DSPA:

Вопросы применения цифровой обработки сигналов

№2

2022

СОДЕРЖАНИЕ

АЛГОРИТМ КОМПЛЕКСНОГО ДИСКРЕТНОГО КОСИНУСНОГО ПРЕОБРАЗОВАНИЯ

Абрамов Валентин Александрович,
Московский Технический Университет Связи и Информатики (МТУСИ), доцент, к.т.н., Москва, Россия
vabramov44@mail.ru

Попов Олег Борисович,
Московский Технический Университет Связи и Информатики (МТУСИ), профессор, к.т.н., Москва, Россия
olegp45@yandex.ru

Чернышева Татьяна Васильевна,
Московский Технический Университет Связи и Информатики (МТУСИ), доцент, к.т.н., Москва, Россия
krba2012@yandex.ru

Кузнецов Петр,
Московский Технический Университет Связи и Информатики (МТУСИ), аспирант, Москва, Россия,
peter.kyznetsov@gmail.com

Аннотация

Основным способом представления звукового сигнала в частотной области является быстрое преобразование Фурье (БПФ). При анализе коротких последовательностей сигнала, БПФ зачастую не может обеспечить необходимой разрешающей способности и точности формирования оценок амплитуды и фазы коэффициентов. Известно и широко используется дискретное косинусное преобразование (ДКП), коэффициенты оценки которого расставлены на шкале частот вдвое чаще. Недостатком ДКП является зависимость оценки от начального фазового положения сигнала, тем не менее ДКП широко используется для устранения статистической избыточности сигналов. Предлагается алгоритм комплексного ДКП обеспечивающий повышение разрешающей способности и точности оценки.

Ключевые слова: *спектральный анализ, быстрое преобразование Фурье, дискретное косинусное преобразование, синтез, субъективно-статистические экспертизы (испытания), помещение прослушивания, фильтрация, сведение.*

Введение

Наиболее эффективные алгоритмы устранения избыточности звукового сигнала реализованы на основе его представления в частотной области [1]. Такое представление позволяет устранить статистическую избыточность, избыточность описания сигнала, так как большинство сигналов включает достаточно ограниченный ряд гармонически связанных спектральных составляющих с известными закономерностями развития на времени существования звукового объекта. Наиболее часто используется для спектрального анализа БПФ, недостатком которого на коротких длительностях соизмеримых с длительностью звуковых объектов является недостаточная разрешающая способность и точность определения фазы. Кроме БПФ известно и широко используется дискретное косинусное преобразование (ДКП), разрешающая способность которого вдвое выше, но нет традиционного представления фазы колебания. Предлагается алгоритм, в котором производится ДКП анализ исходного и ортогонального ему сигнала с устранением боковых лепестков ДКП оценки путем вычитания из нее оценки специально сформированного компенсационного сигнала [2]. Алгоритм позволяет повысить разрешающую способность оценки вдвое относительно БПФ, устранить осцилляцию амплитудной и фазовой оценки и приблизить оценку к точности обеспечиваемой слуховым анализатором человека [3].

Способ формирования комплексной дискретно косинусной оценки спектра сигнала

Во всех случаях формирования требований к точности и разрешающей способности косвенной оценке спектра звукового сигнала, например, с помощью ортогональных преобразований, стараются исходить их возможностей периферического слухового анализатора человека. Проведя анализ существующей литературы по свойствам слуха [1, 2], а также субъективно статистические испытания (ССИ), проведенные на кафедре Телевидения и Звукового вещания МТУСИ, позволили сформировать требования к точности и разрешающей способности оценки спектра [1, 2, 3].

Учитывая нелинейное восприятие слуховым анализатором частоты сигнала, которая воспринимается человеком как «высота», желательно проводить анализ на логарифмической шкале частот, что не соблюдается при использовании дискретного преобразования Фурье (ДПФ) или дискретного косинусного преобразования (ДКП). Учитывая наличие наличия в слуховом анализаторе участка линейного восприятия частоты в диапазоне от 20 до 500 Гц, в его пределах желательно оценивать частоту с точностью около 1,5 Гц. Эти данные получены при восприятии сигнала в зоне максимальной чувствительности слухового анализатора, модулированного с частотой около 4 Гц, что наиболее заметно для слушателя [2]. Для частот выше 500 Гц может быть принята точность оценки порядка 1,5% от абсолютного значения частоты.

Точность определения амплитуды оценки, согласно [2] составлять 0,4 дБ, в зоне максимальной чувствительности слухового анализатора, 2-3,4 кГц при уровне звука 70-96 дБ. На остальных частотах точность может снижаться до 6 дБ на краях диапазона слышимых частот. Не надо удивляться большой величине допустимой ошибки. В большинстве алгоритмов компактного представления и обработки звукового сигнала для перехода в частотную область используется быстрый вариант вычисления ДПФ (БПФ). Авторы алгоритмов не всегда в курсе о так называемой осцилляции амплитудной оценки коэффициентов, изменении амплитуды в зависимости от начальной фаза анализируемого колебания, достигающей 4 дБ, при использовании прямоугольного окна и составляющей 2,8 дБ для окна Наттола.

Обычно не нормируется точность определения фазы. Считается, что слух человеческий к фазе нечувствителен. Это действительно так для начальной фазы колебания. Но при оценке фазовых положений спектральных составляющих в созвучии, чувствительность слуха к смещению фазы достаточно велика. Бали проведены ССИ по оценке заметности изменения фазы гармоник сложного колебания (колокольчик). По результатам ССИ эксперты четко замечали смещение фазы начиная с 8° [3,4].

Труднее всего обеспечить необходимую разрешающую способность анализа. Наиболее заметны ошибки в разрешающей способности при анализе спектра вокализованных звуков, на достаточно большой длительности. Звуковой сигнал в этом случае представляет из себя основной тон и его гармоники. Наиболее низкий основной тон в ариях, исполненных Шаляпиным составляет около 60 Гц. Следовательно, разрешающая способность преобразования должна составлять около 20 Гц, так как между двумя коэффициентами оценки нужно иметь пониженный коэффициент, показывающий их наличие. Кроме того, при близком расположении двух спектральных составляющих в звуковом сигнале человек слышит тон средней частоты, модулированный по амплитуде (биения).

Число отсчетов в каждой выборке при анализе определяет число коэффициентов оценки, а, следовательно, разрешающую способность и точность анализа. Известна высокая информационная ценность так называемых «атак», начальных участков звуковых объектов. Устранение «атак» делает полностью неразборчивым речевой сигнал, в музыкальных исчезает возможность идентификации инструментов, на которых сыграно произведение. Минимальная длительность атаки, измеренная в каналах передачи составляет 5 мс [4]. С учетом статистики звуковых объектов была принята желательной длительность около 8 мс. Для подтверждения обоснованности такого выбора были проведены ССИ по определению заметности подмены плавно меняющегося по частоте и амплитуде тона дискретно меняющимся. По обоим параметрам была получена оценка 9 миллисекунд, С инженерным запасом длительность была принятой 8 мс.

Такая длительность позволяет сохранить и модуляционные характеристики сигнала, весьма существенные для узнаваемости исполнителя или инструмента.

Стандартным требованием к любому способу спектральной оценки является его эффективность – способность отобразить сигнал малым количеством коэффициентов.

С учетом свойств ортогональных преобразований и появлением в ходе анализа боковых лепестков найденных спектральных составляющих предполагается в ходе выполнения алгоритма устранять из входного сигнала найденные составляющие вместе с боковыми лепестками.

Многие способы спектрального анализа не предполагают последующей сборки сигнала по результатам анализа, желательно такую возможность обеспечить [6,7].

Особую трудность вызывает такая сборка сигнала после устранения статистической и психофизической избыточности, сопровождаемой потерями части коэффициентов оценки ортогонального преобразования. Не надо забывать, что ортогональные преобразования это замена непрерывно меняющегося по частоте и амплитуде сигнала дискретно меняющимся и неизменными на длительности каждой выборки. При устранении избыточности, т.е. устранении части стационарных на времени анализа коэффициентов, эффект стационарности становиться заметен слушателю. становиться заметен слушателю.

Наиболее верная оценка параметров спектральной составляющей при ее формировании с помощью ортогональных преобразований будет наблюдаться при совпадении анализируемого колебания и коэффициента оценки. Положение коэффициента на шкале частот фиксировано и определяется длительностью выборки, частотой дискретизации и вариантом ортогонального преобразования. Предлагается кроме анализа исходного сигнала проводить анализ сигналов, сдвинутых (транспонированных) по частоте в пределах бина, частотного разноса между двумя коэффициентами ортогонального преобразования. В этом случае наиболее верные значения характеристик параметров будут определены при наибольшем приближении частоты сигнала к частоте коэффициента оценки. После анализа результата по максимуму амплитуды выбирается наиболее верное значение. Такой подход позволяет анализировать сигнал на нелинейной шкале адаптируя точность анализа к точности слухового анализатора [4] и обеспечивая максимальную эффективность преобразования.

Разработанный алгоритм включает следующие операции:

– формирование ортогонального исходному сигнала, с ошибкой, не превышающей 10^{-5}, в соответствии с алгоритмом, предложенным в [5,6];

– ДКП исходного и ортогонального сигналов;

– формирование компенсирующих сигналов для исходного и ортогонального;

– ДКП компенсирующих сигналов;

– сложение ДКП оценок исходных и компенсирующих сигналов;

– нормирование итоговой оценки.

Кроме того, предусмотрена возможность анализа набора транспонированных исходных сигналов, что позволяет повысить точность и разрешающую способность анализа. Предлагаемый алгоритм позволяет дополнить объективными измерениями субъективную оценку, формируемую по рекомендации [4].

Преобразование Фурье предполагает, что описание сигнала на бесконечной длительности может быть осуществлено с помощью бесконечного числа комплексных синусоидальных сигналов. При цифровой обработке предлагается анализировать сигнал дискретно представленный во времени s(k) и на конечной длительности N дискретных отсчетов, формируя коэффициенты оценки S(n):

$$S(n)=\sum_{k=0}^{N-1}s(k)\left[\cos\left(\frac{2\pi}{N}nk\right)-j\sin\left(\frac{2\pi}{N}nk\right)\right],n=0,1,...,N-1.$$

Фактически анализ производится на бесконечном повторении отрезка сигнала, рис.1. Для устранения разрывов сигнала используются оконные функции, сводящие сигнал к нулю, в начале и конце выборки.

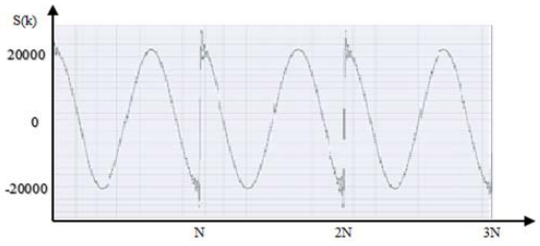

Рис. 1. Осциллограмма последовательности в 3N отсчетов

Чтобы избежать потери информации БПФ приходится производить с перекрытиями по длительности до 50%. Теоретически каждый коэффициент БПФ оценивает энергию сигнала в полосе:

$$df = bF_Д / N,$$

где b – коэффициент увеличения полосы для выбранного окна.

Известно, что каждый коэффициент ДКП оценивает энергию сигнала в полосе вдвое более узкой и не расширяется за счет неиспользуемой оконной функции.

Коэффициенты ДКП L(k), для N отсчетов X(m), рассчитываются по формуле:

$$L_x(0) = \frac{1}{\sqrt{N}} \sum_{m=0}^{N-1} X(m);$$

$$L_x(k) = \sqrt{\frac{2}{N}} \sum_{m=0}^{N-1} X(m) \cos \frac{(2m+1)k\pi}{2N}, k = 1,2,..., N-1,$$

При этом расчет производится на бесконечной длительности зеркально отраженных выборок сигнала (рис. 2). В последовательности отсутствуют разрывы функции, что позволяет не использовать перекрытий и производить анализ в «стык».

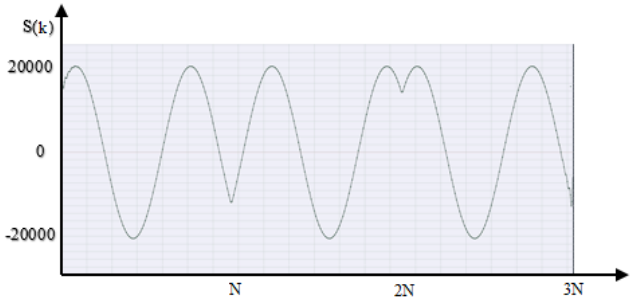

Рис. 2. Осциллограмма сигнала при ДКП анализе

При ДКП анализ производится на длительностях кратных полупериоду колебания (рис. 3).

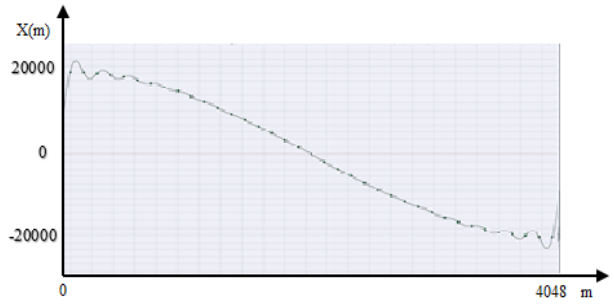

Рис. 3. Значения L (1) при расчете первого коэффициента ДКП

Если анализируемое колебание имеет ту же форму и фазу, вся его энергия будет отображена первым коэффициентом (рис. 4).

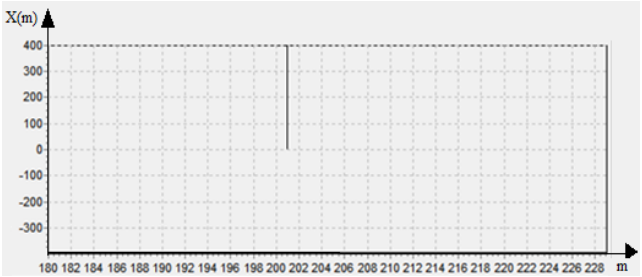

Рис. 4. ДКП оценка первого коэффициента

Для ортогонального сигнала (рис. 5), ДКП оценка содержит большое количество паразитных боковых лепестков (рис. 6). Оценка самого коэффициента нулевая.

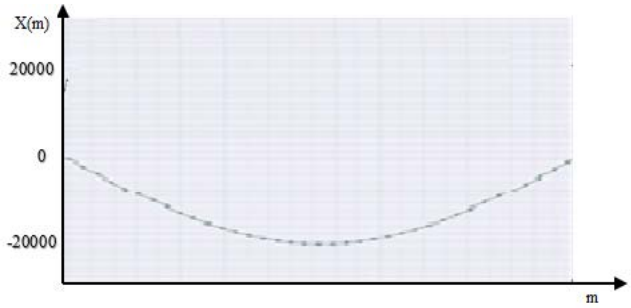

Рис. 5. Значения L(1) ортогонального сигнала

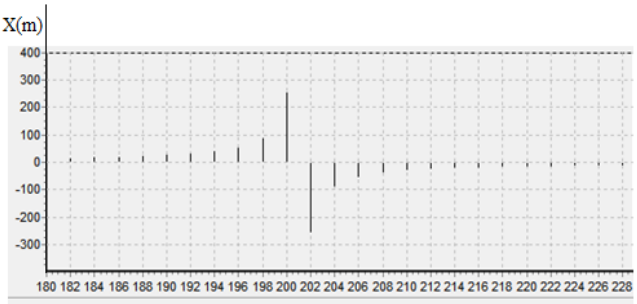

Рис. 6. ДКП оценка ортогонального сигнала

В следующем примере для удобства представления рассмотрен анализ ДКП для 201 коэффициента. Кроме анализируемого сигнала, в соответствии с алгоритмом, предложенным в [2], используется компенсирующий сигнал, оценка ДКП которого приведена на рисунке 7. В дальнейшем можно принять исходный сигнал в виде косинусной составляющей, а ортогональный – синусной.

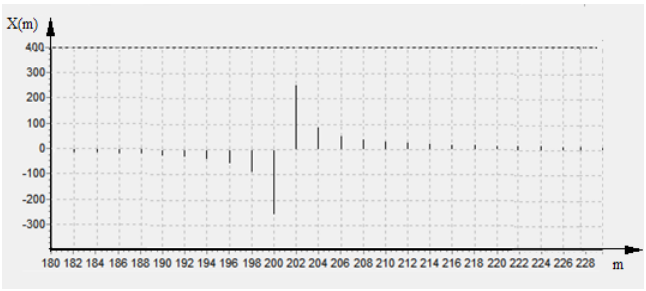

Рис. 7. ДКП оценка компенсирующего сигнала

Как мы видим боковые лепестки ДКП для исходного и ортогонального сигнала противофазны, а информационно значимые синфазны.

Суммирование оценок дает удвоенную амплитудную оценку действительной части и нулевую по мнимой. В данном примере было рассмотрено колебание с нулевой начальной фазой. Ниже рассмотрено колебание с начальной фазой 0,225 рад. (рис. 8-9).

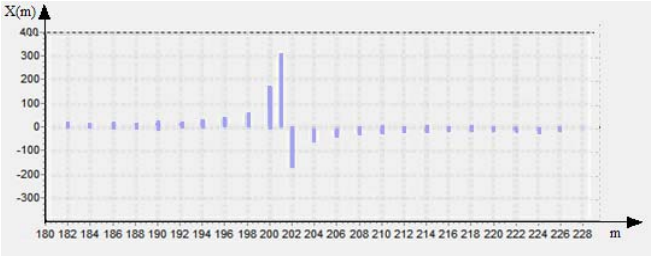

Рис. 8. ДКП оценка колебания с начальной фазой 0,225 рад

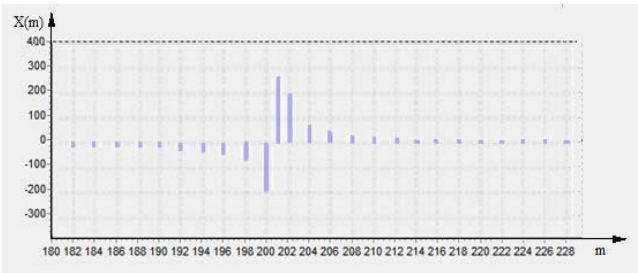

Рис.9. ДКП оценка ортогонального колебания с начальной фазой 0,225 рад.

После сложения с ДКП оценками компенсирующего сигнала получаем комплексную оценку без боковых лепестков (рис. 10-11).

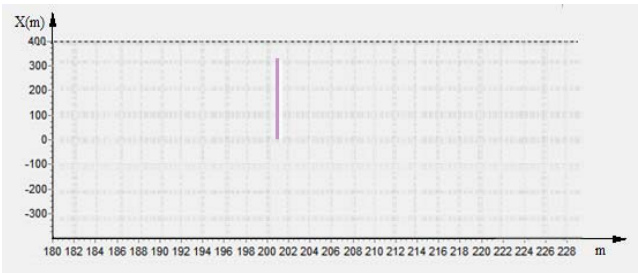

Рис. 10. ДКП оценка действительной части после компенсации

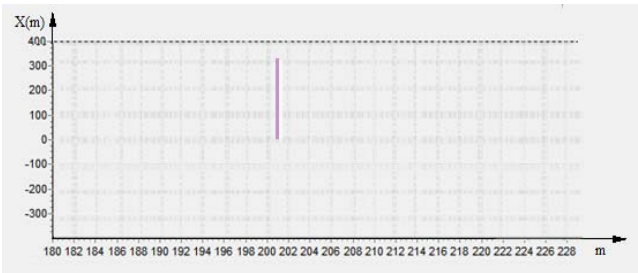

Рис. 11. ДКП оценка мнимой части после компенсации

В приведенных выше примерах рассмотрены сигналы кратные целому числу полупериодов, бинов, на длине выборки. Наиболее общий случай, это несовпадение анализируемого сигнала и частоты бина. На рисунке 12 приведен пример амплитудного спектра комплексного дискретного косинусного преобразования (КДКП) колебания, расположенного между 201 и 202 бином.

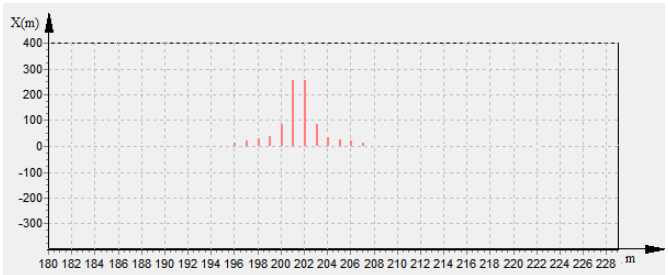

Рис. 12. Амплитудного характеристика оценки спектра КДКП сигнала 201,5 коэффициента

Для сравнения уровня боковых лепестков разработанного алгоритма КДКП и БПФ на рисунке 13 приведена амплитудная характеристика того же по частоте коэфициента полученного при быстром преобразовании Фурье с прямоугольным окном. Для сохранения частотного масштаба длительность выборки N увеличена вдвое.

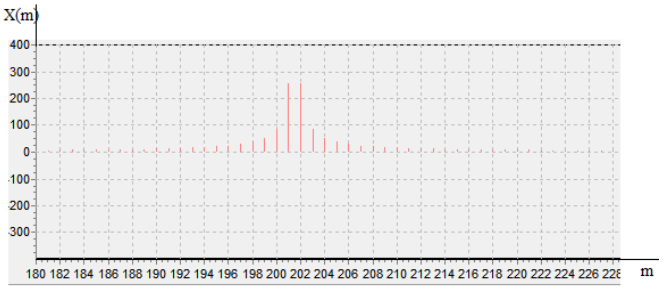

Рис. 13. Амплитудная характеристика спектра оценки БПФ сигнала 202,5 коэффициента

Определить точное значение параметров колебания, частоты, амплитуды, начальной фазы помогает анализ, кроме самого сигнала, набора транспонированных в пределах бина сигналов [7,8]. Точность определения параметров колебания ограничена только количеством используемых сдвигов. Наиболее логично задавать эту точность различной по спектру сигнала с учетом частотных свойств периферического слухового анализатора.

Одним из недостатков БПФ при анализе звукового сигнала является недостаточная разрешающая способность преобразования. Анализ с использованием ДКП и транспонированных сигналов позволяет решить эту проблему. Так становиться возможным распознавание двух спектральных составляющих расположенных в пределах бина по шкале частот. Правда распознавание зависит от начальной фазы колебания, но в худшем случае, при фазовом сдвиге между колебаниями на 0,1 радиан, распознавание осуществляется при разносе частот на 1,1 бин. Пример такого распознавания приведен на рисунке 14.

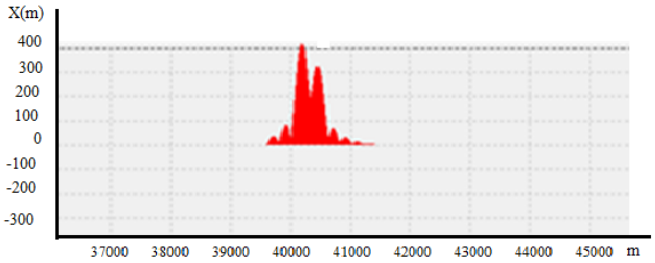

Рис. 14. Амплитудный спектр КДКП при 100 промежуточных сдвигах.

Развитый спектр боковых лепестков приводит к тому, что при анализе реального сигнала слабые спектральные составляющие полностью замаскированы боковыми лепестками более мощных спектральных составляющих. Практически при анализе выявляются составляющие отличающиеся от максимального не более чем на 30 дБ.

Предлагается использовать итерационный вариант анализа спектра. После выявления самой мощной спектральной компоненты это колебание синтезируется во временной области после чего вычитается из исходного сигнала. Далее оценка спектра повторяется в соответствии с разработанным алгоритмом но без самой мощной спектральной составляющей и ее боковых лепестков. На рисунке 15 приведен амплитудный спектр оценки КДКП с устраненной мощной составляющей.

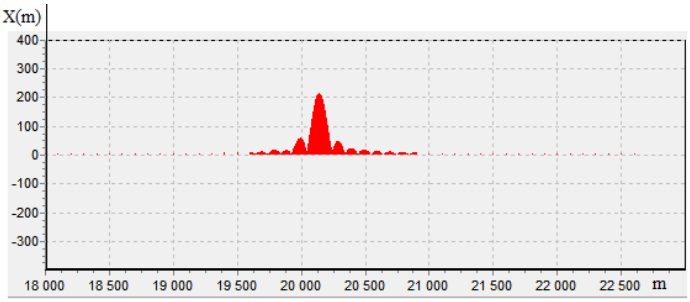

Рис. 15. Амплитудный спектр оценки КДКП с устраненной мощной спектральной составляющей

В таблице 1 приведенной ниже сведены основные результаты оценивания при использовании различных окон; там же приведены оценки сигнала, полученного из набора транспонированных на 64 сдвига по частоте сигналов.

Таблица 1

Результаты сравнения оценок спектра

Параметры	Окна			-	
	Прямоугольное	Треугольное	Хемминга	КДКП	КДКП со сдвигом
Число сдвигов	-				64
Макс. уровень бок. лепестков, дБ	-13	-27	-43	-10	
Скорость спада бок. лепестков, дБ/октава	-6	-12	-18	-6	-6
Когерентное усиление	1,00	0,50	0,54	1,00	1.00
Полоса по уровню 3,0 дБ, бин	0,89	1,28	1,30	0,44	0,44
Паразитная АМ, дБ	3,92	1,82	1,78	1,00	0.0028
Максимальные потери преобразования, дБ	3,92	3,07	3,10	1,00	-36
Точность определения частоты, Бин	1	1	1	0,5	1/64

В таблице использованы следующие характеристики:

- когерентное усиление – оценивает отношение суммы отсчетов сигнала, умноженных на окно, к их сумме (прямоугольное окно);
- паразитная АМ спектра – равна отношению когерентного усиления тона, расположенного посредине между двумя бинами ДПФ, к когерентному усилению тона, совпадающего с одним из бинов ДПФ;
- максимальные потери преобразования сумма максимальных потерь из-за паразитной АМ спектра для данного окна (в дБ) и потерь преобразования, определяемых формой окна;
- максимальный уровень боковых лепестков (по отношению к главному лепестку);
- скорость спадания боковых лепестков (дБ/октава);
- полоса по уровню 3 и 6 дБ.

Разработанный алгоритм КДКП позволяет вдвое увеличить разрешающую способность оценок мгновенного спектра СЗВ по сравнению алгоритмом БПФ. Кроме того, алгоритм КДКП, по сравнению с алгоритмом ДКП, обеспечивает уменьшение осцилляции амплитудной оценки мгновенного спектра СЗВ, и позволяет формировать оценки фазы спектральных составляющих СЗВ.

Проведенное исследование подтвердило перспективность выбранного направления исследований и целесообразность его развития при анализе мгновенных спектров реальных звуковых сигналов на малых временных интервалах. Такой анализ позволит формировать объективную оценку качества передачи и воспроизведения и прогнозировать субъективную оценку качества слушателем [5,6].

Заключение

Разработка алгоритма и его исследование подтвердило перспективность использования комплексного дискретного косинусного преобразования, связанного с повышением точности и разрешающей способности спектрального анализа.

Показано, что вычислительная сложность предложенного алгоритма может быть снижена за счет устранения 50% перекрытий между отдельными выборками, которые были необходимы при реализации спектрального анализа с помощью быстрого преобразования Фурье. Кроме того, в предложенном алгоритме имеется возможность проводить анализ сигнала «встык».

Важным достоинством является отсутствие необходимости использования оконных функций, что позволяет уменьшить ширину основного лепестка и соответственно увеличить разрешающую способность, которая изначально вдвое выше для ДКП по сравнению с БПФ.

Проведенные измерения подтвердили возможность реализации определения значений амплитуды, частоты и фазы спектральных составляющих с точностью приближенной к точности периферического слухового анализатора.

Анализ мгновенных спектров реальных звуковых сигналов на малых временных интервалах позволяет реализовать объективную оценку качества передачи и воспроизведения звукового сигнала [9,10]. Такой анализ на входе и выходе систем, позволяет формировать объективную оценку качества передачи и воспроизведения звуковых сигналов и прогнозировать субъективную оценку качества слушателем, упростив тем самым разработку новых поколений устройств компактного представления, обработки, передачи и звуковоспроизведения [11-14]. На основе рассматриваемого алгоритма можно с высокой точностью контролировать и регулировать не только спектральные характеристики звуковых сигналов, но и аппаратуру звукового вещания и телевидения. Это, в свою очередь, может способствовать увеличению популярности телерадиовещательных программ. Дело в том, что спектральные характеристики звуковых сигналов в существенной степени определяют эмоциональную наполненность программ телерадиовещания, которая в свою очередь, определяет популярность как самих этих программ, так и телерадиовещательных станций. Такое положение способствует росту числа слушателей и зрителей телерадиовещательных станций и росту доходов этих станций. Таким образом, использование предлагаемого алгоритма позволит, в том числе, повысить экономическую эффективность телерадиовещания.

Литература

1. *Mol A.* Theory of Information and Aesthetic Perception, Moscow: Mir, 1966, 211 p.
2. *Цвикер Э., Фельдкеллер Р.* Ухо как приемник информации. М.: Связь, 1965, 104 с.
3. *Abramov V. A., Popov O. B., Chernysheva T. V., Taktakishvili V. G., Ovchinnikov A. A.* On-board Transmission Quality Assessment Using Short Audio Signal // 2021 Systems of Signals Generating and Processing in the Field of on Board Communications, 2021, pp. 1-6, doi: 10.1109/IEEECONF51389.2021.9416085.
4. *Попов О.Б., Рихтер С.Г.* Цифровая обработка и измерения сигналов в трактах звукового вещания. М.: Инсвязьиздат, 2010. 292 с.
5. Патент RU 2573248 C2. Опубликовано 20.01.2016, БИ № 2. Способ измерения спектра информационных акустических сигналов телерадиовещания и устройство для его осуществления. Авторы: Абрамов В.А., Попов О.Б.
6. Патент RU 2756934 C1. Опубликовано 07.10.2021 БИ № 28 Способ и устройство измерения спектра информационных акустических сигналов с компенсацией искажений. Авторы: Абрамов В.А., Попов О.Б., Власюк И.В., Балобанов А.В.
7. *Abramov V. A., Popov O. B., Chernysheva T. V., Peruanski V. O.* Increasing the Accuracy of Sound Signal Spectral Estimation According to the Properties of Hearing Analyzer // 2021 Intelligent Technologies and Electronic Devices in Vehicle and Road Transport Complex (TIRVED) DOI: 10.1109/TIRVED53476.2021 11-12 Nov. 2021
8. ITU-R Rec. BS.1534. Method for the subjective assessment of intermediate quality level of coding systems, June 2001.
9. Исследование заметности искажений в радиовещательных каналах / Под ред. И. Е. Горона. М.: Связьиздат, 1959. 121 с.
10. Патент RU 2731339 C1. Опубликовано 01.09.2020 БИ № 25 Способ и устройство измерения мощности и крутизны нарастания участков нестационарности акустических сигналов. Авторы: Абрамов В.А., Попов О.Б., Тактакишвили В.Г.
11. Патент RU 2691122 C1. Опубликовано 11.06.2019 БИ №17 Способ и устройство компандирования звуковых вещательных сигналов. Авторы: Абрамов В.А., Попов О.Б., Орлов В.Г.
12. Патент RU 2731602 C1. Опубликовано 04.09.2020 БИ № 25 Способ и устройство компандирования с предыскажением звуковых вещательных сигналов. Авторы: Абрамов В.А., Попов О.Б.
13. *Тактакишвили В.Г., Попов О.Б., Абрамов В.А., Борисов А.А.* Методы компактного представления, оценки и обработки звуковых сигналов на основе их комплексного представления // T-Comm: Телекоммуникации и транспорт. 2019. Т. 13. № 2. С. 11-17.
14. *Абрамов В.А., Попов О.Б., Борисов А.А., Черников К.В.* Статистика звукового сигнала, представленного комплексными модулирующими функциями // T-Comm: Телекоммуникации и транспорт. 2017. Т. 11. № 6. С. 29-32.

МНОГОУРОВНЕВАЯ ПОЛИТИКА БЕЗОПАСНОСТИ СЕТИ ПЕРЕДАЧИ ДАННЫХ

Ерохин Сергей Дмитриевич,
Московский технический университет связи и информатики, ректор, к.т.н., Москва, Россия

Пилюгин Павел Львович,
Московский технический университет связи и информатики, ст. научный сотрудник, к.т.н., Москва, Россия
ppl@mail.ru

Аннотация

В статье рассматриваются условия безопасных информационных потоков для различных уровней сетевой модели, а также возможные механизмы для выполнения требований политик безопасности на каждом уровне. Политика безопасности должна обеспечивать согласованный подход для обеспечения безопасности на различных уровнях и для определения состава и архитектуры средств защиты. В статье на основе многоуровневой эталонной модели ISO/OSI и стека протоколов TCP/IP сети передачи данных, предлагается формулировать условия для безопасных информационных потоков между объектами каждого уровня (уровни приложений, транспортный и сетевой), так чтобы политики безопасности этих уровней были взаимосвязаны и дополняли друг друга.

Ключевые слова: *Субъектно-объектная модель, эталонная модель ISO/OSI, стек протоколов TCP/IP, политика безопасности уровня эталонной модели, уровни приложений, транспортный и сетевой.*

Введение

Для большинства типов решений сетевой безопасности понятийный аппарат локализован в пределах одного уровня эталонной сетевой модели ISO/OSI. Даже, когда речь идёт о комплексных и многоаспектных конгломератах, в их составе легко усматриваются отдельные слабо взаимодействующие компоненты, которые оперируют категориями одного вполне определённого уровня. Если и привлекаются понятия соседнего (как правило, более высокого) уровня, например, для криптотуннелирования или для межсетевого экранирования в классе State Inspection, то только в качестве средства, позволяющего определить условия управления механизмом безопасности. Другими словами, в рамках одного механизма безопасности воздействия на разные уровни сетевой модели ISO/OSI, как правило, не «смешиваются». Хотя и неизвестны теоретические ограничения, препятствующие созданию какого-то механизма безопасности, воздействующего на несколько уровней, но практика примеров таких механизмов не знает. Обеспечивая сетевую безопасность удобно опираться на специфическую для каждого уровня аксиоматику, которая определяет некоторое состояние сети как безопасное в терминах данного уровня. Таким образом, есть основания предполагать существование политик безопасности для каждого уровня модели ISO/OSI и утверждать, что эти политики разные.

Многоуровневое представление сети.

Это обстоятельство является некоторым «обременением» для управления безопасностью. Прежде всего, первоначальные требования и постулаты безопасности возникают в сфере семантически значимых информационных агрегатов – документов, сообщений, сведений, измерений и т.п. В сети эти категории отображаются на верхнем (прикладном) уровне, и существование политик на других уровнях вызывает вопрос насчет их адекватности. Кроме того, в процессе управления необходимо сочетать оценивание безопасности и воздействие на безопасность, и хорошо, если эти процедуры выполняются в рамках одного уровня. Если же, например, оценивание, чтобы расширить спектр оцениваемых факторов, будет выполняться на более высоком уровне, то это обязательно потребует согласованности политик соответствующих уровней.

Особенно рельефно это проявляется, когда речь идёт о безопасности критических информационных структур (КИИ), для которых инцидентом является состояние, не допускающее нормальное функционирование самого критического объекта. Поскольку начальный критерий такого состояния находится за пределами собственно КИИ, его признаки в КИИ отображаются, прежде всего, на прикладном уровне. В то же время, оценивание текущего состояния безопасности сетевой КИИ и детектирование компьютерных атак, например, с помощью сигнатур или аномалий трафика, выполняется на базе анализа признаков сетевого и транспортного уровня. Поэтому вопрос о том, насколько атака соответствует инциденту, в условиях КИИ становится далеко не тривиальным. Кроме того, затруднена оценка эффективности использования анализируемых признаков атак при переходе между различными уровнями.

Ещё одно затруднение возникает из-за множественности и несогласованности политик безопасности разных сетевых уровней при реализации асимптотического управления безопасностью КИИ [1]. Архитектура асимптотического управления предполагает коррекцию конфигурации и настроек механизмов защиты как реагирование на возникший или прогнозируемый инцидент. Как уже указывалось, инцидент в КИИ имеет свой паттерн, по крайней мере, на прикладном уровне, и наличие его на других уровнях не обязательно. А нотации настроек и конфигурирования реальных инструментов сетевой безопасности специфицируются параметрами более низких уровней.

Решение всех перечисленных проблем требует методических средств соотнесения политик безопасности различных уровней сетевой модели ISO/OSI.

Рассмотрим цели и политики безопасности уровня приложений для контроля потоков в сети передачи данных. Это позволяет в большинстве случаев обобщить полученные результаты как для контроля семантически значимой информации, так и для управляющей информации – т.е. для защиты от нелегитимных управляющих воздействий, приводящих нарушению целостности и доступности.

Обеспечение безопасности осуществляется контролем доступа к информации, т.е. защитой от НСД. Такие требования по контролю, описанные в политике сетевой безопасности, могут быть реализованы средствами защиты от НСД, которые на разных уровнях модели ISO/OSI могут рассматриваться как средства контроля информационных потоков (далее для простоты будем использовать модель TCP/IP, отметим, однако, что уменьшение числа уровней модели не ограничивает общность проводимых рассуждений). То есть под многоуровневым представлением политики безопасности здесь понимается не разная детализация политики для разных уровней управления в организации или разные уровни защищенности (MLS или модель Biba) [2]. В качестве универсальной модели контроля информационных потоков на различных уровнях стека TCP/IP используется субъектно-объектная модель управления доступом, в терминах которой понимаются понятия субъекта, объекта, доступа и потока, а основным механизмом защиты является монитор безопасности обращений (МБО) [2,3].

Кроме того, сам по себе контроль доступа (априорно или апостериорно) не имеет смысла без предварительной идентификации и аутентификации субъекта, инициирующего доступ. Собственно это и отражено в функциональной модели защиты – AAA (здесь это обозначает authentication, authorization, accounting и далее при описании механизмов защиты, где это не обговорено отдельно, акцент делается на функции аутентификации и авторизации, а акаунтинг (регистрация и проверка) подразумевается как апостериорная проверка (дополнение) авторизации или ее замена).

Субъекты доступа и защищаемые для обеспечения безопасности семантически значимые информационные объекты (документы, сообщения, сведения и пр.) существуют только на уровне приложений вычислительной сети [4]. На этом уровне сети передачи данных рассматриваются информационные обмены между различными сетевыми ресурсами (ассоциированными с людьми устройствами или информационными ресурсами доступными по сети). Соответствующие протоколы устанавливают правила обмена между такими ресурсами, а для обеспечения конфиденциальности и целостности здесь используются механизмы AAA.

Теоретически, можно было бы, используя формализацию субъектно-объектной модели для описания политик безопасности контроля информационных взаимодействий, ограничится рассмотрением только этого уровня, если предположить, что все нижележащие уровни сетевой модели TCP/IP абсолютно безопасны с точки зрения НСД или хотя бы, что информация верхнего уровня изолирована от них (например, криптогомоморфизм по схеме Гентри). Однако первое предположение невозможно, а второе практически пока не реализуемо [5].

В связи с этим необходимо учитывать возможность нарушений, в результате НСД к дейтаграммам (сегментам информации) на транспортном, к пакетам на межсетевом и к кадрам на сетевом уровнях стека TCP/IP. Причём на каждом уровне кроме специфических для этого уровня информационных объектов необходимо рассматривать соответствующих данному уровню субъектов – физические или виртуальные сетевые устройства. Как показано на рисунке 1 субъект уровня приложений реализует свои информационные доступы через один или несколько субъектов транспортного уровня.

Соответственно различные субъекты транспортного уровня могут реализовываться на одном устройстве межсетевого уровня. Собственно сетевой уровень несколько нарушает иерархичность данной сетевой модели, так как на этом уровне взаимодействия ограничены пространством адресов локальной сети. В связи с последним замечанием можно ограничиться рассмотрением уровней L2-L3 и L4, рассматривая особенности L2 по мере необходимости применительно к локальным сетям.

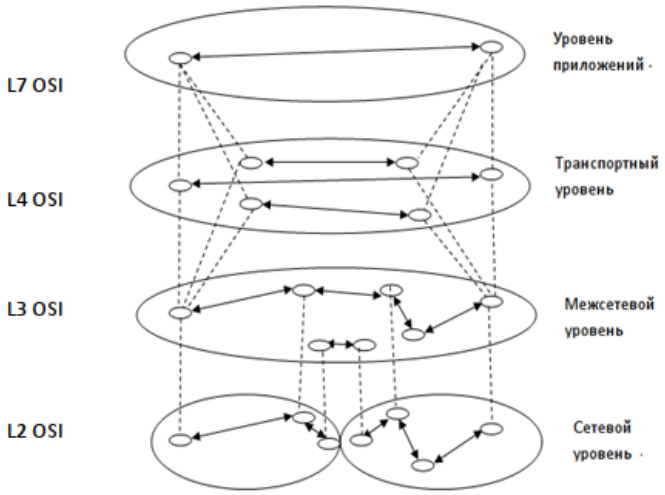

Рис. 1. Иерархия информационных взаимодействий стека TCP/IP

Информационные потоки и политика безопасности уровня приложений

На уровне приложений сетевые ресурсы устанавливают информационное взаимодействие по мере необходимости (по запросу пользователя). В простейшем случае для представления на графе связности сетевых ресурсов такого взаимодействия может быть использована связь точка – точка, т.е. ребро графа. Однако следует учитывать, что взаимодействие ресурсов: «субъект доступа (s) – информационный объект (o)» может использовать одновременно несколько протоколов (хотя это большей степени характерно транспортному уровню как будет показано ниже), например, почта, смс и пр. При этом особенно важно отметить, что информационный объект может сам выступать в роли субъекта доступа и запрашивать необходимую информацию у другого (или других) информационного объекта (такой объект будем обозначать как узел (x)).

Различные взаимодействия удобно описывать, используя рекуррентное определение доступа, как это предложено в теоретико-графовой модели Take-Grant и ее модификациях [6]. В рамках описанных выше понятий уточним формальное описание уровня приложений. Это совокупность узлов (субъектов и/или объектов) $X=\{x_r\}$, где r=1, 2,...n, между которыми может устанавливаться непосредственное информационное взаимодействие или через которые осуществляется трансляция данных для организации информационного взаимодействия (рис. 2).

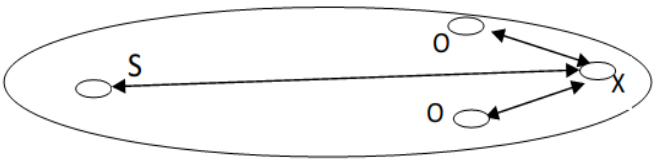

Рис. 2. Информационные взаимодействия уровня приложений

Информационную топологию сети – схему текущих потоков определяет квадратная *nxn* матрица связности Q, двоичные элементы q_{rk} которой означают, что информационная связь от x_r к x_k существует, если $q_{r,k}=1$ или не существует, если $q_{r,k}=0$. Отметим, что такая топология может динамично меняться, так как информационная связь устанавливается по требованию, и матрица связности в этом случае представляет текущую топологию сети. Информационный доступ из узла $x_s \in S$ (субъекта) к узлу $x_o \in O$ (объекту) будет возможен, когда выполняются условия:

$$\exists\ x_s \to x_o \ \textit{если}\ \exists\ \{x_i\}\ \textit{, такие, что}\ q_{i,i+1}=1 \qquad (1)$$

где i=s,s+1,...,o

Здесь x_i выступают в роли транзитных узлов или самостоятельных узлов, распространителей информации, например, как в расширенной модели Take-Grant, кроме того отметим, что это условие существования цепочки доступов описывает только возможность доступа, оно является необходимым, но не достаточным условием. В реальных сетях, например, для почтовых служб, роль таких узлов выполняют почтовые сервера или NS сервера для DNS запросов.

Для реализации монитора безопасности обращений необходим контроль возникающих информационных потоков и блокирование всех нелегитимных потоков (по сути, речь идет о реализации модели с полным перекрытием Клементса-Хоффмана) [7]. Как описывалось выше, для этого монитор безопасности должен использовать механизмы защиты AAA, в которых после успешной аутентификации для авторизации используются политики безопасности (дискреционная, мандатная или др.). Для синхронизации и взаимодействия механизмов обычно используется единый центр (трехсторонние протоколы) аутентификации и авторизации (например, Kerberos) [8].

Обеспечение безопасности информационного потока на этом уровне в по критерию безопасности, предложенному в [4], предполагает, чтобы любой субъект $x_s \in S$ был аутентифицирован и авторизован механизмами защиты (m), а все остальные узлы цепочки доступа либо также были аутентифицированы и авторизованы либо была гарантирована их недоступность сторонним субъектам. Все попытки неавторизованных действий должны пресекаться и регистрироваться.

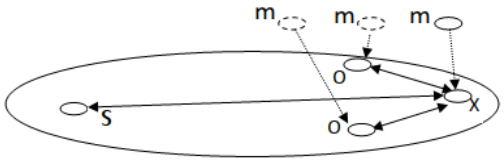

Рис. 3. Контролируемые взаимодействия уровня приложений

Для того чтобы разрешать информационные потоки кроме необходимого условия (возможность реализации потока) необходимо учитывать безопасность такого информационного обмена как достаточное условие его реализации.

Любая политика доступа, основанная на дискреционном, мандатном, ролевом или другом принципе управления доступом [2] должна определять безопасные информационные потоки. В субъектно-объектной моделью функцию проверки соответствия потоков политике P в сети выполняет монитор безопасности [3]. То есть для любой пары информационных объектов (x_s,x_o) должно быть определено «отношение безопасного потока» $x_s \Rightarrow x_o$, которое будет определять политику безопасности сети $P=\{(x_s,x_o)\}$. Такое отношение будет рефлексивно, так как всегда возможен поток «внутри» информационного объекта и транзитивно, потому, что если каждая передача информации безопасна в отдельности, то, очевидно, должна быть безопасна и передача информации по цепочке таких безопасных информационных передач. Однако оно не антисимметрично, так как двунаправленный обмен информацией не только возможен, но часто только такой обмен возможен в вычислительных сетях, можно отметить, что это соответствует свойствам потока, приведённым в [9].

Следовательно, если задана политика *P*, то необходимое и достаточное (с точки зрения обеспечения безопасности) условие существования потока определяется следующим образом:

$$\exists\ x_s \to x_o\ ,\ \textit{тогда и только тогда, когда}\ \exists\ \{x_i\}\ ,$$

$$\textit{такие, что}\ q_{i,i+1}=1\ \text{и}\ x_i \Rightarrow x_{i+1} \qquad (1')$$

где $i=s,s+1,\ldots,o$

Атаки на уровень приложений с нижележащих уровней сети

Описанная выше логически замкнутая модель была бы вполне корректной и безопасной, если предположить, что все нижележащие уровни сетевой модели TCP/IP абсолютно безопасны или изолированы. Однако, существует возможность нарушений безопасности, в результате НСД к дейтаграммам на транспортном, к пакетам на межсетевом и к кадрам на сетевом уровнях стека TCP/IP.

То есть в модели TCP/IP, необходимо учитывать некорректность работы и возможные атаки на незащищенные сетевые протоколы (например, DNS) или средства защиты: механизмы аутентификации (например, атаки повтором) и авторизации (например, возможность троянских коней для дискреционной политики доступа).

Это вызвано тем, что протоколы сетевых я ориентированы, прежде всего, на транспортировку дейтаграмм или пакетов, в результате узлы этого уровня могут быть доступны сторонним узлам, являющимся реализацией (проекцией субъекта) нарушителя, и атакованы ими.

В частности, например, протоколы UDP и IP подвержены атакой имперсонацией, так как не предусматривает проверки атрибутов отправителя. А более надежный протокол TCP может быть подвергнут десинхронизации и в результате также установлен контроль или соединение от имени нарушителя.

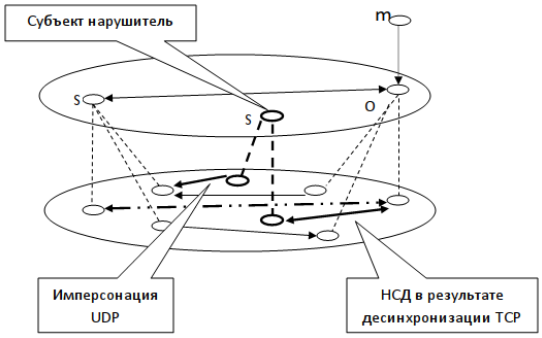

Рис. 4. Атаки на уровень приложений из транспортного уровня

При этом следует отметить, что набор параметров, по которым принимается решение о разрешении или запрещении информационного обмена, ограничен на этих уровнях идентификационными адресными параметрами (которые могут быть подделаны, если предварительно аутентификация не проводилась).

Следовательно, возможность определить легитимный или несанкционированный поток эти протоколы не позволяют. Так они ограничены или не имеют возможности применения механизмов защиты. AAA.

Аутентификация возможна при дополнении протоколов транспортного уровня, однако реализуется она на основе информации уровня приложений (например, TLS использует: пароли, ключи, сертификаты и пр.). Результат такой аутентификации может быть использован как на уровне приложений, так и на транспортном уровне, однако в последнем случае обычно происходит и авторизация – разрешено или запрещено информационное взаимодействие по такому модифицированному или дополнительному протоколу.

Авторизация на транспортном и сетевом уровне в наиболее распространённом механизме защиты (межсетевой экран) основана на идентификаторах без процедуры аутентификации. Это связано с тем, что реализация протокола аутентификации требует установить перед этим информационное взаимодействие, которое также может быть небезопасно в результате атаки на протоколы аутентификации [8].

Однако в корпоративных сетях средства авторизации могут быть расширены возможностями аутентификации (например, по паролю или ключевым параметрам, получаемым из уровня приложений) или более сложными процедурами двухуровневой и трехсторонней аутентификации (например, сеансовые протоколы RADIUS, TACACS или Kerberos).

Акаунтинг на транспортном и сетевом уровне играет более значительную роль, так как регистрация и анализ нарушений или отклонений в работе протоколов позволяют обнаружить попытки НСД (например, средствами IDS и IPS) и оперативно использовать эти данные (рис. 5). Однако следует учитывать, что это вызывает необходимость хранения и обработки больших объемов информации

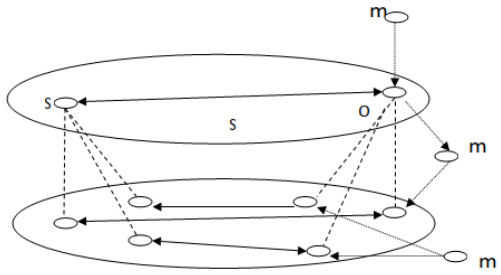

Рис. 5. Контролируемые взаимодействия транспортного уровня

Следовательно, содержательная информация уровня приложений должна быть либо полностью изолирована от нижележащих уровней, через которые технически реализуется информационное взаимодействие (что только отчасти возможно для обеспечения конфиденциальности и проверки целостности информации) либо необходимо обеспечить безопасность этих уровней и соответственно определить политики безопасности для каждого из них.

Чтобы различать субъекты и объекты разных уровней, (если особо оговорено чем идет речь) введем обозначение соответствующего уровня для обозначений в формальной модели: s_i^l - для субъекта, o_i^l - для объекта и x_i^l - для произвольного узла уровня *l*, и пусть *a*-уровень приложений, *t*- транспортный, *m* – межсетевой (интернет), *n* – сетевой Тогда, например, выражение (1') примет вид:

$$\exists x_s^a \to x_o^a \text{, тогда и только тогда когда } \exists \{x_i^a\},$$

$$\text{такие, что } q_{i,i+1}^a = 1 \text{ и } x_i^a \Rightarrow x_{i+1}^a \qquad (1'')$$

где i=s,s+1,…,o

Потоки и политика безопасности транспортного уровня

Субъекты, объекты и узлы транспортного уровня можно представить как реализации (проекции) соответствующих элементов уровня приложений на транспортный уровень сети. Собственно на этом уровне информация (сообщения, документы и пр.) разбиваются на сегменты – дейтаграммы. Эти информационные объекты не анализируются в зависимости от контента и не содержат атрибутов информации верхнего уровня, так как основная задача этого уровня надежная (и/или скоростная) транспортировка, что и обеспечивается протоколами этого уровня.

Например, для стека протоколов TCP/IP проекция (возможно не одна) каждого узла уровня приложений на транспортный представляется сокетом (или набором сокетов) – <адрес:порт>, где адрес это адрес устройства (IP адрес узла межсетевого уровня), а порт (TCP или UDP) идентификатор исполняемого процесса – точки входа в приложение (т.е. одна из проекций конкретного субъекта/объекта уровня приложений).

Формальное описание этого уровня может произведено на основе более простой модели, чем для уровня приложений, так как информационный обмен определяемый протоколами этого уровня (например, UDP или TCP) устанавливается строго между двумя узлами (сокетами) этого уровня, а цепочки из таких информационных обменов определяются существованием цепочек обменов уровня приложений. При этом промежуточный узел уровня приложений должен создавать две проекции: принимающий сокет и передающий сокет для каждого потока.

Таким образом, при формализации транспортного уровня мы также можем рассматривать пары субъект – объект и информационный обмен дейтаграммами между ними. В рамках описанных выше понятий для транспортного уровня приложений будем рассматривать совокупность узлов (субъектов S^t или/и объектов O^t) $X^t=\{x_i^t\}$, i=1, 2, ...m, между которыми может попарно устанавливаться информационное взаимодействие. Топологию сети определяет квадратная mxm матрица связности Q^t, двоичные элементы $q_{r,k}^t$ которой означают, что информационная связь от x_s^t к x_o^t существует, если $q_{s,o}^t=1$ или не существует, если $q_{s,o}^t=0$.

$$\exists\, x_s^t \to x_o^t, \textit{ если } q_{s,o}^t=1 \tag{2}$$

Отметим, что такая информационная топология также может динамично меняться, так как инициируются из уровня приложений, однако доступ одного подключённого к сети транспортного узла к другому также подключённому к сети узлу всегда технически осуществим, если не установлены специальные ограничения или узел в данный момент недоступен.

Можно описать условие потока уровня приложений через информационные обмены транспортного уровня, так как существование такого обмена является необходимым условием потока между двумя узлами уровня приложений. В этом случае для каждой пары (x_i^a, x_{i+1}^a) из выражения (1') должна существовать пара (x_i^t, x_{i+1}^t) из выражения (2), где $\left\{x_i^a\right\}$ имеет для каждого узла x_i^a, по крайней мере две проекции на транспортный уровень $\{(x_{s,i}^t, x_{o,i}^t)\} = \downarrow_t^a \left\{x_i^a\right\}$:

$$\exists\; x_s^a \to x_o^a, \text{ если } \exists \left\{x_i^a\right\}^{\mathrm{t}}, \textit{ такие, что}$$

$$\forall (x_{s,i}^t, x_{o,i+1}^t) \quad q_{i,i+1}^t = 1 \tag{1'''}$$

где $i=s,s+1,...,o$ и $(x_{s,i}^t, x_{o,i}^t) = \downarrow_t^a \left\{x_i^a\right\}$

Важной особенностью для описания потока между узлами транспортного уровня и отображения на него уровня приложений является атрибуция узлов в существующей версии сети с коммутацией пакетов IPv4 (IPv6). Каждому узлу транспортного уровня соответствует сокет - пара <IP_адрес:порт>, где «IP_адрес» это адрес узла (устройства) межсетевого уровня, а «порт» как правило определяет приложение использующее данный узел транспортной сети. При этом в версии IPv4 для определения потока учитывается только значение порта получателя (слушающего порта) $x_{o,i+1}^t$, в темя как для источника соединения $x_{s,i}^t$ значение порта может быть любым. Это обуславливает в том числе возможность использования технологий NAT и PAT в современных сетях для узлов источников соединения. То есть важно существование (доступность) слушающего узла соответствующего конкретному приложению.

Политика безопасности транспортного уровня P^t также как и для уровня приложений определяется множеством разрешенных потоков между узлами транспортного уровня $P^t=\{(x_s^t, x_o^t)\}$. То есть P^t также устанавливает отношение разрешенного потока $x_s^t \Rightarrow x_o^t$. Тогда достаточное условие (2) существования потока может быть представлено в виде:

$$\exists\, x_s^t \to x_o^t, \textit{тогда и только тогда, когда } q_{s,o}^t=1 \text{ и } \exists\, x_s^t \Rightarrow x_o^t \tag{2'}$$

Отметим принципиальное отличие условий существования потока (1') и (2'), заключающееся в том, что выражение (2') не предусматривает никаких цепочек обмена информацией как в (1') так как узлы транспортного уровня суть сетевые ресурсы, которые для $X^t=\{x_i^t\}$ образуют граф всех подключенных и доступных по транспортному протоколу узлов сети. Цепочки образующие информационный поток (распространение информации в уровне приложений) в (1') предполагают, что любой транзитный узел или самостоятельно вещающий и дублирующий информацию узел уровня приложений должен иметь по крайней мере 2-е реализации-проекции: субъект и объект доступа на транспортном уровне как это указано в выражении (1''').

Вместе с тем «контентная ограниченность» протоколов транспортного уровня только параметрами дейтаграмм позволяет использовать средства изоляции от нижележащих уровней для обеспечения конфиденциальности и проверки целостности. Например, протоколы TLS, SSH и другие средства туннелирования транспортного уровня, когда один сетевой протокол инкапсулируется в другой с использованием криптографических методов защиты.

Когда изоляция информации для обеспечения конфиденциальности не используется или в первую очередь должны быть решены проблемы обеспечения целостности и доступности, то должна быть обеспечена безопасность от НСД и определена политика безопасности для сетевого или межсетевого уровней. Так как проблема нарушения безопасности транспортного уровня может возникнуть из любого из этих уровней.

Потоки и политика безопасности межсетевого и сетевого уровня.

Узлы межсетевого уровня обмениваются информацией виде пакетов (IP – пакетов), передаваемых от узла к узлу, а каждый узел сети имеет свой собственный адрес (IP- адрес). Этот адрес позволяет сопоставить любой узел транспортного уровня <IP_адрес:порт> с узлом межсетевого уровня <IP_адрес>. Соответственно для существования потока между узлами транспортного уровня $x_s^t \to x_o^t$, должен существовать поток между узлами $x_s^m \to x_o^m$, где каждый x_i^m проекция соответствующего x_i^t на межсетевой уровень $x_i^m = \downarrow_m^t \left\{ x_i^t \right\}$.

Существование потока $x_s^m \to x_o^m$ связано с двумя взаимоисключающими условиями:

В первом случае x_s^m и x_o^m находятся в одной локальной сети X^m (применительно к локальной сети обозначения X^m и X^n описывают одно и тоже множество узлов, но с разной адресацией, поэтому далее используем эти обозначения в зависимости от того какой адрес подразумевается). Тогда, так как IP адрес устройства однозначно соответствует его аппаратному адресу, то на сетевом уровне в локальной сети все устройства доступны друг другу, за возможно исключением специально установленных ограничений в узлах коммутации (например, VLAN или ограничения на правила потока в программных коммутаторах SDN), т.е. будем полагать, что почти всегда $x_s^m \to x_o^m$ существует.

Во втором случае x_s^m и x_o^m находятся в разных локальных сетях, тогда соответствующие им устройства на сетевом уровне не доступны друг другу (рис.1). Для существования потока $x_s^m \to x_o^m$ в этом случае должен быть организован маршрут из одной локальной сети X_1^m в другую X_n^m, через специальные узлы: шлюзы $x_{tr\_s}^m$ и $x_{tr\_o}^m$ и, если это необходимо, через другие транзитные шлюзы-маршрутизаторы других сетей. Каждый такой шлюз связан со всем узлами своей локальной сети и с шлюзом другой сети (для этой связности они дополнительно образуют свою «межмаршрутизаторную» локальную сеть), что позволяет им предавать пакеты из одной сети в другую (рис. 6).

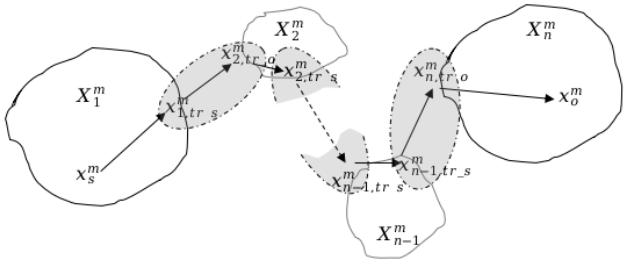

Рис.6. Маршрутизация в глобальной сети

Матрица связности Q^m между узлами x_i^m в общем случае не используется для описания маршрутизации, так как в локальной сети можно считать, что все узлы доступны, а для глобальной сети она не имеет смысла из-за своих размеров, даже если описывать только все узлы маршрутизаторы. Условие существования потока можно в этом случае представить в виде:

$$\exists\ x_s^m \to x_o^m,\ \textit{если}\ \exists\{X_i^m\},\ \textit{где}\ i=1,\ldots,n$$
$$\textit{и}\ x_s^m \in X_1^m,\ x_o^m \in X_n^m\ \text{ и }\ \exists\ x_{i,tr\_s}^m, x_{i,tr\_o}^m \in X_i^m \qquad (3)$$
$$\textit{такие что}\ \exists\, x_{i,tr\_s}^m \to x_{i+1,tr\_o}^m$$

По аналогии с условием (1''') можно описать условие потока на уровне приложений через последовательные отображения субъектов и объектов на транспортный и межсетевой уровни и существование передачи информации в соответствии с условиями (3).

Политика безопасности для локальной сети P^{m+n} предполагает, что ограничения вводятся на потоки между узлами X^m или X^n, что данном случае одно и тоже и позволяет использовать оба вида адресации.

Для реализации атак злоумышленник может в локальной сети просто подключить свое устройство к сети и использовать установленные на нем приложения для взаимодействия с другими узлами сети. Это позволяет организовывать все атаки описанные выше. Кроме того, для имперсонации может использоваться атака на протокол ARP, что позволяет злоумышленнику перехватывать информацию и выдавать себя за другого на межсетевом или транспортном уровне, если они не изолированы или не используют дополнительных средств аутентификации и авторизации.

Для отражения подобных атак механизмами обеспечения безопасности допускается разбиение сети на локальные несвязные сегменты (VLAN), что ограничивает возможности злоумышленника только одним сегментом. Так же в межсетевых экранах допускается идентификация по физическому порту, что может пресечь трафик злоумышленника, если он не должен иметь физического доступа к оборудованию.

Таким образом, для локальной сети можно установить в политике безопасности и контролировать механизмами защиты разрешенные потоки информации, что позволяет ввести для существования потока в локальной сети дополнительное требование - отношение разрешенного потока между узлами сети: $x_s^{m+n} \Rightarrow x_o^{m+n}$.

Политика безопасности межсетевого уровня глобальной сети P^m должна учитывать все политики безопасности P_i^{m+n} для всех X_i^m из (3), которые позволяют организовать информационный поток. Однако если в локальной сети можно предположить, что все узлы управляются (администрируются) из единого центра с соблюдением установленной политики безопасности, то для глобальной сети это не так. В этом случае политика безопасности должна определять и запрещать недоверенные маршруты.

Определим множество доверенных маршрутов D, как маршруты, которые используют транзитные узлы авторизованных локальных сетей (или автономных систем). К таким сетям можно отнести сети с центрами администрирования (субъектами управления), для которых у субъектов уровня приложений определены и согласованы SLA (Service Level Agreement) в отношении обслуживания трафика и правил безопасности (политик P_i^{m+n}):

$$\exists\ x_s^m \to x_o^m\ ,\ \text{если}\ \exists\{X_i^m\} \subseteq \mathrm{D},\ \textit{где}\ i=1,\ldots,n$$
$$\textit{и}\ x_s^m \in X_1^m,\ x_o^m \in X_n^m\ \textit{и}\ \exists\ x_{i,tr\_s}^m, x_{i,tr\_o}^m \in X_i^m \qquad (3`)$$
$$\textit{такие что}\ \exists\, x_{i,tr\_s}^m \to x_{i+1,tr\_o}^m$$

Таким образом, механизмы защиты (межсетевые экраны, правила маршрутизации, протоколы управления маршрутизацией, например, BGP и пр.) должны обеспечивать выполнение политики безопасности P^m.

В случае невозможности определения SLA для транзитных узлов и сетей возможно использование защищенных протоколов межсетевого уровня (PPTP или IPSEC) для обеспечения конфиденциальности и проверки целостности информации.

Акаунтинг на межсетевом и сетевом уровнях так же важен, так как регистрация и анализ нарушений или отклонений в работе протоколов позволяют обнаружить попытки НСД (например, возможна имперсонация на этих уровнях, так как протоколы не предусматривают аутентификацию, так же возможны атаки на маршрутизацию, подмена шлюза или изменение правил маршрутизации), но возможности такой регистрации ограничены при большом объёме передаваемого трафика.

Заключение

Роль политики безопасности заключается не только в отражении и предотвращении возможных атак и их негативных последствий. Часто такую атаку довольно сложно распознать и для этого используются средства обнаружения вторжений, основанные на сигнатурах атак [10] или различные методы обнаружения аномалий [11].

Однако также одним из важнейших критериев выявление событий безопасности и инцидентов является отклонение от принятой политики безопасности, которое многоуровневая политика безопасности позволяет выявлять на каждом уровне эталонной модели ISO/OSI или в стеке протоколов TCP/IP для сетей передачи данных, как это описано выше.

Следовательно, можно предположить, что обеспечение безопасности для сетей передачи данных должно учитывать многоуровневую модель сети и контролировать потоки информации между сущностями каждого уровня. Политики безопасности сети передачи данных должны формулироваться для каждого уровня стека протоколов TCP/IP (или эталонной модели ISO/OSI) и дополнять друг друга.

Это позволяет, во-первых, формулировать требования политики безопасности используя определения сущностей и понятия соответствующего уровня; во-вторых, предотвращать возможные атаки с нижележащих уровней; и, в-третьих, выявить возможные слабости системы защиты и обнаруживать новые потенциально опасные события безопасности.

Литература.

1. *Ерохин С.Д., Петухов А.Н., Пилюгин П.Л.* Управление безопасностью критических информационных инфраструктур. М.: Горячая линия – Телеком, 2021. 240 с.
2. *Девянин П.Н.* Модели безопасности компьютерных систем: Учеб. пособие. М.: Изд. центр «Академия», 2005. 144 с.
3. *Ерохин С.Д., Пилюгин П.Л.* Модель абстрактного сетевого сервиса. Сборник трудов XIV Международной отраслевой научно-технической конференции. Технологии информационного общества. Москва, 2020. С. 245-250.
4. *Конявский В.А, Гадасин В.А.* Основы понимания феномена электронного обмена информацией. Минск, 2004. 327 с
5. *Жиров А.О., Жирова О.В., Кренделев С.Ф.,* Безопасные облачные вычисления с помощью гомоморфной криптографии. Журнал БИТ: безопасность информационных технологий. 2013. С. 6-12.
6. *Петухов А.Н., Пилюгин П.Л.* «Управление конфиденциальностью в децентрализованных сетях с динамической топологией». Труды конференции REDS 2021. С. 356-360.
7. *Хоффман Л.Д.* Современные методы защиты информации./ под ред.В.А.Герасименко. М.: Сов. радио, 1980. 264 с.
8. *Черемушкин А.В.* Криптографические протоколы: основные свойства и уязвимости. Издательство: ИЦ "Академия" 2009. 272 с. ISBN: 978-5-7695-5748-4
9. *Деннинг, Дороти Э.* «Решетчатая модель защищенного информационного потока». Сообщения ACM. 1976. № 19(5). С. 236-243. doi 10.1145 / 360051.360056
10. *Karen Kent Frederick*, Network Intrusion Detection Signatures, http://www.securityfocus.com. December 19, 2001
11. *Шелухин О.И.* Сетевые аномалии. Обнаружение, локализация, прогнозирование. М.: Горячая линия – Телеком 2020. 448 с.

НЕЙРОСЕТЕВОЕ РАСПОЗНАВАНИЕ ВОДОУПОРОВ ПО ЛИТОЛОГИЧЕСКИМ ОПИСАНИЯМ ГЕОЛОГИЧЕСКИХ СЛОЕВ

Палагушин Александр Дмитриевич,
Московский технический университет связи и информатики, магистрант, Москва, Россия
a.palagushin@palagushin.ru

Воронов Вячеслав Игоревич,
Московский технический университет связи и информатики, к.т.н., доцент, Москва, Россия
vorvi@mail.ru

Аннотация

Целью данной работы является создание модели искусственной нейронной сети для распознавания водоупоров по литологическому описанию геологических слоев. В работе рассматривается пример семантического анализа текстовых данных на основе литологического описания пород из каталога скважин. В процессе создания модели используется лемматизация для конструирования признаков с помощью предобученной модели. Проведен анализ полученных результатов и сравнение обученной рекуррентной нейронной сети с модификациями созданной модели и алгоритмом логистической регрессии.

Ключевые слова: *рекуррентная нейронная сеть, лемматизация, водоупор, водопроницаемый слой, литологическое описание геологических слоев, каталог скважин*

Введение

В настоящее время машинное обучение применяется во многих сферах человеческой деятельности и позволяет решать весьма сложные задачи. В частности, стоит отметить высокие возможности машинного обучения в автоматизации технологических процессов и производств для сокращения участия в них человека. Подобные работы ведутся в различных организациях, в том числе на кафедре ИСУиА МТУСИ [1-3].

Машинное обучение согласно кривой Гартнера [4] лишь начинает свое развитие, в частности с развитием квантовых вычислений.

В области гидрогеологии машинное обучение также имеет применение, в том числе для решения задач, связанных с прогнозами загрязнения подземных вод [5], для осуществления прогноза уровней подземных вод [6], а также для поиска и разведки подземных вод [7]. Все вышесказанные примеры позволяют эффективно применять машинное обучение в данной сфере.

В представленной работе нейросетевое распознавание используется в составе NLP (Natural Language Processing). В качестве исходных данных используется каталог скважин, в котором данными для обучения является столбец с литологическими описаниями пород, полученных в результате бурения скважин. Задачей данной работы является создание нейронной сети, распознающей водоупорные слои по текстовому описанию литологических разностей.

Описание исходных данных

В рамках различных гидрогеологических исследований (гидрогеологическая съемка, поиск, разведка и эксплуатация месторождений подземных вод) часто создаются каталоги скважин, включающие информацию о конструкции, назначении, использовании скважин, а также содержащие литологическое описание пород при бурении скважины.

При гидрогеологической съемке, задачей которой является создание комплекта авторских или Государственных гидрогеологических карт, важно корректно составить гидрогеологическую стратификацию изучаемой территории. Гидрогеологическая стратификация включает в себя описание водоносных и водоупорных горизонтов различных возрастов для конкретного исследуемого участка. При создании гидрогеологической стратификации специалистами (гидрогеологами) активно используются ретроспективные (фондовые) данные и результаты полевых работ, содержащиеся в каталоге скважин. Гидрогеологу в рамках создания гидрогеологической стратификации и при построении самой

гидрогеологической карты (для отрисовки границ распространения гидрогеологических горизонтов) важно выделить водоупорные и водоносные слои различных возрастов в каталоге скважин.

Исходный набор данных, используемый для разработанной НС (нейронной сети) представляет из себя таблицу формата Excel с некоторым форматированием, содержащим информацию о скважинах со следующими данными: номер скважины; местоположение скважины, координаты скважины; год бурения, абсолютная отметка скважины, глубина, геологический индекс пород, литологическое описание пород, мощность слоя, глубина залегания эксплуатируемого водоносного горизонта, а также данные о конструкции скважины [8]. Фрагмент исходного набора данных представлен на рисунке 1.

Из вышеуказанной таблицы для создания обучающего и тестового наборов данных в рамках разработанной модели проведена подготовка данных с извлечением столбца с литологическим описанием пород и дальнейшей ручной разметкой данных. В итоговой таблице содержатся 2 столбца – 1-й столбец с литологическим описанием пород, 2-й столбец с номером класса: 0 – водопроницаемый слой, 1 – водоупорный слой. Всего в итоговой выборке 307 примеров. Фрагмент таблицы с размеченной выборкой представлен на рисунке 2.

Архитектура нейронной сети

Архитектура разработанной нейронной сети представлена на рисунке 3.

Первым слоем разработанной нейронной сети является слой Embedding, включающий входной слой и собственно слой Embedding, который преобразует идентификаторы слов во вложения, представляющие собой матрицу векторов исходных данных, в данном случае, индексов слов, таким образом, что слово, похожее на другое слово будет находиться рядом в некотором многомерном пространстве. В разработанной сети на слое вложений используется 128-ми мерное пространство.

№ п/п	№ скважины по каталогу	№ скважины по первоисточнику	Источник обследования	№ акта обследования	Местоположение скважины (привязка к местным ориентирам или адрес)	Долгота по GPS/ГЛОНАСС	Долгота по паспорту	Широта по GPS/ГЛОНАСС	Широта по паспорту	Организация, бурившая скважину	Год бурения	Абсолютная отметка устья скважины по первоисточнику, м	Абсолютная отметка устья скважины по результатам работ 2019 - 2021 гг., м	Глубина скважины по паспорту, м	Геологический индекс пород по первоисточнику	Литологическое описание пород	Мощность слоя, м	Глубина подошвы слоя, м	Индекс эксплуатируемого водоносного горизонта	Мощность эксплуатируемого водоносного горизонта, м	Глубина залегания эксплуатируемого водоносного горизонта, м	Конс Диаметр обсадной колонны, мм
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Гидрогеологические эксплуатационные (действующие и законсервированные), разведочные																						
1	1	06735	1р/1	1	Волгоградская область, Даниловский район, х. Атамановка, 150 м севернее Атамановской СОШ	44°4'40"E	44°04'4"E	50°10'7"N	50°10'15"N	Минеральсоз РФ Госкомцарп "Роснедстрой", АО "Сельхозводстрой"	1993 г.	104.0	104.0	32.0	alQ_{III}	Глина бурая и серая	8.0	8.0				шахта 1000,0
															alQ_{III}	Песок серый разнозернистый	2.0	10.0				
															н.с.	Глина серая	1.0	11.0				
															N_1gr	Песок слюдистый разнозернистый с прослойками песчаника	18.0	29.0	$2(N_{1-2}srv\text{-}zan)$	21,0 (вскрытая)	11.0	324,0
															N_1gr	Глина серая слюдистая	3,0 (вскрытая)	-				
															daQ_{III-IV}	Суглинок желто-бурый, плотный	12.0	12.0				жета 400,0

Рис. 1. Фрагмент каталога скважин из работы [8]

	Литологическое описание пород	CLASS
7	Глина бурая и серая	1
8	Песок серый разнозернистый	0
9	Глина серая	1
10	Песок слюдистый разнозернистый с прослойками песчаника	0
11	Глина серая слюдистая	1
12	Суглинок желто-бурый, плотный	1
13	Песок серый, крупнозернистый	0
14	Глина серовато-зеленая, плотная	1
15	Глина темно-серая до черной, плотная	1

Рис. 2. Фрагмент таблицы с размеченной выборкой

Рис. 3. Архитектура разработанной нейронной сети

Далее, из слоя вложения сигнал поступает на два слоя GRU (Gated Recurrent Unit) по 128 нейронов в каждом,где в последнем слое GRU возвращается только выход последнего временного шага.

Выходной слой представляет собой одиночный нейрон, использующий сигмоидальную функцию активации, который возвращает прогнозируемый класс. Для обучения используется продвинутый алгоритм градиентного спуска Adam [9], функция стоимости Binary cross entropy [10], стремящаяся приблизить распределение прогноза НС к целевому регулируя не только ошибочные предсказания, но и неуверенные, и метрика accuracy, определяющая процент совпадения прогнозируемых значений с истинными [11, 12].

Основные этапы работы программы

Работа программы разделена на блок извлечения и подготовки данных и собственно блок программы, включающий обработку данных и нейросетевое моделирование

В блоке извлечения и подготовки данных используется библиотека pandas для удобства работы с таблицей. На первом этапе подгружается исходная таблица и преобразуется в DataFrame – внутренний табличный формат библиотеки pandas. Далее происходит обработка таблицы DataFrame и подготовка обучающей выборки: индексация полей таблицы, удаление дублирующихся и пустых строк, преобразование типов данных столбцов в правильные форматы данных. После указанных действий выделяется целевой столбец и выгружается в таблицу Excel.

Перед тем, как перейти к следующему блоку работы программы производится ручная разметка данных с присвоением классов каждому примеру.

В блоке обработки данных и работы НС используются следующие библиотеки: pandas, sPacy, NumPy, некоторые модули SciKit-learn, Re, TensorFlow, модуль zip_longest из itertools, модуль pyplot из Matplotlib, а также вспомогательные модули для улучшенной визуализации полосы прогресса обучения.

На первом этапе подгружается набор данных из размеченного файла, разделяется на примеры и ответы и создается обучающая и тестовая выборки средствами SciKit-learn.

Затем импортируется предварительно загруженная языковая модель sPacy «ru_core_news_sm» для русского языка. Языковая модель состоит из нескольких компонент, но в данной работе применяется только лемматизатор, который конвертирует слово в его базовую форму.

После загрузки языковой модели осуществляется лемматизация выборки, затем выборка конвертируется сначала в тензор, а затем в набор данных в формате TensorFlow.

Далее на основе обучающей выборки создается словарь слов, в котором индексом служит количество повторов каждого слова в выборке. Затем визуализируются 20 наиболее часто встречающихся слов.

Затем из словаря удаляются данные о количестве слов, и происходит его преобразование в тензор. Словарь используется для создания идентификаторов слов, представляющих собой целые числа в порядке возрастания. На данном этапе создается таблица из идентификаторов слов с запасом в 1000 слов для случаев, если слово не окажется в таблице. Далее производится кодирование слов и окончательное создание обучающей и тестовой выборки.

После обработки данных и создания окончательных выборок с помощью модуля Keras из пакета TensorFlow создается модель, состоящая из слоя вложения, выполняющего преобразование слов в векторы, 2-х слоев GRU, и последний слой из одного нейрона с сигмоидной функцией активации для классификации. Далее происходит компиляция модели и процесс её обучения.

По завершении обучения модели выводятся графики процесса обучения с указанием точности и матрица ошибок с указанием F-меры, формула которой приведена в выражении (1).

$$precision = \frac{TP}{TP + FP}$$
$$recall = \frac{TP}{TP + TN} \quad , \qquad (1)$$
$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

где precision – точность (не accuracy, применяющейся для оценки модели); recall – полнота; TP – количество истинно положительных примеров; FP – количество ложно положительных примеров; TN – количество истинно отрицательных примеров; F_1 – F-мера.

Точность, матрица ошибок и F-мера рассчитываются для тестовой выборки, что позволяет получить более объективную информацию о качестве разработанной модели.

В завершении работы программы формируется таблица результатов работы НС, содержащая индекс строки из каталога скважин, литологическое описание пород истинный класс и предсказанный класс.

Моделирование и результаты исследований

На этапе извлечения и подготовки данных из таблицы на рисунке 1 создается таблица на рисунке 2.

На следующем этапе происходит собственно моделирование. После разделения загруженных данных размер созданной обучающей выборки составляет 229 примеров (75%), размер тестовой выборки 77 примеров (25%).

Поскольку каждое литологическое описание пород может иметь разное количество слов, при использовании лемматизации каждый пример в выборке приводится к одному размеру, а пустые значения заменяются знаком-заполнителем <pad>.

Далее после создания словаря выводятся первые 20 наиболее частых слов выборке в виде графика частоты встречаемости слов выборке, показанного на рисунке 4.

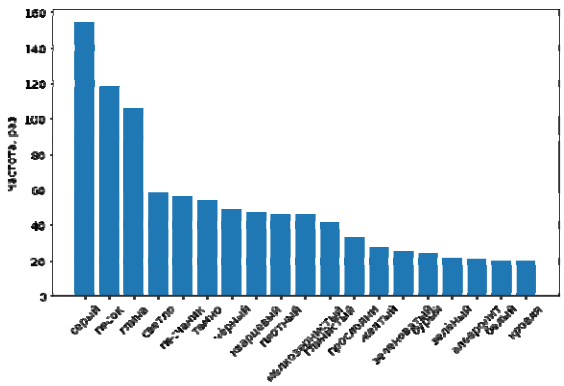

Рис. 4. Частота 20 наиболее часто встречаемых слов в обучающей выборке

На рисунке 4 не показываются пустые слова-заменители, поскольку их наибольшее количество. Исходя из рисунка 4, наиболее часто встречающиеся слова «серый», «песок» и «глина». Также на данном рисунке наблюдается работа лемматизатора – большинство слов, особенно прилагательные, имеют мужской род, то есть своего рода инфинитив (базовая форма). Однако встречаются и слова, которые лемматизатор не затронул, например, слово «прослоями» осталось в изначальном падеже. Это вероятно связано с тем, что лемматизатор из данной языковой модели sPacy не знает указанное слово, что достаточно очевидно, поскольку используемая языковая модель «ru_core_news_sm» построена на новостных текстах.

После обучения выводятся график обучения с указанием точности и матрица ошибок с указанием F-меры, показанные на рисунках 5 и 6.

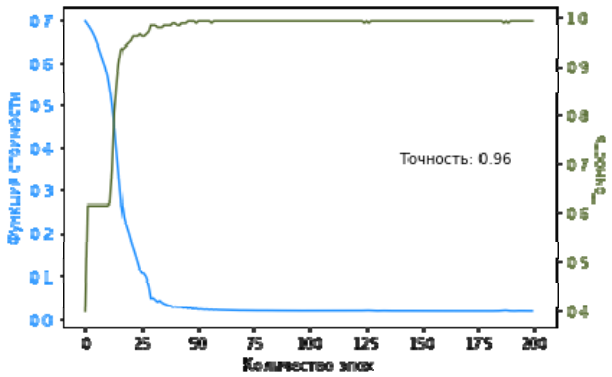

Рис. 5. График обучения нейронной сети

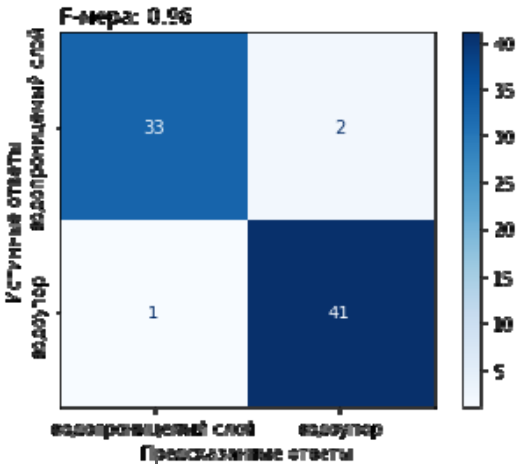

Рис. 6. Матрица ошибок

Исходя из графика обучения, разработанная модель показывает хорошую точность (более 95%) на тестовом наборе. Из рисунка 5 видно, что для получения достаточной точности при обучении модели достаточно около 100 эпох. Анализируя матрицу ошибок, можно сказать, что обученная модель всего 3 из 77 примеров не угадала ответы: 2 примера ложноположительных и 1 пример ложно отрицательный; F-мера составляет 96%, что также указывает на хороший результат.

Фрагмент результатов работы разработанной нейронной сети представлен в таблице 1.

Анализ результатов экспериментальных исследований

Экспериментальные исследования состоят из следующих работ:

- изменение количества эпох;
- изменение функции активации на последнем слое НС;
- изменение рекуррентных слоев на LSTM;
- реализация «мешка» слов вместо векторов слов совместно с логистической регрессией и сравнение результатов с НС.

На рисунках 7 и 8 представлен график обучения и матрица ошибок для НС с количеством эпох равным 80.

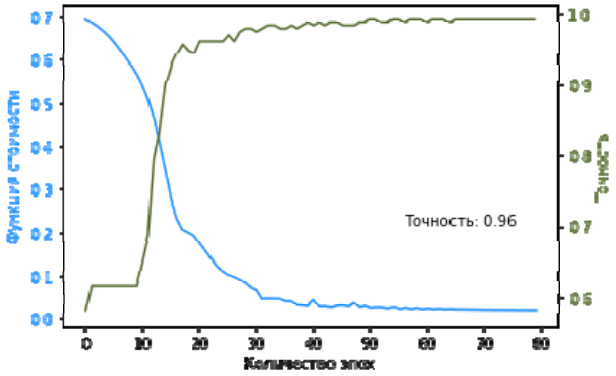

Рис. 7. График обучения нейронной сети (80 эпох)

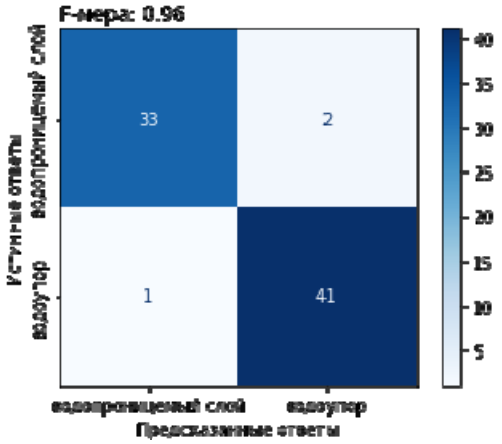

Рис. 8. Матрица ошибок (80 эпох)

Как и предполагалось ранее, для обучения НС с текущей архитектурой достаточно 80 эпох. Модель также показывает хорошую точность и F-меру (>95%) на тестовой выборке.

На рисунках 9 и 10 представлен график обучения и матрица ошибок для НС с измененной функцией активации на гиперболический тангенс (tanh) в последнем слое. Количество эпох – 200.

Таблица 1

Фрагмент результатов работы нейронной сети

Индекс строки из каталога скважин	Литологическое описание пород	Истинный класс	Предсказанный класс
307	Песчаник серый, зеленоватый, кварцевый, тонкозернистый, слабой прочности, прослоями крепкий, на глинистом цементе, в подошве ожелезненный	0*	0
85	Глина серая	1**	1
291	Глина темно-серая, в подошве с голубоватым оттенком, плотная, слюдистая, с обломками раковин моллюсков	1	1
243	Песок белый, ожелезненный, кварцевый, мелкозернистый, глинистый, уплотненный, с прослойками 1-2 см песчаника желто-бурого, ожелезненного	0	0

* - класс 0 – водопроницаемая порода

** - класс 1 - водоупор

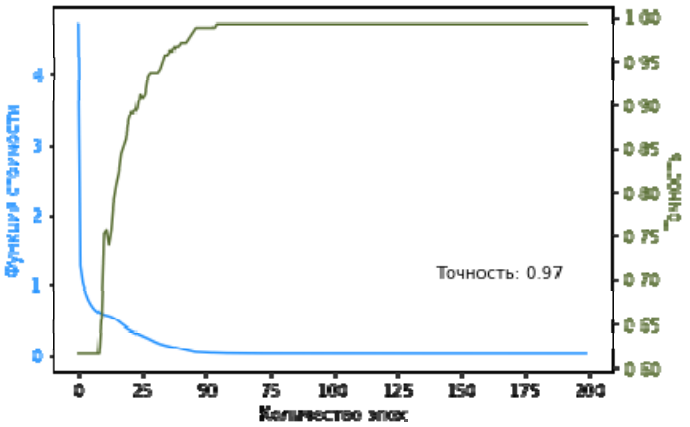

Рис. 9. График обучения нейронной сети с функцией активации tanh в последнем слое

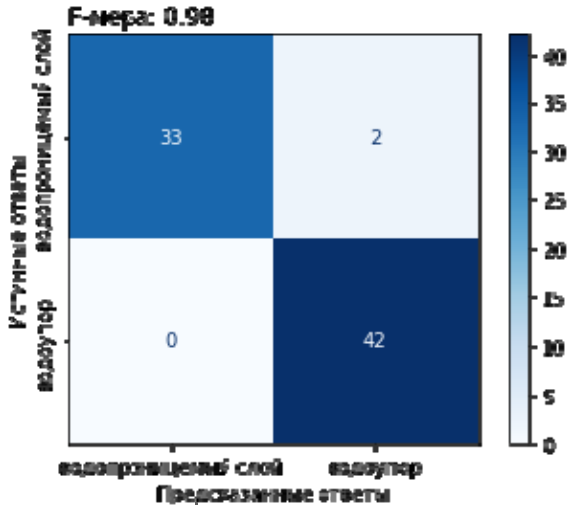

Рис. 10. График обучения модели (слева) и матрица ошибок (справа) с функцией активации tanh в последнем слое

Изменение функции активации на гиперболический тангенс в последнем слое позволило повысить точность модели до 97%, а F-меру до 98% на тестовом наборе данных. При этом модель не дает ложноотрицательных ответов.

На рисунках 11 и 12 представлен график обучения и матрица ошибок для НС, где 2 слоя GRU заменены на 2 слоя LSTM, функция активации на последнем слое не изменена.

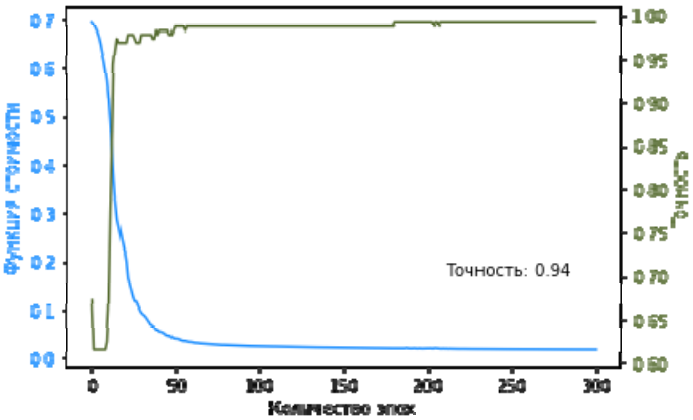

Рис. 11. График обучения со слоями LSTM

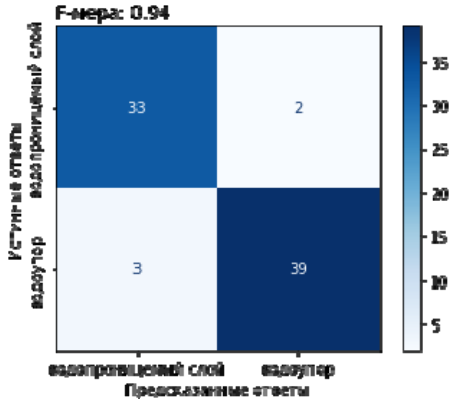

Рис. 12. Матрица ошибок со слоями LSTM

Как видно из рисунков 11 и 12 замена слоев GRU на слои LSTM ухудшила качество модели на тестовой выборке до 94%. Данная НС дает 2 ложноположительных примера и 3 ложноотрицательных вместо одного ложноотрицательного примера в начальной НС. Кроме этого, для данной НС недостаточно 80 и 200 эпох для обучения, поэтому количество эпох для данной модели увеличено до 300.

Для реализации логистической регрессии используется модель «мешка слов», при которой создается разреженная матрица, где хранятся только ненулевые элементы для экономии памяти ПК. Каждая строка разреженной матрицы является одним примером, а столбцы представляют собой слова (признаки). В матрице в каждом примере указывается частота встречаемости каждого слова (признака) в конкретном примере. Существенным недостатком «мешка слов» является полное игнорирование последовательности слов в примерах [13].

В качестве токенизатора в данной модели используется та же языковая модели sPacy – «ru_core_news_sm». Для реализации модели логистической регрессии и решетчатого поиска для нахождения лучшего параметра регуляризации используется пакет SciKit-Learn. На рисунке 13 показана матрица ошибок с F-мерой и точностью для тестового набора данных.

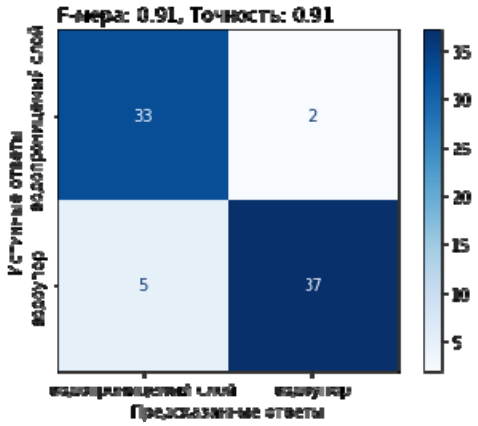

Рис. 13. Матрица ошибок для модели логистической регрессии

Как видно из рисунка 13 точность модели на тестовом наборе данных хуже, чем при модели RNN – модель дает 5 ложноотрицательных ответов (вместо одного, как у НС) и 2 ложноположительных примера. Точность модели и F-мера на тестовом наборе составляет 91%, что в целом говорит о хорошем результате, но НС в данном случае имеет большее преимущество по качеству модели.

Заключение

Использование рекуррентной нейронной сети при семантическом анализе текста, в частности, при определении водоупоров по литологическому описанию геологических слоев в скважинах, дает хорошие результаты, что доказывается высокой точностью (более 95%) на тестовом наборе данных разработанной модели НС. При этом использование слоев НС архитектуры управляемого рекуррентного

блока (GRU) показывает лучший результат, чем при более сложной архитектуре долговременной краткосрочной памяти (LSTM). Использование логистической регрессии с моделью «мешка слов» также дает неплохой результат, однако точность при этом ниже, чем у НС (около 90%).

Использование модели НС, определяющей водоупор по литологическому описанию геологических слоев имеет пользу для специалистов, работающих в области гидрогеологии, в частности, для гидрогеологов, составляющих стратификационную таблицу при подготовке комплекта Государственных гидрогеологических карт.

В качестве продолжения данной работы можно усовершенствовать разработанную модель для более детальной классификации, в частности, для разделения водоупоров на относительные и абсолютные, для определения водонасыщенности слоев, водоносных и напорных горизонтов и др. Для удобства визуализации результатов можно реализовать подсветку или выделение строк в исходной таблице с каталогом скважин, а также реализовать разработанную НС в виде приложения с графическим интерфейсом.

Литература

1. *Ван Цзи, Воронов В. И.* Анализ результатов компьютерной томографии головного мозга с помощью сверточной нейронной сети // DSPA: Вопросы применения цифровой обработки сигналов. 2020. № 1, Том 10. С. 32-40.

2. *Мартыненко Э. В., Воронов В. И.* Применение нейронных сетей для семантической классификации текстовой информации на русском языке (на примере оперативных сводок системы МВД России) // Технологии информационного общества: материалы XIII Международной отраслевой научно-технической конференции. М.: Издательский дом Медиа паблишер, 2019. С. 449-452.

3. *Иванов А. В., Воронов В. И.* Применение алгоритма машинного обучения для разработки автоматизированной системы интеллектуального управления стендом проверки авиационных генераторов // Технологии информационного общества: сборник трудов XV Международной отраслевой научно-технической конференции «Технологии информационного общества». М.: Издательский дом Медиа паблишер, 2021. С. 308-311.

4. *Meghan, Rimol* Gartner Identifies Key Emerging Technologies Spurring Innovation Through Trust, Growth and Change. Gartner: [сайт]. — URL: https://www.gartner.com/.

5. *Ali El, Bilali* Groundwater quality forecasting using machine learning algorithms for irrigation purposes / Bilali Ali El. — Текст: электронный // ScienceDirect : [сайт]. URL: https://www.sciencedirect.com/.

6. *Sahoo, S.* Machine learning algorithms for modeling groundwater level changes in agricultural regions of the U.S. AGU Advancing Earth and Space science: [сайт]. URL: https://agupubs.onlinelibrary.wiley.com/ (дата обращения: 09.01.2022).

7. *Eslam, A. H.* Groundwater Prediction Using Machine-Learning Tools MDPI. URL: https://www.mdpi.com/.

8. *Кокорева С. В., Балашов В. А., Севтинова Е. Б., Смусенок О. И.* Отчет о результатах работ по объекту «Гидрогеологическое доизучение масштаба 1:200 000 по группе листов на территории Российской Федерации в 2020-2022 гг.». Том IV. Гидрогеологическое доизучение масштаба 1: 200 000 листа M-38-XV (Котово). М.: ФГБУ "Гидроспецгеология, 2021.

9. Введение в алгоритм оптимизации Адама для глубокого обучения. URL: https://www.machinelearningmastery.ru/adam-optimization-algorithm-for-deep-learning/ (дата обращения: 09.01.2022).

10. Настройка функции потерь для нейронной сети на данных сейсморазведки. URL: https://habr.com/ru/company/ods/blog/488852/ (дата обращения: 09.01.2022).

11. Dommaraju, Goutham Keras' Accuracy Metrics / Goutham Dommaraju. URL: https://towardsdatascience.com/keras-accuracy-metrics-8572eb479ec7 (дата обращения: 09.01.2022).

12. *Орельен, Ж.* Прикладное машинное обучение с помощью Scikit-Learn, Keras и TensorFlow: концепции, инструменты и техники для создания интеллектуальных систем. 2-е изд. М., СПб.: ООО «Диалектика», 2020. 1040 с.

13. *Мюллер Андреас*, Гвидо Сара. Введение в машинное обучение с помощью Python: руководство для специалистов по работе с данными. М.: Вильямс, 2017. 480 с.

МОДЕЛИРОВАНИЕ СИСТЕМЫ MIMO В РЕЖИМЕ BEAMFORMING ДЛЯ РАЗНОГО ЧИСЛА ПОТОКОВ ДАННЫХ

Панкратов Денис Юрьевич,
МТУСИ, к.т.н., доц., Москва, Россия
dpankr@mail.ru

Сердюков Александр Александрович,
МТУСИ, Москва, Россия
camper867@ya.ru

Хасанн Диаа Мохамад,
МТУСИ, Москва, Россия
diaaahassan159@gmail.com

Аннотация

В статье рассматривается моделирование алгоритмов передачи и приёма информации в системе MIMO с разным числом потоков и конфигурацией антенн 4x2 с использованием режима направленной передачи. Приводится методика и результаты компьютерного моделирования системы MIMO в режиме направленной передачи. Показано влияние различного числа потоков на помехоустойчивость системы MIMO.

Ключевые слова: *MIMO, сингулярное разложение, Beamforming, компьютерное моделирование, направленная передача.*

Введение

В настоящее время широко распространены системы 5G, в которых используется технология MIMO (Multiple Input Multiple Output). Системы MIMO, в свою очередь, используют технологию прекодирования, одним из видов которой является режим Beamforming. Режим известен давно и представляет собой технологию направленной передачи с подстраивающимся массивом антенн [1]. Благодаря этому можно сформировать соответствующую диаграмму направленности в область определённого абонента. Это достигается с помощью весовых коэффициентов, тем самым увеличивая скорость передачи данных и помехоустойчивость системы. Этот режим используется во многих стандартах, таких как 802.11n и 802.11ac [2, 3].

Данная статья направлена на изучение и моделирование систем MIMO в режиме Beamforming. Конкретно будет рассмотрена система MIMO с конфигурацией 4x2 и будет проведено моделирование помехоустойчивости этой системы в зависимости от разного числа потоков информации. В качестве результата моделирования будет приведён график помехоустойчивости этой системы и будут сделаны соответствующие выводы.

Технология направленной передачи и ее преимущества

При направленной передаче концентрация энергии в определённую сторону формируется за счёт весовых коэффициентов антенной решетки, излучающей передаваемые сигналы [1]. Получаются лепестки в диаграмме направленности в нужных направлениях, а в других направлениях антенны не излучают сигналы. Для настройки антенн информация о нахождении абонентской станции передаётся по каналу обратной связи от самой абонентской станции [4, 5].

Например, на рисунке 1 изображена базовая станция с 4 передающими антеннами и абонентская станция с 2 приемными антеннами. В этом случае конфигурация системы будет 4x2. В идеальном случае лучшие параметры системы MIMO будут тогда, когда передача осуществляется через несколько параллельных потоков, при этом сложность обработки сигналов на приёмной стороне значительно уменьшается [5].

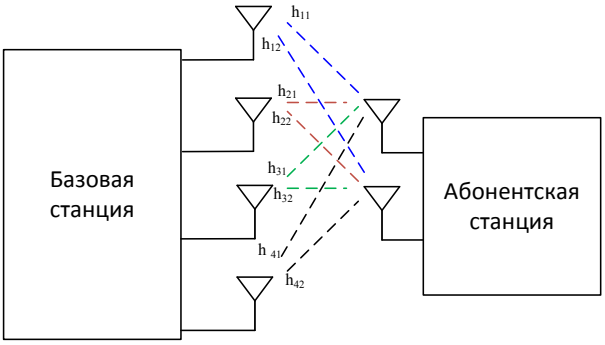

Рис. 1. Упрощенная структурная схема системы MIMO с конфигурацией 4x2

Математическая модель системы MIMO описывается следующим образом [5]:

$$\mathbf{y} = \sqrt{\rho / M}\,\mathbf{HVs} + \mathbf{n}\,, \tag{1}$$

где ρ – отношение сигнал/шум на приемной стороне; M – число передающих антенн; $\mathbf{V}$ – весовая матрица размерностью $M \times M$, $\mathbf{s}$ – вектор передаваемой информации.

Технологии обеспечения MIMO в стандартах Wi-Fi

Впервые наибольшую эффективность системы MIMO проявили в стандарте 802.11n за счёт совместного использования с пространственным мультиплексированием и работы на 40МГц [6]. Также, благодаря использованию агрегации кадров и усовершенствованию протокола подтверждения блоков, повысилась эффективность управления доступом к среде передачи. Это позволило значительно улучшить скорость передачи данных по сравнению со стандартами 802.11a и 802.11g [6].

Устойчивость повышается по своей сути за счет увеличенного пространственного разнесения, обеспечиваемого использованием нескольких антенн. Пространственно-временное блочное кодирование в качестве опции дополнительно повышает надежность системы MIMO, как и быстрая адаптация канала - механизм для быстрого отслеживания меняющихся условий канала. Поправки к стандарту 802.11n вводят улучшения режима направленной передачи (Beamforming) как на уровне PHY (Physical layer), так и на уровне и MAC (Media Access Control) для повышения помехоустойчивости [6].

Ряд других улучшений дает дополнительные преимущества [6, 3]. В PHY они включают более короткий защитный интервал, который может использоваться при определенных условиях канала. Однако, для некоторых новых режимов нет обратной совместимости с существующими устройствами 802.11a и 802.11g. На уровне MAC добавлены новые протоколы для улучшения производительности для определенных шаблонов трафика, позволяя станции передавать часть выделенных ресурсов передачи другой станции и, таким образом, сокращать общие накладные расходы.

Теперь несколько слов про частотные диапазоны и пространственные потоки. Первое поколение устройств 802.11n обычно работает только в диапазоне 2.4 ГГц, с двумя пространственными потоками и шириной канала 40 МГц. При использовании короткого защитного интервала устройства первого поколения могут достигать скорости передачи данных порядка 300 Мбит/с. В устройствах второго поколения появляется два диапазона: 2.4 и 5 ГГц. Они также достигают скорости 300 Мбит/с, но некоторые из них включают дополнительные приемные антенны для дополнительного разнесения на приеме. В устройствах следующих поколений увеличили число передающих антенн для поддержки трёх и четырех пространственных потоков, что позволяет обеспечить скорости порядка 450 Мбит/с и 600 Мбит/с [6, 2]. Дальнейший рост числа передающих антенн привел к превышению порога скоростей 1 Гбит/с [3, 7].

Многочисленные дополнительные функции в 802.11n и 802.11ac означают, что для обеспечения сосуществования и взаимодействия требуется расширенная сигнализация возможностей устройства. Например, поддерживает ли устройство определенные функции PHY, такие как преамбула формата зеленого поля или функции MAC, например, возможность участвовать в обмене протоколами в об-

ратном направлении. Обзор функций, добавленных к MAC в 802.11n и 802.11ac [2, 3], приведен на рисунке 2.

Наличие широкого канала (40 МГц или больше) создает ряд проблем сосуществования сетей разных стандартов. Поскольку при работе с широким каналом используются два или более канала по 20 МГц, необходимы механизмы для смягчения воздействия на соседние станции с полосой 20 МГц, работающие независимо на любом из этих каналов [6].

Сосуществование разных сетей в первую очередь достигается за счет тщательного выбора каналов, то есть выбора пары каналов, у которых мало или нет активного соседнего трафика. С этой целью поправки в стандартах 802.11n и 802.11ac добавляют требования к сканированию для обнаружения присутствия активных соседних станций, чтобы обеспечить возможность перехода станций на другие каналы, если соседняя станция с полосой канала 20 МГц начинает свою передачу [3].

Признавая растущую важность портативных устройств, в 802.11n добавили метод планирования доступа к каналу для эффективной поддержки большого количества станций, называемый энергосберегающим множественным опросом [6].

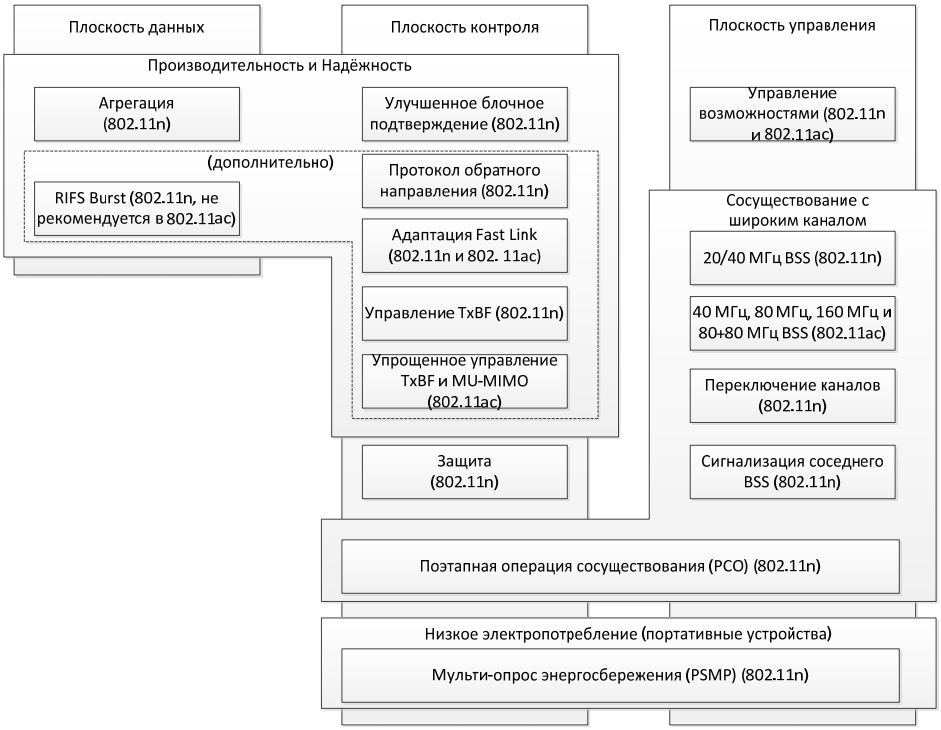

Рис. 2. Улучшения 802.11n и 802.11ac на уровне MAC [2,3]

Методы, представленные в стандарте 802.11n для повышения помехоустойчивости, также используются в 802.11ac. Однако в каждый из методов были внесены изменения, чтобы упростить работу и уменьшить количество дополнительных опций. В отличие от 802.11n, сигнал 802.11ac имеет только одну преамбулу. Подобно преамбуле в стандарте 802.11n, преамбула VHT (Very High Throughput) содержит унаследованную совместимую часть преамбулы для совместимости с устройствами 802.11a/n. Кроме того, преамбула VHT поддерживает как однопользовательскую, так и многопользовательскую работу [3].

В 802.11ac уровень MAC включает усовершенствования в агрегировании и улучшенную работу в условиях широкого частотного канала посредством динамического и статического механизма сигнализации. Включена расширенная операция энергосбережения с присутствием идентификатора станции в заголовке на уровне PHY, а также механизм для сигнализации рабочего режима станции (ширина полосы канала и количество радиочастотных цепей для приема).

Представлен новый протокол зондирования для поддержки режимов Beamforming и MIMO для случая многих станций [5]. Поскольку увеличенная ширина канала расходует значительную часть доступного спектра в диапазоне 5 ГГц, существуют новые правила сканирования для обнаружения перекрывающейся работы станций с требованиями и рекомендациями по выбору канала [6].

Пространственные преобразования в системе MIMO с различным числом потоков данных

Для передачи данных в системе MIMO используют различные матрицы пространственного отображения, примеры которых можно найти в стандарте 802.11n [2]. Существуют специальные типы этих матриц: **Q**

а) Матрица прямого отображения: $\mathbf{Q}_k$ – это диагональная матрица комплексных значений, которая может принимать две формы:

1) $\mathbf{Q}_k = \mathbf{I}$ (единичная матрица)

2) Матрица CSD (Cyclic Shift Diversity), в которой диагональные элементы представляют собой циклические сдвиги сигналов во временной области.

3) Матрица косвенного отображения может быть произведением матрицы CSD и квадратной унитарной матрицы.

4) Матрица пространственного расширения - произведение матрицы CSD и квадратной матрицы, сформированной из ортогональных столбцов. В данной работе используется именно эта матрица.

Пространственное расширение может быть выполнено путём дублирования некоторых потоков данных для формирования пространственных потоков [2], при этом каждый поток нормализуется с помощью коэффициента: $\sqrt{N_{Streams}/N_{TX}}$, где $N_{Streams}$ – число потоков, а N_{TX} – число передающих антенн.

Согласно стандарту 802.11n [2] передача данных в режиме Beamforming может осуществляться с различным числом потоков, что достигается за счет применения специальных матриц отображения **G**. Например, для разного числа антенн и потоков эта матрица имеет вид, представленный в таблице 1.

Таблица 1

Вид матрицы отображения в зависимости от разного числа антенн и потоков данных

Число антенн, N_{TX}	Число потоков, $N_{Streams}$	Матрица **G**
2	1	$\frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\end{bmatrix}^T$
4	1	$\frac{1}{2}\begin{bmatrix}1 & 1 & 1 & 1\end{bmatrix}^T$
4	2	$\frac{1}{\sqrt{2}}\begin{bmatrix}1 & 0\\ 0 & 1\\ 1 & 0\\ 0 & 1\end{bmatrix}$

Технология Beamforming с применением сингулярного разложения

В технологии направленной передачи используется преобразование SVD (Singular Values Decomposition). Оно заключается в разложении матрицы, которое показывает её геометрическую структуру [6]. Сингулярным разложением матрицы **H** является преобразование вида:

$$\mathbf{H} = \mathbf{UDV}', \tag{2}$$

где **D** – диагональная матрица с сингулярными числами матрицы **H**, которые лежат на её диагонали, в то время как все остальные элементы – нули; матрицы **U** и **V** – унитарны и состоят из левых и правых сингулярных векторов соответственно; **V'** – сопряжённо-транспонированная матрица к матрице **V**.

Структурная схема системы MIMO с использованием преобразования SVD приведена на рисунке 3. Данная схема справедлива и для большего числа антенн, изменятся только размеры матриц.

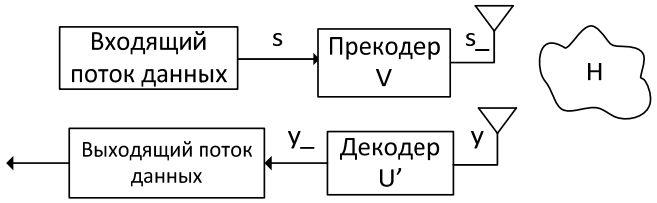

Рис. 3. Упрощённая структурная схема системы MIMO с использованием SVD

Математическая модель такой системы будет выглядеть следующим образом: входящий поток данных преобразуется в прекодере с помощью матрицы $\mathbf{V}$:

$$\mathbf{s}\_ = \mathbf{V}\mathbf{s}\_, \tag{3}$$

где $\mathbf{s}$ – вектор входящих данных.

Пространственные потоки проходят через радиоканал с матрицей $\mathbf{H}$ и на приемной стороне выглядят следующим образом:

$$\mathbf{y}\_ = \mathbf{H}\mathbf{s}\_ \tag{4}$$

В этом примере не учитывается влияние шума, но учитывается соотношение унитарных матриц:

$$\mathbf{U}'\mathbf{U} = \mathbf{V}'\mathbf{V} = \mathbf{I} \tag{5}$$

С учетом выражений (2) и (3) с несколькими преобразованиями после прохождения через декодер информационный поток имеет вид:

$$\mathbf{y}\_svd = \mathbf{U}'\mathbf{y}\_ = \mathbf{U}'\mathbf{U}\mathbf{D}\mathbf{V}'\mathbf{V}\mathbf{s}\_ = \mathbf{D}\mathbf{s}\_ \tag{6}$$

С учетом выражений (2) и (3) с несколькими преобразованиями после прохождения через декодер информационный поток имеет вид:

$$\mathbf{D}^{-1}\mathbf{y}\_svd = \mathbf{D}^{-1}\mathbf{D}\mathbf{s}\_ = \mathbf{s}\_ \tag{7}$$

В идеальном случае благодаря сингулярному разложению матрицы канала **H** можно получить исходный поток данных, компенсируя все влияния канала. Более подробный пример SVD-преобразования см. в [9].

Структурные схемы систем MIMO с конфигурацией 4x2 с различным числом потоков и компьютерное моделирование

Согласно [10] общая структура моделирования системы радиосвязи может быть представлена следующим образом:

1) Генератор случайных воздействий и формирование потоков данных.
2) Моделирование передачи по радиоканалу и обработка потоков данных.
3) Интерпретация результатов моделирования в виде графиков.

На рисунках 4 и 5 представлены структурные схемы системы MIMO с конфигурацией 4x2 с разным числом потоков.

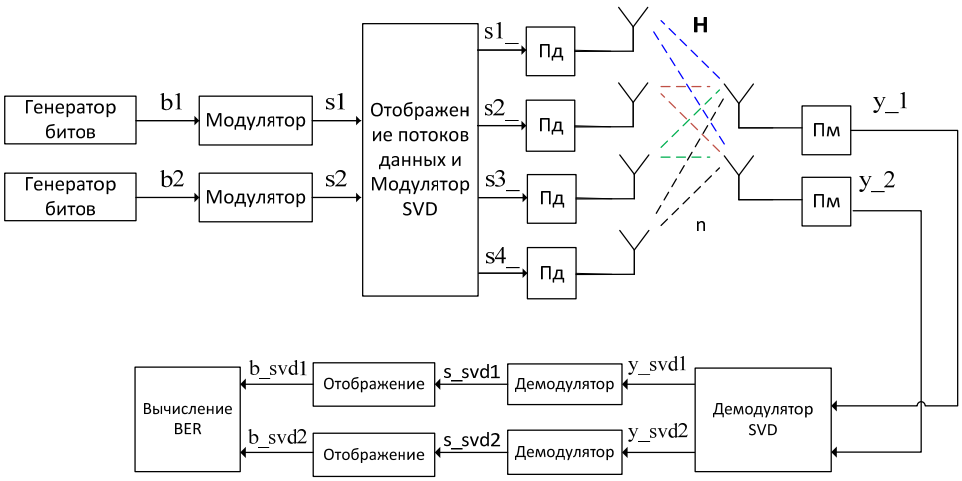

Рис. 4. Структурная схема системы MIMO с конфигурацией 4x2 для 2 потоков

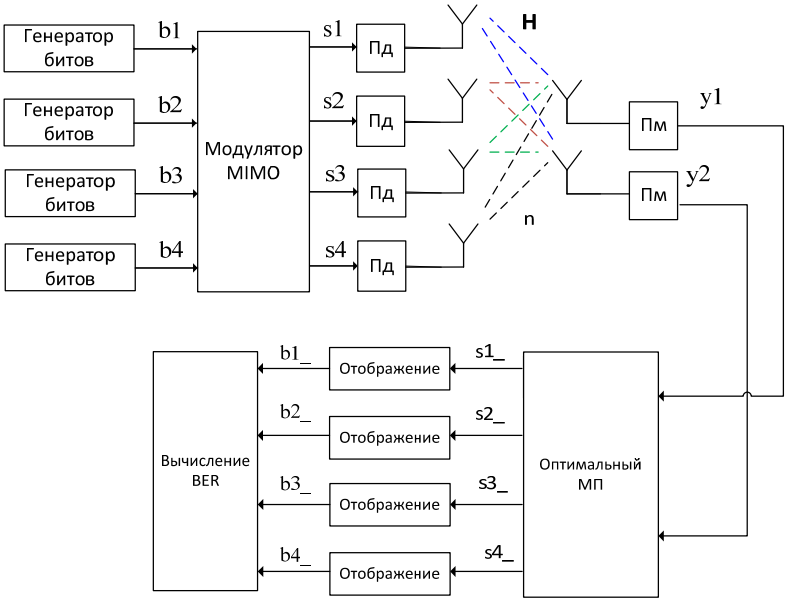

Рис. 5. Структурная схема системы MIMO с конфигурацией 4x2 для 4 потоков

Алгоритм моделирования этих систем будет одинаковый, за исключением числа имеющихся потоков, и выглядит следующим образом:

Таблица 2

Алгоритм моделирования системы MIMO

1	Ввод начальных данных – число испытаний	L
2	Начало цикла по ОСШ в дБ от 1 до 15	SNR
3	Обнуление исходного числа ошибок	sum
4	Вычисление среднеквадратичного отклонения для текущего значения ОСШ	sigma
5	Начало цикла по числу испытаний	L
6	Генерация 2 или 4 потоков данных	b
7	Отображение потоков данных в информационные потоки символов	b, s
8	Генерация комплексного гауссовского случайного вектора белого шума	n
9	Генерация матрицы канала	H
10	Сингулярное разложение матрицы канала	U, D, V
11	Отображение потоков символов в пространственные потоки	s, G

12	Модификация информационных векторов символов	G, V, s
13	Передача модифицированного вектора информационных символов через канал с учетом АБГШ	s_, H, n
14	Обработка принятого информационного потока с помощью Beamforming	y_, U',y_svd, D^{-1}
15	Отображение информационных векторов в биты	s_svd
16	Определение факта наличия ошибок	f
17	Подсчет общего числа ошибок	sum
18	Завершение цикла п.5	
19	Вычисление коэффициента ошибок	BER_SVD
20	Завершение цикла п.2	
21	Построение графика зависимости коэффициента битовых ошибок для 2 или 4 потоков данных	

Математическое описание содержит в себе объяснения для каждого функционального блока из структурной схемы. Изначально входной поток данных в виде битов в блоке модулятора отображается в вектор информационных символов. Затем символы модифицируются в блоке модулятора SVD. В схеме с 2 потоками используется специальная матрица отображения потоков данных в пространственные потоки:

$$\mathbf{s}_{-4x1} = \mathbf{V}_{4x4}\mathbf{G}_{4x2}\mathbf{s}_{2x1}, \tag{8}$$

где $\mathbf{G}$ – матрица отображения потоков данных, которая имеет вид: $\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$.

Модифицированный информационный поток данных передаётся по радиоканалу с учётом шума $\mathbf{n}$ в канале с матрицей $\mathbf{H}$, состоящей из коэффициентов передачи. Таким образом, на демодулятор SVD поступает информационный вектор:

$$\mathbf{y}_{-2x1} = \sqrt{\rho \cdot M}\mathbf{H}_{2x4}\mathbf{s}_{-4x1} + \mathbf{n}_{2x1},$$

В самом блоке демодулятора SVD принимаемый информационный поток $\mathbf{y}\_$ обрабатывается с помощью матрицы $\mathbf{U}'$ и осуществляется разделение сигналов:

$$\mathbf{y\_svd}_{2x1} = \mathbf{U}'_{2x2}\,\mathbf{y}_{-2x1},$$

В блоке демодулятора происходит обратное отображение вектора $\mathbf{y\_svd}$ информационных символов в поток данных $\mathbf{s\_svd}$:

$$\mathbf{s\_svd}_{2x1} = G^{T}_{4x2}\mathbf{D}^{T}{}_{2x4}\mathbf{y\_svd}_{2x1} \tag{9}$$

Затем потоки символов отображаются обратно в биты $\mathbf{b\_svd}$ и сравниваются с изначальными битами $\mathbf{b}$. Если совпадения не обнаружено, то регистрируется ошибка.

В схеме на рисунке 5 входной поток данных в виде битов также в блоке модулятора отображается в вектор информационных символов и передаётся сразу в радиоканал, без использования технологии SVD. Математическое описание такой системы будет выглядеть следующим образом:

$$\mathbf{y}_{2x1} = \sqrt{\rho / M}\mathbf{H}_{2x4}\mathbf{s}_{4x1} + \mathbf{n}_{2x1}$$

Для демодуляции применяется демодулятор, оптимальный по критерию максимального правдоподобия (Maximum Likelihood – ML) [5], где осуществляется перебор всех комбинаций вектора информационных символов $\mathbf{s}_{-4x1}$, которые берутся из специальной матрицы.

Искомый вектор $\mathbf{s}_{-4x1}$ оценки, оптимальной по критерию ML, отображается в биты $\mathbf{b}\_$, которые сравниваются с исходными битами, и в конечном блоке происходит регистрирование ошибок. Также возможно применение других, менее сложных алгоритмов демодуляции [11].

Результаты моделирования

В качестве результатов моделирования приведён график зависимости коэффициента битовых ошибок (BER) от отношения сигнал/шум (SNR). Число испытаний при моделировании – 5000. На рисунке 6 представлены результаты компьютерного моделирования.

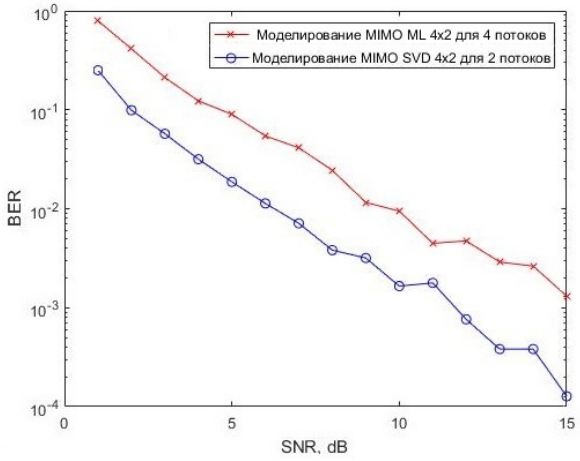

Рис. 6. Помехоустойчивость системы MIMO с конфигурацией 4x2 с разным числом потоков

Заключение

По рисунку 6 видно, что система MIMO с 4 потоками уступает в помехоустойчивости системе с 2 потоками. Например, по уровню BER=10^{-2} выигрыш системы MIMO с двумя потоками составляет примерно 3 дБ. Эта разница объясняется тем, что одинаковое число антенн принимает разное количество информации, из-за чего большее число потоков обработать сложнее, и возникает большее число ошибок при передаче. В стандарте 802.11n приведены примеры различных матриц отображения для разного числа потоков, одна из которых использовалась в работе [2].

Таким образом, представленную методику моделирования [10] можно использовать для моделирования системы MIMO в режиме Beamforming при разном числе антенн и информационных потоков. Также данная работа показывает, что при моделировании системы MIMO можно использовать не только технологию Beamforming с применением сингулярного разложения, как в работе [9]. В будущем планируется расширить полученные результаты для системы Massive MIMO, в которой будет использоваться большое число антенных элементов, что характерно для миллиметровых диапазонов (сверхвысокие частоты) [1, 4, 11-23].

Литература

1. *Perahia, E. and Pottie, G. J.* Adaptive antenna arrays and equalization for indoor digital radio. International Conference on Communications, June 23–27, Dallas, TX. – 1996.
2. IEEE Standard Wireless Local Area Networks, IEEE 802.11n – 2009.
3. IEEE Standard Wireless Local Area Networks, IEEE Std 802.11ac Draft5.0 – 2013.
4. *Montojo, Juan, et* al. Fundamentals of 5G Communications: Connectivity for Enhanced Mobile Broadband and Beyond. McGraw-Hill Education. – 2021.
5. *Бакулин М.Г., Крейнделин В.Б., Панкратов Д.Ю.* Технологии в системах радиосвязи на пути к 5G. М.: Горячая линия – Телеком, 2018.
6. *Eldad Perahia, Robert Stacey,* Next Generation Wireless LANs: 802.11n and 802.11ac, 2nd Edition, Cambridge. 2013.
7. IEEE Standard Wireless Local Area Networks, IEEE 802.11ad – 2012.
8. *Foschini, G. J., and Gans, M. J.* On the limits of wireless communications in a fading environment when using multiple antennas. Wireless Personal Communications. 1998.
9. *Панкратов Д.Ю., Сердюков А.А.* Моделирование системы MIMO в режиме Beamforming // DSPA: Вопросы применения цифровой обработки сигналов. 2021. Т. 11. № 2. С. 12-21.
10. *Панкратов Д.Ю., Степанова А.Г.* Компьютерное моделирование технологии MIMO для систем радиосвязи // T-Comm: Телекоммуникации и транспорт. 2018. Т. 12. № 12. С. 33-37.

11. *Bakulin M.G., Kreindelin V.B., Pankratov D.Y., Stepanova A.G.* Applying a new approximation to demodulation in massive MIMO systems // 2021 Wave Electronics and its Application in Information and Telecommunication Systems, WECONF 2021 - Conference Proceedings, 2021.

12. *Бакулин М.Г., Крейнделин В.Б., Панкратов Д.Ю.* Анализ пропускной способности канала MIMO в условиях замираний // Системы синхронизации, формирования и обработки сигналов. 2018. Т. 9. № 2. С. 13-20.

13. *Крейнделин В.Б., Старовойтов М.Ю.* Повышение помехоустойчивости системы связи MIMO с пространственным мультиплексированием методом додетекторного сложения // T-Comm: Телекоммуникации и транспорт. 2017. Т. 11. № 4. С. 4-13.

14. *Бакулин М.Г., Крейнделин В.Б.* Проблема повышения спектральной эффективности и емкости в перспективных системах связи 6G // T-Comm: Телекоммуникации и транспорт. 2020. Т. 14. № 2. С. 25-31.

15. *Крейнделин В.Б., Резнёв А.А.* Матрица пространственно-временного кода высокой размерности типа "Голден" // T-Comm: Телекоммуникации и транспорт. 2018. Т. 12. № 6. С. 34-40.

16. *Бакулин М.Г., Крейнделин В.Б., Панкратов Д.Ю.* Исследование вероятностных моделей радиоканала MIMO с учетом взаимной корреляции передающей и приемной сторон с помощью компьютерного моделирования // REDS: Телекоммуникационные устройства и системы. 2017. Т. 7. № 1. С. 64-68.

17. *Бакулин М.Г., Крейнделин В.Б., Панкратов Д.Ю.* Алгоритмы нелинейной фильтрации двоичной ЛРП со случайной задержкой и случайной начальной фазой // Системы синхронизации, формирования и обработки сигналов. 2019. Т. 10. № 2. С. 45-51.

18. *Бакулин М.Г., Крейнделин В.Б., Панкратов Д.Ю.* Методы приема псевдослучайных последовательностей в системах радиосвязи // REDS: Телекоммуникационные устройства и системы. 2018. Т. 8. № 1. С. 108-112.

19. *Крейнделин В.Б., Григорьева Е.Д.* Анализ быстрого алгоритма умножения матриц и векторов для банка цифровых фильтров // T-Comm: Телекоммуникации и транспорт. 2021. Т. 15. № 1. С. 4-10.

20. *Бакулин М.Г., Бен Режеб Т.Б.К., Крейнделин В.Б., Смирнов А.Э.* Способы минимизации объёма передаваемой информации в обратном канале многоантенных систем MIMO // T-Comm: Телекоммуникации и транспорт. 2021. Т. 15. № 3. С. 17-24.

21. *Бакулин М.Г., Крейнделин В.Б., Панкратов Д.Ю.* Применение технологии MIMO в современных системах беспроводной связи разных поколений // T-Comm: Телекоммуникации и транспорт. 2021. Т. 15. № 4. С. 4-12.

22. *Бакулин М.Г., Крейнделин В.Б., Панкратов Д.Ю.* Исследование вероятностных моделей радиоканала MIMO с учетом взаимной корреляции передающей и приемной сторон с помощью компьютерного моделирования // REDS: Телекоммуникационные устройства и системы. 2017. Т. 7. № 1. С. 64-68.

23. *Крейнделин В.Б., Григорьева Е.Д.* Реализация банка цифровых фильтров с пониженной вычислительной сложностью // T-Comm: Телекоммуникации и транспорт. 2019. Т. 13. № 7. С. 48-53.

НЕЙРОСЕТЕВЫЕ МЕТОДЫ В ЗАДАЧЕ СЕНТИМЕНТ-АНАЛИЗА

Рябыкин Алексей Сергеевич,
Московский Авиационный Институт, студент, бакалавр, Москва, Россия
ras.unlucky@gmail.com

Сухов Егор Аркадьевич,
Московский Авиационный Институт, доцент кафедры 802, Москва, Россия
sukhov.george@gmail.com

Аннотация

Современный мир, содержащий уже слабо измеримый объем данных и отличающийся либерализмом общественного сознания, срастив эти две парадигмы, породил невозможность охвата потока мнений в формате ручного анализа. Востребованность в автоматизации сентимент-анализа заволакивает ныне немалое количество сфер: от информационной безопасности в лице выявления угроз и выбросов агрессии, экономической науки в качестве анализа фондовых рынков путем мониторинга настроений до определения лояльности потребителей в продукт-менеджменте. Очевидным является желание получить качественную модель, способную изменяться вместе с языком, совмещающую рациональность разработки и точность. Модели, основанные на правилах, предельно ясно, будут лучшими относительно последнего пункта, однако разработка и последующие изменения потребуют долгой деятельности и постоянной модерации соответственно. Вследствие этого, востребованными являются реализации, способные обучаться с хорошим качеством и легко дообучаться при лингвистических мутациях. В данной работе произведена попытка поиска, реализации и сравнения нейросетевых моделей глубокого обучения, подходящих под поставленные для них требования. Рассмотренные модели были реализованы как по отдельности, так и в конкатенации. Лучшим результатом работы является конечная мета-модель, получившая значение метрики accuracy = 0.965 на тестовом наборе данных.

Ключевые слова: *сентимент-анализ, векторное представление слов, свертка, рекуррентная сеть, градиентный спуск, обратное распространение ошибки, рекуррентный блок*

Введение

Решение задачи семантического анализа необходимо содержит предварительную обработку текста, в лучшем случае, включающую:

1) приведение текста к единому регистру (обычно, нижнему);
2) токенизацию текста – разбиение блоков текста на более мелкие атомарные единицы: n-граммы, слова, ныне актуально – предложения или даже абзацы;
3) морфологическую разметку: с целью различать омонимы, несущие отличные семантические роли. Например, «гладь» существительное, значащее ровную поверхность и, одновременно с этим, глагол в повелительном наклонении «гладь». В английском языке, который выбран для анализа методов нейросетевого моделирования в этой статье, подобное встречается реже, однако имеет место быть: “can” в смысле мочь, уметь (глагол) и “can” как консервная банка (существительное);
4) лемматизацию (приведение слова к лемме)**.** Этот процесс выполняется для оптимального сокращения необходимого словаря, редуцируя слова до их корней;
5) векторизацию слова: представление токенов в качестве численных векторов для последующей обработки.

Лишь первая из приведенных выше задач является очевидной и простой. Остальные же требуют творческих путей решений, однако, хоть и стоят упоминания, подобные подзадачи будут рассмотрены в статье лишь косвенно. В ней акцент сфокусирован исключительно на задаче классификации, как бинарной (определение полярности текста), так и многоклассовой (когда речь заходит о ранжировании мнений) методами нейросетевого моделирования. Рассмотрены варианты решения подобных проблем методами глубокого обучения: CNN (convolution neural networks 1 dimension, одномерные

сверточные нейронные сети), RNN (recurrent neural networks, рекуррентные нейронные сети), в частности, LSTM (long short-term memory, долгая краткосрочная память), GRU (gated recurrent unit, управляемый рекуррентный блок). Рассмотрены их положительные стороны и слабые места, способы борьбы с последними. Кроме того, применен метод ансамблирования Stacking к обученным моделям для получения более уверенного и лучшего результата с использованием современных способов прототипирования моделей (PyTorch и TensorFlow) и с применением программно-аппаратной архитектуры параллельных вычислений Nvidia CUDA. Используемыми данными послужили рецензии на кино с платформы IMDB (в случае бинарной классификации выборка составлялась на основе оценки: 10 – отзыв положительный, 1 – отзыв отрицательный), отзывы с сайта YELP и комментарии социальной сети Twitter. Язык исследования: английский. Для адаптации результатов под работу с другими языками, необходима другого рода предобработка данных, сами архитектуры моделей сохранят свою работоспособность. Итоговые результаты исследований предоставлены в таблице 1.

Таблица 1

Результаты исследований

Данные		Модели			
		CNN	LSTM	GRU	META
IMDB (2)	Train	0.9963	0.9982	0.9751	-
	Val	0.8991	0.88	0.9612	-
	Test	0.8996	0.8765	0.9353	-
Twitter (3)	Train	0.93	0.8431	0.9937	-
	Val	0.8542	0.8343	0.9605	-
	Test	0.8771	0.841	0.9574	-
Yelp (2)	Train	0.5538	0.9172	0.9670	-
	Val	0.534	0.9762	0.9562	-
	Test	0.613	0.945	0.9459	-
Merged (2)	Score	-	-	-	0.965

Сверточные нейронные сети

Созданные для задач компьютерного зрения сверточные нейронные сети [1, 2, 3] были весьма продуктивно применены и в других сферах глубокого обучения, в том числе и NLP (natural language preprocessing, обработка естественного языка). Использование их достигло апогея в ответ на ограниченность применения полносвязных сетей для обработки данных с сеточной топологией [5].

Определение: Пусть $u, v : \mathbb{R}^n \to \mathbb{R}$ – две функции, интегрируемые относительно меры Лебега на пространстве $\mathbb{R}^n$. Тогда функция $u * v : \mathbb{R}^n \to \mathbb{R}$ называется сверткой и определяется соотношением:

$$(u * v)(x) := \int_{\mathbb{R}^n} u(y)v(x-y)dy,$$

где y иногда называют «давностью». При этом функция u называется входом (input), функция v – ядром (kernel).

Результат свертки именуется картой признаков (feature map).

В задаче классификации текстов будем рассматривать одномерную свертку в силу векторного представления текстовой информации. Для $n = 1$ формула примет вид:

$$(u * v)(x) := \int_{-\infty}^{\infty} u(y)v(x-y)dy.$$

Однако, так как задача классификации текстов представляет собой конечный, дискретный случай, окончательно формула принимает следующий вид:

$$(u * v)(x) := \sum_{y=0}^{K} u(y)v(x-y),$$

где K – размер ядра.

Для иллюстрации операции свертки рассмотрим функции u, v как сигналы, Рис. 1. Сдвигая ядро v относительно входного сигнала u и посчитав «сходство» фрагмента сигнала с ядром (в случае векторов, речь идет о скалярном произведении) для каждого смещения, получим результат свертки.

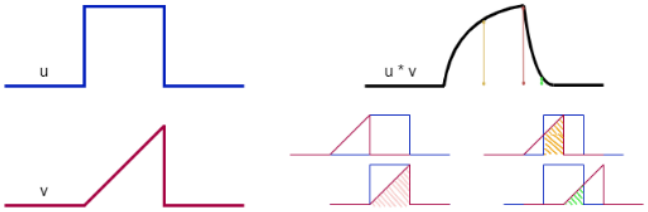

Рис. 1. Пример операции свертки для двух сигналов

Для изображений характерно матричное представление, каждому элементу которого соответствует значение пикселя. На Рис.2. приведен случай с изображением размерности 5×5.

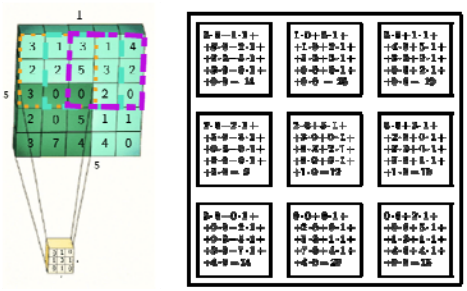

Рис. 2. Пример операции свертки для изображения 5x5x1 с ядром свертки 3x3x1

Ядро свертки (размерности $3 \times 3 \times 1$) прикладывается к каждому фрагменту изображения с задаваемым шагом (stride). На рисунке 2 приведен пример со значением шага, равным единице. Видно, что каждый нейрон связан лишь с частью пикселей, если речь идет о первом сверточном слое или нейронов, для последующих слоев (это свойство называют разреженной связностью). Необходимым становится обучение лишь значений ядер свертки, что значительно меньше, чем в случае традиционных нейронных сетей (полносвязных).

Как уже было сказано ранее, слова (после векторизации) представляют собой плоские вектора. Поэтому для задач обработки естественного языка применяются одномерные сверточные нейронные сети. Токены, представленные в векторной форме, конкатенируются в матрицу текста. Затем выбирается k строк матрицы, вытягиваются в плоский вектор, к которому «прикладывается» (берется скалярное произведение) одномерное ядро размерности $k \cdot s$, где s – размерность вектора токена. Этот процесс проиллюстрирован на рисунках 3 и 4.

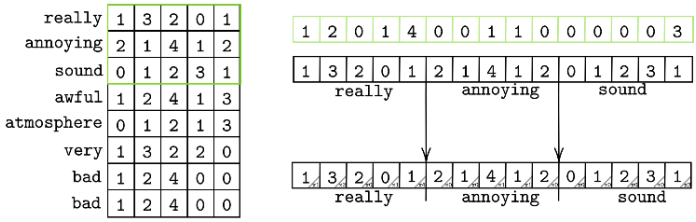

Рис. 3. Применение свертки для извлечения признаков из текстов

Двигаясь таким образом по матрице текста, можно получить столбец значений – результатов операции одномерной свертки. Для получения большего количества столбцов, которые, в свою очередь, образуют матрицу признаков, используется несколько ядер.

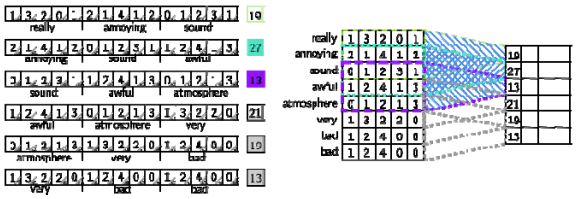

Рис. 4. Применение свертки для извлечения признаков из текстов

После применения операции свертки как для изображений, так и для других структур, в том числе и текстов, возможно применение слоя субдискритизации (Pooling, подвыборки). Пулинг работает подобно сверточному слою, однако вместо скалярного произведения применяет другие функции на фрагменты, например, возвращает максимальный или средний выходы в прямоугольной окрестности (max-пулинг [10] и average-пулинг соответственно).

Max-Pooling:

$$f_{MP}(x) = \max_i x_i.$$

Average-Pooling:

$$f_{AP}(x) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Логика употребления слоя пулинга заключается в том, что выявленные сверточным слоем закономерности избыточны в своей подробности, значит, возможно уплотнение до менее подробного результата. Однако, если в задаче необходимо сохранять точную пространственную информацию, применение пулинга по всем признакам существенно увеличивает ошибку модели, что исследовано экспериментально [7]. Сверточный слой и слой пулинга, следующий за ним, образуют сверточный блок. Чем больше количество сверточных блоков, тем более абстрактной или высокоуровневой становится карта признаков. Заключим, что сверточная нейронная сеть достаточно успешно решает задачу извлечения признаков из данных. Экстраполируем этот вывод на одномерную плоскость для работы с текстами: сверточные нейронные сети, извлекая закономерности в текстовых данных, способны различать контекст. Причем ширину контекста можно приблизительно идеализировать шириной рецептивного поля (или пятна восприятия). На рисунке 5 ширина рецептивного поля на первом слое равна трем, а ширина рецептивного поля на втором слое равна четырем, что позволяет расширить контекст. Однако всё же речь идет о, с точки зрения языка, маленьких паттернах, словосочетаниях и коротких фразах. Для масштабного и более полного моделирования языка необходимо делать очень глубокие сети, что неминуемо приводит к увеличению числа обучаемых параметров.

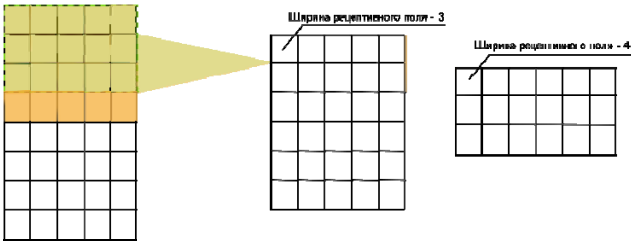

Рис. 5. Ширина рецептивных полей

Существуют некоторые специальные способы увеличения контекста, которые являются спорными и требуют аккуратного использования: ранее уже описанный пулинг и прореживание (dilation) [8]. Последнее заключается в применении свертки к фрагменту, из которого удалена часть элементов. Например, возьмем k строк матрицы текста не подряд, а через одну, но значит это то, что мы пропускаем каждое второе слово, что не может однозначно хорошо сказаться на поставленной задаче. Для решения задачи классификации после череды сверточных блоков (сверточный слой или сверточный слой + пулинг) конечную карту признаков любой размерности, в зависимости от предыдущих слоев, необходимо растянуть в вектор и подать в полносвязный слой с функцией активации, например, SoftMax (для классификации n-классов, $n \in \mathbb{N}$) или Sigmoid (для бинарной классификации), чтобы

получить вероятности классов. Получим формулу размерности карты признаков, учитывая всевозможные преобразования, после применения сверточного слоя или пулинга:

$$L_{\text{вых}} = \frac{L_{\text{вх}} + 2 \cdot \text{padding} - \text{dilation} \cdot (\text{kernel\_size} - 1) - 1}{\text{stride}} + 1,$$

где padding (замощение) – параметр, используемый в случае, когда есть необходимости в сохранении размерности при выполнении операции свертки или пулинга: недостающие элементы для этого заполняются нулями.

Итого примерный вид архитектуры сверточной нейронной сети для задачи классификации текстов изображен на рисунке 6.

В таблице 2 приведены метрики качества accuracy исследуемых архитектур на этапах обучения, валидации и тестирования моделей.

Таблица 2

Метрика accuracy на разных этапах обучения

Данные	Модели								
	CNN (1 слой)			CNN (3 + 3 пулинг)			CNN (5+5)		
	Train	Val	Test	Train	Val	Test	Train	Val	Test
IMDB (2)	**0.9996**	0.8757	0.889	0.997	**0.8994**	0.8765	0.9963	0.8991	**0.8996**
Twitter (3)	0.913	0.8338	0.8447	**0.945**	0.8491	0.8312	0.93	**0.8542**	**0.8771**
Yelp (2)	0.5312	0.5671	0.5431	0.5538	0.534	**0.613**	**0.5652**	**0.5843**	0.548
Epoch	5			5			5		

Таблица 3 содержит сравнение современных фреймворков для обучения нейронных сетей PyTorch и Tensorflow.

В ней приведены длительности обучения моделей.

На рисунке 9 показана матрица ошибок для классификации на три класса.

Таблица 3

Затраченное время на разных фреймворках

Данные	Модели						Объем
	CNN (1 слой)		CNN (3+ 3 пулинг)		CNN (5 + 5)		
	TF	PyTorch	TF	PyTorch	TF	PyTorch	
IMDB (2)	222	81	307	100	383	132	50000
Twitter (3)	385	313	497	433	754	532	75000
Yelp (2)	100421	120212	220053	189429	315371	435341	560000

Таблица 4 посвящена однослойной модели на датасете IMDB и отражает качество с помощью метрик precision (точность), recall (полнота), F1-score (F1-мера).

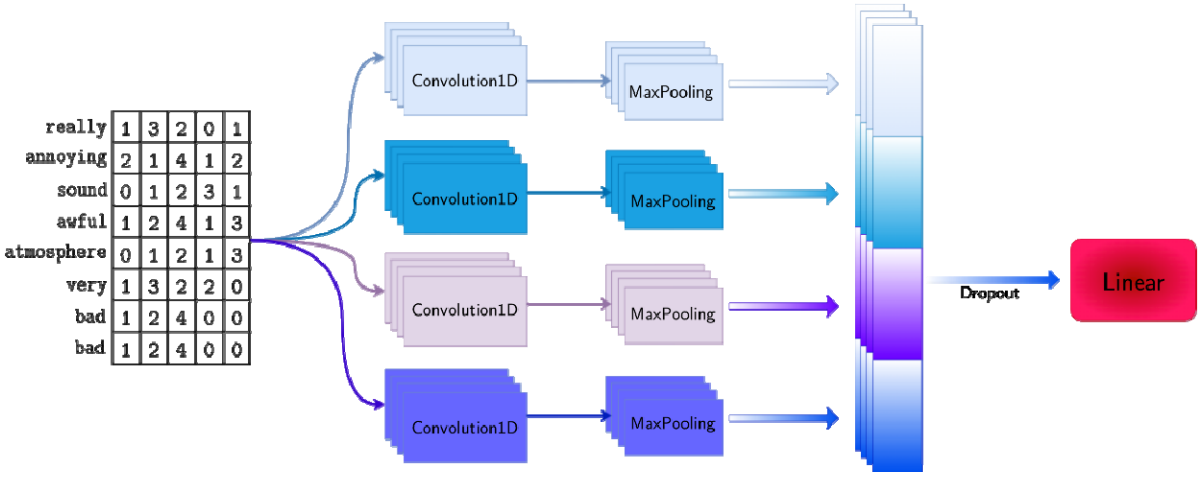

Рис. 6. Архитектура сверточной нейронной сети
(4 сверточных слоя, 4 слоя субдискритизации, линейный слой с дропаут)

Таблица 4

Метрики качества для датасета IMDB

	precision	recall	F1-score	support
negative	0.87	0.93	0.9	7490
positive	0.93	0.86	0.89	7510
accuracy			0.9	15000
macro avg	0.9	0.9	0.9	15000
weighted avg	0.9	0.9	0.9	15000

На рисунках 7. и 8 изображены поведения ошибки и метрики качества в зависимости от выбора размера пакета и оптимизатора, соответственно.

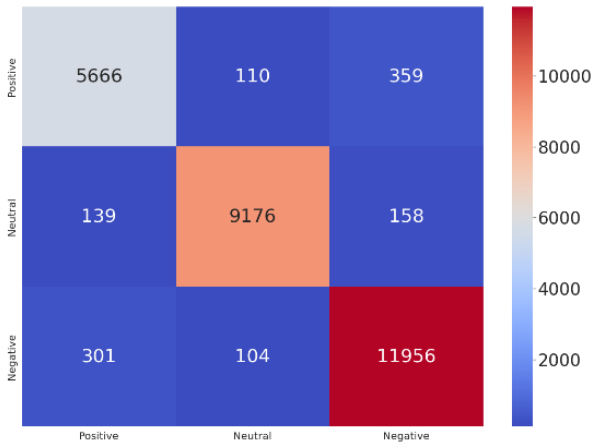

Рис. 7. Confusion Matrix для многоклассовой классификации

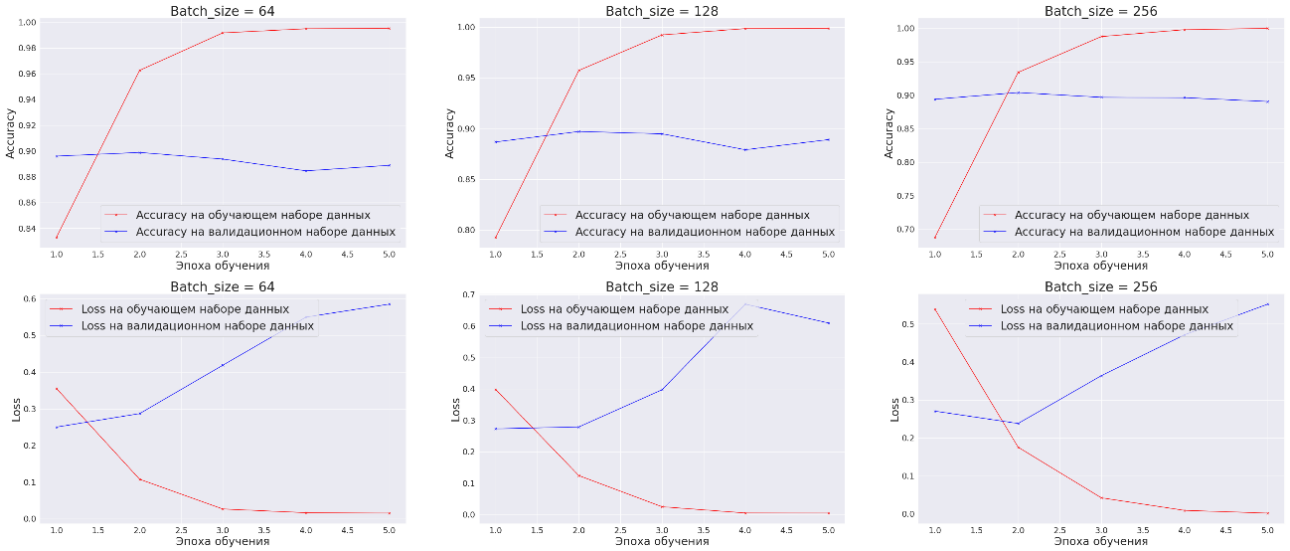

Рис. 8. Поведения accuracy и loss в зависимости от batch size

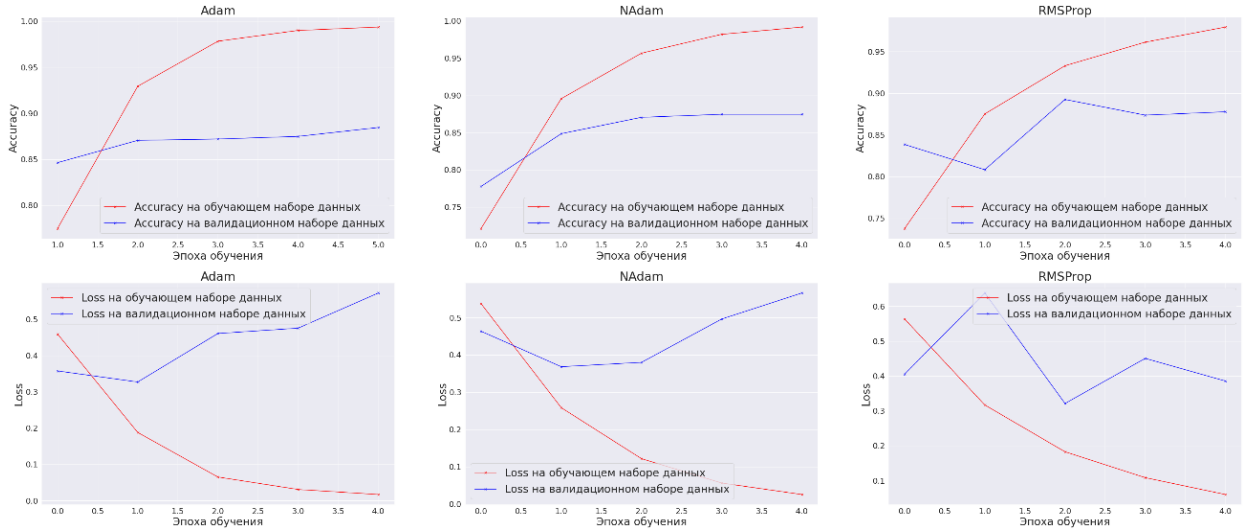

Рис. 9. Поведения accuracy и loss в зависимости от оптимизатора

Рекуррентные нейронные сети

Ширина рецептивных полей сверточных нейронных сетей позволяет определять частичный контекст текстовой информации: прореживание и пулинг дают возможность увеличить ширину рецептивных полей, однако, использование их может привести к росту ошибки в отдельных задачах. Это не мешает отлично справляться с задачей классификации текстов одномерным сверточным сетям, что доказывают, в первую очередь, результаты, несильно уступающие моделям, написанным на основе правил. Но всё становится значительно хуже при постановке задачи моделирования языка. К тому же, сверточные нейронные сети, как и традиционные, требуют входные данные фиксированной длины, что накладывает ограничение на плоское пространство текстов.

В среднем, такие ограничения не оказывают критического влияния, но объемные тексты могут нести важные для классификации детали в той своей части, которая не может быть проанализирована из-за ограничения. В случае сверточных нейронных сетей, текст позиционировался как пространственная структура. В случае рекуррентных нейронных сетей [4,11] следует рассматривать его как последовательность, информация об элементе которой находится в зависимости от предшествующих ему элементов, другими словами, текст имеет направленное повествование. Подобная структура присуща не только текстам, но также и аудиосигналам или временным рядам.

Основная идея рекуррентных нейронных сетей зиждется на введении скрытых состояний, и определении новых состояний через предыдущие. Таким образом, происходит латентное запоминание информации. То есть, в отличие от других нейронных сетей (сверточных, полносвязных), рекуррентная определяет своё состояние не только за счет входных данных, но и за счет предыдущих состояний. Таким образом, типичный слой сети может быть развернут в ленту состояний Рис. 10.

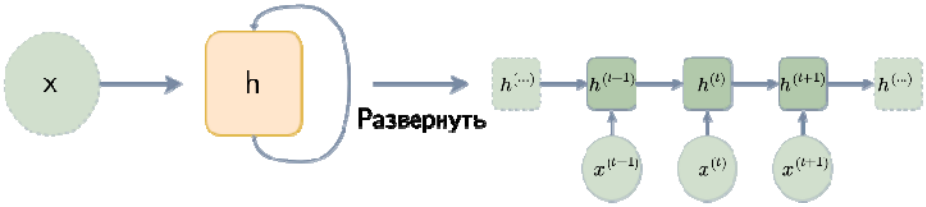

Рис. 10. Развернутая рекуррентная сеть без выходов

На каждом шаге рекуррентной сети:

1. Прочитать очередной элемент последовательности x^t, применить к нему линейное преобразование:

$z^t = W_{\text{input}} \cdot x^t$;

2. Вычислить новое значение состояния, исходя из старого:

$h^t = \text{act}(\text{W}_{\text{hidden}} \cdot h^{t-1} + z^t)$;

3. Выйти из рекуррентного шага:

$$y^t = W_{\text{output}} \cdot h^t.$$

Обучаемыми параметрами являются веса $W_{\text{input}}, W_{output}, W_{\text{hidden}}$, участвующие в линейных преобразованиях.

Нулевое состояние сети может задаваться разными способами: исходя из экспериментальных показателей, лучше всего использовать малодисперсный шум.

Архитектуры рекуррентных сетей, изображенные на рисунке 11 разнятся в соответствии с количеством блоков и выходов из них. Например, для задачи машинного перевода используется Many-to-Many архитектура. В поставленной задаче классификации используется архитектура Many-to-One.

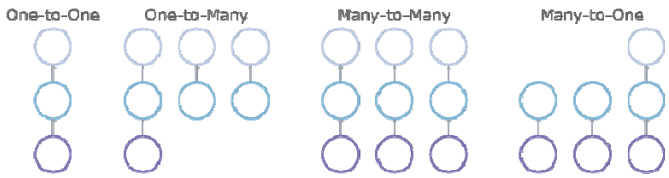

Рис. 11. Архитектуры рекуррентных нейронных сетей

Рекуррентные нейронные сети не могут подвергнуться параллельным вычислениям, поскольку происходит последовательная обработка данных, как следствие, обучение проходит ощутимо дольше, чем для других типов нейронных сетей. Но основной проблемой является то, что во время обучения возникают проблемы сходимости: затухание или взрыв градиента [9].

Затухание и взрыв градиента

Рассмотрим прямой проход (forward pass) по классической рекуррентной нейронной сети (Vanilla RNN) на рисунке 12.

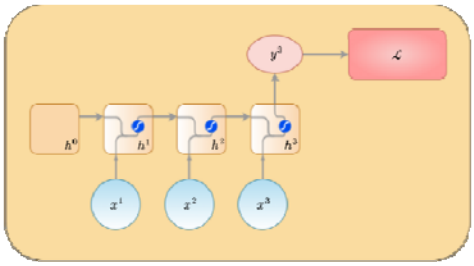

Рис. 12. Vanilla RNN

Введем обозначения:

Входная последовательность $x^t \in \mathbb{R}$, скрытое состояние $h^t = f(w \cdot h^{t-1} + x^t)$, где f – нелинейная, дифференцируемая функция активации, $w \in \mathbb{R}$ параметр функции перехода; выход $y^t = g(h^t)$, функция потерь (функция ошибки, функционал качества) $\mathcal{L}(y^t)$.

1. $h^0 \in \mathbb{R}$;
2. $h^1 = f(w \cdot h^0 + x^1)$;
3. $h^2 = f(w \cdot h^1 + x^2) = f(w \cdot f(w \cdot h^0 + x^1) + x^2)$;
4. $h^3 = f(w \cdot h^2 + x^3) = f(w \cdot f(w \cdot h^1 + x^2) +$

$+x^3) = f(w \cdot f(w \cdot f(w \cdot h^0 + x^1) + x^2) + x^3)$.

Выполним обратный проход (back propagation through time [12]):

$$\mathcal{L}\left(y^3\right) = \mathcal{L}\left(g\left(h^3\right)\right) = \mathcal{L}\left(g\left(f\left(w \cdot h^2 + x^3\right)\right)\right) =$$
$$= \mathcal{L}\left(g\left(f\left(w \cdot f\left(w \cdot h^1 + x^2\right) + x^3\right)\right)\right).$$

Применяя градиентный спуск, нам понадобится производная по параметрам:

$$\frac{\partial \mathcal{L}\left(y^3\right)}{\partial w} = \frac{\partial \mathcal{L}\left(y^3\right)}{\partial g} \cdot \frac{\partial g}{\partial w} = \frac{\partial \mathcal{L}\left(y^3\right)}{\partial g} \cdot \frac{\partial g}{\partial f} \cdot \frac{\partial f\left(w \cdot h^2 + x^3\right)}{\partial w}$$

$$\frac{\partial f\left(w \cdot h^2 + x^3\right)}{\partial w} = f'\left(w \cdot h^2 + x^3\right) \cdot \left(w \cdot h^2 + x^3\right)' = f_3' \cdot \left(w \cdot h^2\right)' =$$
$$= f_3' \cdot \left(w' \cdot h^2 + w \cdot h^{2'}\right) = f_3' \cdot \left(h^2 + w \cdot f\left(w \cdot h^1 + x^2\right)'\right) =$$
$$= f_3' \cdot \left(h^2 + w \cdot f_2' \cdot \left(h^1 + w \cdot f\left(w \cdot h^0 + x^1\right)\right)\right) =$$
$$= \sum_{i=1}^{3} \left(h^{i-1} \cdot w^{3-i} \prod_{j=i}^{3} f_j' \right).$$

Именно это умножение является причиной затухания или взрыва градиента. Одной из часто употребляемых функций активации является гиперболический тангенс, производная которого лежит в диапазоне $0 < \tanh' < 1$.

Умножение большого числа таких значений приведет к затуханию градиента (vanishing gradient) [12]. Борьба с этим – ограничение количества скрытых состояний, что, по сути, ограничивает последовательности. Основная идея этих сетей теряет смысл. Эту проблему можно интерпретировать как заполненность последовательности маловажными в рамках задачи токенами, которые требуют маленькие веса, участвующие в этом произведении и, как следствие, ухудшают оптимизацию (происходит затухание градиента).

С другой стороны, если модуль произведения больше единицы, то произойдет взрыв градиента (переполнение и падение точности). Борьбой с этим является ввод фиксированной границы градиента, до которой снижается градиент, ее превысивший. Затухающий градиент так же связан с долгосрочными зависимостями [13]. Любые флуктуации в краткосрочных зависимостях будут подавлять долгосрочные. В случае, если модель способна представить долгосрочные зависимости, градиент долгосрочного взаимодействия будет экспоненциально ниже краткосрочного. [5]

LSTM (Long short-term memory)

Борьба между мощностью и обучаемостью рекуррентных нейронных сетей неразрывно связана с поведением градиента при их обучении. В 1990-е годы было предложено немало способов решения этой проблемы, среди которых временные задержки внутри состояний [14], временные постоянные. Но наиболее часто ныне используемым решением стала архитектура LSTM [4]. Сеть LSTM, изображенная на Рис. 13., использует концепцию вентильных нейронных сетей (gated Recurrent Neural Networks), основанную на идее контролированного выбора путей между состояниями. Путей, на которых не обнуляются и не устремляются в бесконечность производные. Веса связей скрытых состояний в такой структуре могут изменяться на каждом временном (или пространственном, зависит от структуры анализируемых данных) шаге.

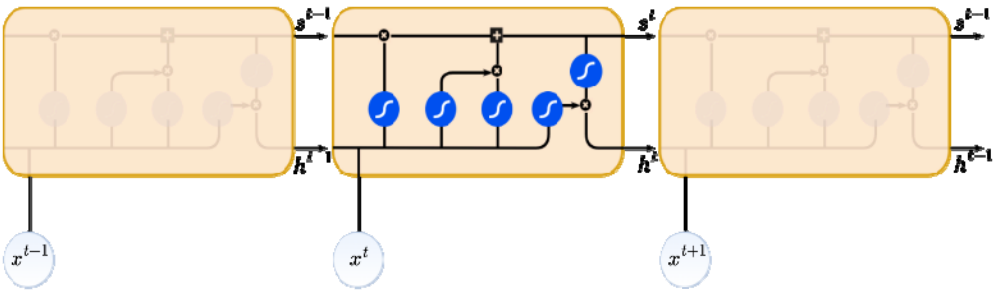

Рис. 13. Срез развернутого LSTM слоя.

Идея вентилей состоит в том, чтобы контролировать вектор значений за счет его умножения на вектор шлюза (вентиля), который управляет потоком ошибки. Целью является сохранение постоянного объема потока ошибки.

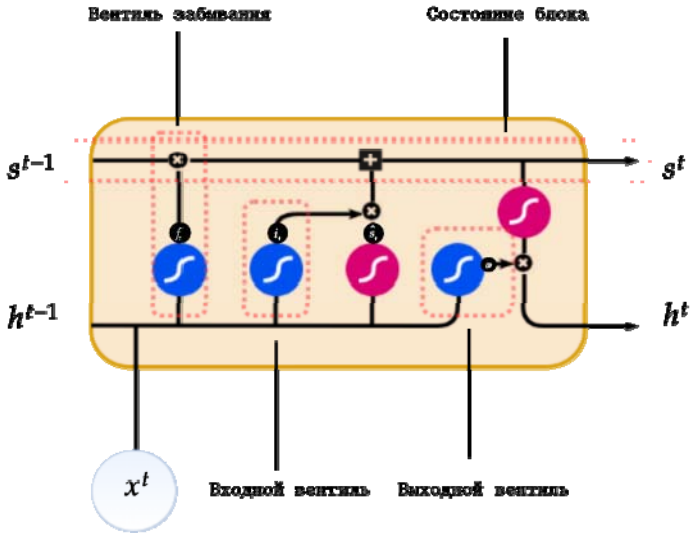

Рис. 14. LSTM блок.

Запишем формальный прямой ход LSTM блока, изображенного на рисунке 14, в момент времени t:

- Конкатенируем h^{t-1} и x^t. Применяем линейные преобразование и нелинейную функцию активации:

$$f_t = \sigma\left(W_f \cdot \left[h^{t-1}, x^t\right] + b_f\right);$$
$$i_t = \sigma\left(W_i \cdot \left[h^{t-1}, x^t\right] + b_i\right);$$
$$\tilde{s}^t = \mathrm{act}\left(W_{\tilde{s}} \cdot \left[h^{t-1}, x^t\right] + b_{\tilde{s}}\right);$$
$$o_t = \sigma\left(W_o \cdot \left[h^{t-1}, x^t\right] + b_o\right);$$

act – функция активации (например, $\tanh$);

- Получим сигнал как линейную комбинацию:

$$s^t = f_t \cdot s^{t-1} + i_t \cdot \tilde{s}^t;$$

Выходим из рекуррентного шага:

$$h^t = o^t \cdot \tanh(s^t).$$

f^t – вентиль забывания (forgetting gate), обучаемый за счет весов W_f и сдвига b_f, определяющий насколько нужно забыть состояние с прошлого блока. i_t – входной вентиль (input gate), обучаемый за счет весов W_i и сдвига b_i, ответственный за чувствительность ко входу:

определяет, насколько важно для запоминания полученное новое состояние $\tilde{s}^t$. Состояние $\tilde{s}^t$ в рекуррентных сетях без вентильной парадигмы было бы выходом блока. o^t – выходной вентиль (output gate), применяемый для увеличения точности аппроксимации нейронной сети за счет нелинейности функции активации. За счет того, что сигмоида принимает значения в диапазоне $0 < \sigma < 1$, все значения вентилей находятся в этом же диапазоне, что позволяет управлять амплитудой значений состояний, не меняя знак и контролируя объем потока ошибки. Внутренние преобразования с новыми и старыми состояниями позволяют классифицировать их важность, как следствие настроить запоминание и забывание того, что нужно для решения задачи и бесполезно, соответственно. LSTM достаточно успешно представляет долгосрочные зависимости, моделирует язык, позволяет решать широкий спектр задач, но не лишена проблем. За счет появления новых внутренних обучаемых весов и сдвигов, увеличивается количество обучаемых параметров в 4 раза по сравнению с обычной рекуррентной нейронной сетью. LSTM часто переобучается и может показывать плохие результаты на новых данных.

В таблице 5 приведены метрики качества accuracy исследуемых архитектур на этапах обучения, валидации и тестирования моделей. Таблица 7 посвящена модели, состоящей из одного LSTM блока, на датасете IMDB и отражает качество с помощью метрик precision (точность), recall (полнота), F1-score (F1-мера).

Таблица 5

Метрика accuracy на разных этапах обучения

Данные	Модели								
	LSTM (1 блок)			LSTM (2 блока)			CNN + LSTM (1 слой + 1 пулинг + 1 блок)		
	Train	Val	Test	Train	Val	Test	Train	Val	Test
IMDB (2)	0.9968	0.8809	0.8573	0.9533	**0.8834**	0.8673	**0.9982**	0.88	**0.8765**
Twitter (3)	**0.855**	0.84	0.778	0.8431	0.8343	**0.841**	0.814	**0.8542**	0.8175
Yelp (2)	**0.9172**	**0.9472**	**0.945**	0.91	0.864	0.8621	-	-	-
Epoch	13			10			10		

Таблица 6 содержит сравнение современных фреймворков для обучения нейронных сетей PyTorch и Tensorflow. В ней приведены длительности обучения моделей.

Таблица 6

Затраченное время на разных фреймворках.

Данные	Модели						Объем
	LSTM (1 блок)		LSTM (2 блока)		LSTM + CNN		
	TF	PyTorch	TF	PyTorch	TF	PyTorch	
IMDB (2)	225	201	8400	8513	1800	1680	50000
Twitter (3)	541	650	28327	31652	3366	3254	75000
Yelp (2)	432002	542331	479032	492151	-	-	560000

Таблица 7

Метрики качества для датасета IMDB

	precision	recall	F1-score	support
negative	0.96	0.91	0.93	7490
positive	0.93	0.86	0.89	7510
accuracy			0.93	15000
macro avg	0.94	0.93	0.93	15000
weighted avg	0.94	0.93	0.93	15000

В результате исследований были разобраны различные способы оптимизации и размеры пакетов (batch size). Результаты исследований во время обучения отображены на рисунках 15 и 16.

На рисунке 17 отображена матрица ошибок для трехклассовой классификации.

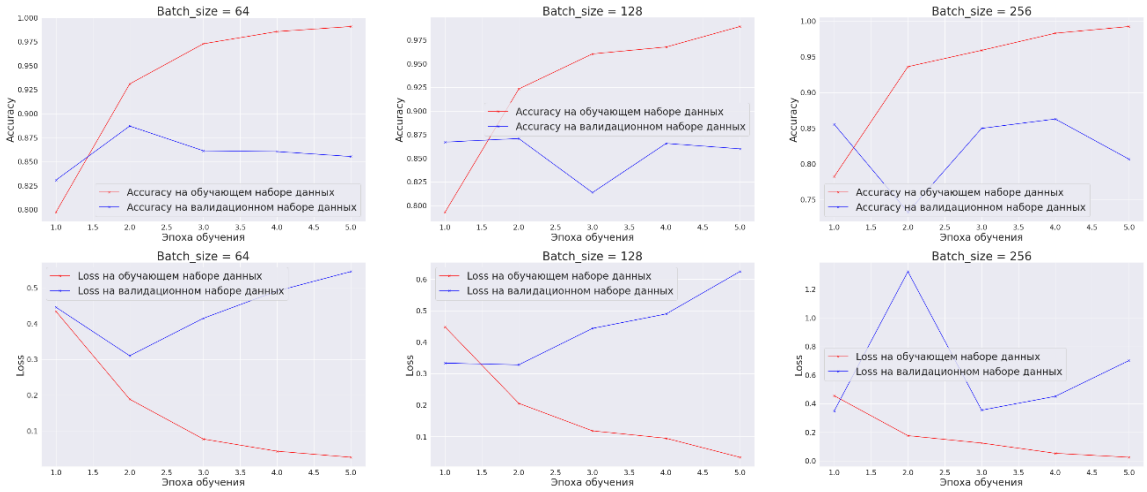

Рис. 15. Поведения accuracy и loss в зависимости от batch size

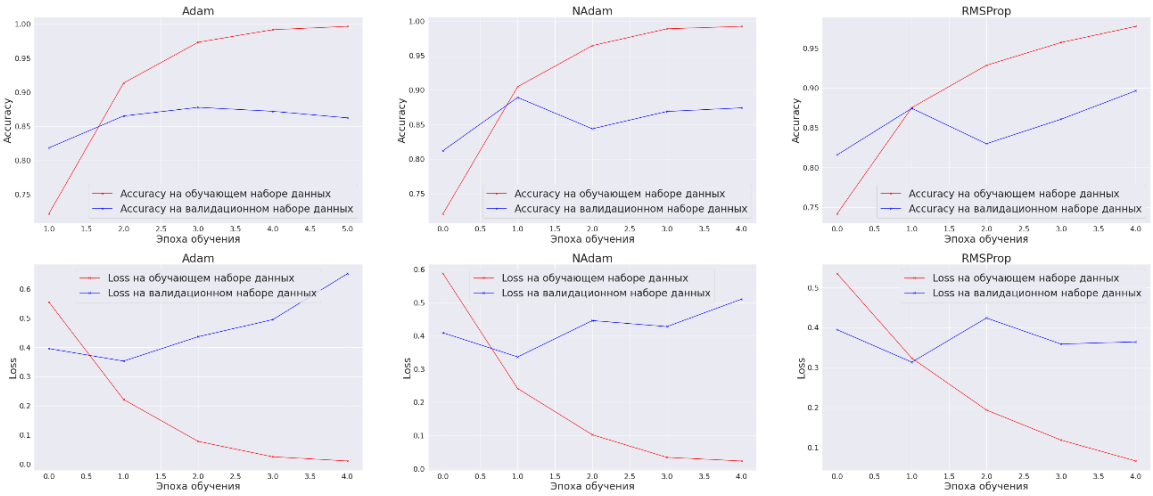

Рис. 16. Поведения accuracy и loss в зависимости от оптимизатора

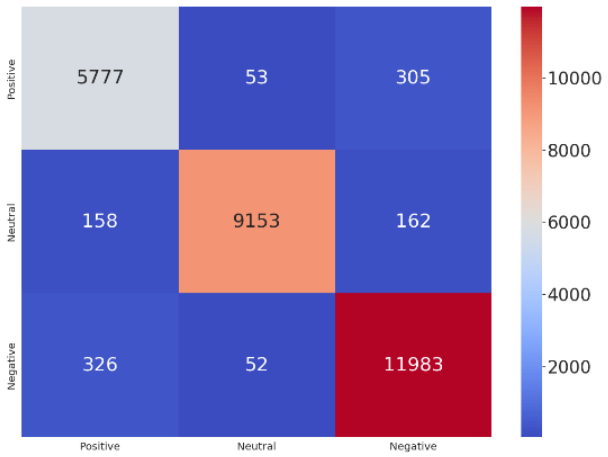

Рис. 17. Confusion Matrix для многоклассовой классификации

GRU (Gated Recurrent Unit)

Уменьшить количество обучаемых параметров без сильной потери качества удалось сети GRU (gated recurrent unit) [15]. Основная идея остается той же – контроль постоянства объема потока ошибки. В сети GRU один вентиль управляет и коэффициентом забывания, и решением об обновлении блока состояния:

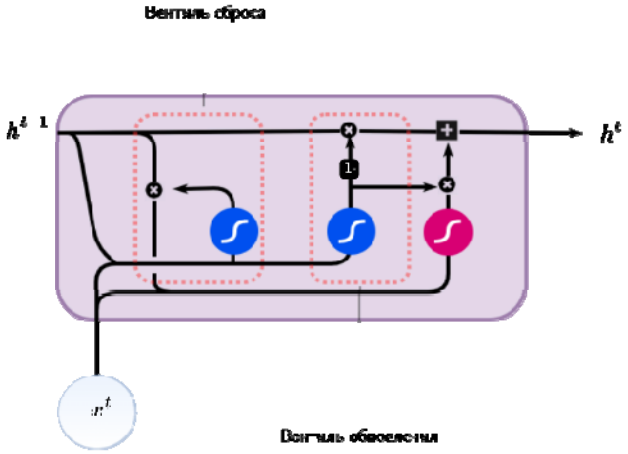

Рис. 18. GRU блок

Запишем формальный прямой ход GRU блока на рисунке 18 в момент времени t:

- Конкатенируем h^{t-1} и x^t. Применяем линейное преобразование и нелинейную функцию активации:

$$u_t = \sigma\left(W_u \cdot \left[x^t, h^{t-1}\right]\right);$$
$$r_t = \sigma\left(W_r \cdot \left[x^t, h^{t-1}\right]\right);$$
$$\tilde{h}^t = \text{act}\left(W_{\tilde{h}} \cdot \left[x^t, r_t \cdot h^{t-1}\right]\right);$$

act – нелинейная функция активации (например, tanh);

- Получим сигнал как линейную комбинацию:

$$h^t = \left(1 - u_t\right) \cdot h^{t-1} + u_t \cdot \tilde{h}_t$$

- Выходим из рекуррентного шага:

u_t – вентиль обновления (update gate), решающий на каждом состоянии, оставить предыдущее значение или обновить.

r_t – вентиль «сброса» (reset gate), позволяющий

контролировать поведение обновленного состояния. В результате, количество обучаемых параметров сократилось на треть. Такой подход значительно сокращает обучение без весомых потерь в мощности.

В таблице 8 приведены метрики качества accuracy исследуемых архитектур на этапах обучения, валидации и тестирования моделей.

Таблица 8

Метрика accuracy на разных этапах обучения

Данные	Модели								
	GRU (1 блок)			GRU (2 блока)			CNN + GRU (1 слой +5 пулинг + 1 блок)		
	Train	Val	Test	Train	Val	Test	Train	Val	Test
IMDB (2)	0.9751	**0.9612**	**0.9353**	0.9903	0.8763	0.8931	**0.9969**	0.8851	0.9184
Twitter (3)	0.8788	0.8338	0.8447	**0.9937**	0.9605	0.9574	0.93	**0.8542**	**0.8771**
Yelp (2)	**0.9670**	**0.9562**	**0.9459**	0.954	0.943	0.9126	**-**	**-**	-
Epoch	5			5			5		

Таблица 9 содержит сравнение фреймворков PyTorch и Tensorflow по времени обучения.

Таблица 9

Затраченное время на разных фреймворках.

Данные	Модели						Объем
	GRU (1 блок)		GRU (2 блока)		GRU + CNN		
	TF	PyTorch	TF	PyTorch	TF	PyTorch	
IMDB (2)	189	204	569	642	244	183	50000
Twitter (3)	546	498	759	802	650	663	75000
Yelp (2)	433200	4542233	679036	592431	-	-	560000

Таблица 10 содержит прочие метрики: precision (точность), recall (полнота), F1-score (F1-мера) для модели с одним GRU блоком на датасете IMDB.

Таблица 10

Метрики качества для датасета IMDB

	precision	recall	F1-score	support
negative	0.93	0.97	0.95	7490
positive	0.97	0.93	0.95	7510
accuracy			0.95	15000
macro avg	0.95	0.95	0.95	15000
weighted avg	0.95	0.95	0.95	15000

На рисунке 21 изображена матрица ошибок для классификации отзывов на три класса для модели с одним блоком GRU.

На рисунке 19 и 20 представлены зависимости метрики качества accuracy и функции ошибки (бинарная кросс-энтропия) от размеров пакетов (batch size) и выбора оптимизатора.

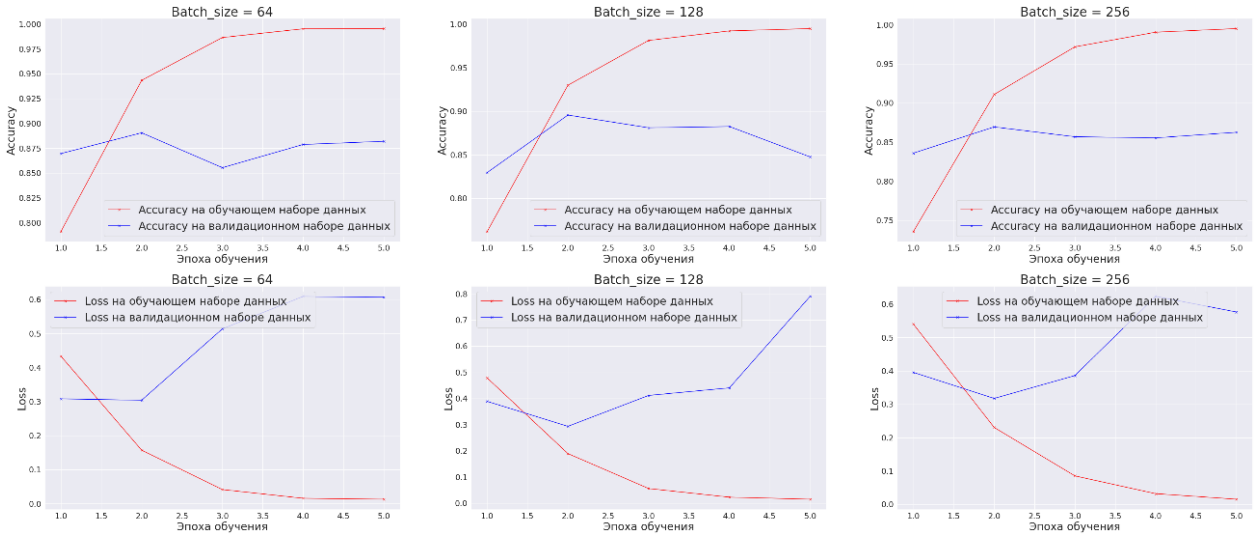

Рис. 19. Поведения accuracy и loss в зависимости от batch size

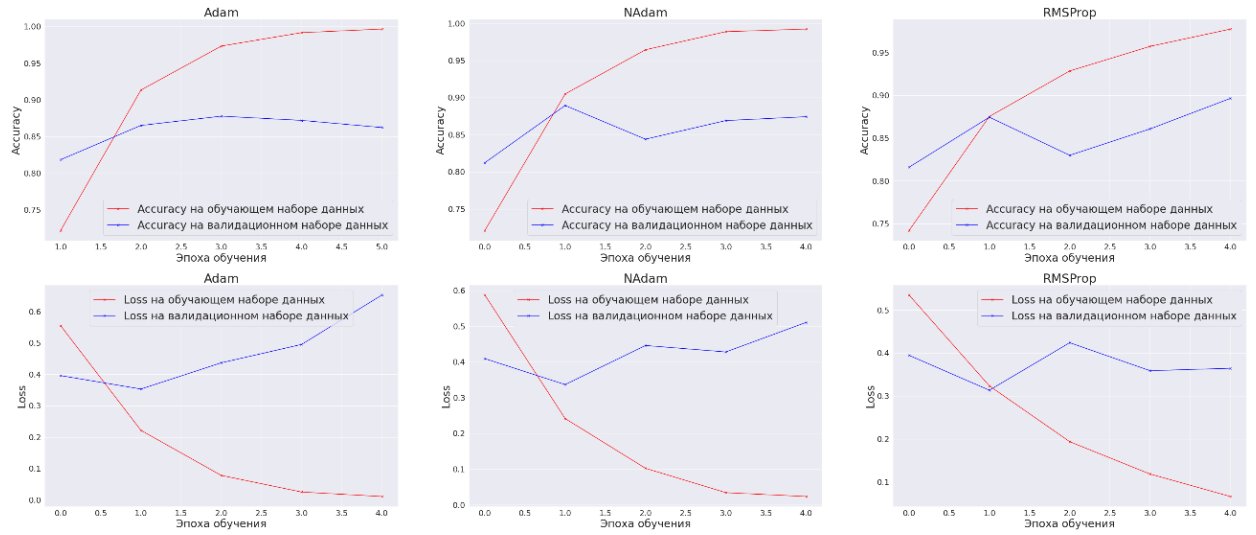

Рис. 20. Поведения accuracy и loss в зависимости от оптимизатора

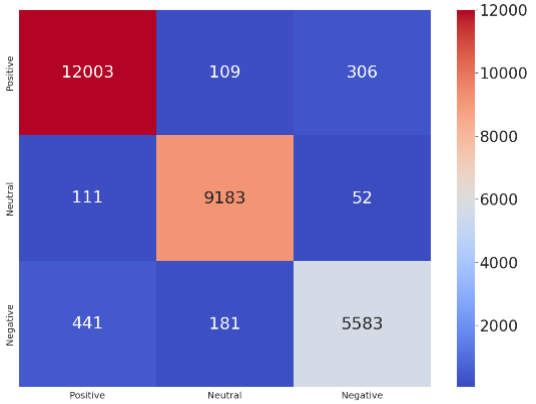

Рис. 21. Confusion Matrices для многоклассовой классификаций соответственно

Stacking

Полученные обученные модели показывают хорошие результаты, однако их можно дополнительно улучшить с помощью технологий стэкинга. Идея заключается в создании метамодели, которая будет по результатам мета-признаков моделей давать конечное предсказание. Обучение моделей для создания мета-модели проходило по принципу K-Foldation (рис. 22).

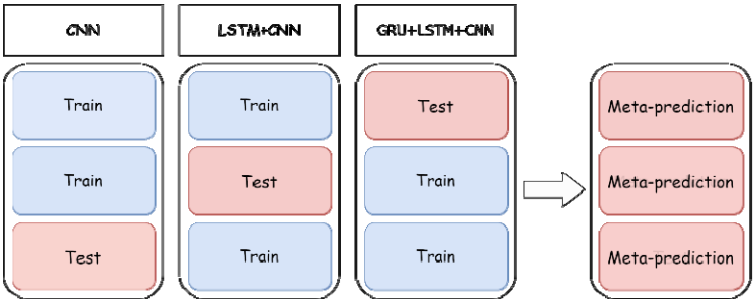

Рис. 22. K-foldation.

Тренировочный датасет объемом 1.5 миллионов отзывов (результат присоединения датасетов Amazon и IMDB), был разбит на три части, на двух из которых каждая модель обучалась, затем у нее убирался полносвязный слой с sigmoid функцией активации, чтобы получить признаковое описание начальных данных сложной структуры в виде табличной структуры, а на третьей части каждая модель давала предсказания без пересечения между собой. Мета-предсказания, будучи табличными данными, были адресованы в классификаторы Random Forest, GBM, SVM, обученные по принципу кросс-валидации (рис. 23.).

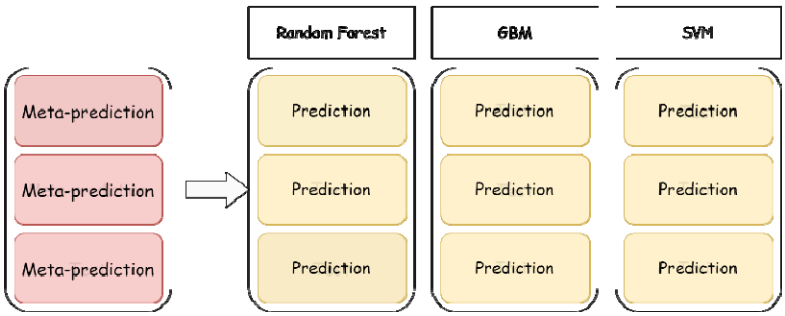

Рис. 23. Meta-Models.

Конечный результат получается усреднением мета-предсказаний мета-моделей (рис. 24).

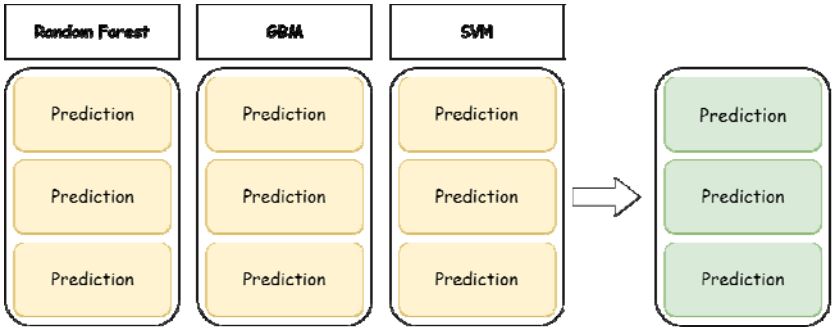

Рис. 24. Prediction

Полученная мета-модель работает несколько лучше других обученных алгоритмов. Итоговая метрика качества accuracy = 0.965.

Заключение

В этой статье для решения задачи классификации текстов были реализованы архитектуры: одномерные сверточные нейронные сети, сети LSTM и GRU. Кроме того, рассмотрен метод ансамблирования моделей, несколько повысивший качество обученных алгоритмов. Обучение моделей происходило с применением технологий регуляризации Batch Normalization [6] и Dropout [8]. Обученные модели показывают отличные результаты, но они не лишены индивидуальных проблем. Одномерные сверточные нейронные сети способны моделировать контекст шириной в несколько слов, словосочетаний, предложений, чтобы увеличить ширину контекста, необходимо делать сети намного более глубокими, что увеличивает количество обучаемых параметров. Это неминуемо приводит к увеличению длительности обучения.

Для рекуррентных нейронных сетей нет проблемы контекста, поскольку они способны латентно запоминать прошедшую через их состояния информацию, однако проблемы затухающего градиента ухудшают или даже делают невозможным обучение. LSTM решает эту проблему, но имеет в 4 раза больше параметров, чем традиционные рекуррентные сети, что замедляет обучение. Как следствие, можно сделать вывод, что самым стабильным по качеству классификатором текстов можно считать GRU, не уступающий по качеству LSTM, однако превосходящий его по скорости обучения. Рассмотренное ансамблирование способно сделать результаты моделей еще более уверенными.

Литература

1. *Y. LeCun and Y.Bengio*, Convolutional networks for images, speech and time-series. Brain Theory and Neural Networks, 1995.
2. *Y. LeCun, K. Kavukcuoglu and C. Farabet*. Convolutional networks and applications in vision. In Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on, pages 253-256. IEEE, 2010.
3. *Y. LeCun, B.Boser, J.S. Denker, D. Henderson, R.E.Howard, W. Hubbard, L.D. Jackel,* et al. Handwritten digit recognition with a back-propagation network. In Advances in neural information processing systems, 1990.
4. *J. Schmidhuber*, “Deep learning in neural networks: an overview”, Neural networks, vol. 61, pp. 29-33, 2015
5. “Deep learning” by Ian Goodfellow, Yoshua Bengio, Aaron Courville, pages 283-308, 2017.
6. *Ioffe, S. and Szegedy. C.* Batch normalization: Acceleration deep network training by reducing internal covariate shift, 2015.
7. *Szegedy, C., Liu, W., Jia, Y., Sermanet, P. Reed, S. Anguelov, D. Erhan, D. Vanhoucke, V. and Rabinovich A*. Going deeper with convolutions. Technical report, 2014.
8. *Srivastava N., Hinton G., Krizhevsky A, Sutskever I., Salakjutdinov R. Dropout*: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 2014, 15, pp. 1929-1958.
9. *B. Hammer*, On the Approximation Capability of Recurrent Neural Networks. Neurocomputing, 31(1-4):107-123, 2000.
10. *Zhou, Y. and Chellappa R*. Computation of optical flow using a neural network. In Neural Networks, 1988.
11. Sepp Honchreiter, Jurden Shmidhuber, Long Short-Term Memory. Neural Comput 1996; 9 (8), pp. 1735-1780.
12. *R. J. Williams and D. Zipser*. Gradient-based Learning Algorithms for Recurrent Networks and Their Computational Complexity. In Y. Chauvin and D. E. Rumelhart, editors, Back-propogation: Theory, Architectures and Applications, pages 433-486. Lawrence Erlbaum Publishers, 1995.
13. *S. Hochreiter. Y. Bengio, P. Frasconi and J. Schmidhuber*. Gradient Flow in Recurrent Nets: the Difficulty of Learning Long-term Dependencies. In S.C. Kremer and J.F. Kolen, editors, A Field Guide to Dynamical Recurrent Networks. IEEE Press, 2001a.
14. *K. J. Lang, A. H. Waibel and G. E. Hinton*. A Time-delay Neural Network Architecture for Isolated Word Recognition. Neural Networks, 3(1):23-43, 1990.
15. *Jozefowicz R., Zaremba W. and Sutskever I.* An empirical evaluation of recurrent network architectures. In ICML2015, 2015.

АНАЛИЗ ТЕХНОЛОГИЙ БЕСКОНТАКТНОЙ АВТОМАТИЧЕСКОЙ ИДЕНТИФИКАЦИИ ПАЦИЕНТОВ

Тимощук Юлия Сергеевна,
Московский технический университете связи и информатики, студент группы БСУ1801, Москва, Россия
iul.tim2012@yandex.ru

Маклачкова Виктория Валентиновна,
Московский технический университете связи и информатики, ст. преп. кафедры СИТиС, Москва, Россия
v.v.maklachkova@mtuci.ru

Аннотация
Ошибки идентификации пациентов представляют один из самых серьезных факторов риска безопасности пациентов в медицинских учреждениях. Одним из вариантов снижения данного риска является автоматическая идентификация. Производится сравнительный анализ существующих методов для автоматической идентификации пациентов. Особое внимание уделяется тем характеристикам, которые являются наиболее важными при выборе конкретной технологии.

Ключевые слова: *Идентификация, штрих-код, RFID, биометрия, биометрические данные, риски, персональные данные, медицинские учреждения.*

Введение

Одной из целей при обеспечении безопасности пациента является его идентификация. Неправильная идентификация является одной из важнейших причин инцидентов, подвергающих опасности здоровье, а зачастую и жизнь, людей. Ошибки в идентификации могут встречаться на любом этапе получения человеком медицинской помощи: начиная с диагностики и заканчивая самим процессом лечения. Подобные ошибки происходят, в частности, в таких ситуациях, когда пациент находится в бессознательном состоянии (ввиду принятия седативных препаратов или по иным причинам), в состоянии нарушенной ориентации или у пациента есть физические или психические отклонения, что чревато ошибками при проведении таких жизненно важных мероприятий, как прием лекарств, переливание крови, клинические испытания, хирургические операции и целый ряд других медицинских процедур.

Использование автоматической идентификации снижает подобные риски. Автоматическая идентификация - набор методов, предназначенных для автоматизации ввода данных. Подобные системы могут обеспечить быструю и надежную идентификацию пациента, а также удаленное управление историей болезни пациента и мгновенный доступ к ней. Хотя основной целью любой системы такого рода является повышение надежности идентификации пациента, обеспечение быстрого доступа к клинической информации является крайне полезным качеством [23,24]. Кроме того, система должна включать аспекты безопасности и конфиденциальности обрабатываемых медицинских данных [26].

Оптические методы идентификация объектов

В настоящее время широко распространено использование оптических методов идентификации. К ним относятся технология штрихового кодирования.

Популярность штрих-кодов можно объяснить их низкой стоимостью производства и простотой печати. Штрих-коды достаточно надежны (при условии их целостности), их вероятность ошибки один к двум миллионам. Также штрих-коды на данный момент являются стандартом для идентификации чего-либо во всем мире, что долгое время давало им значительное преимущество перед более молодыми технологиями как RFID.

Штрих-коды могут быть как линейными, так и двухмерными (рисунок 1). Среди линейных популярных стандартов являются EAN, UPC, Code39, Code128, Codabar, Interleaved 2 of 5 [2], однако их

недостатком является малый объем закодированной информации по сравнению с двухмерными кодами.

Двухмерные отличаются бо́льшим объемом данных. Среди них самыми распространенными стандартами являются PDF 417, MaxiCode, DataMatrix, Aztec Code [2] и QR-код.

Рис. 1. Пример линейного и двухмерных штрих-кодов

Технологию штрих-кодов уже давно применяют в медицине – в частности, в таких сферах как фармацевтика, для защиты от возможности подделки лекарств, для идентификации в лабораторной диагностике, а также для идентификации людей. Системы верификации, к примеру, переливания крови со штрих-кодом подтверждают личность пациента, а также могут отображать заказы на переливание, отслеживать препараты крови и вести учет переливания. Такие системы с использованием штрих-кодов устраняют возможность человеческой ошибки при использовании браслетов пациентов и делают процесс более безопасным и эффективным [4].

Среди преимуществ штрих-кодов можно выделить:

- дешевизна, поскольку для их изготовления требуются только чернила и материал для печати (бумага или пластик);
- работоспособность штрих-кодов не зависит от материала, на котором они расположены;
- всегда есть возможность считать штрих-код с любого устройство, имеющего камеру.

Однако, у данной технологии довольно много недостатков:

- для считывания штрих-кода сканерам необходима прямая видимость штрих-кода;
- чтобы считать штрих-код, сканер должен находиться довольно близко;
- штрих-коды не имеют возможности чтения/записи, по этой причине для изменения информации код придется создать и напечатать заново;
- нет способа защиты информации в штрих-коде;
- печатный идентификатор достаточно легко повредить, что сделает штрих-код нечитаемым;
- не могут хранить большой объем данных, что ограничивает их применение в качестве идентификатора;
- при увеличении объема информации, кодируемого двухмерным штрих-кодом, увеличивается площадь необходимой поверхности для печати;
- относительно медленная скорость чтения.

На данный момент эта технология является достаточно распространенной технологией автоматической идентификации в области здравоохранения в России: одним из последних нововведений стало использование QR-кодов для подтверждения наличия прививки от COVID-19 или перенесенного заболевания [14], кроме этого, в больнице вводят идентификационные браслеты, с нанесенными на них штрих-кодами для обеспечения безопасности пациентов [15].

Радиочастотный метод идентификации

RFID (Radio Frequency IDentification) – технология радиочастотной автоматической идентификации, принцип действия которой основан на электромагнитном излучении. Стандартная система RFID состоит из трех компонентов: RFID-метка, которая прикрепляется к объекту для идентификации,

RFID-считыватель, который запрашивает данные с метки, и программное обеспечение для обработки и хранения информации, полученной от считывателей.

RFID-системы способны работать на низких, высоких, сверхвысоких или микроволновых частотах. С повышением частоты уменьшается дальность считывания, но повышается скорость передачи данных, что позволяет реализовать сложные протоколы данных [3]. По методу используемых источников питания RFID-системы делятся на пассивные и активные. Пассивные метки действуют на небольшой дальности и, помимо этого, имеют почти неограниченный срок службы, поскольку в качестве источника питания используется электромагнитное поле RFID-считывателя. Активные RFID-метки имеют собственный источник питания, благодаря чему достигается большая дальность считывания и возможность реализации энергозависимой памяти [1].

В качестве идентификатора пациента может выступать умный браслет с пассивной меткой RFID, который может быть отсканирован для получения информации о пациенте. В качестве альтернативного варианта для пассивных меток с небольшим объемом памяти возможна идентификация на основе запроса и получения правильных медицинских данных из различных существующих информационных систем здравоохранения, что, в свою очередь, потенциально может минимизировать ошибки при работе с пациентами и повысить безопасность всей системы [5,11].

Среди положительных сторон технологии RFID можно выделить следующее:

- высокая скорость считывания;
- устойчивость к разрывам и влаге;
- возможность использования меток на средней и большой дальности;
- возможность хранения большого объема данных;
- при наличии надежного протокола шифрования возможно хранение части конфиденциальной информации на самой метке;
- возможность отслеживания метки и данных на ней в режиме реального времени;
- небольшой размер метки.

В качестве отрицательных сторон можно отметить [1]:

- высокая цена развертки RFID-систем;
- сложность в обеспечении конфиденциальности пациентов при использовании низкочастотных меток;
- возможность существования проблем при использовании низкочастотных меток с медицинским оборудованием, чувствительным к радиочастотам;
- для активных RFID-меток требуется источник питания, требующий замены раз в определенный период;
- вероятность ложноположительных и ложноотрицательных срабатываний.

Ввиду высокой цены развертки RFID-системы, она не так широко распространена в области здравоохранения, однако в России осуществляются попытки внедрения для идентификации пациентов [16].

Методы идентификации по биометрическим данным

Третьей группой методов идентификации можно выделить биометрическую идентификацию, когда изучаются уникальные признаки, которые в большей или меньшей степени не поддаются изменению.

Сейчас активно используются такие параметры, как:

- отпечатки пальцев – сравнение существующего отпечатка с приложенного к сканеру производят корреляционным методом (попиксельное сравнение), по узору (алгоритм с разбивкой на области и сравнение узоров, описанных синусоидальной волной) и по особым точкам (выделение точек края отпечатка и ветвления и последующее сравнение корреляционным методом);
- голос – для сравнения голоса и идентификации говорящего используется голосовой отпечаток (или спектрограмма), который создается путем анализа частоты, продолжительности и амплитуды звука;
- структура лица - существует несколько групп методов распознавания лиц: метод сравнения на графах, сканирование уникальных характеристик, а также метод с использованием нейронных сетей [6,7];

- радужная оболочка глаза – с помощью сегментации находят область радужной оболочки глаза, откуда извлекается фазовая информация [8,9];
- сетчатка глаза – для данного способа биометрической идентификации с помощью инфракрасного метода сканируются кровеносные капилляры сетчатки [9];
- ладонь – идентификация может проводится разными методами: методом прямых измерений (сравнение фиксированных параметров руки с помощью векторов) [10], методом с использованием силуэта руки (сравниваются контуры руки), а также методом идентификации по рисунку вен ладони [10];
- почерк – для данного способа, помимо формы букв, также анализируются такие характеристики, как давление при письме, наклон и специальные координаты. Этот метод популярен в банках и других финансовых учреждениях.

Помимо перечисленных методов биометрической идентификации, могут использоваться комбинированные системы, использующие несколько биометрических параметров.

Преимуществами биометрической идентификации являются:

- биометрические данные всегда с человеком, в отличии от документов или других идентификаторов;
- биометрические данные достаточно сложно подделать;
- методы биометрической идентификации просты в использовании.

Недостатками биометрической идентификации являются:

- сложности в опознавании детей, поскольку многие параметры, использующиеся для идентификации у детей подвержены изменениям;
- высокий шанс ложноположительного или ложноотрицательного срабатывания;
- биометрические данные не подлежат замене при утечке данных;
- цена развертки биометрической системы может быть очень высокой.

Биометрия в России в основном используется в системах контроля доступа, а также в финансовой отрасли (бесконтактные платежи). Кроме единой биометрической системы, предполагаемой для использования в других отраслях, помимо финансовой, биометрическая идентификация практически никак не представлена в сфере здравоохранения. Одной из ранних попыток внедрения биометрии в больницы является проект 2002 года по реализации информационной системы и обеспечением её защиты по биометрическим данным (по отпечатку пальца) [17], однако, в данном случае производится идентификация персонала.

Возможные риски автоматической идентификации

Поскольку описанные методы предполагается использовать в здравоохранительных учреждениях, задача построения систем автоматической идентификации усложняется из-за наличия некоторых общих факторов:

- **Риск ошибочного считывания** – в области медицины крайне важны высокоточные данные, поскольку часто от них зависит жизнь пациента. Возможные ошибки в идентификации пациентов сильно снижают точность данных и по этой причине больницы, использующие автоматическую идентификацию, должны быть готовы к возможным ошибкам.
- **Обеспечение конфиденциальности пациентов** – данный риск возникает при хранении личной информации пациента на идентификаторе, что открывает злоумышленнику шанс получить доступ к конфиденциальным данным. Также у уникального идентификатора, привязанного к определенному человеку, существует риск установления незаконной слежки [1,19,20].
- **Риск потери идентификатора** – подобный риск наиболее вероятен при использовании штрих-кода или метки RFID, кроме того, при печати на легко деформируемом материале (таком как бумага) есть вероятность случайного уничтожения идентификатора.
- **Несовместимость ПО с медицинской информационной системой (МИС)** – для любой системы автоматической идентификации требуется собственное программное обеспечение, но оно может быть несовместимо с больничной МИС, что ограничивает использование части методов и может содержать угрозу конфиденциальности данных.

В области здравоохранения также наблюдается тенденция использования территориально-распределенной облачной инфраструктуры хранения данных пациентов [18,25], что также может не-

сти в себе потенциальный риск раскрытия личных данных пациента и усложняет процесс идентификации рисков с целью разработки в дальнейшем способов и средств по их управлению [1,22,26-28].

Заключение

Был проведен анализ трех групп методов идентификации личности пациента. Краткое сравнение их параметров отображено в таблице 1.

Таблица 1

Сравнение параметров технологий автоматической идентификации

Параметры	RFID	Штрих-код	Биометрические данные
Прямая видимость	Нет	Да	Да
Расстояние считывания	Большое	Очень низкое	Низкое
Легкость создания идентификатора	Нет	Да	Нет
Объем данных	до 32 КБ (MIFARE DESFire EV2 [12])	до 3 КБ (QR code, Version 30-L [13])	Нет
Скорость считывания	Очень быстро	Низкая	Низкая
Влияние внешней среды на читаемость (грязь и т.д.)	Нет	Высокое	Высокое
Стоимость	Средняя	Низкая	Очень высокая

На данный момент в сфере здравоохранения для автоматической идентификации используются одномерные и двухмерные штрих-коды DataMatrix [21]. Также, во многих процессах, требующих идентификации, таких как хирургическая операция или прием лекарственных препаратов, возможно использование RFID-меток в качестве идентификаторов с неменьшей эффективностью ввиду их легкости считывания независимо от положения идентификатора (в частности, изгиб браслета с длинным штрих-кодом мешает считыванию), а также отсутствию необходимости прямой видимости. Однако, применение технологии RFID сильно ограничено из-за высокой цены интеграции.

В мировой практике внедряются различные методы идентификации пациентов, каждый из которых сопровождается своим набором возможностей и проблем и не дает единого решения со 100% совпадением. Ожидается, что объем медицинских данных будет продолжать расти, как и потребность включения различной информации в электронную медицинскую карту. Кроме того, возникает необходимость в объединении электронных записей, обмене и совместном использовании данных. Без уникальных, однозначных идентификаторов объединить новые потоки данных в медицинской карте будет все сложнее.

Применение технологий автоматической идентификации в здравоохранении открывает широкие возможности для повышения безопасности пациентов и может намного снизить вероятность медицинских и человеческих ошибок в больницах и других медицинских учреждениях. Однако, не стоит забывать про существующие риски, общие для всех методов, в том числе и такие серьезные как вероятность ошибки при считывании или утечки конфиденциальных данных пациентов, которые могут серьезно ограничивать внедрение автоматической идентификации.

Литература

1. *Тимощук Ю.С., Маклачкова В.В.* Риски применения RFID-технологии в медицинских учреждения // Телекоммуникации и информационные технологии. 2021. Т. 8. № 2. С. 80-84.
2. *Белышев Д. В., Гулиев Я. И.* Использование технологий штрих-кодирования в медицинских информационных системах – Программные системы: теория и приложения". Переславль-Залесский. Vol. 2, 2009. P. 71-96.
3. *Mier, J., Jaramillo-Alcázar, A., & Freire, J. J.* (2019). At a Glance: Indoor Positioning Systems Technologies and Their Applications Areas. Explorations in Technology Education Research. P. 483-493.
4. *Dea M. Hughes, Agrawal, A.* (Ed.). Patient Identification – Patient Safety – Springer Science+Business Media New York 2014. P. 3-18.
5. *Moutaz Haddara, Anna Staaby,* RFID Applications and Adoptions in Healthcare: A Review on Patient Safety Procedia Computer Science. Vol. 138 2018. P. 80-88.
6. *Dzhangarov A. I., Suleymanova M. A., Zolkin A. L.* Face recognition methods IOP Conference Series: Materials Science and Engineering 2020.
7. *Jafri, Rabia, Arabnia, Hamid. A* Survey of Face Recognition Techniques . JIPS. 5. 2009.

8. *Daugman J.* How iris recognition works // IEEE Transactionson Circuits and Systems for Video Technology. 2004. Vol. 14, no. 1. P. 21-30.

9. *Павельева Е. А.* Обработка и анализ изображений на основе использования информации о фазе // Компьютерная оптика, 42:6 (2018). С. 1022-1034.

10. *Panchal P., Bhojani R.* An Algorithm for Retinal Feature Extraction Using Hybrid Approach – Procedia Computer Science. Vol. 79. 2016. P. 61-68.

11. *Riplinger L, Piera-Jiménez J, Dooling JP.* Patient Identification Techniques - Approaches, Implications, and Findings. Yearb Med Inform. 2020 Aug;29(1), pp. 81-86.

12. MIFARE DESFire EV2 contactless multi-application IC – Product short data sheet 2019 // URL:https://www.nxp.com/docs/en/data-sheet/MF3DX2_MF3DHX2_SDS.pdf (дата обращения 25.12.2021).

13. ISO/IEC 18004:2015 Information technology — Automatic identification and data capture techniques — QR Code bar code symbology specification // URL: https://www.iso.org/standard/62021.html (дата обращения 25.12.2021).

14. Законопроекты об использовании QR-кодов в общественных местах и на транспорте внесены в Госдуму // Правительство Российской Федерации, официальный сайт (дата обращения 27.12.2021).

15. Еще одна больница в Приморье внедрила браслетную систему для пациентов // URL:https://www.primorsky.ru/news/186465/ (дата обращения 27.12.2021).

16. RFID-браслеты «Микрона» для идентификации пациентов протестировали в подмосковной клинике // URL: https://mikron.ru/company/press-center/news/2137/ (дата обращения 27.12.2021).

17. Биометрия в больнице // URL:https://www.itweek.ru/infrastructure/article/detail.php?ID=63007 (дата обращения 27.12.2021).

18. *Maklachkova V. V., Dokuchaev V. A., Statev V. Y. Risks* identification in the exploitation of a geographically distributed cloud infrastructure for storing personal data // 2020 International Conference on Engineering Management of Communication and Technology, EMCTECH 2020 - Proceedings, Vienna, 20-22 октября 2020 года. Vienna, 2020. P. 9261541 – DOI 10.1109/EMCTECH49634.2020.9261541.

19. *Dokuchaev V. A., Maklachkova V. V., Statev V. Yu*. Classification of personal data security threats in information systems // T-Comm. 2020. Vol. 14. No 1. P. 56-60. DOI 10.36724/2072-8735-2020-14-1-56-60.

20. *Докучаев В. А*., *Маклачкова В. В., Статьев В. Ю*. Требования к информационным системам при работе с «цифровым образом» субъекта // III Научный форум телекоммуникации: теория и технологии ТТТ-2019 : Материалы XXI Международной научно-технической конференции, Казань, 18-22 ноября 2019 года. Казань: Казанский государственный технический университет им. А.Н. Туполева, 2019. С. 296-297.

21. ГОСТ Р 58502-2019. Информатизация здоровья. Термины и определения: Утвержден и введен в действие Приказом Федерального агентства по техническому регулированию и метрологии от 29 августа 2019 г. N 572-ст. URL: https://docs.cntd.ru/document/1200167695(дата обращения: 25.01.2021).

22. *Гадасин Д. В., Шведов А. В., Клыгина О. Г., Гадасин Д. Д.* Реализация платформы туманных вычислений для предоставления сервисов IoT // REDS: Телекоммуникационные устройства и системы. 2021. Т. 11. № 2. С. 65-75.

23. *Назаров М. Д., Шведов А. В.* Корреляция атрибутов соглашения об уровне обслуживания с основными параметрами QoS в корпоративных сетях // Телекоммуникации и информационные технологии. 2020. Т. 7. № 2. С. 73-79.

24. *Шведов А. В., Назаров М. Д.* Зависимость показателей эффективности функционирования корпоративных сетей связи от показателей качества обслуживания (QoS) // Технологии информационного общества: Сборник трудов XIV Международной отраслевой научно-технической конференции, Москва, 18-19 марта 2020 года. М.: Издательский дом Медиа паблишер, 2020. С. 302-304.

25. *Докучаев В. А., Шведов А. В*. Классификация показателей надежности корпоративных цифровых платформ // Актуальные проблемы и перспективы развития экономики : труды XIX Всероссийской с международным участием научнопрактической конференции, Симферополь-Гурзуф, 15-17 октября 2020 года. Симферополь: ИП Зуева Т. В., 2020. С. 28-29.

26. *Shvedov A. V., Nazarov M. J.* Methods for Improving the Efficiency of Information and Communication Networks // 2020 International Conference on Engineering Management of Communication and Technology (EMCTECH), 2020, pp. 1-5, doi:10.1109/EMCTECH49634.2020.9261563.

27. *Докучаев В.А., Маклачкова В.В., Статьев В.Ю.* Цифровизация субъекта персональных данных // T-Comm: Телекоммуникации и транспорт. 2020. Т. 14. № 6. С. 27-32.

28. *Pavlov S.V., Dokuchaev V.A., Maklachkova V.V., Mytenkov S.S.* Features of supporting decision making in modern enterprise infocommunication systems // T-Comm: Телекоммуникации и транспорт. 2019. Т. 13. № 3. С. 71-74.