**МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ  
РОССИЙСКОЙ ФЕДЕРАЦИИ**

**Московский авиационный институт  
(национальный исследовательский университет)**

Факультет прикладной математики и физики

Кафедра вычислительной математики и программирования

**МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ  
на тему:**

ОПТИМИЗАЦИЯ (РЕДУКЦИЯ) ПОЛНОСВЯЗНЫХ НЕЙРОННЫХ СЕТЕЙ

01.04.02 «Прикладная математика и информатика»

|  |
| --- |
| **Выполнил:**  магистрант  Рожлейс Иварс Андрисович |
|  |
| **Научный руководитель:**  ст. преп.  Аносова Наталья Павловна |
|  |
| **Рецензент:**  *звание, должность?*  *ФИО?* |

Москва, 2020

**Оглавление**

[Список условных обозначений 3](#_Toc38631680)

[Введение 4](#_Toc38631681)

[Глава 1. Что такое долговременная краткосрочная память? 8](#_Toc38631682)

[Список литературы 9](#_Toc38631683)

[Приложение 10](#_Toc38631684)

# Список условных обозначений

LSTM – долгая краткосрочная память (англ. Long short-term memory, LSTM)

W – матрица весовых коэффициентов слоя

# Введение

Процесс оптимизации заключается в том, чтобы улучшить какое-нибудь свойство. Приведи примеры параметров нейронной сети... Как правило в многопараметрических системах повышение качества одного параметра сопровождается понижением качества другого параметра. Иными словами, оптимизация без потерь как правило очень редко достижима, поэтому часто приходится иметь дело с компромиссным решением – улучшая одно свойство неизбежно ухудшается другое.

Идея редукции или, иными словами, упрощения нейронной сети заключается в том, чтобы избавиться от нейронов, которые вносят минимальное воздействие на итоговый результат работы сети, и лишних связей между ними. Некоторые нейроны могут оказаться абсолютно бесполезными и не влиять на итоговый результат работы сети, некоторые могут оказывать лишь минимальное воздействие. [1] Поэтому в соответствие с некоторым компромиссным соглашением между степенью упрощения и правдивостью результата работы сети можно отсечь те коэффициенты, которые вносят минимальный вклад в итоговый результат работы сети.

Процесс редукции нейронный сети можно описать как процесс оптимизации сети по критерию избыточности. Помимо процесса упрощения, который можно сказать работает по типу сверху-вниз, можно также отметить противоположный вариант оптимизации – наращивание сети, который работает по типу снизу-вверх. Нейроны и даже целые нейронные слои могут добавляться в сеть по мере требования. [2] В данной работе будет рассмотрен только процесс редукции нейронных сетей.

Нейроны в сети как правило группируются во взаимодействующие между собой нейронные слои, образуя таким образом различные топологии нейронных сетей. Возможно, стоит добавить различные типы топологий или привести ссылку. [3] В качестве математической модели взаимодействия между нейронами применяется матричная алгебра. Каждый нейрон имеет степень влияния или весовой коэффициент в пределах своего слоя. Весовые коэффициенты всех нейронов в пределах одного слоя образуют матрицу весовых коэффициентов этого слоя. Таким образом отсечение нейронов происходит посредством манипуляций с матрицей весовых коэффициентов слоя.

Наивным алгоритмом редукции нейронной сети может считать алгоритм, при котором малые коэффициенты принимаются равными нулю тем самым исключая нейроны с малыми весами из обработки. С учётом возможной оптимизации алгоритмов работы с разреженными матрицами, такой метод вполне может повысить быстродействие сети. [4] Тем не менее как показывает практика кажущийся в первую очередь верным, такой способ может быть редко применён, так как малые весовые коэффициенты не всегда оказывают малое влияние на итоговый результат нейронной сети. [5]

Более верным методом можно считать метод оптимального прореживания нейронных сетей OBD (англ. Optimal Brain Damage) предложенный Яном ЛеКуном (Yann LeCun) в 1990 году. Данный метод использует лучший критерий чем абсолютные значения весовых коэффициентов, а именно учитывает чувствительность нейронов к вариациям весовых коэффициентов других нейронов. [6] Этот алгоритм говорит, что безопасно можно принять равными нулю весовые коэффициенты только тех нейронов чья чувствительность к вариациям оказывается минимальной. Дальнейшее развитие этого метода получило название OBS (Optimal Brain Surgeon) и было предложено Бабаком Хассиби (Babak Hassibi) и Дэвидом Шторком (David G. Stork). В каком году? Этот алгоритм подстраивает значение зафиксированного параметра во время варьирования другого. [7]

Ещё одним методом оптимизации является метод упрощения сети с использованием штрафной функции. Идея метода заключается в том, чтобы организовать этап обучения сети таким образом, чтобы спровоцировать автоматическую минимизацию весовых коэффициентов маловажных нейронов тем самым по итогу обучения можно отбросить те нейроны чьё влияние на сеть минимальное. Иными словами, получаем автоматическое контрастирование сети ещё на этапе обучения. [8]

Выше представленные методы имеют ряд ограничений. Так, например, наивный метод является неточным и в ряде случаев приводит к ухудшению результатов работы нейронной сети. Далее опиши почему OBD и OBS плохо.

В последнее время часто применяется метод сингулярного разложения (англ. SVD Singular Value Decomposition) матриц весовых коэффициентов влияние которого и будет рассмотрено в данной работе.

Задача упрощения нейронных сетей имеет несколько важных следствий.

1. В силу того, что упрощённая сеть имеет меньше нейронов и меньше весовых коэффициентов, то, как следствие нужно производить меньше вычислений чтобы получить результат. Таким образом повышается быстродействие сети.
2. «Слабые» (по тому или иному критерию) нейроны оказывают меньший эффект воздействия на итоговый результат работы сети, однако вносят определённый уровень шума. Процесс контрастирования сети позволяет в определённой степени снизить шумовое воздействие «слабых» нейронов на результат работы сети. Таким образом результат освобождается от шумового воздействия.
3. Процесс упрощения сети подразумевает собой избавление от некоторых нейронов без существенной потери качества сети, что аналогично процессу сжатия информации с потерей данных. Итоговый размер сети становится меньше. В определённых ситуациях уменьшение объёма данных необходимых для хранения матриц весовых коэффициентов может оказаться полезным если, к примеру сеть реализуется на машине с небольшим объёмом памяти.
4. Прореживание нейронной сети также повышает её классифицирующую способность. [9]

Почему это нужно?

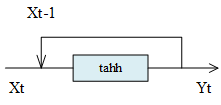
Перед упрощением сети возможно провести также контрастирование её коэффициентов так чтобы в результате обучения сети сильные нейроны становились сильнее, а слабые слабее чтобы тем самым облегчить процесс выкидывания отдельных нейронов из сети и повысить эффективность упрощения.

# Часть 1. Что такое долгая краткосрочная память?

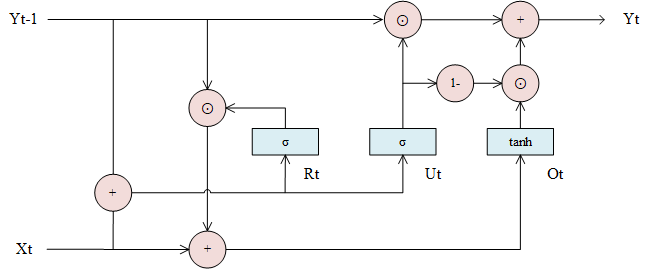
Долгая краткосрочная память (англ. Long short-term memory, LSTM) это архитектурная разновидность рекуррентных нейронных сетей, которая была предложена Зеппом Хохрайтером и Юргеном Шмидхубером в 1997 году. [10] В основе работы рекуррентных нейронных сетей лежит принцип обратной связи, когда результат работы сети зависит от выходного значения посчитанного на предыдущем шаге. LSTM это архитектурное решение проектирования рекуррентной сети в том смысле что в состав этой сети входят несколько полносвязнных слоя. Почему LSTM называют ячейкой?

Управляемый рекуррентный блок (англ. Gated Recurrent Network, GRU) ещё одна схожая архитектура рекуррентных нейронных сетей, предложенная Кёнхёном Тё в 2014 году. [11]

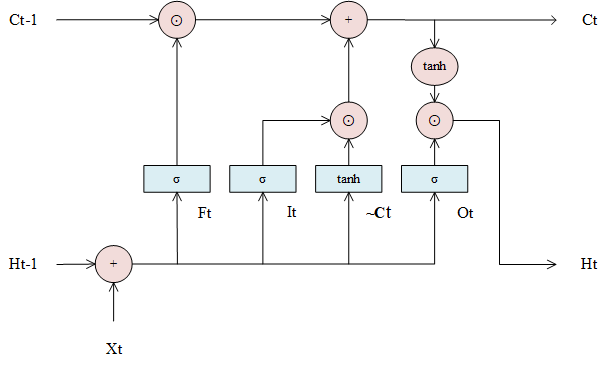
Ниже представлена схематическая диаграмма простой рекуррентной нейронной сети. Прямоугольником показан один слой сети с активационной функцией в качестве гиперболического тангенса. Стрелками показано направление потока данных.



Как видно из рисунка результат работы сети Xt+1 одновременно зависит как от входного значения Xt так и результата Xt+1.



В данной работе будет рассмотрена оптимизация именно LSTM ячйеки.



На картинке покажи, чем они похожи и чем отличаются.

Напиши в чём преимущество использовать одну архитектуру вместо другой (LSTM vs GRU). Напиши, что в этой работе исследуется именно LSTM архитектура. Причины выбора именно LSTM можно не указывать. GRU в отличие от LSTM имеет два фильтра вместо трёх что делает её более грубой. Если нужна большая точность, то стоит использовать LSTM. [12]

В конце думаю нужно добавить, что умеет LSTM ячейка. Например, предсказывать временные ряди и т.д.

# Список литературы

1. А
2. А
3. А
4. А
5. А
6. А
7. А
8. А
9. А
10. Hochreiter S., Schmidhuber J. Long short-term memory // Neural Computaion, 1997
11. Cho K., van Merrienboer B., Gulcehre C., Bahdanau D., Bougares F., Schwenk H., Bengio Y. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation // 2014

Добавь ещё ссылки на статьи.

# Приложение

А