Министерство образования и науки Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего профессионального образования

«Московский физико-технический институт (Государственный университет)»

Факультет радиотехники и кибернетики

Кафедра теоретической и прикладной информатики

**Исследование качества классификации электронных писем нейронными сетями в зависимости от входных параметров**

**Выпускная квалификационная работа на степень бакалавра
Направление подготовки: 03.03.01 Прикладные математика и физика**

Выполнил:

Студент 211 группы Соболев Константин Викторович

Научный руководитель:

 Соболев Артемий Анатольевич

**Москва 2016**

## Содержание

[Аннотация 4](#_Toc454872018)

[Введение 5](#_Toc454872019)

[1. Постановка задачи 7](#_Toc454872020)

[2. Метрики 8](#_Toc454872021)

[1. Функция потерь 8](#_Toc454872022)

[2. Функция точности 8](#_Toc454872023)

[3. Обучающая выборка 9](#_Toc454872024)

[4. Предлагаемое решение 10](#_Toc454872025)

[1. Архитектура сверточной нейронной сети 10](#_Toc454872026)

[2. Полносвязный слой 11](#_Toc454872027)

[3. Сверточный слой 12](#_Toc454872028)

[4. Субдискретизирующий слой 12](#_Toc454872029)

[5. Dropout слой 13](#_Toc454872030)

[6. Модели использования сверточной нейронной сети для классификации текстов с использованием кодирования слов. 13](#_Toc454872031)

[7. Обучение нейронной сети 14](#_Toc454872032)

[8. Обработка данных, получаемых от нейронных сетей 15](#_Toc454872033)

[9. Метод градиентного спуска 16](#_Toc454872034)

[5. Общее описание 18](#_Toc454872035)

[6. Перевод текста в вектор фиксированной длины 19](#_Toc454872036)

[7. Эксперименты 20](#_Toc454872037)

[1. Набор данных 20](#_Toc454872038)

[2. Описание экспериментов и полученные данные 21](#_Toc454872039)

[8. Вывод 25](#_Toc454872040)

[Дальнейшая работа 25](#_Toc454872041)

[Список литературы 26](#_Toc454872042)

## Аннотация

Целью данной исследовательской работы является создание прототипа классификатора почтовых писем на основе нейронной сети. Особенностью текстового данного классификатора является использование метаданных письма, как части входных данных нейронной сети.

В ходе данной работы построен новый классификатор почтовых писем, работающий на основе сверточной нейронной сети. Было проведено исследование зависимости точности классификатора от размера обучающей выборки и от добавления различных типов метаданных в модель классификатора.

## Введение

С ростом текстовой информации в цифровом виде задача автоматической классификации приобретает все большую актуальность. Например, такая задача возникает при классификации почтовых писем.

Ежедневно отправляется более сотни миллионов электронных писем. Часть из них носит вредоносный характер, и успешно отсеивается системами фильтрации. Но даже после фильтрации спама на наиболее активные постовые аккаунты приходит порядка сотен почтовый писем каждый день. Возникает необходимость классификации почтовых писем по категориям.

После преобразования почтового письма в векторный вид, данная задача решается методами машинного обучения. В рамках данной задачи предлагается использовать искусственную нейронную сеть в качестве алгоритма классификации.

Данная задача отличается от классической задачи классификации текста присутствием множества метаданных в письме. Учтя их можно значительно повысить точность классификатора. Например, если владелец почтового аккаунта получает множество писем на определенную тему от одного и того же адресата, то новое письмо, полученное от него с большой вероятностью будет относиться к тому же классу, что и предыдущие письма. Так же не стоит недооценивать важность использования заголовка письма в механизме классификации. Заголовок зачастую содержит саму тему письма. При классификации его стоит учитывать, как отдельную от текста сущность и давать ему больший вес, влияния на результат классификатора.

Целью данной работы является получение количественной оценки улучшения производительности классификатора в зависимости от использования различных метаданных в процессе классификации.

В итоге был разработана программа на Python, в котором реализованы функции загрузки и выгрузки данных, алгоритм классификации, использующий сверточную нейронную сеть.

Ниже дается более конкретная постановка задачи и описание исходных данных. Также в деталях описывается архитектура классификатора.

Затем описываются проведенные эксперименты по повышению точности классификации. Дается обзор закономерностей, выявленных в ходе решения задачи. В конце приводится метод, позволяющий максимально повысить точность.

## Постановка задачи

Имеется множество почтовых писем (обучающая выборка) и множество классов , каждому документу из приписан класс из [2]. Требуется на основе этих данных построить процедуру автоматической классификации почтовых писем, которая помимо текста самого письма будет учитывать метаданные письма.

Помимо текста самого письма почтовые письма имеют множество метаданных, такие как автор письма, список получателей, тема письма и так далее. В рамках данной задачи предлагается использовать метаданные письма наравне с текстом письма, чтобы увеличить качество классификации.

В качестве алгоритма классификации предлагается использовать искусственную нейронную сеть, которая будет обучена на обучающей выборке .

Качество полученной процедуры оценивается на основе тестовой выборки, в которую входят почтовые письма, для которых заранее известны соответствующие им классы. Документы из тестовой выборки подаются на вход классификатору, результаты сравниваются с заранее известным значением класса. Функции оценки качества будем называть метриками. Метрики качества классификации подробнее описаны ниже.

## 2. Метрики

В качестве метрик нейронной сети использовались две функции: функция потерь и функция точности. Первая функция имеет первостепенную важность, она характеризует близость ответа классификатора на тестовой выборке к оригинальным значениям классов экземпляров тестовой выборки задача нашего классификатора сводится к уменьшению функции потерь. Вторая функция отображает процент верных ответов нашего классификатора и является наглядной характеристикой точности классификатора.

### Функция потерь

В качестве функции потерь использовалась функция перекрестной энтропии, которая задается формулой [4]:

, где – среднее значение j –го элемента вектора значений класса .

### Функция точности

Пусть контрольная выборка состоит из объектов. Из них объектов классифицируются правильно. Тогда доля правильных ответов вычисляется по формуле [5]:

## Обучающая выборка

В качестве обучающей выборки используется открытая база данных почтовых писем американской энергетической компании Enron, которая была выложена в открытый доступ после того, как она стала банкротом. Преимуществом обучающей выборки является то, что письма рассортированы по сотрудникам компании. Письма многих сотрудников в свою очередь были вручную рассортированы по категориям, что дает нам размеченную обучающую выборку. Максимальное число писем, приходящихся на одного сотрудника составляет порядка 12000 писем, что является достаточным для обучения нейронной сети.

## Предлагаемое решение

В качестве алгоритма классификации используется сверточную нейронную сеть, которая будет обучена на обучающей выборке . Архитектура сверточных нейронных сетей была предложена Яном Лукуном [7] и нацелена на эффективное распознавание изображений. Свое название архитектура сети получила из-за наличия операции свертки, суть которой заключается в том, что каждый фрагмент изображения умножается на матрицу (ядро) свертки поэлементно, а результат суммируется и записывается в аналогичную позицию изображения. Данный тип нейронной сети входит в состав технологий глубинного обучения и является одной из лучших архитектур, распознающих изображение. Успех применения сверточных сетей к классификации изображений привел к множеству попыток использовать данный метод к другим задачам. В последнее время их стали активно использовать в задаче классификации текстов. Сверточная нейронная сет имеет ряд преимуществ в сравнении с остальными нейронными сетями:

* Концептуально простые схемы обучения, что приводит к высокой производительности при работе с большим объемом данных
* Гораздо меньшее количество настраиваемых параметров, в сравнении с полносвязными нейронными сетями
* Удобное распараллеливание вычислений, а, следовательно, возможность реализации алгоритмов работы и обучения на графических процессорах

### Архитектура сверточной нейронной сети

Сверточная нейронная сеть обычно представляет собой чередование сверточных слоев, субсидирующих слоев и при наличии полносвязных слоев на выходе. Все три вида слоев могут чередоваться в произвольном порядке [7][1].

В сверточном слое нейроны, которые используют одни и те же веса, объединяются в карты признаков, а каждый нейрон карты признаков связан с частью нейронов предыдущего слоя. При вычислении сети получается, что каждый нейрон выполняет свертку некоторой области предыдущего слоя (определяемой множеством нейронов, связанных с данным нейроном).

Пример архитектуры сверточной нейронной сети представлен на Рис.1



Рис.1 Архитектура сверточной нейронной сети

### Полносвязный слой

Слой, в котором каждый нейрон соединен со всеми нейронами на предыдущем уровне, причем каждая связь имеет свой весовой коэффициент. На Рис.2 Показан пример полносвязного слоя.



Рис.2 Полносвязный слой

- весовые коэффициенты.

 - функция активации.

### Сверточный слой

В отличие от полносвязного, в сверточном слой нейрон присоединен лишь с ограниченным количеством нейронов предыдущего уровня, то есть сверточный слой аналогичен применению операции свертки, где используется лишь матрица весов небольшого размера (ядро свертки), которую «двигают».

На Рис.3 Показан пример сверточного слоя с ядром свертки размера 3 х 3.



Рис.3 Сверточный слой

### Субдискретизирующий слой

Слои этого типа выполняют уменьшение размерности (обычно в несколько раз). Это можно делать различными способами, но зачастую используется метод выбора максимального элемента – вся карта признаков разделяется на ячейки, из которых выбираются максимальное по значению.

На рис.4 Показан пример субдискретизирующего слоя с методом выбора максимального элемента.



Рис.4 Субдискретизирующий слой

### Dropout слой

Dropout слой (dropout регуляция) [8] способ борьбы с переобучением в нейронных сетях, обучение в которых обычно производят стохастическим градиентным спуском, случайно выбирая некоторые объекты из выборки. Dropout регуляция заключается в изменении структуры сети: каждый нейрон выбрасывается с некоторой вероятностью . По такой прореженной сети производится обучение, для оставшихся весов делается градиентный шаг, после чего все выброшенные нейроны возвращаются в нейронную сеть. Таким образом, на каждом шаге стохастического градиента мы настраиваем одну из возможных архитектур сети, где под архитектурой мы понимаем структуру связей между нейронами, а через N обозначаем суммарное число нейронов. При тестировании нейронной сети нейроны уже не выбрасываются, но выход каждого нейрона умножается на - благодаря этому на выходе нейрона мы будем получать математическое ожидание его ответа по всем архитектурам. Таким образом, обученную с помощью dropout – регуляции нейронная сеть можно рассматривать как результат усреднения сетей.

### Модели использования сверточной нейронной сети для классификации текстов с использованием кодирования слов.

Подход был описан в статье. В данном подходе каждому слову в тексте сопоставляется вектор фиксированной длины, затем из полученных векторов для каждого объекта выборки составляется матрица, которая аналогично изображениям подается на вход сверточной нейронной сети. На рис. Приведен пример сверточной нейронной сети с использованием кодирования слов.



Рис. Посимвольный подход

Для экспериментов в статье [9] была реализована нейронная сеть с одним сверточным, одним субдискритезирующим и одним полносвязным слоем. Данная нейронная сеть использовалась для классификации текстов небольшого размера.

### Обучение нейронной сети

Для обучения нейронных сетей использовался алгоритм обратного распространения ошибки [11]:

1. Инициализируем веса связей нейронной сети малыми случайными значениями,
2. Повторить N раз:

Для всех d от 1 до m:

1. Подать на вход сети и посчитать выходы каждого узла.
2. Для всех :
3. Для каждого уровня l, начиная с предпоследнего:

Для каждого узла j уровня l вычислить:

1. Для каждого ребра сети
2. Выдать значения .

Где – коэффициент инерциональности для сглаживания скачков при перемещении по поверхности целевой функции.

### Обработка данных, получаемых от нейронных сетей

Отличительной чертой нашей задачи классификации почтовых писем является использование метаданных почтового письма. Для каждого типа метаданных используется отдельная сверточная нейронная сеть, обученная на размеченном наборе соответствующих метаданных письма.

Каждая нейронная сеть на выходе выдает числовое значение, соответствующее классу письма, к которому его относит нейронная сеть. Так как выдаваемые числовые значения для различных типов метаданных письма могут отличаться, то возникает задача поиска алгоритма оптимального учета выходных данных всех выходов нейронных сетей.

В нашей задаче выходным значением классификатора является линейная суперпозиция выходов нейронных сетей с весами . В векторном представлении выходам нейронной сети соответствует вектор с коэффициентами , а весам соответствует вектор весов c коэффициентами . Таким образом выходное значение классификатора является скалярным произведением вектора весов и вектора выходных значений нейронных сетей.

, где с – множество классов

Таким образом задача поиска алгоритма оптимального учета выходных данных всех выходов нейронных сетей сводится с поиску оптимального значения весовых коэффициентов. Данная задача решается с помощью метода градиентного спуска.

### Метод градиентного спуска

Данный метод используется для поиска минимума функции , записывая в виде [10]:

Основная идея этого метода заключается в осуществлении оптимизации в направлении наискорейшего спуска, задаваемом антиградиентом функции :

На каждой итерации данного алгоритма происходит ствиг в сторону минимума функции.

На рис.6 Показана графическая иллюстрация движения в сторону минимума:



Рис.6 Геометрическая интерпретация метода градиентного спуска

Критерием остановки алгоритма является следующее неравенство:

Применяя данный метод, находим оптимальные значения коэффициентов.

## Общее описание

При загрузке почтового письма в прототип происходит разбиение письма на метаданные. После разбиения на метаданные, письмо проходит через систему фильтрации: происходит удаление лишних данных и преобразование оставшихся в векторный вид. Затем вектор данных подается на вход нейронной сети. Сеть определяет к какому классу относится письмо, и подает выдает соответствующее ему численное значение. Численные значения выходов нейронной сети суммируются с весовыми коэффициентами, образуя в сумме финальное значение класса, к которому классификатор отнес почтовое письмо.

На рис.7 изображена полная схема прототипа:



Рис.7 Схема прототипа

## Перевод текста в вектор фиксированной длины

Перед переводом текста в векторный вид, он проходит процесс предобработки. Процесс предобработки состоит из нескольких частей: фильтрация (удаление знаков препинания, слов, не несущих смысл, например, артиклей), замена верхних на нижние регистры, стемминг (замена слова его смысловой основой, например, cats заменяется на cat).

Для текста письма в данной модели используется вектор текста фиксированной размерности – 480 слов (средняя длина письма набора данных после предобработки). Если письмо меньшего размера, то оно дополняется до полного набора пустыми словами. Если больше, то нарезается на части. Каждая часть письма будет использоваться, как отдельный пример класса, соответствующего письму.

Каждому слову ставится в соответствие его номер в общем словаре слов, составляющем 18765 слов. Таким образом каждому тексту ставится в соответствие вектор цифр размерности 480 слов.

## Эксперименты

### Набор данных

Эксперименты проводились для наборов почтовых писем трех сотрудников компании ENRON с самым большим числом входящих писем, с учетом того, что их письма были распределены по тематическим директориям. Размер набора данных составляет порядка 4000 экземпляров почтовых писем.

Таблица 1: Статистика по набору данных писем ENRON после удаления папок без тем и папок маллых размеров.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Пользователь | Количество папок | Количество сообщений | Размер наименьшей папки (сообщений) | Размер наибольшей папки (сообщений) | Размер наименьшего сообщения (слов) | Размер наибольшего сообщения (слов) |
| Beck-s | 101 | 1971 | 3 | 166 | 45 | 2620 |
| Farmer-d | 25 | 3672 | 5 | 1192 | 43 | 3507 |
| Kaminski-v | 41 | 4477 | 3 | 547 | 44 | 7885 |
| Kitchen – l | 47 | 4015 | 5 | 715 | 47 | 46296 |
| Lokay – m | 11 | 2489 | 6 | 1159 | 45 | 4456 |
| Sanders – r | 30 | 1188 | 4 | 420 | 55 | 19331 |
| Williams – w3 | 18 | 2769 | 3 | 1398 | 49 | 2287 |

### Описание экспериментов и полученные данные

В наших экспериментах использовалась 3 типа данных письма: текст письма, тема письма и список получателей. Набор данных разбивается на 2 группы – обучающая выборка (80% электронных писем) и тестовая выборка (20% электронных писем).

На первом этапе наших исследований рассмотрим зависимость точности классификации от размера обучающей выборки нейронной сети. В ходе исследований были получены следующие зависимости:

Из графика видно, что наибольшую точность имеет классификатор по тексту письма. Причиной этого является большая размерность вектора, получаемого из письма. Таким образом при большем наборе признаков мы получаем большую точность.

Экспериментально было установлено, что при классификации по списку получателей хорошо получается классифицировать письма, отправляемые группе из нескольких получателей (точность порядка 85%). Это можно проиллюстрировать тем, что если письмо было отправлено одной группе получателей, то следующее письмо той же группе будет относиться к тому же классу с большой вероятностью.

На втором этапе исследований рассматривались различные комбинации метаданных письма в классификаторе. В ходе исследований были получены следующие зависимости:

Результаты исследований размещены в таблице 2:

Таблица 2. Результаты исследований

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Текст | Тема | Получатели | Текст + тема | Получатели + тема | Получатели + текст | Получатели + текст + тема |
| Точность на максимальном размере набора данных | 76,5% | 59,4% | 42,1% | 80,1% | 66,9 % | 80,3% | 81,7% |

Использование списка получателей и темы письма увеличивает точность классификатора на 7,5% по сравнению с классификацией только по теме письма. Такой большой прирост в точности классификации является следствием того, что классификация только списку получателей и только темам письма имеют низкую точность и не очень большую корреляцию между правильными ответами

Использование списка получателей и текста письма увеличивает точность классификатора на 3,6% по сравнению с классификацией только по тексту письма.

Использование темы письма и темы письма увеличивает точность классификатора на 3,8% по сравнению с классификацией только по тексту письма.

Совместное использование всех трех типов данных дает выигрыш в 5,2% по сравнению с классификацией только по тексту письма.

## Вывод

С помощью нашего исследования было успешно показана состоятельность использования различных метаданных в задаче классификации почтовых писем. За счет использования всех трех типов данных удалось добиться выигрыша в 5,2%.

## Дальнейшая работа

Задачей данной работы было лишь показать возможность получения прироста точности классификации в рамках задачи классификации нейронной сетью, но не ставилось цели создания классификатора с максимально возможной точностью

В дальнейшем планируется расширить список метаданных, с которыми работает классификатор, а также использовать нейронную сеть в комбинации с другими методами классификации с целью повышения предсказания. Таким образом планируется создать прототип классификатора на языке Python, работающий с максимально возможной точностью.

После получения завершенного прототипа почтового классификатора, планируется реализовать его алгоритмы в работающем почтовом клиенте.

## Список литературы

1. Krizhevsky, A. Imagenet classification with deep convolutional neural networks / Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton // NIPS. — 2012. — 1106 -1114 p.
2. Kim, Y. Convolutional neural networks for sentence classification / Yoon Kim // IEMNLP. — 2014. — Sep. — 1746 -1751 p.
3. John Duchi Elad Hazan, Y. S. Adaptive subgradient methods for online learning and stochastic optimization / Yoram Singer John Duchi, Elad Hazan // JMLR. — 2011.
4. Damashek, M. Gauging similarity with n-grams: Language-independent categorization of text / Marc Damashek // Science, New Series. — 1995.
5. Efficient estimation of word representations in vector space / Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean // ICLR. — 2013.
6. McCulloch, W. S. A logical calculus of the ideas immanent in nervous activity / Warren S. McCulloch, Walter Pitts // Springer New York. — 1943.
7. Yann LeCun Leon Bottou, Y. B. Gradient-based learning applied to document recognition / Yoshua Bengio Yann LeCun, Leon Bottou, Patrick Haffner // IEEE. — 1998.
8. Воронцов, К. В. Курс лекций по машиному обучению / К. В. Воронцов. — 2015.
9. Pennington, J. Glove: Global vectors for word representation / Jeffrey Pennington, Richard Socher, Christopher D // EMNLP. — 2014. — 1532 -1543 p.
10. Акулис И.Л. Математическое программирование в примерах и задачах. – М.: Высшая школа, 1986. – С. 298-311.
11. Гробань А.Н., Россиев Д.А., Нейронные сети на персональном компьютере. – Новосибирск: Наука, 1996. – 276 с.