ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

**Факультет Санкт-Петербургская школа
физико-математических и компьютерных наук**

Носивской Владислав Дмитриевич

ОБНАРУЖЕНИЕ СДВИГА РАСПРЕДЕЛЕНИЯ ДЛЯ ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ

Выпускная квалификационная работа - БАКАЛАВРСКАЯ РАБОТА
по направлению подготовки *01.03.02 Прикладная математика и информатика*
образовательная программа «Прикладная математика и информатика»

Рецензент
к.т.н., Principal Engineer, Huawei

_____

Платонов Алексей Владимирович

Руководитель
д.ф. - м.н., проф., департамент информатики

_____

Новиков Борис Асенович

Консультант
к.ф. - м.н., Laboratory Director, Huawei

_____

Кураленок Игорь Евгеньевич

Консультант
Руководитель службы, Яндекс.Облако

_____

Ершов Василий Алексеевич

Санкт-Петербург 2023

Машинное обучение и технологии искусственного интеллекта все активней начинают применяться для решения практических задач. При всей своей мощи конвейеры машинного обучения бывают очень хрупки из-за получения неожиданных данных. В ходе обучения моделей машинного обучения используется ограниченное число данных, формирующих домен модели. При получении данных вне домена точность предсказаний может резко падать. Основная сложность детектирования сдвига распределений в реальных данных заключается в большой размерности примеров и отсутствия их разметки. Предыдущие исследования используют выходы последнего слоя глубокой нейронной сети для оценки уверенности модели в своем предсказании. Далее оценки уверенности для конкретных примеров агрегируют и сравнивают с распределением оценок на данных внутри домена. Данный подход, во-первых, не использует данные из внутреннего представления глубокой нейронной сети, которое может содержать дополнительную информацию для поиска сдвига. Во-вторых, накладывает ограничения на исходную задачу модели — в работах рассматривают только классификаторы, как наиболее подходящие для такого анализа. Поэтому мы предлагаем новый метод анализа, учитывающий внутреннее представление модели, а также универсальный относительно архитектуры глубокой нейронной сети.

Ключевые слова: сдвиг распределения, обнаружение аномалий.

Machine learning and artificial intelligence technologies are becoming increasingly applied in practical problem-solving. Despite their power, machine learning pipelines can be very fragile due to unexpected data. During the training of machine learning models, a limited number of data, forming a model domain, are used. When data outside the domain are acquired, the accuracy of predictions can significantly decrease. The main difficulty in detecting distribution shift in real data lies in the high dimensionality of examples and their lack of labeling. Previous research has used the outputs of the last layer of a deep neural network to assess the model's confidence in its prediction. Then, the confidence values for specific examples were aggregated and compared with the distribution of confidence estimates on domain data. Firstly, this approach does not use the data from the internal representation of a deep neural network, which may contain additional information for identifying shifts. Secondly, it imposes limitations on the original task of the model – previous studies only consider classifiers as models suitable for such analysis. Therefore, we are suggesting a new analysis method that considers the internal representation of the model, and is also universal regarding the architecture of the deep neural network.

Keywords: distribution shift, out of domain detection.

Содержание

Введение

Актуальность и релевантные работы

Технологии на основе машинного обучения повсеместно вошли в нашу жизнь как части программного обеспечения сервисов. Многие привычные для современного общества технологии основаны на методах машинного обучения: в социальных сетях [32]; поисковые системы (Google, Yandex) используют машинное обучения для ранжирования результатов [14]; рекомендательные системы помогают в подборе музыки, фильмов; голосовые технологии используются для автоматизации колл-центров и для других приложений [13].

При этом системы на основе машинного обучения все еще являются частью программного обеспечения. Поэтому, как и у классической разработки ПО, методологии тестирования является важной компонентой. Стандарты тестирования новых версий ПО хорошо изучены и задокументированы. На сегодняшний день систематические подходы к тестированию и оценке качества систем на основе машинного обучения развиты не так хорошо и являются активной областью исследований. Исследователи строят метрики оценки качества в таких областях, как распознавание речи [20], машинный перевод [5] и других. Изучаются методы поиска единичных аномалий в данных [9], а также методы по выявлению сдвигов в распределениях данных [26].

Современные нейронные сети являются широко используемым инструментов в системах с использованием машинного обучения. Однако они часто не устойчивы к небольшим колебаниям данных [23, 17]. В связи с этим, необходимо исследование новых методов выявления сдвига распределения в данных, так как это может являться критически важным фактором для надежной эксплуатации моделей машинного обучения.

Цель и задачи

Значительная часть работ в области определения аномалий и сдвигов распределений исследует поведение распределения последнего слоя нейронной сети [15, 18, 19, 28, 26, 10]. Данный подход отлично себя зарекомендовал для случая классификации изображений с помощью сверточных нейронных сетей.

В этом случае на выходе нейронной сети вектор значений имеет размерность равную числу классов, а также интерпретируемые значения — вероятности принадлежности к соответствующему классу. Однако в других случаях выход нейронной сети может иметь менее удобные свойства для обнаружения аномалий. Ключевая идея работы состоит в рассмотрении всех слоев нейронной сети с целью извлечения сигнала о сдвиге распределения. Мы предполагаем, что данный подход поможет расширить область применимости методов обнаружения сдвига на большее число архитектур нейронных сетей и предметных областей.

Данная работа ставит своей целью разработать метод обнаружения сдвигов распределения для нейронных сетей, который бы удовлетворял следующим требованиям:

- Независимость относительно архитектуры нейронной сети.
- Независимость от предметной области.
- С вычислительной точки зрения метод должен позволять тестировать на сдвиг распределения большие потоки данных.

Для достижения этой цели ставятся следующие задачи:

- Разработать метод извлечения сигнала о сдвиге распределения из внутренних выходов нейронной сети.
- Провести сравнительный анализ мощности метода для классификаторов в случае картиночных коллекций данных.
- Провести сравнительный анализ мощности метода для трансформеров.

Достигнутые результаты

- Предложен новый метод обнаружения сдвига распределения в нейронной сети, удовлетворяющий поставленным требованиям.

- Предложен универсальный метод генерации аномальной выборки на основе эталонной.
- Для классификаторов и картиночных коллекций данных достигнуты статистически равные результаты относительно классических методов при уровне значимости 0.05.
- Для трансформеров и задачи машинного перевода достигнут прирост средней мощности обнаружения сдвига на 0.24 при уровне значимости 0.05.
- Проанализировано распределение сигнала об аномальности по слоям нейронной сети для случаев сверточных нейронных сетей и кодировщиков трансформеров.

Структура работы

В главе 1 проводится обзор методов обнаружения сдвигов распределения в общем случае и применительно к моделям машинного обучения, а также приводится схема обобщения методов поиска единичных аномалий на случай сдвига распределения.

В главе 2 представлено описание разработанного метода обнаружения сдвигов распределения для нейронных сетей на основе анализа внутренних представлений модели.

В главе 3 представлено сравнение разработанного метода с существующими аналогами на задаче классификации изображений, а также на задаче машинного перевода.

В главе 4 представлены результаты дополнительных экспериментов по анализу влияния выбора аномальной выборки на мощность, а также анализ распределения сигнала об аномальности по слоям нейронной сети.

В заключении проводится анализ проделанной работы, полученных результатов, а также возможных направлений для дальнейших исследований в этой области.

Глава 1. Обзор литературы

1.1 Обнаружение сдвигов распределений случайных величин

Сдвиги в распределениях случайных величин — это давно изучаемая задача. На данный момент хорошо изучена методика проверки гипотезы о совпадении распределений для выборок из независимых одинаково распределенных случайных величин. В зависимости от допускаемых предположений или априорном знании о природе данных, обычно применяют один из классических статистических критериев.

Двух-выборочный критерий Стьюдента [33] применяется при предположении, что при изменении распределения случайный величины, изменится и среднее значение величины. У критерия есть ограничение в виде предположения, что сравниваемые случайные величины распределены нормально.

При этом в реальной жизни такое предположение не всегда может быть выполнено, поэтому в случае его отсутствия, применяется непараметрических критерий Уилкоксона [34]. Так как критерий не требует дополнительных предположений, в общем случае его мощность ниже, чем у критерия Стьюдента.

Данные тесты исследуют простые изменения в распределениях. На практике, распределение может изменяться произвольным образом, поэтому часто необходимо учитывать всю форму распределения.

Для этого чаще всего используются критерий согласия Колмогорова-Смирнова [31] для непрерывных случайных величин или критерий согласия хи-квадрат [25] для дискретных случайных величин.

Рассмотренные тесты применяются для определения сдвигов в распределениях одномерных случайных величин. На практике нам часто необходимо работать с многомерными данными. Например, данные в машинном обучении обычно кодируются в виде векторов большой размерности.

Самым простым способом обобщения одномерных критериев на многомерные данные является применение критериев к каждой компоненте независимо. Если наблюдается сдвиг хотя бы в одной из компонент, считается, что сдвиг в данных произошел. При этом важно учитывать эффекты, возникающие при проверке множественных гипотез: в уровень статистической

значимости необходимо вносить поправку. Примером часто используемой поправки является поправка Бонферрони [4].

Альтернативой может служить применение специализированных критериев для работы с многомерными распределениями. Среди таких критериев часто используемым может являться критерий Т-квадрат Хотеллинга [16]; Критерий является обобщением критерия Стьюдента. Из достоинств теста можно выделить дешевую вычислительную сложность. Недостатки теста те же, что у критерия Стьюдента — чувствительность только к сдвигам среднего значения. В связи с этим разрабатывается большое число тестов для применения в многомерных пространствах [1].Однако чаще всего такие тесты имеют высокую вычислительную сложность, что делает их неприменимыми к задачам машинного обучения, где чаще всего размерности данных слишком большие, чтобы применять такие критерии эффективно.

1.2 Сдвиги распределения в данных машинного обучения

В качестве примеров сдвигов распределений для случая изображений можно использовать различные синтетические фильтры. В работе Failing Loudly [26] предложена методология сравнения различных методов обнаружения сдвигов на наборе изменений классических картиночных коллекций данных MNIST и CIFAR-10. Во-первых, в качестве аномалий предлагается использовать наложение гауссовского шума на изображения с тремя степенями интенсивности. Во-вторых, предлагается изменение изображений с помощью поворотов и приближений также с тремя степенями интенсивности. Дополнительно предлагается тестирование на Adversarial Attack [2], с изменением баланса классов (оставляют только изображения, принадлежащие одному и тому же классу или наоборот выбрасывают из данных один из классов). Также авторы предлагают разделение сдвигов на опасные и безопасные. Деление происходит на основе изменения целевой метрики модели при подмешивании измененных данных в стандартные тестовые. Такой набор сдвигов может служить хорошим синтетическим тестов для оценки методов обнаружения сдвига.

Основная проблема исследований по разработке методов обнаружения

сдвигов распределений заключается в том, что они ограничиваются синтетическими изменениями классических наборов данных с небольшой размерностью, которые мало похожи на реальные данные. В связи с этим были выпущены две работы проекта Shifts [30, 29]. В них авторы презентуют несколько наборов данных со сдвигами распределений из реальной жизни. Например, набор данных для предсказания погоды со сдвигами по климатическому типу; набор для перевода с английского языка на русский со сдвигом из языкового корпуса новостных изданий в корпус интернет-форума, а также наборы для предсказания траектории беспилотного автомобиля, обнаружения рассеянного склероза и текущего энергопотребления торговых судов. Также авторы предоставляют обученные модели машинного обучения на этих наборах данных, кроме классических сверточных нейронных сетей присутствуют модели на основе механизма трансформеров и модели на основе метода градиентного бустинга.

1.3 Задача поиска аномалий

Широкое развитие в последнее время получили метод обнаружения единичных аномалий. Это тесно связанная с выявлением сдвигов распределений задача. Основное отличие находится в гранулярности, с которой рассматриваются данные.

В случае определения сдвига данных работа производится с выборками из некоторой генеральной совокупности. Мы предполагаем, что при сдвиге данных новые выборки начинаются вести себя статистически другим образом. При создании сервиса на основе машинного обучения часто происходит сбор дополнительных данных, после этого полезно понимать, произошло ли существенное изменение за время, прошедшее с прошлого обучения модели или нет.

В случае определения аномалии мы смотрим на единичные объекты или группы объектов внутри выборки. Постановка проблемы звучит иначе: выделяется ли данная группа объектов по сравнению со всей выборкой? Наличие таких объектов может означать, что генеральная совокупность представляет собой смесь распределений, в которой есть как типичные для модели данные, так и аномалии.

Таким образом, концепции одновременно и похожие, и независимые. Сдвиг в распределении данных может произойти без появления аномалий (например, из-за изменения пропорций классов). И также единичные аномалии не обязательно ведут за собой сдвиг распределения. В качестве примера из жизни можно представить завод, который производит некоторую деталь. Обнаружение аномалий — задача обнаружения единичных бракованных деталей, обнаружение сдвига распределения — обнаружение повышения процента брака во всей продукции завода.

Данная задача давно изучалась в контексте линейных моделей [8] и временных рядов [11]. В последнее время всё большее распространение получают глубокие нейронные сети, поэтому сейчас ведется активная разработка методов обнаружения сдвигов для таких моделей.

1.4 Методы поиска аномалий и их обобщение для сдвигов

Для решения задачи поиска аномалий часто используются методы, использующие выходы нейронной сети как источник сигнала для классификации объектов на аномальные и не аномальные. Из таких методов можно выделить MSP [15] — агрегация выходов с помощью максимума логарифмов вероятностей или метод MaxLogit [28] на основе максимума логитов. Данные методы часто используются в качестве базовых решений, так как дают хорошую точность на большом наборе тестовых сценариев, а также являются легковесными по вычислениям и непараметрическими.

Лучшее качество во многих тестах показывает метод ODIN [19]. Метод является обучаемым под набор данных с примерами аномалий, поэтому может быть нестабилен при появлении незнакомого типа аномалий. Данный недостаток предлагается преодолеть с помощью обобщения метода ODIN [10]. Однако ODIN и его обобщение для применения требуют несколько проходов по нейронной сети, включая тяжеловесный обратный проход, поэтому в практическом смысле методы полезны для точечного изучения конкретных примеров, а не для тестирования большого потока данных.

Методы поиска единичных аномалий можно простым алгоритмом обобщить на случай поиска сдвига распределений. Обычно, такие методы в качестве

индикатора аномальности выдают некоторую статистику, в случае наличия выборки, мы можем собрать статистику по каждому объекту, а дальше сравнить две выборки с помощью применения критерия согласия к двух наборам статистик. Таким способом авторы Failing Loudly [26] получили метод для обнаружения сдвигов распределений для задачи классификации. При решении этой задачи методами глубоких нейронных сетей на выходе модели получается вектор по размерности, соответствующий числу классов. Авторы предлагают подсчитывать статистику из критерия согласия Колмогорова-Смирнова [31] для каждой компоненты вектора по отдельности. Далее результаты объединяются с помощью поправки Бонферрони [4]. Таким образом получается определение метода, называемого далее BBSD (Black Box Shift Detection).

1.5 Методология оценки мощности методов обнаружения сдвига

Основная сложность оценки точности выявления сдвигов распределения является зависимость от нескольких параметров: размера выборки и доли аномальных примеров в выборке. В статье [26] предложена следующая методология сравнения: на каждом типе сдвига по отдельности оценивается мощность для следующих размеров выборки: $\{10, 20, 50, 100, 200, 500, 1000, 10000\}$.

Для фиксированного размера выборки мощность усредняется по трем долям аномальных примеров в выборке: $\{0.1, 0.5, 1.0\}$. При фиксации двух параметров мощность оценивается по нескольким случайным выборкам. Количество случайных выборок поровну распределяется между всеми участвующими коллекциями данных: в статье используются две коллекции CIFAR-10 и MNIST.

Дополнительно следует отметить, что авторы статьи разделили сдвиги на вредоносные и безвредные на основе влияния на точность предсказаний исходной модели. Вредоносные сдвиги выделяются при сравнении красным, а безвредные выделяются зеленым.

Такой метод позволяет более точно определить слабые и сильные стороны метода, а также применимость в разных ситуациях. При использовании в реальных условиях может иметь большой смысл использование разных методов, в зависимости от размера выборки. Например, метод может давать лучшее

качество на маленьких выборках, но быть неприменимым на больших выборках из-за высокой вычислительной сложности.

В итоге получается следующая таблица:

Таблица 1: Пример сравнения методов Univariate BBSDs и Multivariate UAE

Test	Shift	Number of samples from test							
		10	20	50	100	200	500	1,000	10,000
Univariate BBSDs	**s_gn**	0.00	0.00	0.03	0.03	0.07	0.10	0.10	0.10
	m_gn	0.00	0.00	0.10	0.13	0.13	0.13	0.23	0.37
	l_gn	0.17	0.27	0.53	0.63	0.67	0.83	0.87	1.00
	s_img	0.00	0.00	0.23	0.30	0.40	0.63	0.70	0.93
	m_img	0.30	0.37	0.60	0.67	0.70	0.80	0.90	1.00
	l_img	0.30	0.50	0.70	0.70	0.77	0.87	0.97	1.00
	adv	0.13	0.27	0.40	0.43	0.53	0.77	0.83	0.90
	ko	0.00	0.00	0.07	0.07	0.07	0.33	0.40	0.70
	m_img+ko	0.13	0.40	0.87	0.93	0.90	1.00	1.00	1.00
	oz+m_img	0.67	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Multivariate UAE	s_gn	0.03	0.03	0.03	0.03	0.03	0.07	0.07	–
	m_gn	0.03	0.03	0.03	0.03	0.17	0.27	0.30	–
	l_gn	0.50	0.57	0.67	0.70	0.80	0.90	1.00	–
	s_img	0.17	0.20	0.27	0.30	0.40	0.47	0.63	–
	m_img	0.23	0.33	0.37	0.40	0.47	0.60	0.70	–
	l_img	0.30	0.30	0.37	0.47	0.60	0.77	0.87	–
	adv	0.03	0.20	0.27	0.27	0.33	0.40	0.40	–
	ko	0.10	0.13	0.13	0.13	0.17	0.17	0.30	–
	m_img+ko	0.20	0.30	0.37	0.53	0.54	0.63	0.87	–
	oz+m_img	0.27	0.63	0.77	1.00	1.00	1.00	1.00	–

1.6 Выводы

- Обнаружение сдвигов распределения это давно изучаемая задача для одномерных и многомерных случайных величин, а также для линейных моделей и временных рядов. В контексте глубоких нейронных сетей разработка методов решения этой задачи ведется особенно активно в течение последних нескольких лет.

- При сравнении методов обнаружения сдвигов для глубоких нейронных сетей важно использовать не только синтетические сдвиги на данных небольшой размерности, а также по возможности использовать собранные исследователями коллекции сдвигов из области практического применения методов машинного обучения.
- Методы поиска единичных аномалий могут применяться для задачи определения сдвигов распределения через обобщение с помощью критерия согласия двух распределений.

Глава 2. Метод

2.1 Описание метода

Ключевая идея нового метода заключается в рассмотрении не только последнего слоя нейронной сети, как это предлагается в других работах, а также включение всех слоев или части слоев нейронной сети в рассмотрение. Мы предполагаем, что сигнал об аномальности можно извлечь не только из последнего слоя. Также такой подход может помочь достичь универсальности относительно архитектуры сети, так как концентрация на последнем слое вынуждает рассматривать в основном классификаторы.

При извлечении сигнала из выходов нейронной сети можно рассматривать либо все выходы как единое целое, либо как независимые датчики. В первом случае необходимо изучать поведение многомерного распределения выходов. Основная проблема заключается в том, что статистические критерии в случае многомерных данных либо обладают большой точностью, но вычислительно очень тяжеловесные [1], либо чувствительны только к конкретным видам сдвигов [16]. Поэтому в нашей работе мы будем придерживаться второго подхода и рассмотрим все выходы сети как независимые датчики, которые будут сигнализировать о появлении сдвига распределения. Так как у современных нейронных сетей число выходов оценивается как минимум десятками миллионов, нам необходимо семплировать только часть из них. Далее обучим классификатор, который обобщит показания всех единичных датчиков.

Для обучения классификатора необходимо иметь негативные примеры — выборки со сдвинутым распределением. Для удовлетворения требованию универсальности относительно предметной области необходимо избавить пользователя от необходимости самостоятельно собирать аномальную выборку. Поэтому метод должен иметь возможность автоматической генерации аномальных данных по предоставленной эталонной выборке. Стоит отметить, что предоставление собственной аномальной выборки может позитивно повлиять на мощность классификатора, поэтому в случае практического применения может иметь смысл дополнительное исследование возможных аномалий и

формирование аномальной выборки, исходя из особенностей данных, которые могут встретиться модели.

Итоговая схема подготовки метода получается следующей:

- Из всех выходов модели семплируется подмножество.
- Каждый выход модели преобразуется в единичный датчик обнаружения сдвига.
- Эталонная выборка преобразуется в аномальную.
- С помощью эталонной и аномальной выборки обучается классификатор сдвига распределения.

Далее во время инференса нейронной сети каждый датчик передает в классификатор собственные показания и на основе информации со всех датчиков выносится вердикт: имеет ли место сейчас сдвиг распределения.

2.2 Преобразование эталонной выборки в аномальную

В качестве общей идеи, применимой к любой предметной области, мы предлагаем использовать перемешивание признаков. Этот процесс будет похож на способ оценки важности признаков в алгоритме градиентного бустинга [6]. Такой подход имеет большой потенциал обобщения на любые домены, а также имеет важное свойство: распределение перемешанного признака не изменяется. Это позволит в процессе обучения классификатора отдавать большую важность тем единичным датчикам, которые реагируют на изменение общей картины, а не конкретного признака. Приведем несколько примеров работы такой схемы.

Во-первых, рассмотрим случай картиночных коллекций данных. Это классическая предметная область при изучении обнаружения аномалий. Ключевой идеей будет перемешивание квадратных областей изображения. Алгоритм следующий:

- Выбираем случайную картинку из коллекции данных.

- Случайно генерируем значение стороны квадрата.
- Выбираем случайную позицию квадрата выбранного размера на картинке.
- Выбираем вторую случайную картинку из коллекции данных.
- Меняем соответствующий квадрат первой картинки на квадрат из второй картинки.

Каждая полученная с помощью этого алгоритма картинка помещается в аномальную выборку.

Во-вторых, рассмотрим задачу машинного перевода. В этой предметной области также обнаружение аномалий играет важную роль [35]. Для нас особенный интерес область представляет из-за широкого распротранения моделей на основе механизма трансформера в ней. Также коллекция данных вместе со сдвинутым распределением представлена в статье. В качестве предмета перемешивания мы предлагаем использовать токены в предложениях из коллекции данных. Получается следующий алгоритм:

- В предложении выбирается подмножество токенов с заданной вероятностью P
- Выбранные токены случайно перемешиваются

Параметр вероятности P позволяет регулировать степень внесения аномального шума в данные. Например, для коллекции данных машинного перевода Shifts [30] при выборе $P = 0.15$ падение метрики BLEU [5] сравнимо с падением метрики при использовании предложенной авторами коллекции со сдвигом распределения.

Таким образом, мы предлагаем базовую идею, на основе которой можно строить различные алгоритмы генерации аномальной выборки на основе эталонной, а также обобщать схему на новые предметные области.

2.3 Выходы нейронной сети как единичные датчики

Изменение в распределении каждого выхода нейронной сети может нести информацию о появлении сдвига распределения в потоке данных. Рассмотрим случай сверточной нейронной сети и картиночной коллекции данных. В этом случае при прохождении одной картинки через слои нейронной сети, каждый выход генерирует одно значение. Будем следить за изменением распределения каждого выхода относительно распределения на этом выходе при прохождении эталонной выборки. В качестве критерия согласия для двух распределений выберем критерий Колмогорова-Смирнова [31]. Выбор обусловлен сравнением критериев Колмогорова-Смирнова и Хи-Квадрат [25] в статье [26], которое показало, что лучшее качество достигается при использовании первого критерия. Таким образом, мы получили обобщенный случай метода BBSD, основное отличие в том, что в данном методе используется только последний слой, а мы предлагаем расширить число рассматриваемых выходов.

Заметим, что выбор критерия Колмогорова-Смирнова сужает число доступных выходов нейронной сети. Это связано со случаями, когда распределение выхода нейронной сети не является непрерывным. Например, классическая функция активации ReLU имеет большую концентрацию значений в нуле. Выбрать только подходящие выходы можно с помощью следующей процедуры:

- Посчитаем распределение выхода на эталонной выборке
- Сгенерируем несколько выборок из эталонного распределения
- Рассмотрим P-Значения критерия Колмогорова-Смирнова на этих выборках для выхода
- Проверим распределение этих P-Значений на равномерность с помощью еще одного критерия Колмогорова-Смирнова
- В случае выполнения гипотезы равномерности — допускаем датчик

Здесь мы пользуемся свойством P-Значения — эта случайная величина имеет равномерное распределение при выполнении нулевой гипотезы и всех

остальных предположений [22]. Мы искусственно выполняем нулевую гипотезу и проверяем выполнение остальных предположений — для теста Колмогорова-Смирнова это непрерывность распределения.

Выше мы предполагаем, что один пример из коллекции данных соответствует одному значению на выходе нейронной сети. Это предположение неверно, например, в случае решения задачи Sequence-to-Sequence, которая возникает в машинном переводе. В этом случае один пример представляет собой несколько токенов, каждый из которых проходит через нейронную сеть. Таким образом, один пример соответствует некоторому множеству значений на одном выходе нейронной сети. Для решения этой проблемы перейдем от использования чистого выхода нейронной сети к использованию P-Значению принадлежности данного значения выхода к распределению значений этого выхода на эталонной выборке. Этот переход позволит агрегировать P-значения для всех токенов с помощью геометрического среднего, получая таким образом одно значение на выходе нейронной сети для одного примера. В качестве функции агрегации было выбрано среднее геометрическое, так как тогда при появлении нескольких редких токенов в одном предложении, это сильно отразится на итоговом значении, даже если остальные токены встречаются очень часто.

Таким образом, после семплирования случайных выходов из нейронной сети, каждый рассматривается как единичный датчик на основе критерия согласия. Для каждой выборки картинок один датчик генерирует значение, которое может быть использовано для построения единого классификатора.

2.4 Агрегация датчиков

В статье [26] агрегация нескольких P-значений от критериев согласия происходит с помощью поправки на множественную проверку гипотез. Авторами была использована поправка Бонферрони. Данный подход является крайне простым для реализации и использования, а также не требует обучения, однако основная проблема метода — при увеличении числа гипотез мощность процедуры сильно уменьшится. Этот эффект не достигается на коллекциях данных CIFAR-10 и MNIST, которые используются в статье, так как число

гипотез в метода равняется числу классов, а в обеих коллекциях представлены 10 классов. В случае, например, использования коллекции ImageNet-1k, число классов будет равняться уже 1000, что может начать негативно влиять на мощность. Мы же намерены использовать потенциально десятки тысяч выходов нейронной сети в качестве датчиков, поэтому этот метод агрегации нам не подходит.

В качестве метода агрегации мы предлагаем использовать метод градиентного бустинга. Данный подход обеспечит нам выполнение требования к производительности нашего решения, так как современные библиотеки позволяют выстроить применение моделей градиентного бустинга на видеокартах, существенно ускорив вычисления. Также данный подход позволит уменьшить число необходимых для работы метода выходов, так как после обучения модели можно воспользоваться механизмом определения важности признаков и оставить только определенное число самых важных выходов. В качестве библиотеки для реализации алгоритма градиентного бустинга была выбрана библиотека CatBoost [7].

2.5 Выводы

- Предложен метод генерации аномальной выборке на основе эталонной, универсальный относительно предметной области, а также приложены два примера работы метода для случая изображений и задачи машинного перевода.
- Разработана методология рассмотрения внутреннего выхода нейронной сети как единичного датчика обнаружения сдвигов распределения.
- Предложен метод агрегации информации с единичных датчиков для создания единого классификатора сдвигов распределения.

Глава 3. Эксперименты

3.1 Описание схемы экспериментов. Failing Loudly

Необходимо сравнить метод с BBSD, в качестве схемы сравнения выбрана соответствующая схема из статьи [26]. Сравнение происходит с применением следующих картиночных коллекций данных для решения задачи классификации:

- **CIFAR-10:** Содержит 10 классов. Изображения цветные размера 32×32 пикселя. Данный датасет широко используется для сравнения методов обаружения аномалий, однако из-за маленького размера картинок и небольшого числа классов не позволяет делать выводы о применинении методов на реальных данных.

- **MNIST:** Коллекция для распознавания рукописных цифр. Изображения черно-белые размера 20×20 пикселя. Коллекция также характеризуется простотой для современных сверточных нейронный сетей, уровень ошибки ниже 0.3%.

Данный набор коллекций с одной стороны не позволяет обобщать результаты на случаи практического применения, с другой стороны большая часть профильных статей используют такой набор для сравнения. В статье [26] предлагается накладывать на изображения следующий набор аномалий:

- **Adversarial (*adv*)**: Примеры с состязательной атакой с помощью FGSM [12].

- **Knock-out (*ko*)**: Изображения без нулевого класса (дисбаланс классов).

- **Gaussian noise (*gn*)**: Нанесение гауссовского шума со стандартным отклонением $\delta \in \{1, 10, 100\}$ (обозначаются как *s_gn*, *m_gn* и *l_gn*).

- **Image (*img*)**: Нанесение комбинации случайных поворотов, отражений, приближений с тремя степенями интенсивности (обозначаются как *s_img*, *m_img* и *l_img*).

- **Image + knock-out (*m_img+ko*)**: Нанесение *m_img* и дополнительно отбрасывание нулевого класса.

- **Only-zero + image (*oz+m_img*)**: Нанесение *m_img* только и отбрасывание всех классов, кроме нулевого.

Для вычисления мощности метода для фиксированного сдвига и размера выборки используется 30 запусков. По 10 запусков с разными долями аномальных примеров в выборке: $\{0.1, 0.5, 1.0\}$. В качестве нейронной сети использовалась обученная авторами статьи сверточная нейронная сеть ResNet-18. Уровень значимости во всех запусках $\alpha = 0.5$. Размеры выборок $s \in \{10, 20, 50, 100, 200, 500, 1000, 10000\}$.

Сеть ResNet-18 содержит примерно 130 тысяч выходов для CIFAR-10 и примерно 110 тысяч выходов для MNIST. Все выходы были использованы для обучения классификатора.

3.2 Эксперимент. Failing Loudly

В первую очередь получим результаты метода BBSD, который получил лучшие результаты при сравнении разных методов в статье [26]. Результаты можно видеть на таблице 2.

Таблица 2: Результаты метода BBSD

Test	Shift	Number of samples from test							
		10	20	50	100	200	500	1,000	10,000
BBSD	**s_gn**	0.02	0.03	0.05	0.05	0.02	0.08	0.05	0.12
	m_gn	0.07	0.02	0.03	0.07	0.02	0.12	0.20	0.55
	l_gn	0.30	0.58	0.65	0.70	0.78	0.80	0.90	1.00
	s_img	0.10	0.17	0.42	0.48	0.52	0.60	0.67	0.88
	m_img	0.42	0.45	0.68	0.70	0.77	0.87	0.90	1.00
	l_img	0.53	0.65	0.65	0.75	0.85	0.92	0.95	1.00
	adv	0.32	0.45	0.53	0.57	0.73	0.82	0.85	0.93
	ko	0.02	0.05	0.02	0.03	0.12	0.38	0.45	0.70
	m_img+ko	0.28	0.30	0.50	0.58	0.70	0.68	0.87	0.97
	oz+m_img	0.45	0.63	0.70	0.72	0.78	0.92	0.98	1.00

Стоит отметить, что оригинальный метод хорошо справляется практически со всеми сдвигами — мощность меньше 0.5 при размере выборки 10000 достигается только на гауссовском шуме низкой интенсивности (*s_gn*). Основные проблемы у метода возникают именно с определением гауссовского шума низкой и средней интенсивности, а также со свдигом *ko*, который эмулирует дисбаланс классов.

Средняя мощность по таблице для BBSD метода — 0.51.

Сравним результаты метода BBSD с нашим. Для этого проведем такие же эксперименты и отразим на таблице 3 разницу между мощностью нашего метода и метода BBSD. Отметим зеленым цветом улучшения в мощности превышающие порог значимости, а красным цветом ухудшения в мощности превышающие порог значимости.

Таблица 3: Сравнение метода BBSD с предложенным методом

Test	Shift	Number of samples from test							
		10	20	50	100	200	500	1,000	10,000
BBSD vs. Ours	**s_gn**	+0.08	-0.02	0.00	0.00	0.00	-0.02	0.02	0.05
	m_gn	0.03	0.02	+0.10	0.02	+0.07	+0.13	+0.13	+0.07
	l_gn	+0.25	+0.08	+0.10	0.05	0.05	+0.10	+0.08	0.00
	s_img	+0.07	0.03	0.00	0.03	-0.05	-0.05	0.02	0.03
	m_img	-0.02	0.05	0.05	-0.03	0.02	-0.02	-0.03	0.00
	l_img	-0.03	-0.02	0.05	0.05	0.00	0.00	0.03	0.00
	adv	-0.02	-0.07	-0.08	-0.03	-0.08	-0.02	-0.02	-0.07
	ko	-0.02	0.00	+0.10	0.03	0.05	-0.08	-0.10	-0.05
	m_img+ko	0.02	+0.07	-0.15	-0.03	-0.07	+0.08	-0.05	-0.05
	oz+m_img	-0.05	-0.05	-0.10	0.02	-0.02	-0.08	-0.10	0.00

Заметим, что наш метод немного улучшает качество на гауссовском шуме (*gn*): разница в среднем на $+0.06$. В среднем разница в мощности составляет $+0.007$, что позволяет говорить о статистическом равенстве результатов нашего метода и метода BBSD при сравнении на классификаторах изображений

— главном поле исследований методов обнаружения сдвигов распределения в контексте нейронных сетей.

3.3 Описание схемы экспериментов. Shifts

В качестве архитектуры нейронной сети, отличной от сверточной, мы выбрали архитектуру на основе механизма трансформеров, так как современные текстовые [24], голосовые [27] и другие модели основаны на этом механизме. Для сравнения с другими методами мы взяли методологию из статьи [26], в которой представлена обученная нейронная сеть для перевода в языковой паре Английский-Русский. В статье используются следующие коллекции данных:

- Для обучения нейронной сети был использован свободно распространяемый корпус текстов WMT‘20 En-Ru. Эта коллекция сконцентрирована на новостных и государственных данных, в них использован формальный язык и соблюдены орфографические и грамматические нормы языков.
- В качестве тестовых данных без сдвига распределения используется корпус новостных данных Newstest‘19 En-Ru.
- В качестве тестовых данных со сдвигом используется корпус текстов с интернет-форума Reddit, подготовленный для WMT‘19 Robustness Challenge [21].

В качестве нейронной сети использовалась обученная авторами сеть Big-Tranfsormer [3]. Данная архитектура состоит кодировщика и декодировщика. В своих экспериментах мы будем исследовать выходы только кодировщика. Такой выбор был сделан из-за особенностей процесса декодирования: для декодирования одного предложения используется лучевой поиск, в процессе которого декодировщик применяется множество раз к токенам предложения и некоторые ветви отсекаются. В связи с этим необходимы дальнейшие исследования по разработке метода извлечения сигнала об аномальности с учетом особенностей лучевого поиска, а также влияния ширины луча на итоговое качество предсказаний сдвигов распределения. Кодировщик имеет более классическое устройство: каждый токен проходит через слои один раз. В

отличие от случая картинок, в текстах одно предложение содержит несколько токенов, каждый из которых соответствует множеству выходов нейронной сети. В параграфе 2.3 описан способ работы с данными подобной природы.

В качестве базового решения для сравнения мы предлагаем использовать обобщение на выборки метода поиска аномалий из статьи [26]. Схема работы метода следующая:

- При декодировании предложения используется лучевой поиск с шириной 5
- В результаты работы поиска получается 5 гипотез с назначенными им вероятностями
- В качестве статистики аномальности по приему используется арифметическое среднее полученных вероятностей
- На основе распределения статистики строится бинарный классификатор

Для обобщения метода на случай выборок воспользуемся схемой из параграфе 1.4. В итоге метод получается схожим с частным случаем BBSD, когда на последнем слое один выход.

В качестве выходов нейронной сети для нашего метода были выбраны все выходы с финальных слоев каждого блока кодировщика, а также все выходы с финальных слоев каждого механизма внимания с каждого блока кодировщика. Итоговое число выходов, используемых при обучении метода — 12288.

3.4 Эксперимент. Shifts

Оценка мощности работы полученного базового решения в задаче машинного перевода приведена на таблице 4.

Отметим относительно высокую мощность базового решения — мощность впервые становится выше 0.5 уже на размере выборки 50. Данный факт может говорить о том, что предложенный сдвиг распределения является очень

Таблица 4: Результаты обобщенного метода OOD

Shift	Number of samples from test							
	10	20	50	100	200	500	1,000	10,000
Reddit	0.13	0.23	0.53	0.60	0.67	0.70	0.73	1.00

явным для нейронной сети и легко приводит к смещению распределения вероятности предсказания.

Рассмотрим теперь результаты предложенного метода на таблице 5.

Таблица 5: Результаты предложенного метода

Shift	Number of samples from test							
	10	20	50	100	200	500	1,000	10,000
Reddit	0.57	0.60	0.73	0.77	0.87	0.97	1.00	1.00

Сразу отметим существенное улучшение результатов — мощность выше 0.5 достигается на минимальной размере выборки 10. Более наглядно рассмотрим разницу в мощности на таблице 6.

Таблица 6: Разница в мощности между предложенным методом и обобщением OOD

Shift	Number of samples from test							
	10	20	50	100	200	500	1,000	10,000
Reddit	+0.43	+0.37	+0.20	+0.17	+0.20	+0.27	+0.27	0.00

Видим существенное улучшение результатов относительно базового решения. В среднем улучшение мощности составило 0.24. Данный результат подтверждает гипотезу об универсальности предложенного подхода, а также позволяет говорить о существенном улучшении результатов, относительно обобщения классического OOD подхода в случае трансформеров в решении задачи машинного перевода.

3.5 Выводы

- Предложенный метод показывает статистически равные результаты относительно классических методов обнаружения сдвигов в случае классификации изображений: средняя разница мощности составляет $+0.05$ при уровне значимости 0.05.
- Предложенный метод существенно улучшает результаты обобщения OOD в случае трансформеров: средняя разница в мощности составляет $+0.24$ при уровне значимости 0.05.

Глава 4. Дополнительные эксперименты

4.1 Влияние выбора аномальной выборки на мощность метода

В основной части экспериментов в качестве аномальной выборки использовалась выборка, полученная с помощью метода преобразования эталонных примеров в аномальные, описанная в параграфе 2.2. Такая схема выбрана для доказательства универсальности подхода относительно предметных областей, а также для упрощения практического использования метода — исследователям не требуется собирать дополнительные данные. Однако в случае потенциального применения подхода в реальной жизни команды исследователей могут улучшать мощность обнаружения сдвигов с помощью сбора собственной выборки аномальных данных. Так как при эксплуатации моделей машинного обучения чаще всего происходит непрерывная разметка дополнительных данных, выборка может формироваться специально с помощью тех же инструментов разметки.

Для исследования влияния аномальной выборки на мощность метода рассмотрим схему экспериментов со статьей [26], описанную в параграфе 3.1. В данной схеме есть 10 типов сдвигов. Предположим, что исследователи могут угадать один из сдвигов, который может произойти при эксплуатации модели. В качестве двух сдвигов для рассмотрения выберем гауссовский шум (*gn*) и комбинацию поворотов, отражений, приближений (*img*).

Сравним предложенный метод с BBSD в случае, если для обучения классификатора используется наложение гауссовского шума, а не подход с перемешиванием признаков. Результаты можно видеть на таблице 7

Таблица 7: Результаты сравнения методов при обучении на сдвиге *gn*

Test	Shift	Number of samples from test							
		10	20	50	100	200	500	1,000	10,000
BBSD vs. Ours (gn)	**s_gn**	+0.47	+0.42	+0.48	+0.55	+0.57	+0.52	+0.55	+0.72
	m_gn	+0.42	+0.40	+0.47	+0.43	+0.57	+0.43	+0.42	+0.25
	l_gn	+0.32	+0.13	+0.15	+0.15	+0.08	+0.07	0.05	0.00
	s_img	+0.07	-0.07	-0.25	-0.35	-0.42	-0.28	-0.25	-0.18
	m_img	-0.02	-0.07	-0.15	-0.13	-0.08	-0.10	-0.10	-0.08
	l_img	-0.02	-0.12	0.02	0.02	-0.02	-0.02	-0.03	-0.02
	adv	+0.22	0.02	-0.03	-0.02	-0.22	-0.13	-0.17	-0.08
	ko	+0.15	0.05	0.00	0.03	-0.02	-0.22	-0.22	-0.20
	m_img+ko	-0.02	-0.05	-0.12	-0.20	-0.22	-0.03	-0.12	-0.10
	oz+m_img	0.02	-0.10	-0.10	0.02	-0.05	-0.02	-0.03	0.00

Сразу можно заметить существенный рост мощности при обнаружении гауссовского шума, что напрямую исходит из выбора аномальной выборки для обучения классификатора. Средняя разница в мощности для гауссовского шума составила $+0.36$. При этом мощность обнаружения остальных сдвигов настолько же существенно снизилась. Средняя разница в мощности на всех методах кроме гауссовского шума составила -0.08. В итоге средняя разница в мощности составила $+0.05$ при уровне значимости 0.05.

Данные результаты могут свидетельствовать о том, что выходы нейронной сети, особенно остро реагирующие на появление гауссовского шума являются очень специфичными для данного типа сдвига. Также это говорит о необходимости проверять выбор аномальной выборки на разных видах сдвигов при использовании в реальной жизни, так как существует риск составления специфичной выборки, ведущей к переобучению классификатора под конкретный сдвиг.

При этом если рассмотреть в качестве аномальной выборки комбинацию поворотов, отражений, приближений, то результаты получатся иными, подробнее на таблице 8.

Таблица 8: Результаты сравнения методов при обучении на сдвиге *img*

Test	Shift	Number of samples from test							
		10	20	50	100	200	500	1,000	10,000
BBSD vs. Ours (img)	**s_gn**	+0.12	0.05	+0.10	+0.20	+0.28	+0.28	+0.37	+0.38
	m_gn	+0.12	+0.32	+0.35	+0.35	+0.35	+0.33	+0.32	+0.17
	l_gn	+0.25	+0.20	+0.12	+0.08	+0.08	+0.10	+0.07	0.00
	s_img	+0.18	+0.30	+0.10	+0.12	+0.17	+0.13	+0.17	+0.12
	m_img	0.03	+0.15	0.03	+0.07	0.02	0.03	0.02	0.00
	l_img	0.03	+0.07	+0.07	+0.07	+0.10	0.02	0.05	0.00
	adv	+0.08	0.02	0.02	0.02	0.00	0.03	0.00	0.05
	ko	+0.08	0.02	+0.07	0.02	0.02	-0.12	-0.12	-0.02
	m_img+ko	-0.05	0.03	-0.03	-0.03	0.00	0.03	-0.05	0.00
	oz+m_img	0.02	-0.02	0.00	0.03	0.05	0.00	0.00	0.00

Видно, что использование данного типа аномальной выборки практически равномерно улучшает результаты. Единственное слабое место – сдвиг *adv*. Средняя разница в мощности составила $+0.09$ при уровне значимости 0.05.

Заметим, что результаты на обнаружении *gn* при обучении на *img* изменились в среднем на $+0.2$, а результаты при обнаружении *img* при обучении *gn* в среднем упали на -0.11. Данный эффект свидетельствует о несимметричности переноса мощности обнаружения с одного сдвига на другой и потенциально требует дополнительного исследования. В частности, представляет интерес выделение свойств выходов нейронной сети, которые позволяют улучшать определение и *gn*, и *img*, а также их отличия от выходов, которые существенно улучшают определение *gn* и при этом ухудшают определение *img*.

4.2 Анализ распределения сигнала по слоям нейронной сети

Так как мы предложили метод, анализирующий внутренние слои нейронной сети, представляет интерес исследование распределения сигнала об аномальности по слоям. В литературе используются методы, использующие информацию с последнего слоя нейронной сети, предполагая, что именно в нем содержится основная часть сигнала. Проведем анализ основываясь на схеме из параграфа 3.1.

Для исследования распределения сигнала построим тепловые карты на основе рассчета значимости признаков для градиентного бустинга. Далее сгруппируем признаки по блокам нейронной сети ResNet-18, отдельно вынесем последний слой. При использовании для получения аномальной выборки схему из параграфа 2.2, получается тепловая карта, представленная на рисунке 1.

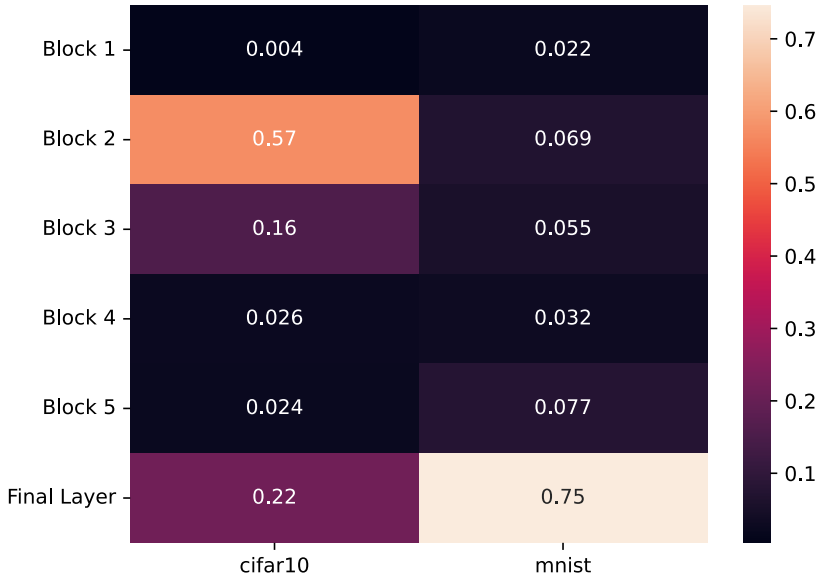

Рис. 1: Тепловая карта распространения сигнала

Обратим внимание на различия между двумя коллекциями данных – у CIFAR-10 большая часть сигнала сконцентрирована на втором блоке, у MNIST большая часть сигнала сконцентрирована на последнем слое. Этот эффект проводит различия между этими двумя коллекциями данных, которые авторами статьи [26] отмечаются как похожие, поэтому происходит усреднение мощности по этим коллекциям. Требуются дополнительные исследования для объяснения высокой значимости второго блока для анализа сдвигов на CIFAR-10.

Тепловая карта для случая обучения на гауссовском сдвиге представлена на рисунке 2.

Отметим снова значительные отличия между датасетами. Для CIFAR-10 картина осталась близкой к прошлому случаю, однако значимость последнего слоя перенеслась на 3-5 блоки, общая картина с доминированием второго

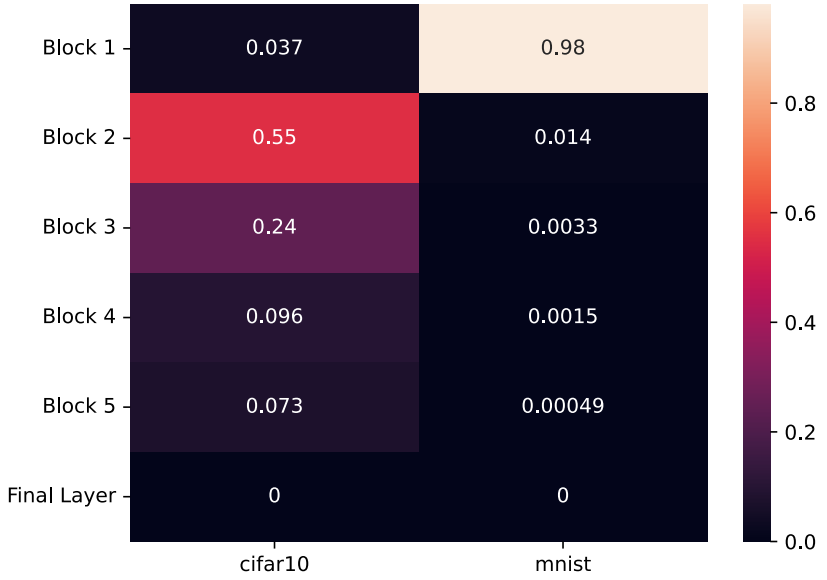

Рис. 2: Тепловая карта распространения сигнала (обучение на *gn*)

блока не изменилась. Однако для MNIST картина целиком перевернулась — практически весь вес отдан первому слою. Мы предполагаем, что это связано с особенностью изображений в MNIST — это цифры на черном фоне, наложение гауссовского шума, в отличие от остальных типов сдвигов, взаимодействует напрямую на значение пикселя в точке. Поэтому первый блок, в котором происходит первичная обработка пикселей начинает доминировать. Для случая CIFAR-10 такого изменения на пикселях недостаточно для доминирования первого блока — изображения имеют более сложную структуру.

Тепловая карта для случая обучения на поворотах, отражениях, приближениях представлена на рисунке 3.

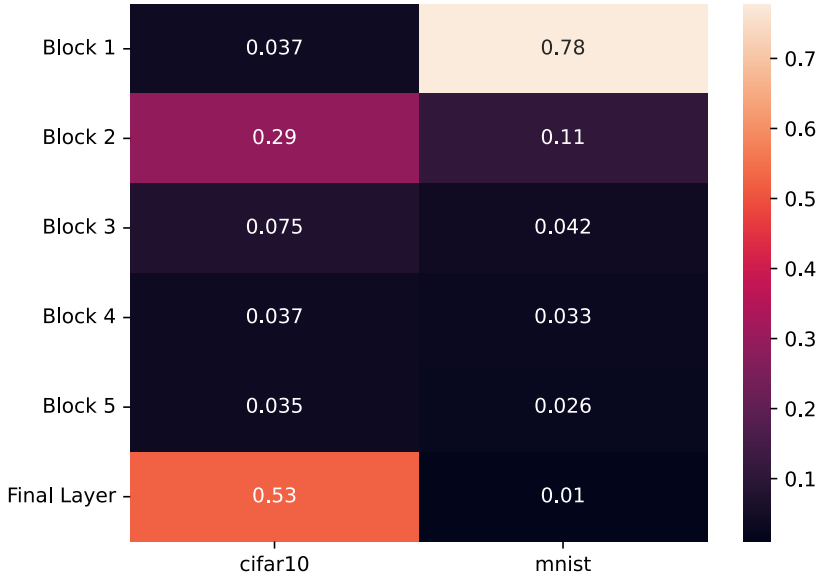

Рис. 3: Тепловая карта распространения сигнала (обучение на *img*)

Обучение на *img* впервые привело к доминированию последнего слоя в случае CIFAR-10, второй блок перешел на второе место по значимости. Для случая MNIST ситуация похожа на обучение на *gn*, снова большая часть значимости ушла на первый блок. Возможно, здесь снова играет роль структура изображений — белые цифры вписаны в прямоугольники и угловые части всегда черные, однако при поворотах, приближениях и отражениях это свойство может измениться.

Также интерес для анализа представляет схема из параграфа 3.4. В кодировщике трансформера из схемы содержится 6 блоков. Тепловую карту для кодировщика можно увидеть на рисунке 4

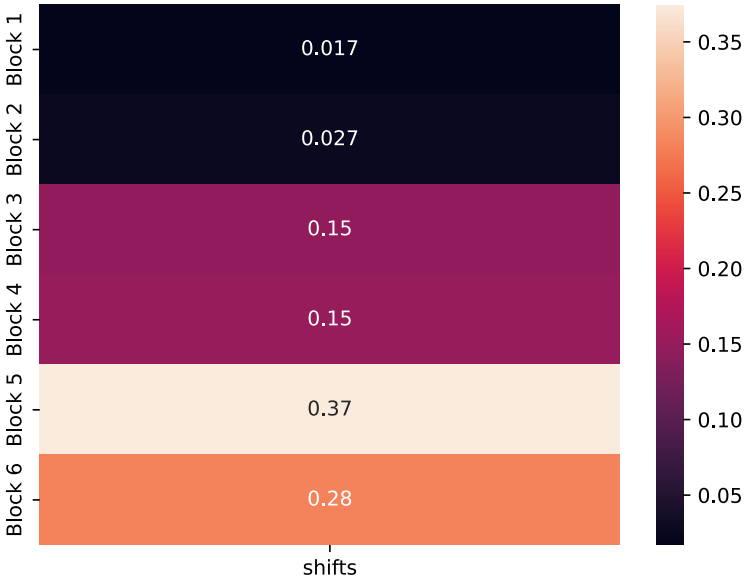

Рис. 4: Тепловая карта распространения сигнала для Shifts

Заметим, что сигнал увеличивается от начала к концу нейронной сети, но увеличение происходит не каждый блок, а примерно каждые два блока. Для обучения классификатора в этой схеме использовались выходы нейронной сети двух типов — выходы блоков механизма внимания, а также последние слои блоков трансформера. Рассмотрим тепловые карты для двух типов выходов по отдельности. Тепловая карта для выходов механизма внимания на рисунке 5, для выходов последних слоев блоков на рисунке 6

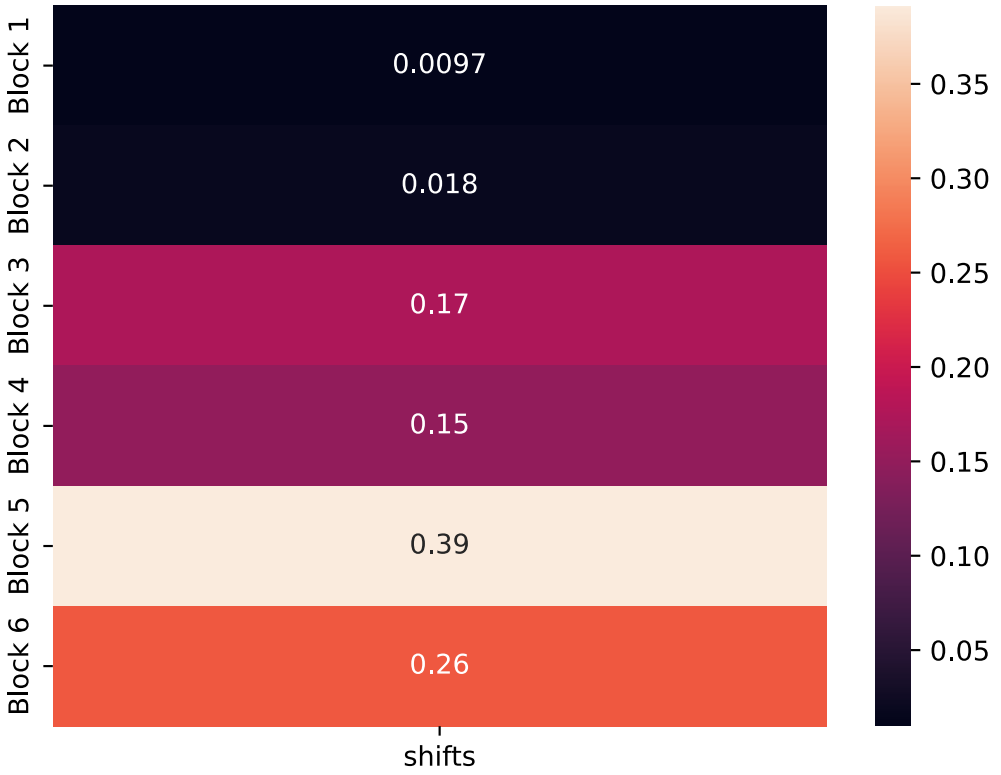

Рис. 5: Тепловая карта распространения сигнала для Shifts (механизм внимания)

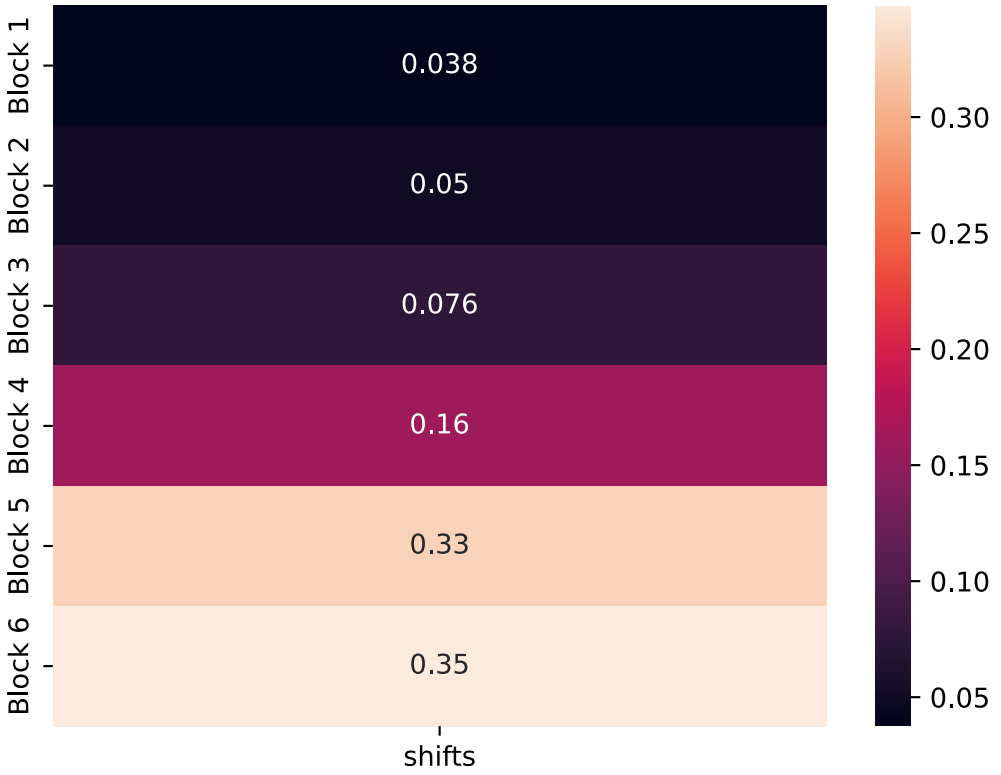

Рис. 6: Тепловая карта распространения сигнала для Shifts (последние слои блоков)

Для финальных блоков слоев заметна тенденция монотонного накопления сигнала от начала к концу сети. Для блоков механизма внимания наблюдается повышение степени содержания сигнала скорее каждые два блока сети. Соотношение между содержанием сигнала у этих двух категорий выходов можно видеть на рисунке 7

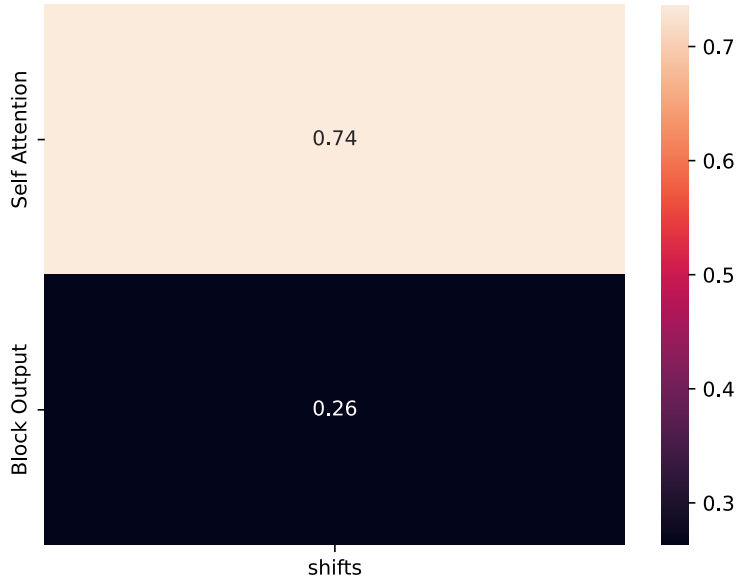

Рис. 7: Сравнение двух категорий выходов для Shifts

Видно, что выходы механизма внимания значительно доминируют и несут основную часть сигнала об аномальности.

4.3 Выводы

- Выбор аномальной выборки имеет существенное влияние на мощность полученного классификатора. При использовании в реальной жизни для повышения мощности метода требуется дополнительное исследование возможных сдвигов с учетом специфики нейронной сети и потока данных, с которой эта нейронная сеть сталкивается.

- Требуется тестирование выбранной аномальной выборки на разных типах сдвигов, так как некоторые типы выборок могут приводить к переобучению классификатора.

- Выбор аномальной выборки может улучшать результаты в схеме тестирования из параграфа 3.1 до $+0.09$ в среднем при уровне значимости в 0.05.

- Для CIFAR-10 предположение о накоплении нейронной сетью информации об аномальности от начала к концу оказывается не верным. Для

MNIST доминирует в разных случаях либо первый, либо последний слой. Данные наблюдения приводят к выводу, что коллекции данных отличаются по своим свойствам и может быть некорректно подсчитывать мощность усредняя показания на них.

- Для кодировщика трансформера большую часть сигнала несут блоки механизма внимания, а также есть тенденция к накоплению информации об аномальности от начала к концу сети.
- Возможны дополнительные исследования по извлечению сигнала об аномальности из блоков механизма внимания, учитывая специфику вычислений в данном блоке.

Заключение

Главным результатом работы является новый метод обнаружения сдвига распределения в данных к глубокой нейронной сети. Достигнута независимость метода относительной архитектуры сети и предметной области. При сравнении с другими методами достигнут либо паритет, либо улучшение результатов.

Для применения метода к любому набору данных предложен метод получения аномальной выборки по эталонной на основе перемешиваний факторов. Приведены примеры работы метода для случая изображений и текстов. Дополнительные эксперименты показывают, что выбор аномальной выборки существенно влияет на результаты. В случае практического применения метода для увеличения мощности следует собирать дополнительные наборы аномальных данных, исходя из предположений о природе аномалий, которые могут встретиться модели во время работы.

При сравнении метода с менее универсальным методом BBSD для классификаторов на датасетах CIFAR-10 и MNIST получены статистически равные результаты. Это позволяет утверждать, что метод может соперничать с методами, приспособленными к конкретной архитектуре сети или предметной области. При тестировании обучения на сдвигах из тестового множества достигается преимущество по мощности в среднем до $+0.09$ при уровне значимости 0.05.

Для подтверждения универсальнсти метода относительно архитектуры нейронной сети проведено сравнение на датасете машинного перевода из статьи Shifts. Авторский метод обобщен на выборки. При сравнении обобщенного метода с предложенным получено улучшение результатов: в среднем мощность увеличивается на $+0.24$ при уровне значимости 0.05.

Результаты экспериментов показывают, что выходы нейронной сети являются источниками сигнала для определения сдвигов распределения. Другие методы агрегации этого сигнала требуют исследования и могут обладать большей мощностью. Также получены знания о распределении сигнала относительно сети. Определены места концентрации большей части сигнала: для случая сверточных сетей и CIFAR-10 это второй блок ResNet-18, для MNIST это последний слой сети, для случая энкодера трансформера это выходы

блоков механизма внимания.

Таким образом, в рамках данной работы предложены техники извлечения информации о сдвигах распределения из выходов слоев нейронной сети. В дальнейшей перспективе необходимо исследовать другие способы агрегации, влияние выбора аномальной выборки для обучения, а также извлечение дополнительного знания о природе распространения сигнала об аномальности сквозь слои сети. Также необходимо отметить, что представляет интерес разработка библиотеки, реализующей предложенные техники, а также другие современные методы определения сдвигов данных.

Список литературы

Список литературы

1. A kernel two-sample test / A. Gretton [и др.] // Journal of Machine Learning Research. — 2012. — 1 марта. — Т. 13, № 1. — С. 723—773. — DOI: 10.5555/2188385.2188410. — URL: http://is.tuebingen.mpg.de/fileadmin/user_upload/files/publications/2012/gretton12a.pdf.
2. A survey on adversarial attacks and defences / A. Chakraborty [и др.] // CAAI Transactions on Intelligence Technology. — 2021. — 1 марта. — Т. 6, № 1. — С. 25—45. — DOI: 10.1049/cit2.12028. — URL: https://doi.org/10.1049/cit2.12028.
3. Attention is All you Need / A. Vaswani [и др.] //. Т. 30. — Cornell University, 12.06.2017. — С. 5998—6008. — URL: https://arxiv.org/pdf/1706.03762v5.
4. *Bland M.*, *Altman D. G.* Statistics notes: Multiple significance tests: the Bonferroni method // BMJ. — 1995. — 21 янв. — Т. 310, № 6973. — С. 170. — DOI: 10.1136/bmj.310.6973.170. — URL: https://doi.org/10.1136/bmj.310.6973.170.
5. BLEU / K. Papineni [и др.] //. — 06.07.2002. — DOI: 10.3115/1073083.1073135. — URL: https://doi.org/10.3115/1073083.1073135.
6. *Breiman L.* Random Forests // Machine Learning. — 2001. — 1 окт. — Т. 45, № 1. — С. 5—32. — DOI: 10.1023/a:1010933404324. — URL: https://doi.org/10.1023/a:1010933404324.
7. CatBoost: unbiased boosting with categorical features / L. O. Prokhorenkova [и др.] //. Т. 31. — 03.12.2018. — С. 6639—6649. — URL: https://papers.nips.cc/paper/7898-catboost-unbiased-boosting-with-categorical-features.pdf.

8. *Cook R. D.* Detection of Influential Observation in Linear Regression // Technometrics. — 1977. — 1 февр. — Т. 19, № 1. — С. 15. — DOI: 10.2307/1268249. — URL: https://doi.org/10.2307/1268249.

9. Deep Learning for Anomaly Detection / G. Pang [и др.] // ACM Computing Surveys. — 2021. — 5 марта. — Т. 54, № 2. — С. 1—38. — DOI: 10.1145/3439950. — URL: https://doi.org/10.1145/3439950.

10. Generalized ODIN: Detecting Out-of-Distribution Image Without Learning From Out-of-Distribution Data / Y. Hsu [и др.] //. — 14.06.2020. — DOI: 10.1109/cvpr42600.2020.01096. — URL: https://doi.org/10.1109/cvpr42600.2020.01096.

11. *Golyandina N.*, *Zhigljavsky A.* Singular Spectrum Analysis for Time Series. — 01.01.2013. — DOI: 10.1007/978-3-642-34913-3. — URL: https://doi.org/10.1007/978-3-642-34913-3.

12. *Goodfellow I.*, *Shlens J.*, *Szegedy C.* Explaining and Harnessing Adversarial Examples //. — 20.03.2015. — URL: https://ai.google/research/pubs/pub43405.

13. *Graves A.*, *Mohamed A.*, *Hinton G. E.* Speech recognition with deep recurrent neural networks //. — 26.05.2013. — DOI: 10.1109/icassp.2013.6638947. — URL: https://doi.org/10.1109/icassp.2013.6638947.

14. *Gulin A.*, *Kuralenok I.*, *Pavlov D.* Winning The Transfer Learning Track of Yahoo!'s Learning To Rank Challenge with YetiRank. — 2011. — 26 янв. — URL: http://proceedings.mlr.press/v14/gulin11a/gulin11a.pdf.

15. *Hendrycks D.*, *Gimpel K.* A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks //. — 04.11.2016. — URL: https://dblp.uni-trier.de/db/conf/iclr/iclr2017.html#HendrycksG17.

16. *Hotelling H.* The Generalization of Student's Ratio. — Springer Science+Business Media, 01.08.1931. — С. 54—65. — DOI: 10.1007/978-1-4612-0919-5_4. — URL: https://doi.org/10.1007/978-1-4612-0919-5_4.

17. Intriguing properties of neural networks / C. Szegedy [и др.] //. — 01.01.2014. — URL: http://datascienceassn.org/sites/default/files/Intriguing%20Properties%20of%20Neural%20Networks_0.pdf.

18. *Lakshminarayanan B.*, *Pritzel A.*, *Blundell C.* Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles //. Т. 30. — Cornell University, 01.01.2017. — С. 6402—6413. — URL: https://arxiv.org/pdf/1612.01474.

19. *Liang S.*, *Li Y.*, *Srikant R.* Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks //. — 01.01.2018. — URL: https://openreview.net/pdf?id=H1VGkIxRZ.

20. Meaning Error Rate / L. Gordeeva [и др.] //. — 14.08.2021. — DOI: 10.1145/3447548.3467372. — URL: https://doi.org/10.1145/3447548.3467372.

21. *Michel P.*, *Neubig G.* MTNT: A Testbed for Machine Translation of Noisy Text //. — 01.09.2018. — DOI: 10.18653/v1/d18-1050. — URL: https://doi.org/10.18653/v1/d18-1050.

22. *Murdoch D. J.*, *Tsai Y.*, *Adcock J. L.* P-Values are Random Variables // The American Statistician. — 2008. — 1 авг. — Т. 62, № 3. — С. 242—245. — DOI: 10.1198/000313008x332421. — URL: https://doi.org/10.1198/000313008x332421.

23. On calibration of modern neural networks / C. F. Guo [и др.] //. — 17.07.2017. — С. 1321—1330. — URL: http://proceedings.mlr.press/v70/guo17a/guo17a.pdf.

24. *O.* GPT-4 Technical Report // arXiv (Cornell University). — 2023. — 15 марта. — DOI: 10.48550/arxiv.2303.08774. — URL: http://arxiv.org/abs/2303.08774.

25. *Pearson K.* X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling // The London, Edinburgh and Dublin philosophical magazine and journal of science. — 1900. — 1 июля. — Т. 50, № 302. — С. 157—175. — DOI:

10.1080/14786440009463897. — URL: https://doi.org/10.1080/14786440009463897.

26. *Rabanser S.*, *Günnemann S.*, *Lipton Z. C.* Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift //. T. 32. — Cornell University, 01.01.2019. — С. 1396—1408. — URL: https://arxiv.org/pdf/1810.11953.pdf.

27. Robust Speech Recognition via Large-Scale Weak Supervision / A. Radford [и др.] // arXiv (Cornell University). — 2022. — 6 дек. — DOI: 10.48550/arxiv.2212.04356. — URL: http://arxiv.org/abs/2212.04356.

28. Scaling Out-of-Distribution Detection for Real-World Settings / D. Hendrycks [и др.] // arXiv (Cornell University). — 2019. — 25 нояб. — DOI: 10.48550/arxiv.1911.11132. — URL: http://arxiv.org/abs/1911.11132.

29. Shifts 2.0: Extending The Dataset of Real Distributional Shifts / A. Malinin [и др.] // arXiv (Cornell University). — 2022. — 30 июня. — DOI: 10.48550/arxiv.2206.15407. — URL: http://arxiv.org/abs/2206.15407.

30. Shifts: A Dataset of Real Distributional Shift Across Multiple Large-Scale Tasks / A. Malinin [и др.] // arXiv (Cornell University). — 2021. — 15 июля. — DOI: 10.48550/arxiv.2107.07455. — URL: http://arxiv.org/abs/2107.07455.

31. *Smirnov N.* Table for Estimating the Goodness of Fit of Empirical Distributions // Annals of Mathematical Statistics. — 1948. — 1 июня. — Т. 19, № 2. — С. 279—281. — DOI: 10.1214/aoms/1177730256. — URL: https://doi.org/10.1214/aoms/1177730256.

32. *Stone Z.*, *Zickler T.*, *Darrell T.* Autotagging Facebook: Social network context improves photo annotation //. — 23.06.2008. — DOI: 10.1109/cvprw.2008.4562956. — URL: https://doi.org/10.1109/cvprw.2008.4562956.

33. *S.* The Probable Error of a Mean // Biometrika. — 1908. — 1 марта. — Т. 6, № 1. — С. 1. — DOI: 10.2307/2331554. — URL: https://doi.org/10.2307/2331554.

34. *Wilcoxon F.* Individual Comparisons by Ranking Methods // Biometrics bulletin. — 1945. — 1 дек. — Т. 1, № 6. — С. 80. — DOI: 10.2307/3001968. — URL: https://doi.org/10.2307/3001968.

35. *Xiao T.*, *Gomez A. N.*, *Gal Y.* Wat zei je? Detecting Out-of-Distribution Translations with Variational Transformers // arXiv (Cornell University). — 2021. — 4 мая. — DOI: 10.48550/arxiv.2006.08344. — URL: http://arxiv.org/abs/2006.08344.