждый.

Ошибка в методе обратного распространения определяется по следующей формуле:

$$\delta_j = y' * \sum_k \delta_k * \omega_{kj} \quad (4)$$

Здесь δ_j — ошибка элемента с индексом j;k — индекс, соответствующий слою, который возвращает ошибку; ω_{kj} — весовой коэффициент, y' — производная логистической функции.

Алгоритм метода обратного распространения ошибки следующий:

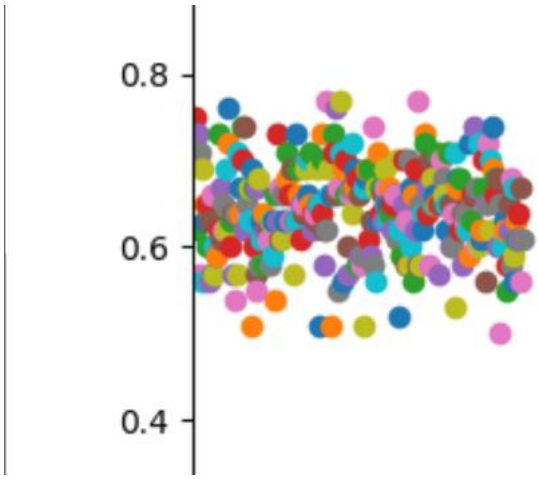
1. инициализировать синаптические веса маленькими случайными значениями;
2. выбрать очередную обучающую пару из обучающего множества; подать информацию на вход сети;
3. вычислить выход сети;
4. вычислить разность между выходом сети и требуемым выходом;
5. подкорректировать веса сети для минимизации ошибки;
6. повторять шаги с 2 по 5 для каждого элемента обучающего множества до тех пор, пока ошибка на всем множестве не достигнет приемлемого уровня.

Результаты

Выходной слой нейронной сети является определяющим для определения языка. Если значение первого нейрона больше, чем значение второго, то язык — английский; в обратном случае же наоборот. Точность нейронной сети является отношением числа верных решений к общему числу слов, где верные

решения — случаи, в которых значения нейронов совпадают с кодом языка, заданным при генерации слова.

Как на малой выборке слов (до 100 слов), так и на большой выборке нейронная сеть доказывает свою эффективность (около 70% точности) (рис. 4,5).



(Рис. 4) Результаты на малой выборке

(Рис. 5) Результаты на большой выборке

Предложения по практическому использованию результатов

Использовать данную нейронную сеть на практике можно в вопросах перевода. Более того, увеличив количество выходных нейронов и переобучив нейронную сеть, можно научить её определять не только эти два языка, но и любые языки, символы которых можно представить в формате Unicode. Данные же по точности нейронной сети можно в дальнейшем использовать для улучшения результативности представленной программы.

Выводы

Данная работа доказывает возможность использования искусственных нейронных сетей в лингвистике, что представляет собой опору для дальнейшего использования этой технологии в вопросах, связанных с естественными языками и их анализом.

Заключение

Итогом работы являетсясозданная искусственная нейронная сеть, которая может определять принадлежность слов к одному из приведенных естественных языков. На практике это означает, что нейросеть справляется со своей задачей, но по-прежнему необходимы дальнейшие исследования с целью получения наибольшей точности и достижения наибольшей эффективности.

Список использованных источников

1. Горбань А. Н. – Обучение нейронных сетей (Москва, СП ПараГраф, 1990)

2. Барский, А.Б. Логические нейронные сети: моногр. / А.Б. Барский. - М.: Бином. Лаборатория знаний / Интернет-Университет Информационных Технологий (ИНТУИТ), 2017.

3. Круглов, В.В. Искусственные нейронные сети. Теория и практика / В.В. Круглов, В.В. Борисов. - М.: Горячая линия - Телеком; Издание 2-е, стер., 2002. - 382 с.

4. Нейронныесети. Statistica Neural Networks. Методология и технологии современного анализа данных. - М.: Горячая линия - Телеком, 2008. - 392 с.

5. Документация языка Python (https://www.python.org/doc/)