**МИНОБРНАУКИ РОССИИ**

**ФГБОУ ВО «СГУ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»**

**ДОКЛАД**

студента 5 курса

факультета компьютерных наук и информационных технологий

Афанасенко Кирилла Павловича

Прогнозирование временных рядов нейронными сетями

кафедра теоретических основ компьютерной безопасности и криптографии

курс *пятый*

семестр *девятый*

Саратов

2022

Введение

Нейронные сети обладают широким спектром возможностей. Они способны решать задачи классификации, кластеризации, распознавания образов, анализ паттернов, принятие решений, прогнозирование. В данной работе будут рассмотрены некоторые способы прогнозирования временных рядов с помощью искусственных нейронных сетей.

**Рассматриваемые нейронные сети**

Основные виды нейронных сетей использующихся для прогнозирования временных рядов:

Рекуррентные нейронные сети

Нейронные сети долгой краткосрочной памяти

Нейронные сети на основе управляемого рекуррентного блока

Свёрточные нейронные сети

**Рекуррентные нейронные сети**

Рекуррентные нейронные сети очень похожи на обычные нейронные сети, в том смысле, что они так же организованы по слоям нейронов, формирующих три слоя – входной, рабочий и выходной. Каждая связь между нейронами имеет свой вес.

Отличие заключается в присвоении нейронов определённому временному шагу. Нейроны в рабочем слое так же направлены по оси коррелирующей со временем. При этом соединены нейроны в рабочем слое исключительно однонаправленно по этой оси с нейроном на следующем временном шаге, а входные и выходные нейроны соединены только с рабочим слоем с таким же шагом. Так как для каждого шага времени присущ свой рабочий нейрон или группа нейронов в каждый временной шаг осуществляются вычисления только нейрона соответствующего этому шагу.

Веса рёбер при этом для каждого шага времени фиксированы. Результат рабочих нейронов определяется результатом работы в предыдущий шаг времени. Выходной вектор при обучении сравнивается с вектором реальных значений после чего функция потерь минимизируется путём многократного прямого применения нейронной сети и вычисления градиента по различным параметрам функции активации. Каждый градиент при этом вычисляется для элементов одного и того же временного шага с одними и теми же значениями полученными из предыдущего шага.

Преимущества рекуррентных нейронных сетей:

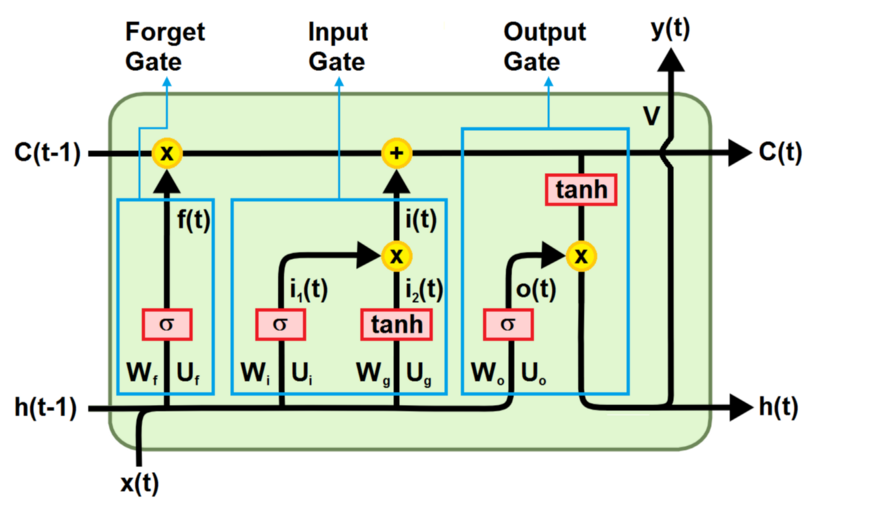
1. Эффективность работы мало зависит от пробелов в значениях.
2. Может находить сложные взаимосвязи во временных рядах.
3. Неплохой результат предсказания при большом количестве шагов.
4. Может смоделировать последовательность значений, где каждое следующее с высокой долей вероятности зависит от предыдущих.

Недостатки:

1. При работе на длинном временном ряде может возникнуть проблема градиентного голода или переполненности – то есть слишком малого изменения параметров или слишком хаотичного.
2. Неспособность учитывать множество прошлых значений.
3. Сложность параллелизации обучения.

**Нейронные сети долгой краткосрочной памяти**

Изначально нейронные сети долгой краткосрочной памяти были созданы для преодоления проблемы градиентного голода. Вместо рабочего слоя используется. Понимание сети долгой краткосрочной памяти значительно облегчается с использованием схемы соответствующей временному шагу



Основной его идеей является наличие состояния ячейки (верхняя горизонтальная линия) и тройного фильтра (анг. gate).

Первым фильтром является фильтр забывания. Информация из предыдущего состояния рабочего слоя изначально поступает именно в него и пропускается через сигмоидальную функцию. Результат близкий к 0 означает, что информацию можно забыть, в то время как близкий к 1 – что информацию необходимо сохранить.

Входной фильтр применяется для обновления состояния ячейки. Текущее состояние и предыдущее состояние пропускаются через сигмоидальную функцию и функцию сжатия данных. После этого результаты прохождения значений через эти функции попеременно перемножаются, а результирующий вектор направляется в обновляющий механизм.

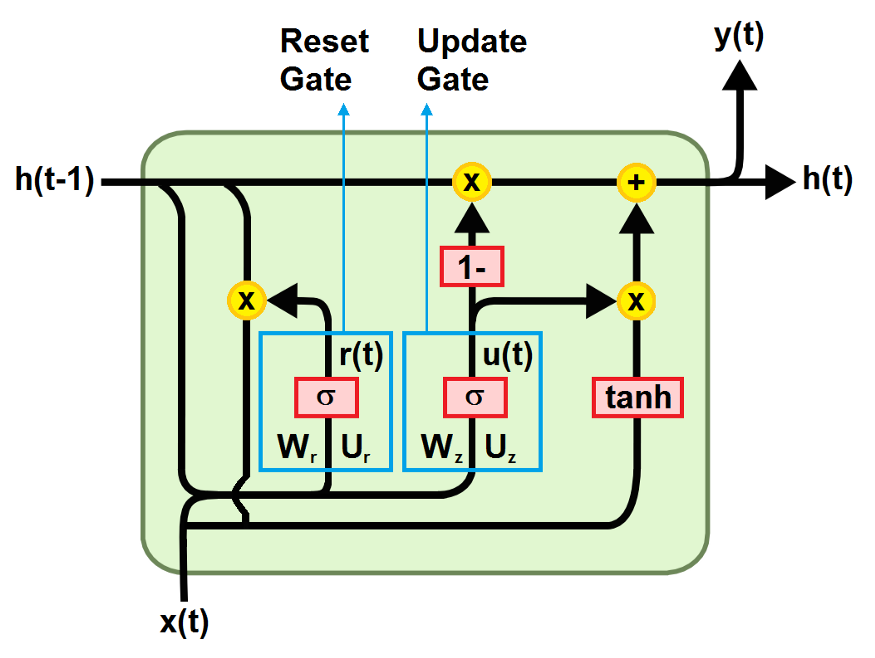
Сам механизм обновления заключается в поэлементном умножении состояния ячейки из прошлого временного шага на вектор фильтра забывания, после чего полученный вектор поэлементно складывается с вектором полученного из входного фильтра.

Выходной фильтр в свою очередь определяет значения поступающие в следующий блок соответствующий шагу времени. Предыдущее состояние пропускается через сигмоидальную функцию, а состояние ячейки через функцию сжатия, после чего результаты этих функций поэлементно умножаются и передаются в следующий рабочий блок.

Не смотря на такой механизм работы, эта нейронная сеть всё ещё не лишена проблем рекуррентных сетей, хотя и сильно ослабляет проблему градиентов за счёт механизма забывания.

**Управляемый рекуррентный блок**

Управляемый рекуррентный блок – похож на нейронные сети долгой краткосрочной памяти, но использует другие методы в решении проблемы градиентов.



Вместо трёх фильтров он использует всего два – перезапуска и обновления (анг. Reset и Update).

Фильтр перезапуска определяет, каким образом будут объединены входные данные и память предыдущего временного шага. Сначала находится их взвешенная векторная сумма, после чего результат пропускается через сигмоидальную функцию.

Фильтр обновления позволяет определить какая информация из текущей должна быть отправлена на следующий временной шаг. Его формирование аналогично фильтру перезапуска, но имеет другие коэффициенты взвешенной векторной суммы.

Текущей памятью в данном случае выступает элемент получаемый поэлементным умножением векторов: поступившего из предыдущего шага с фильтром перезапуска, а затем с вектором обратным фильтру обновления. После чего результат поэлементно складывается с поступившими на вход данными, пропущенными через функцию сжатия и поэлементно умноженными на результат фильтра обновления.

**Свёрточные нейронные сети**

Сеть состоит из большого количества слоёв. После начального слоя (входного изображения) сигнал проходит серию свёрточных слоёв, в которых чередуется собственно свёртка и субдискретизация (пулинг). Чередование слоёв позволяет составлять «карты признаков» из карт признаков, на каждом следующем слое карта уменьшается в размере, но увеличивается количество каналов. На практике это означает способность распознавания сложных иерархий признаков. Обычно после прохождения нескольких слоёв карта признаков вырождается в вектор или даже скаляр, но таких карт признаков становятся сотни. На выходе свёрточных слоёв сети дополнительно устанавливают несколько слоёв полносвязной нейронной сети (перцептрон), на вход которому подаются оконечные карты признаков.

Слой свёртки

Нейроны слоя свёртки, преобразуемые по нескольким выходным каналам

Слой свёртки — это основной блок свёрточной нейронной сети. Слой свёртки включает в себя для каждого канала свой фильтр, ядро свёртки которого обрабатывает предыдущий слой по фрагментам (суммируя результаты поэлементного произведения для каждого фрагмента). Весовые коэффициенты ядра свёртки (небольшой матрицы) неизвестны и устанавливаются в процессе обучения.

Особенностью свёрточного слоя является сравнительно небольшое количество параметров, устанавливаемое при обучении. Так например, если исходное изображение имеет размерность 100×100 пикселей по трём каналам, а свёрточный слой использует фильтры c ядром 3×3 пикселя с выходом на 6 каналов, тогда в процессе обучения определяется только 9 весов ядра, однако по всем сочетаниям каналов, то есть 9×3×6=162, в таком случае данный слой требует нахождения только 162 параметров, что существенно меньше количества искомых параметров полносвязной нейронной сети.

Слой активации

Скалярный результат каждой свёртки попадает на функцию активации, которая представляет собой некую нелинейную функцию. Слой активации обычно логически объединяют со слоем свёртки (считают, что функция активации встроена в слой свёртки). Функция нелинейности может быть любой по выбору исследователя, традиционно для этого использовали функции типа гиперболического тангенса или сигмоиды. Однако в 2000х годах была предложена и исследована новая функция активации — ReLU (сокращение от англ. rectified linear unit), которая позволила существенно ускорить процесс обучения и одновременно упростить вычисления (за счёт простоты самой функции), что означает блок линейной ректификации, вычисляющий функцию. То есть по сути это операция отсечения отрицательной части скалярной величины.

Пулинг или слой субдискретизации

Слой пулинга представляет собой нелинейное уплотнение карты признаков, при этом группа пикселей (обычно размера 2×2) уплотняется до одного пикселя, проходя нелинейное преобразование. Наиболее употребительна при этом функция максимума. Преобразования затрагивают непересекающиеся прямоугольники или квадраты, каждый из которых ужимается в один пиксель, при этом выбирается пиксель, имеющий максимальное значение. Операция пулинга позволяет существенно уменьшить пространственный объём изображения. Пулинг интерпретируется так: если на предыдущей операции свёртки уже были выявлены некоторые признаки, то для дальнейшей обработки настолько подробное изображение уже не нужно, и оно уплотняется до менее подробного. К тому же фильтрация уже ненужных деталей помогает не переобучаться. Слой пулинга, как правило, вставляется после слоя свёртки перед слоем следующей свёртки.

Кроме пулинга с функцией максимума можно использовать и другие функции — например, среднего значения или L2-нормирования. Однако практика показала преимущества именно пулинга с функцией максимума, который включается в типовые системы.

В целях более агрессивного уменьшения размера получаемых представлений, всё чаще находят распространение идеи использования меньших фильтров или полный отказ от слоёв пулинга.

После нескольких прохождений свёртки изображения и уплотнения с помощью пулинга система перестраивается от конкретной сетки пикселей с высоким разрешением к более абстрактным картам признаков, как правило, на каждом следующем слое увеличивается число каналов и уменьшается размерность изображения в каждом канале. В конце концов, остаётся большой набор каналов, хранящих небольшое число данных (даже один параметр), которые интерпретируются как самые абстрактные понятия, выявленные из исходного изображения.

Эти данные объединяются и передаются на обычную полносвязную нейронную сеть, которая тоже может состоять из нескольких слоёв. При этом полносвязные слои уже утрачивают пространственную структуру пикселей и обладают сравнительно небольшой размерностью (по отношению к количеству пикселей исходного изображения).

**Способы применения**

Наиболее распространённые способы применения предсказания временных рядов:

- Планирование бизнес развития

- Инженерный контроль – как наиболее пример здесь можно привести прогнозирование температур реакторов.

- Тренды криптовалют

- Финансовые рынки

- Одна из наиболее актуальных на сегодняшний день областей – прогнозирование распространения болезней.

- Распределение ресурсов

- Обработка сигналов

- Аналитика спортивных состязаний

- Статистика

- Прогнозирование погоды

**Заключение**

Не смотря на это, предсказание временных рядов не является самой эффективной техникой применения нейронных сетей. Она довольно успешна в предоставленных выше примерах, но на самом деле является довольно ограниченной и не способной к обработке многомерных объёмов данных. При её применении, важно понимать – какой конечный результат из каких данных мы пытаемся получить.

Список использованной литературы:

[1] <https://towardsdatascience.com/time-series-forecasting-with-deep-learning-and-attention-mechanism-2d001fc871fc>

[2] <https://www.timescale.com/blog/what-is-time-series-forecasting/>

[3] <https://medium.com/analytics-vidhya/time-series-forecasting-a-complete-guide-d963142da33f>

По состоянию на 21.12.2022