НИУ ВШЭ Санкт-Петербург
Санкт-Петербургская школа физико-математических и компьютерных наук
Промышленное программирование

Енгалыч Глеб Артурович

Анализ градиента нейронной сети для поиска аномалий в данных

Научный руководитель:
Куралёнок Игорь Евгеньевич,
к.ф.-м.н., руководитель сервисов машинного обучения Яндекс.Облако

Ершов Василий Алексеевич,
руководитель подразделения,
Яндекс.Облако

Рецензент:
Малинин Андрей Алексеевич,
к.ф.-м.н., старший исследователь,
Яндекс

Санкт-Петербург
2021

Оглавление

Введение

Современные нейронные сети используются практически во всех областях жизни человека: от рекомендации друзей в социальных сетях до медицинских обследований. Люди так привыкли полагаться на алгоритмы искусственного интеллекта, что уже не могут представить свою жизнь без них.

Всё машинное обучение работает в предположении о том, что данные для обучения, для тестирования и для применения взяты из одного и того же распределения. К сожалению, в процессе применения модели это предположение может нарушаться, что приводит к необъяснимым последствиям. Особенно, такие нарушения опасны в областях, где требуется быстро и точно принимать решения: медицина, финансы, self-driving cars.

Нарушение этого предположения может вызвать сдвиг распределения, сдвиги могут характеризоваться, например,

- добавлением новых, не виденных ранее моделью классов;
- появлением прецедентов уже известных классов, но в другой текстуре или форме;
- добавлением шума или искажения к прецеденту.

Так как методы поиска аномалий ещё плохо изучены, то системы машинного обучения часто просто игнорируют сдвиги распределения и продолжают работать в штатном режиме, не представляя, что ответы на запросы могут быть невалидными. С другой стороны, если в систему машинного обучения добавить возможность поиска аномалий, то можно отлавливать такие нарушения, чтобы сделать работу системы более точной и понять причину происхождения некорректных данных.

В главе 1 представлен обзор предметной области. В конце главы формулируются цель и задачи работы.

Глава 2 содержит краткий обзор существующих методов поиска аномалий, с которыми я сравниваю разработанный мной метод.

В главе 3 подробно описан предложенный градиентный метод и приведены эксперименты, доказывающие конкурентоспособность данного метода по сравнению с другими решениями.

Глава 4 представляет собой набор дополнительных экспериментов, ставящих целью проанализировать полученные результаты и заглянуть внутрь предложенного метода.

В главе 5 приводится обсуждение теоретических аспектов метода: временная сложность работы и детальный анализ градиента и сравнение градиентного метода с популярным алгоритмом для поиска аномалий в данных – ODIN.

В заключительной главе 6 приводится заключение, а также список направлений для будущих исследований и потенциальные негативные последствия, которые может нести предложенное решение.

Глава 1

Обзор предметной области и формулировка цели работы

1.1. Поиск аномалий в данных. Постановка задачи

Пусть исходная задача это задача классификации изображений на C классов. Тогда нейронная сеть по прецеденту x возвращает номер предсказанного класса $c \in \{1, 2, \ldots, C\}$, а метод поиска аномалий в процессе предсказания выдаёт значение $t \in \{0, 1\}$, ноль, если прецедент обыкновенный, единичка – если аномальный. Таким образом, метод поиска аномалий это бинарный классификатор.

В академической среде такие методы принято оценивать следующим образом: берутся два датасета – inlier (тестовая часть датасета, на котором обучалась нейронная сеть, то есть, датасет с данными из того же самого распределения) и outlier (датасет, целиком состоящий из аномальных данных, например, классов, к которым в процессе обучения у нейронной сети не было доступа), эти два датасета смешиваются и получается один тестовый датасет, на котором тестируется метод как бинарный классификатор – аномальный пример, или нет.

Метрики, по которым можно оценивать качество метода это любые метрики для бинарной классификации. В данной работе используются: площадь под Receiver-Operating-Characteristic кривой, площадь под Precision-Recall кривой (в двух вариантах: когда аномалии трактуются как положительный класс и когда аномалии трактуются как негативный класс), False Positive Rate классификатора при 95% True Positive Rate.

1.2. Требования к методу

Методы поиска аномалий с практической точки зрения должны быть универсальны и легко встраиваемы в готовые системы. Аналитик в процес-

се решения исходной задачи не должен думать о том, какой метод оценки неопределённости стоит применить. Этот вопрос должен быть поднят после получения готового решения, поэтому мной выдвигаются следующие требования к разрабатываемому методу:

- применимость к любой архитектуре нейронной сети;
- метод никак не должен влиять на процесс обучения модели для исходной задачи;
- метод никак не должен модифицировать архитектуру нейронной сети, используемой для решения исходной задачи;
- метод должен использовать как можно меньше outlier-данных для настройки используемых гиперпараметров;
- метод не должен требовать обучения каких-либо дополнительных моделей, допускается лишь подбор гиперпараметров на валидационной выборке.

Метод поиска аномалий, удовлетворяющий всем выдвинутым требованиям, легко может быть встроен в production-системы и параллельно с решением исходной задачи давать оценку, насколько можно и можно ли вообще доверять предсказанию для конкретного прецедента.

1.3. Мотивация к использованию градиентных методов

В последние несколько лет в научном сообществе стала популярной тема интерпретации нейронных сетей и развитие теории, связанной с моделями глубокого обучения: сходимостью, масштабируемостью и способностями к обобщению.

Данная работа была вдохновлена двумя методами:

1. Метод Neural Tangent Kernel [1]:

- исследует свойства градиентов по весам модели бесконечно широких нейронных сетей;
- с точки зрения NTK норма градиента нейронной сети определяет такие свойства моделей, как ожидаемая точность предсказания, неопределённость предсказания, аномальность, etc.
- работы в области NTK исследуют возможность замены градиентного спуска для нейронной сети на явные формулы для линейных моделей, порождаемых бесконечно широкими нейронными сетями.

Моя работа – попытка неявно применить идеи NTK к обычным нейронным сетям.

2. Influence functions [2]:

   - авторы статьи отвечают на вопрос: как изменится качество обученной модели для прецедента z_{test}, если из обучающей выборки исключить прецедент z?
   - для ответа на этот вопрос используется классический метод из статистики – influence functions, использующий первую и вторую производную нейронной сети по параметрам модели:

     $$-\nabla_\theta L(z_{test}, \hat{\theta})^T H_{\hat{\theta}}^{-1} \nabla_\theta L(z, \hat{\theta}),$$

     где L – функция потерь, а $H_{\hat{\theta}}^{-1}$ – обратная матрица Гессе для обученных весов, т.е. $H_{\hat{\theta}} = \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta^2 L(z_i, \hat{\theta})$

     По сути, influence function это градиент модели в точке z_{test}, скалярно домноженный на оптимизационный шаг Ньютона-Рафсона.
   - Обращение матрицы Гессе для нейронной сети – трудоёмкая задача, поэтому авторы статьи выносят способы упростить вычисления в отдельную секцию статьи.

Моя работа – попытка обобщить influence function на задачу поиска аномалий с одним упрощением: игнорируется информация о производной второго порядка модели по весам.

1.4. Цель исследования и поставленные задачи

Данная работа ставит целью разработку и исследование градиентого метода для поиска аномалий (то есть, оценки неопределённости) в данных, который улучшит или повторит результаты существующих методов. Для этого ставятся следующие задачи:

- Применить градиентный анализ к поиску аномалий в данных, то есть придумать, как извлечь полезную для оценки неопределённости предсказания информацию из градиента по весам модели.
- Разработать и имплементировать модификации к предложенному методу.
- Провести сравнительный анализ качества работы метода относительно других существующих решениях.
- Провести анализ полученного метода, чтобы разобраться в принципах его работы и описать границы применимости.

Глава 2

Обзор существующих решений

2.1. Используемые обозначения

В этой главе и далее используются следующие обозначения и понятия:

- C – количество классов, на которые требуется классифицировать прецеденты x в исходной задаче.
- $\theta \in \mathbb{R}^d$ – веса нейронной сети. Веса нейронной сети обладают структурой слоёв, но в этой работе для простоты нотации θ представляется как плоский вектор. Значком $\hat{\theta}$ будут обозначаться обученные веса.
- $h(x, \theta)$ – нейронная сеть. x – прецедент, θ – веса нейронной сети, архитектура сети следует из контекста.
- $S(x)$ – softmax-активация, которая вычисляется по формуле:

$$S_c(x) = \frac{\exp\left(h_c(x,\theta)\right)}{\sum\limits_{i=1}^{C} \exp\left(h_i(x,\theta)\right)} \tag{2.1}$$

  Также, в работе используется вариант активации softmax с температурой T: $S(x, T)$, данная активация вычисляется следующим образом:

$$S_c(x,T) = \frac{\exp\left(h_c(x/T,\theta)\right)}{\sum\limits_{i=1}^{C} \exp\left(h_i(x/T,\theta)\right)} \tag{2.2}$$

- $g(x)$ – скоринговая функция. Она принимает на вход прецедент и возвращает вещественное число – меру неопределённости для прецедента x. Значения скоринговой функции в данной работе никак не нормируются, поэтому их можно сравнивать между собой лишь в одинаковом сеттинге экспериментов, значения из разных экспериментов

не сравнимы друг с другом. Скоринговую функцию следует использовать следующим образом: пусть задан некоторый порог $\delta \in \mathbb{R}$, тогда если $g(x) \geq \delta$, то прецедент x объявляется аномалией, иначе – прецедент x объявляется обыкновенным прецедентом, сэмплированным из того же распределения, на котором обучалась исходная нейронная сеть.

Стоит отметить, что в некоторых методах скоринговая функция g принимает на вход не только прецедент, но и дополнительные параметры, необходимые для работы метода, для простоты обозначений все аргументы после первого будут обозначать именно дополнительные параметры, например: $g(x, \varepsilon, T)$.

2.2. Maximum Softmax Probability

Этот метод в данной работе может быть обозначен как baseline или MSP. Он описан в статье [3], одной из первых статей по теме поиска аномалий. Метод очень прост: берётся обученная нейронная сеть, а в качестве скоринговой функции используется отрицательный максимум softmax-вероятностей, то есть:

$$g(x) = -\max_i S_i(x) \tag{2.3}$$

Maximum Softmax Probability метод интуитивно понятен – если softmax-вероятность максимального класса достаточно высока, то, значит, модель уверена в каком-то классе, значит, прецедент вероятно принадлежит исходному распределению.

2.3. ODIN

Метод ODIN [4] является модификацией baseline метода, к которому добавили препроцессинг. Препроцессинг состоит из двух частей:

- Обычную softmax-активацию заменили на softmax с температурой T:

$$S_c(x, T) = \frac{\exp\left(h_c(x/T, \theta)\right)}{\sum\limits_{i=1}^{C} \exp\left(h_i(x/T, \theta)\right)} \tag{2.4}$$

- Скоринговая функция применяется не к исходному прецеденту x, а к прецеденту с возмущением:

$$x_p = x - \varepsilon \text{sign}(-\nabla_x \log S_{c'}(x, T)), \tag{2.5}$$

где ε – величина вносимой пертурбации, а c' – предсказанный нейронной сетью класс.

В качестве скорингвой функции используется softmax с температурой T от возмущённого прецедента x_p:

$$g(x, \varepsilon, T) = -\max_i S_i(x_p, T) \tag{2.6}$$

2.4. Maximum Logit

Этот метод описан в статье [5]. Maximum Logit является вариацией baseline метода. В качестве скоринговой функции используется не отрицательный максимум softmax-вероятности, а отрицательный максимум логита нейронной сети, то есть ненормированного выхода:

$$g(x) = -\max_i h_i(x) \tag{2.7}$$

Этот метод сохраняет скорость работы и простоту baseline метода и, по утверждению авторов статьи, лучше масштабируется на реальные задачи.

2.5. Ансамблевые методы

Ансамблевые методы на данный момент считаются самыми надёжными методами для поиска аномалий. Ансамбль требует несколько обученных нейронных сетей и агрегирует предсказания от каждой из них. Предложенный в данной работе метод использует два прямых прохода и два обратных прохода по нейронной сети, что по времени работы эквивалентно четырём прямым проходам, поэтому при сравнении используются ансамбли из четырёх моделей.

Ансамбль даёт оценку неопределённости по шести разным мерам, из которых выбирается лучшая по качеству на тестовой выборке:

- confidence
- entropy of expected
- expected entropy
- mutual information
- EPKL
- MKL

Подробную информацию о принципах работы ансамблей и об используемых оценках неопределённости можно найти в [6] и [7].

2.6. Градиентный метод

В научной литературе уже описан один градиентный метод для поиска аномалий [8], но у этого метода есть существенные недостатки:

- метод использует в качестве скоринговой функции $g(x)$ нейронную сеть, которую авторы статьи отдельно обучают;

- обучение дополнительной нейронной сети требует большого количества outlier данных, то есть, предложенный метод не может быть масштабирован на большие объёмы данных;

- метод протестирован на датасете CIFAR-10, который никак не отражает качество метода на задачах хоть сколько-то близким к реальным;

- авторы статьи не предоставили исходные коды программ для экспериментов и детали обучения нейронной сети, которую используют для скоринга. Из-за этого метод, описанный в статье, невозможно воспроизвести.

Как было отмечено в последнем пункте – этот градиентный метод невозможно воспроизвести, поэтому в своей работе я не сравниваюсь с ним. В своей работе я попытался избежать всех вышеперечисленных недостатков.

Глава 3

Градиентный метод

3.1. Архитектура ResNet

В данной секции приводится краткое описание архитектуры нейронной сети ResNet [9], являющейся хорошим бейзлайном для классификации изображений.

Проблема затухания градиента [10] стала очень серьёзной преградой для развития глубоких нейронных сетей, одним из её решений стали residual connections. См. рис. 3.1.

Архитектура сетей ResNet основана на последовательном соединении Res-блоков, например, в ResNet-18 используются 4 таких блока и fully connected слой для классификации, что в сумме даёт 18 слоёв. См. рис. 3.2.

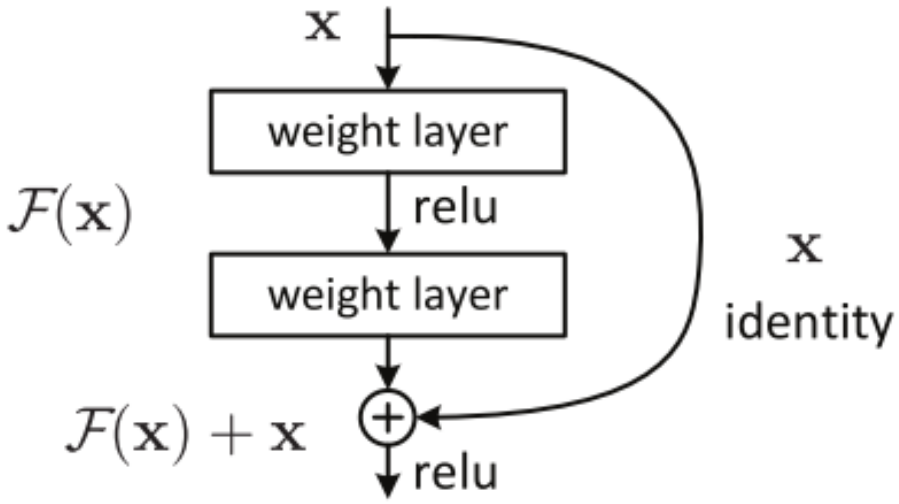

Рис. 3.1. Схематичное изображение Res-блока

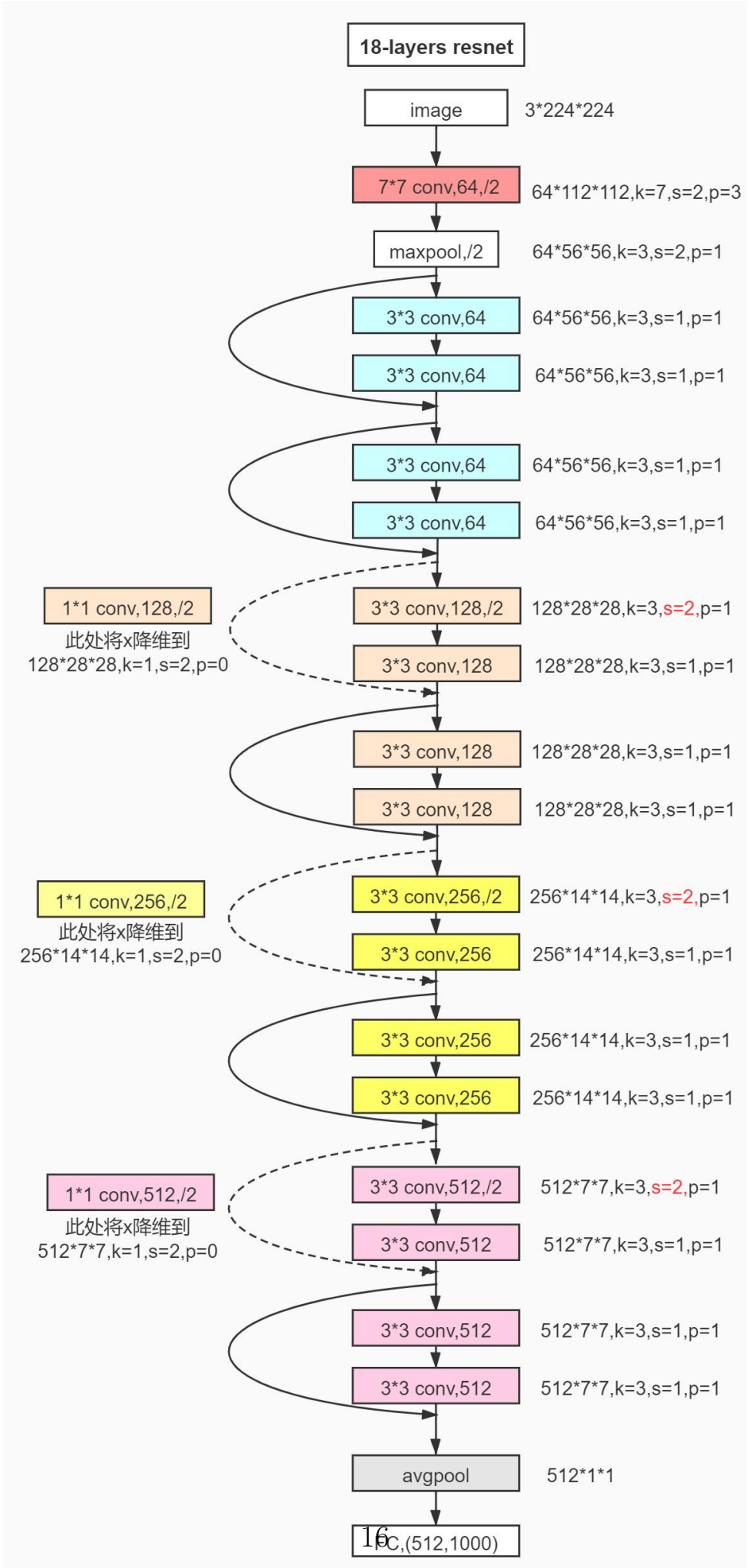

Рис. 3.2. Архитектура нейронной сети ResNet-18

3.2. Описание градиентного метода

При решении задачи многоклассовой классификации чаще всего используются функция потерь категориальная кросс-энтропия:

$$-\sum_{c=1}^{C} y_{o,c} \log p_{o,c}, \tag{3.1}$$

где $p_{o,c}$ это предсказанная нейронной сетью h для прецедента x_o вероятность класса c, а y_o это one-hot вектор для прецедента, то есть, вектор, на всех позициях которого стоят нули, кроме одной, которой соответствует истинный класс прецедента, на которой стоит единица.

Евклидову норму градиента по весам модели можно использовать в качестве скоринговой функции для метода поиска аномалий:

$$g(x) = ||\nabla_\theta \left(-\sum_{c=1}^{C} y_{o,c} \log p_{o,c} \right) ||_2 \tag{3.2}$$

Причём важно подчеркнуть несколько моментов:

- на этапе применения у нас нет доступа к настоящим меткам классов, поэтому в one-hot векторе y_o единственная единица будет стоять на позиции предсказанного моделью класса c';
- вместо евклидовой нормы можно попробовать взять любую другую векторную функцию от градиента, но в данном исследовании используется лишь евклидова норма;
- также, можно взять не просто градиент, а рассмотреть поворот и масштабирование градиента, но в данном исследовании используется лишь градиент без преообразований.

Идеально обученная модель переводит все точки из генеральной совокупности в стационарную точку с точки зрения весов $\nabla_\theta \ell(\mathbf{x}, \theta)|_{\theta^*} = 0$. Важно подчеркнуть, что градиентный скоринг работает в двух простых и естественных предположениях:

- ядро этого преобразования совпадает с генеральной совокупностью;
- наблюдаемое преобразование отличается от идеального на Гауссовскую величину: $\nabla_\theta \ell(\mathbf{x}, \theta)|_{\hat{\theta}} = \nabla_\theta \ell(\mathbf{x}, \theta)|_{\theta^*} + \delta, \delta \sim \mathcal{N}(0, E)$.

Предложенный градиентный скоринг можно модифицировать с помощью препроцессинга, идентичного препроцессингу из алгоритма ODIN, описанного в секции 2.3:

- Вместо softmax-активации используется softmax с температурой T, как в формуле 2.4.
- Скоринговая функция считается не от исходного прецедента x_o, а от возмущённого прецедента x_p, вычисляемого по формуле 2.5.

Комбинируя обе части препроцессинга и градиентный скоринг, получаем вычислительную формулу скоринга для модифицированного градиентного метода:

$$g(\mathbf{x}_o, \hat{\theta}, T) = \left\| \nabla_\theta \log \frac{\max\limits_{c'} \exp(h_{c'}(\mathbf{x}_p, \theta)/T)}{\sum\limits_{c=1}^{C} \exp(h_c(\mathbf{x}_p, \theta)/T)} \right|_{\hat{\theta}} \Bigg\|_2 \tag{3.3}$$

где c' – предсказанный класс, T – температура, $\hat{\theta}$ – обученные веса, а x_p – возмущённое изображение, которое имеет вид:

$$\mathbf{x}_p = \mathbf{x}_o - \varepsilon \text{sign} \left(-\nabla_{\mathbf{x}} \log \frac{\max\limits_{c'} \exp(h_{c'}(\mathbf{x}, \theta)/T)}{\sum_{c=1}^{k} \exp(h_c(\mathbf{x}, \theta)/T)} \Bigg|_{\mathbf{x}_o} \right) \tag{3.4}$$

Если значение скоринговой функции $g(\mathbf{x}_o, \hat{\theta}, T)$ больше некоторого выбранного порога δ, то изображение объявляется аномалией, иначе – обычным изображением из распределения, совпадающего с тем, на котором обучалась модель.

Далее модифицированный градиентный метод, то есть, градиентный скоринг с препроцессингом будет во всех экспериментах называться градиентным методом. В секции 4.1 исследовано влияние препроцесинга на качество работы метода.

3.3. Описание сеттинга экспериментов

Данная глава следует методологии проведения экспериментов из секции 1.1. Тестовая часть датасета, на котором обучалась нейронная сеть, трактуется как inlier-данные. Из таблицы 3.1 выбирается подходящий датасет как outlier-данные. Параметры для используемых детекторов аномалий подбираются на валидационных выборках так, чтобы максимизировалась метрика ROC-AUC на валидационной выборке. Оптимальные параметры были выбраны для оценки качества детекторов аномалий на тестовом множестве. Список нейронных сетей, на основе которых строились детекторы аномалий, приведён в таблице 3.2

Таблица 3.1. Список использованных датасетов

Датасет	#train	#val	#test	Аномалия для	Ссылка
CIFAR-10	50000	500	9500	—	[11]
ImageNet-1k	$\approx 1.28 * 10^6$	500	49500	—	[12]
tiny ImageNet	—	500	9500	CIFAR-10	[12]
ImageNet-O	—	500	1500	ImageNet-1k	[13]
ImageNet-R	—	500	29500	ImageNet-1k	[14]

Таблица 3.2. Список использованных нейронных сетей

Архитектура	Датасет	Число параметров	Точность (test)	Ссылка
Wide-ResNet-28-10	CIFAR-10	36,479,194	96.2	[15]
ResNet-18	ImageNet-1k	11,689,512	69.7	[9]
DenseNet-161	ImageNet-1k	28,681,000	77.1	[16]
ResNet-50	ImageNet-1k	25,557,032	76.1	[9]

Параметры для двух методов, требующих препроцессинга – ODIN и предложенного градиентного метода, были найдёны с помощью поиска по гриду.

Грид для ε представляет собой 21 равноотстоящую точку от 0 до 0.004, температура T выбирается из множества $\{1, 2, 5, 10, 20, 50, 100, 200, 500, 1000\}$. Для обоих алгоритмов используются одинаковые гриды. Важно отметить, что ровно этот же грид используется в оригинальной статье про алгоритм ODIN.

В работе рассматривается два вида экспериментов:

- полная end-to-end схема, в которой исследователь обучает модель с нуля. Эти эксперименты проводились с ResNet-50, обученной на ImageNet-1k, и Wide-ResNet-28-10, обученной на датасете CIFAR-10. Для работы ансамблевых методов требуется несколько обученных нейронных сетей одной архитектуры, поэтому сравнение с ансамблями проводится лишь для этой группы экспериментов

- эксперименты с предобученными моделями ResNet-18 и DenseNet-161, обученными на ImageNet-1k, из PyTorch [17]. Эти эксперименты отличаются простотой воспроизводимости, так как доступ к данным обученным весам имеется у любого исследователя в мире.

Во всех проведённых экспериментах в качестве значений метрик приводится среднее с погрешностью, равной стандартному отклонению, по бутстрап статистике, посчитанной на 100 повторениях каждого эксперимента.

Приведём краткое описание датасетов, на которых обучались нейронные сети для решения исходной задачи классификации:

- **CIFAR-10:** состоит из 10 классов: самолёт, автомобиль, птица, кошка, олень, собака, лягушка, лошадь, корабль, грузовик. Каждое изображение имеет размер 32×32 пикселя, все изображения цветные. Этот датасет является очень популярным для проверки качества решения разнообразных задач, связанных с обработкой изображений.

Но, к сожалению, датасет слишком прост и никак не может продемонстрировать, как будет вести себя метод в реальных условиях применения.

- **ImageNet-1k:** состоит из 1000 классов, например, страус, воздушный змей, фартук, банжо, саксофон, футбольный мяч. Этот датасет всё ещё является настоящим вызовом для исходной задачи классификации. В области детекции аномалий существует небольшое число работ, в которых приводятся эксперименты с этим датасетом, например, [5, 18].

Исходные коды экспериментов и градиентного метода можно найти по ссылке: `https://github.com/herrbilbo/GRODIN`.

3.4. Эксперимент. CIFAR-10 vs tiny ImageNet

Приведём краткое описание датасета tiny ImageNet. Этот датасет является подмножеством датасета ImageNet-1k, все изображения были уменьшены до размера 64×64 пикселей. Он состоит из 200 классов исходного датасета ImageNet-1k, по 50 изображений на класс. Этот датасет был изначально собран как альтернатива огромному датасету ImageNet-1k, на которой нейронные сети смогут обучаться и применяться со скоростью сравнимой с той, которую предоставляет датасет CIFAR-10.

Для того, чтобы tiny ImageNet можно было использовать как outlier данные относительно CIFAR-10 его требуется сжать до размера 32×32 пикселя, по размеру изображений в CIFAR-10. Сжатая версия датасета была взята из статьи про алгоритм ODIN.

В таблице 3.3 представлены результаты детекции аномалий в сеттинге: inlier-данные – CIFAR-10, outlier-данные – tiny ImageNet. Предложенный градиентный метод по всем четырём метрикам превосходит прочие решения.

Таблица 3.3. Wide-Resnet-28-10. CIFAR-10 vs tiny-ImageNet.

	MSP	ODIN	Gradient	ML
ROC-AUC ↑	0.954 ± 0.001	0.966 ± 0.001	**0.982** ± 0.001	0.965 ± 0.001
FPR at 95% TPR ↓	0.146 ± 0.005	0.203 ± 0.013	**0.087** ± 0.005	0.133 ± 0.005
PR-AUC (outlier +) ↑	0.946 ± 0.002	0.972 ± 0.001	**0.981** ± 0.001	0.965 ± 0.001
PR-AUC (outlier -) ↑	0.956 ± 0.001	0.954 ± 0.002	**0.981** ± 0.001	0.962 ± 0.001

3.5. Эксперимент. ImageNet vs ImageNet-O

Датасет ImageNet-O [13] иллюстрирует сдвиг распределения, добавляющий новые, ранее не виденные классы. Он был собран из датасета ImageNet-21k, из которого оставили лишь классы, которых нет в ImageNet-1k. Среди всех этих изображений были отобраны те, на которых ошибается MSP детектор аномалий, работающий на основе сети ResNet-50. Среди отобранных изображений вручную были отобраны 2000 изображений из соображений качества картинки. Этот датасет – настоящий вызов для методов поиска аномалий.

См. рис. 3.3, 3.4 с примерами изображений из ImageNet-O.

Рис. 3.3. Слева – фотография мозгового коралла (ImageNet-1k), справа – фотография осиного гнезда (ImageNet-O), которая была принята сетью ResNet-50 за мозгового коралла

Рис. 3.4. Слева – фотография цифровых часов (ImageNet-1k), справа – фотография кнопок на клавиатуре (ImageNet-O), которая была принята сетью ResNet-50 за цифровые часы

Результаты экспериментов по отделению ImageNet от ImageNet-O приведены в таблицах 3.4, 3.5, 3.6.

Таблица 3.4. ResNet-18. ImageNet vs ImageNet-O.

	MSP	ODIN	Gradient	ML
ROC-AUC ↑	0.483 ± 0.007	0.624 ± 0.006	**0.799** ± 0.006	0.599 ± 0.005
FPR at 95% TPR ↓	0.87 ± 0.007	0.802 ± 0.015	**0.648** ± 0.024	0.791 ± 0.013
PR-AUC (outlier +) ↑	0.026 ± 0.0	0.04 ± 0.001	**0.126** ± 0.005	0.034 ± 0.001
PR-AUC (outlier -) ↑	0.972 ± 0.001	0.983 ± 0.0	**0.991** ± 0.0	0.979 ± 0.0

Таблица 3.5. DenseNet-161. ImageNet vs ImageNet-O.

	MSP	ODIN	Gradient	ML
ROC-AUC ↑	0.484 ± 0.006	0.556 ± 0.006	**0.754** ± 0.006	0.548 ± 0.007
FPR at 95% TPR ↓	0.91 ± 0.007	0.896 ± 0.01	**0.817** ± 0.02	0.9 ± 0.009
PR-AUC (outlier +) ↑	0.026 ± 0.0	0.032 ± 0.001	**0.101** ± 0.006	0.031 ± 0.001
PR-AUC (outlier -) ↑	0.972 ± 0.0	0.977 ± 0.001	**0.988** ± 0.0	0.976 ± 0.001

Таблица 3.6. ResNet-50. ImageNet vs ImageNet-O.

	MSP	ODIN	Gradient	ML	Ensemble
ROC-AUC ↑	0.474 ± 0.006	0.601 ± 0.008	**0.765** ± 0.007	0.573 ± 0.007	0.614 ± 0.006
FPR at 95% TPR ↓	0.874 ± 0.009	0.884 ± 0.013	**0.744** ± 0.02	0.848 ± 0.01	0.785 ± 0.012
PR-AUC (outlier +) ↑	0.025 ± 0.0	0.038 ± 0.001	**0.107** ± 0.005	0.032 ± 0.001	0.042 ± 0.002
PR-AUC (outlier -) ↑	0.972 ± 0.001	0.98 ± 0.001	**0.989** ± 0.0	0.978 ± 0.001	0.982 ± 0.0

Предложенный градиентный метод на всех трёх рассмотренных архитектурах нейронных сетей показывает результаты лучше чем прочие рассмотренные методы.

3.6. Эксперимент. ImageNet vs ImageNet-R

Датасет ImageNet-R представляет собой сдвиг распределения, не добавляющий новые классы, но меняющий текстуру или форму уже известных модели классов. Датасет состоит из 200 классов оригинального ImageNet-1k, но в которых вместо реальных фотографий объекта представлены разнообразные поделки, рисунки, комиксы, аниме, кадры из мультфильмов, игрушки, украшения, etc. Этот сдвиг распределения достаточно сложен, например, в ImageNet-R присутствуют компьютерные рисунки футбольных мячей, которые даже человеку сложно отличить от фотографий. Также, в классе 'футбольный мяч' присутствуют изображения, на которых футбольный мяч не является главным элементом композиции. См. примеры изображений из ImageNet-R в рис. 3.5.

Можно поставить вопрос: является ли компьютерный рисунок футбольного мяча, практически не отличимый от настоящего, аномалией относительно класса 'футбольный мяч'? Для ответа на этот вопрос стоит более подробно исследовать формулировку задачи детекции аномалий и понятие сдвига распределения.

Рис. 3.5. Примеры изображений из датасета ImageNet-R, принадлежащих классу 'футбольный мяч'

Результаты экспериментов по отделению ImageNet от ImageNet-R приведены в таблицах 3.7, 3.8, 3.9.

Таблица 3.7. ResNet-18. ImageNet vs ImageNet-R.

	MSP	ODIN	Gradient	ML
ROC-AUC ↑	0.776 ± 0.002	**0.84** ± 0.001	0.831 ± 0.001	0.838 ± 0.001
FPR at 95% TPR ↓	0.699 ± 0.004	0.581 ± 0.005	0.647 ± 0.005	**0.569** ± 0.006
PR-AUC (outlier +) ↑	0.658 ± 0.003	**0.749** ± 0.002	**0.749** ± 0.002	0.73 ± 0.002
PR-AUC (outlier -) ↑	0.846 ± 0.001	0.891 ± 0.001	0.879 ± 0.001	**0.892** ± 0.001

Таблица 3.8. DenseNet-161. ImageNet vs ImageNet-R.

	MSP	ODIN	Gradient	ML
ROC-AUC ↑	0.781 ± 0.002	**0.843** ± 0.001	0.828 ± 0.002	0.838 ± 0.001
FPR at 95% TPR↓	0.703 ± 0.005	0.654 ± 0.006	0.683 ± 0.006	**0.63** ± 0.006
PR-AUC (outlier +) ↑	0.676 ± 0.003	**0.79** ± 0.002	0.761 ± 0.002	0.747 ± 0.002
PR-AUC (outlier -) ↑	0.848 ± 0.001	0.883 ± 0.001	0.873 ± 0.001	**0.884** ± 0.001

Таблица 3.9. ResNet-50. ImageNet vs ImageNet-R.

	MSP	ODIN	Gradient	ML	Ensemble
ROC-AUC ↑	0.803 ± 0.001	0.858 ± 0.001	0.855 ± 0.001	**0.863** ± 0.001	0.856 ± 0.001
FPR at 95% TPR↓	0.661 ± 0.005	0.59 ± 0.005	0.631 ± 0.005	**0.558** ± 0.005	0.617 ± 0.006
PR-AUC (outlier +) ↑	0.707 ± 0.002	**0.803** ± 0.002	0.796 ± 0.002	0.786 ± 0.002	0.793 ± 0.002
PR-AUC (outlier -) ↑	0.865 ± 0.001	0.899 ± 0.001	0.894 ± 0.001	**0.904** ± 0.001	0.896 ± 0.001

Предложенный градиентный метод на датасете ImageNet-R не смог показать такие же хорошие результаты, как на датасете ImageNet-O. Возможно, это связано с тем, что у ImageNet-R сдвиг распределения по своей природе отличается от сдвига ImageNet-O.

Требуется более детальное исследование датасета ImageNet-R, чтобы делать какие-либо выводы в этих экспериментах.

Глава 4

Дополнительные эксперименты

4.1. Влияние предобработки изображений на качество работы

Ключевым вкладом этой работы является введение градиентного скоринга, описанного в секции 3.2, а препроцессинг является лишь модификацией. Но встаёт закономерный вопрос: как влияет препроцессинг (в частности, конкретные его части) на работу метода?

Напомним, что препроцессинг состоит из двух частей: softmax с температурой T и ε-пертурбации. Каждую из этих частей можно убрать из метода, установив $T = 1$ или же $\varepsilon = 0$, проведя перебор параметров по гриду лишь по другому из параметров.

Таким образом, получаются четыре метода: без препроцессинга, только с температурой, только с пертурбацией, градиентный метод с полным препроцессингом. Эти методы были апробированы на датасете ImageNet-O с использованием нейронной сети ResNet-50. См. результаты эксперимента в таблице 4.1

Таблица 4.1. Влияние препроцессинга. ResNet-50. ImageNet vs ImageNet-O.

	No preprocessing	T-scaling only	ε-perturbation only	Full scheme
ROC-AUC ↑	0.517 ± 0.013	0.664 ± 0.007	0.517 ± 0.013	**0.765** ± 0.007
FPR at 95% TPR ↓	0.864 ± 0.007	0.824 ± 0.014	0.864 ± 0.007	**0.744** ± 0.02
PR-AUC (outlier +) ↑	0.03 ± 0.001	0.051 ± 0.002	0.03 ± 0.001	**0.107** ± 0.005
PR-AUC (outlier -) ↑	0.976 ± 0.001	0.984 ± 0.0	0.976 ± 0.001	**0.989** ± 0.0

Можно отметить следующее:

- В случае только ε-пертурбации метод вырождается в метод без препроцессинга. С увеличением коэффициента пертурбации ε качество работы метода падает.

- Внесение шума через температуру T в softmax-активацию более важно чем внесение ε-пертурбации.
- Обе части препроцессинга важны для достижения наилучшего результата.

Более подробное изучение влияния параметров препроцессинга на работу приведено в секции 4.3.

4.2. Влияние блоков ResNet на качество работы

Как было отмечено в секции 3.1, нейронные сети архитектуры ResNet состоят из блоков. Количество блоков определяет глубину сети и, соответственно, её обобщающую способность. В силу процесса обучения сети с помощью алгоритма обратного распространения ошибки [19] чем ближе блок к выходам, тем большую порцию сигнала он получит. Эта проблема называется затухание градиента, один из способов борьбы с ней – добавление Residual connections, впервые появившееся в сетях ResNet.

Целью данного эксперимента является проверить, как сильно теряется информация, заключённая в градиент блока, с удалением от выходов сети.

Для этого была взята сеть ResNet-18 и на примере задачи ImageNet vs ImageNet-O замерялось качество детекции аномалий у следующих конфигураций градиентного метода:

- в скоринговой учитывается лишь градиент по весам θ_i, принадлежащих первому Res-блоку;
- в скоринговой учитывается лишь градиент по весам θ_i, принадлежащих первому и второму Res-блокам;
- в скоринговой учитывается лишь градиент по весам θ_i, принадлежащих первому, второму и третьему Res-блокам;

- в скоринговой учитывается лишь градиент по весам θ_i, принадлежащих первому, второму, третьему и четвёртому Res-блокам, то есть градиент не считается лишь по fully-connected слою;
- в скоринговой учитывается градиент по всем весам сети.

Результаты эксперимента представлены в таблице 4.2.

Таблица 4.2. Последовательное добавление Res-блоков. ResNet-18. ImageNet vs ImageNet-O.

	1	1, 2	1, 2, 3	1, 2, 3, 4	1, 2, 3, 4, fc
ROC-AUC ↑	0.725 ± 0.008	0.754 ± 0.007	0.775 ± 0.005	0.798 ± 0.008	**0.799** ± 0.006
FPR at 95% TPR ↓	0.739 ± 0.017	0.702 ± 0.025	0.687 ± 0.02	**0.647** ± 0.022	0.648 ± 0.024
PR-AUC (outlier +) ↑	0.066 ± 0.01	0.083 ± 0.003	0.102 ± 0.007	**0.126** ± 0.005	**0.126** ± 0.005
PR-AUC (outlier -) ↑	0.988 ± 0.001	0.989 ± 0.0	0.99 ± 0.001	**0.991** ± 0.0	**0.991** ± 0.0

Легко видно, что с последовательным добавлением Res-блоков растёт качество детекции аномалий. Также, видно, что добавление или удаление fully-connected слоя не влияет на качество модели.

Таким образом, мы приходим к выводу, что важно брать градиент именно всей модели, нельзя обойтись лишь градиентом нескольких Res-блоков.

4.3. Устойчивость гиперпараметров

В этой секции исследуются оптимальные значения параметров ε и T, найденные в ходе перебора параметров по гриду на валидационной выборке. В таблицах 4.3 и 4.4 приведены оптимальные значения параметров для градиентного метода и ODIN для разных датасетов и архитектур нейронных сетей.

Таблица 4.3. Оптимальные значения параметров. ImageNet-O.

	ResNet-18	ResNet-50	DenseNet-161
Gradient	$\varepsilon = 0.0026,\ T = 5$	$\varepsilon = 0.0024,\ T = 5$	$\varepsilon = 0.0014,\ T = 5$
ODIN	$\varepsilon = 0.004,\ T = 1000$	$\varepsilon = 0.004,\ T = 100$	$\varepsilon = 0.0,\ T = 5$

Таблица 4.4. Оптимальные значения параметров. ImageNet-R.

	ResNet-18	ResNet-50	DenseNet-161
Gradient	$\varepsilon = 0.001,\ T = 2$	$\varepsilon = 0.001,\ T = 2$	$\varepsilon = 0.001,\ T = 2$
ODIN	$\varepsilon = 0.0004,\ T = 500$	$\varepsilon = 0.004,\ T = 200$	$\varepsilon = 0.004,\ T = 100$

Можно заметить, что для градиентного метода оптимальная температура остаётся неизменной в рамках одной и той же задачи. То есть, при смене архитектуры нейронной сети можно не настраивать температуру T, а лишь донастроить величину пертурбации ε. Для более детального исследования этого эффекта были построены хитмпапы перебора параметров. См. рисунки 4.1, 4.2, 4.3, 4.4.

По горизонтальной оси отложено перебираемое значение температуры T, по вертикальной оси отложены значения величины пертурбации ε. На пересечении осей цветом изображается значение метрики ROC-AUC соответствующего разделения для валидационной выборки.

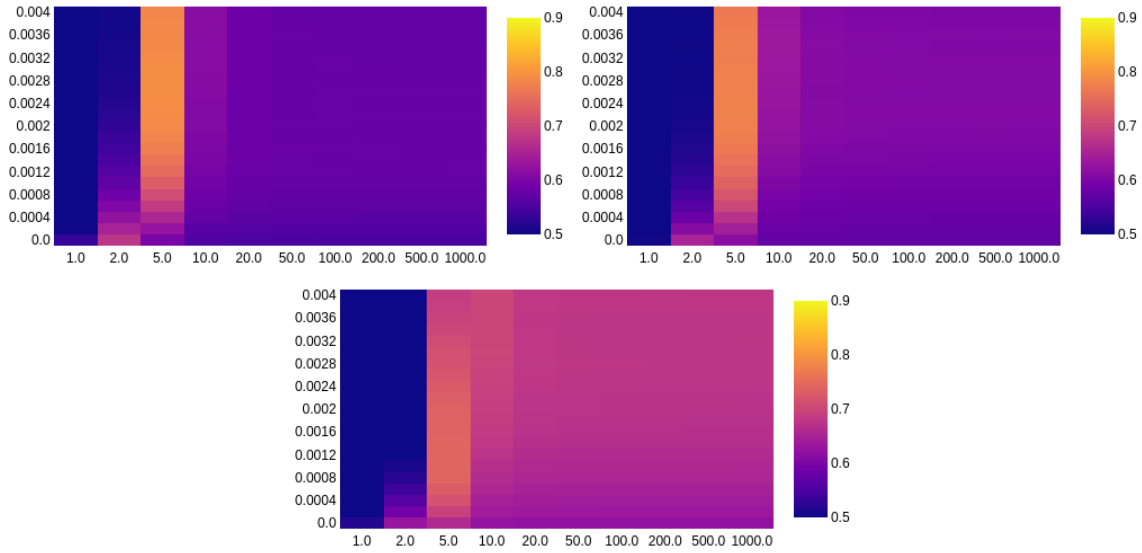

Рис. 4.1. Хитмапы для градиентного метода на датасете ImageNet-O: ResNet-18, ResNet-50, DenseNet-161 (слева направо, сверху вниз)

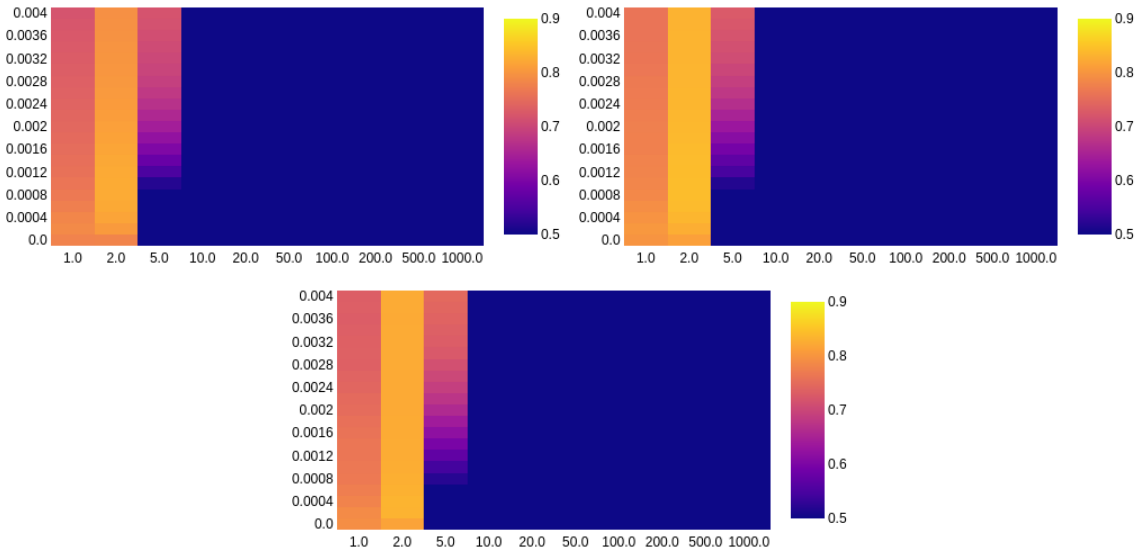

Рис. 4.2. Хитмапы для градиентного метода на датасете ImageNet-R: ResNet-18, ResNet-50, DenseNet-161 (слева направо, сверху вниз)

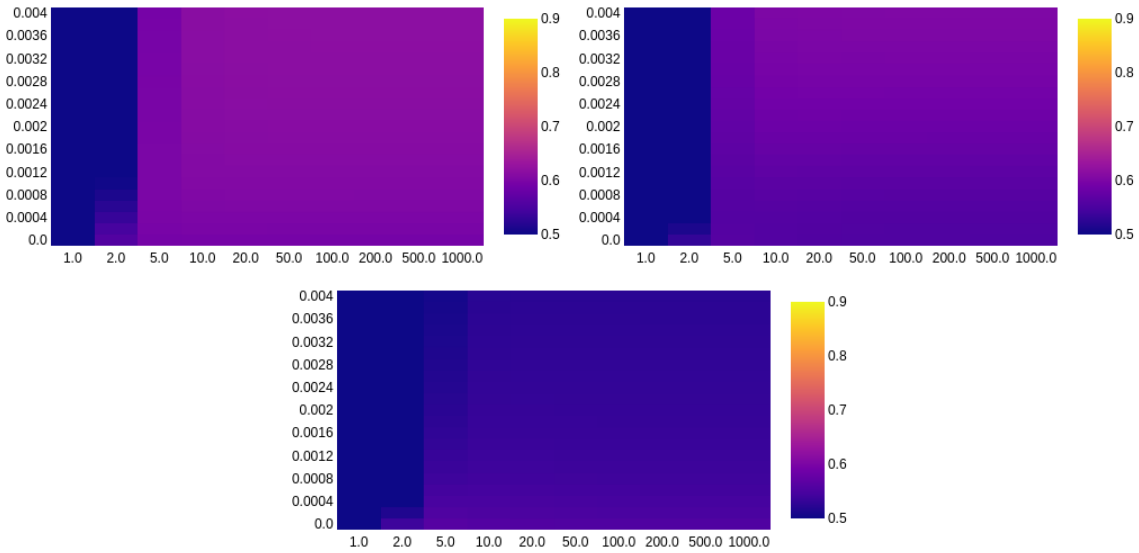

Рис. 4.3. Хитмапы для ODIN на датасете ImageNet-O: ResNet-18, ResNet-50, DenseNet-161 (слева направо, сверху вниз)

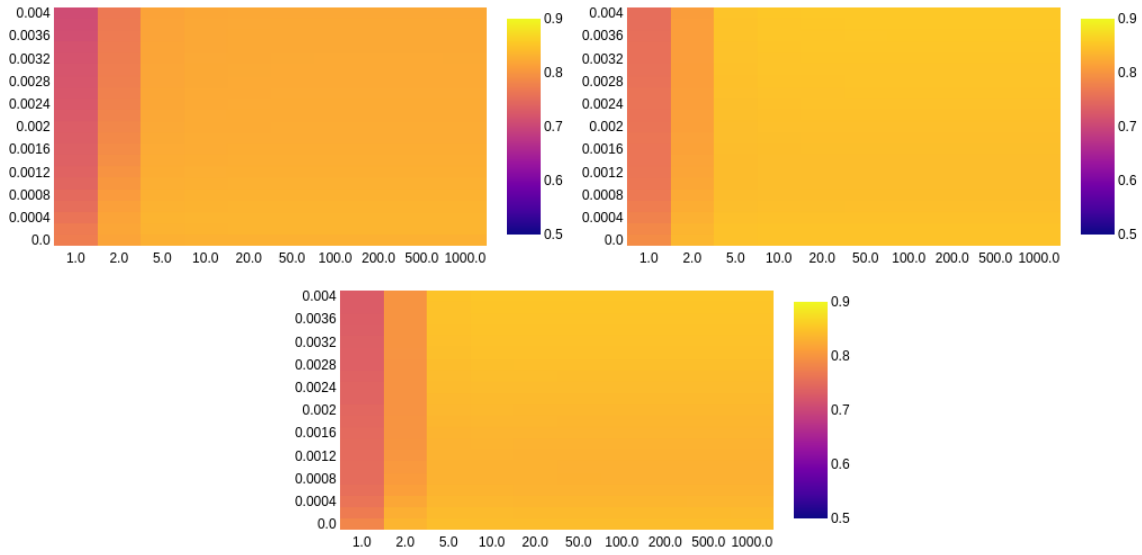

Рис. 4.4. Хитмапы для ODIN на датасете ImageNet-R: ResNet-18, ResNet-50, DenseNet-161 (слева направо, сверху вниз)

По приведённым хитмапам видно, что предложенный градиентный метод обладает в некотором смысле инвариантом: при фиксированной задаче детекции аномалий параметры хорошо переносятся с одной архитектуры на другую, то есть, оптимальное значение температуры T не меняется.

Аналогичное утверждение неверно для алгоритма ODIN: при смене архитектуры нейронной сети может меняться оптимальная температура, что видно по переменам яркости при фиксированном значении ε по вертикали и варьируемом значении температуры T по горизонтали.

4.4. Масштабируемость метода на большие объёмы данных

Для исследования масштабируемости предложенного градиентного метода на большие объёмы данных был собран дополнительный датасет, который для простоты обозначений будет называться ImageNet-G.

Он был собран из изображений из 18041 классов датасета ImageNet-21k, которых нет в датасете ImageNet-1k. Из каждого класса было взято по 2 изображения, в итоге получилось 36082 изображения.

На этом датасете в аналогичном предыдущим экспериментам сеттинге

сравниваются градиентный метод и ODIN, основанные на модели ResNet-50. 500 изображений из датасета ImageNet-G были использованы в качестве валидационной выборки для подбора параметров. Результаты детекции аномалий приведены в таблице 4.5.

Таблица 4.5. ResNet-50. ImageNet vs ImageNet-G.

	ODIN	Gradient
ROC-AUC ↑	0.789 ± 0.002	**0.792** ± 0.002
FPR at 95% TPR ↓	0.73 ± 0.005	**0.7** ± 0.005
PR-AUC (outlier +) ↑	**0.742** ± 0.002	0.727 ± 0.002
PR-AUC (outlier -) ↑	0.819 ± 0.002	**0.828** ± 0.002

Оба метода практически одинаково справились с поставленной задачей. Стоит отметить, что датасет ImageNet-G идеологически похож на датасет ImageNet-O. Сдвиг распределения, вносимый датасетом ImageNet-G, добавляет новые ранее не виденные классы. Возможно, такого резкого отрыва градиентного метода от ODIN как в ImageNet-O в этом эксперименте не наблюдается из-за того, что ImageNet-O специально отбирался, чтобы обманывать методы поиска аномалий.

4.5. Оценка надёжности предсказания с помощью градиента

У предложенного градиентного метода есть ещё одно применение, связанное с оценкой неопределённости. До этого была исследована лишь задача детекции аномалий, но норму градиента функции потерь можно также использовать в качестве меры надёжности предсказания.

Гипотеза: чем больше норма градиента у прецедента, тем с большей вероятностью модель на этом прецеденте ошибётся в исходной задаче классификации.

Для проверки этой гипотезы вычислим нормы градиентов сети ResNet-50

для всех изображений в тестовой выборке ImageNet-1k. Отсортируем все прецеденты по неубыванию нормы градиента и сгруппируем их в бины по 500 прецедентов в каждом. Для каждого бина посчитаем с помощью бутстрап-статистики медиану и 1-перцентиль точности предсказания для всех прецедентов, попавших в текущий бин или любой из бинов, находящихся правее.

Полученные значения точности предсказания представлены на рисунке 4.5.

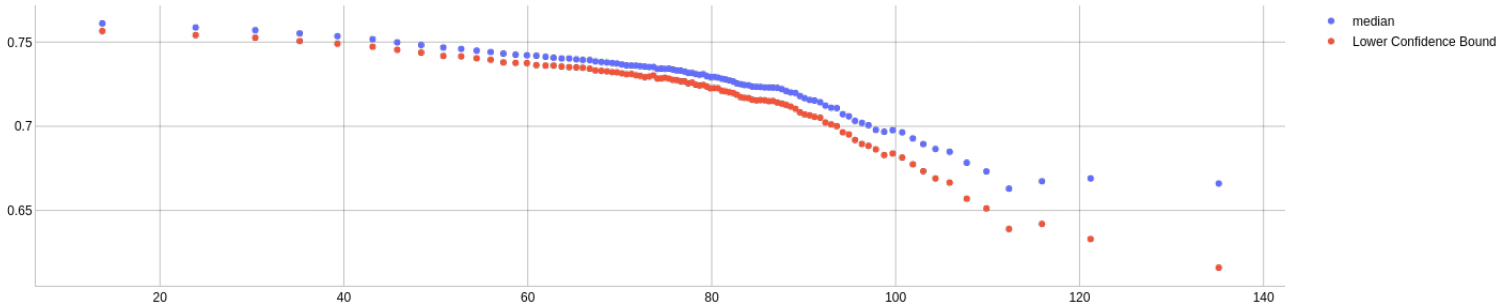

Рис. 4.5. Синие точки обозначают медиану бутстрап-выборки, красные – 1-перцентиль

Таким образом, гипотеза подтвердилась: с увеличением нормы градиента падает надёжность предсказания. Например, разница между первым и последними бинами составляет приблизительно 7% точности, что является серьёзным ухудшением качества решения исходной задачи классификации.

Глава 5

Анализ градиентного метода

5.1. Анализ метода и ограничения для работы

Структура предложенного градиентного метода совпадает со структурой ODIN, но поведение алгоритма отличается во многом. В этой секции проводится более детальный анализ сигнала, используемого методом, чтобы понять разницу между градиентным методом и ODIN.

Идеально обученная модель переводит все точки из генеральной совокупности в стационарную точку с точки зрения весов $\nabla_\theta \ell(\mathbf{x}, \theta)|_{\theta^*} = 0$. Важно подчеркнуть, что градиентный метод работает в двух простых и естественных предположениях:

- ядро этого преобразования совпадает с генеральной совокупностью;
- наблюдаемое преобразование отличается от идеального на Гауссовскую величину: $\nabla_\theta \ell(\mathbf{x}, \theta)|_{\hat{\theta}} = \nabla_\theta \ell(\mathbf{x}, \theta)|_{\theta^*} + \delta, \delta \sim \mathcal{N}(0, E)$.

Для начала стоит подробно расписать выражение для градиента функции потерь. Для упрощения обозначений $\hat{c}$ обозначает предсказанный моделью класс.

$$\begin{aligned}\nabla_\theta \ell(\mathbf{x}, \hat{\theta}) &= \nabla_\theta \log \frac{\max\limits_{c'} \exp\left(h_{c'}(\mathbf{x}, \theta)/T\right)}{\sum\limits_{c=1}^{k} \exp\left(h_c(\mathbf{x}, \theta)/T\right)}\Bigg|_{\hat{\theta}} = \\ &= \frac{1}{1 + \sum\limits_{c \neq \hat{c}} \exp\left((h_c(\mathbf{x}, \hat{\theta}) - h_{\hat{c}}(\mathbf{x}, \hat{\theta}))/T\right)} \left(\sum_{c \neq \hat{c}} e^{((h_c(\mathbf{x}, \hat{\theta}) - h_{\hat{c}}(\mathbf{x}, \hat{\theta}))/T)} \frac{\partial h_{\hat{c}}(\mathbf{x}, \theta) - h_c(\mathbf{x}, \theta)}{\partial \theta}\Bigg|_{\hat{\theta}} \right)\end{aligned} \tag{5.1}$$

Полученное выражение разбивается на два множителя (для простоты обозначений – S и G). Первая часть S – скаляр, равный вероятности предсказанного класса. Вторая часть G – значение градиента всех голов сети, взвешенное на отношение предпочтительности предсказанного класса.

S это скоринговая функция (с точностью до знака) метода ODIN, то есть, максимум softmax-вероятности. G это дополнительная информация, которая позволяет градиентному методу использовать информацию о пространстве NTK.

См. рисунок 5.1, на котором изображено значение $||G(x, \hat{\theta}, T)||_2$ в случае $T = 1$ для inlier и outlier данных. Эта информация сама по себе даёт хорошее разделение для прецедентов, и позволяет градиентному методу обгонять ODIN в случае, когда выполнены предположения, необходимые для работы градиентного метода, как например, у датасета ImageNet-O. Как обсуждалось в секции 3.6 датасет ImageNet-R немного другой: один и тот же класс в исходной задаче классификации разделён между inlier и outlier данными. Например, футбольные мячи встречаются как в виде фотографий (трактуются как inlier) и в виде компьютерных изображений (трактуются как outlier). Модель хорошо обучена на исходной задаче, поэтому её предсказание на картинке с футбольным мячом – футбольный мяч, так как прецедент действительно является футбольным мячом в обоих случаях, а, значит, лежит в ядре NTK, но определение in-domain и out-of-domain для этого датасета вступает в противоречие с нашим предположением. В таком случае градиентный метод не имеет доступа к дополнительной информации от компоненты G, что объясняет, почему результаты детекции аномалий у метода ODIN и градиентного метода на датасете ImageNet-R почти совпадают (секция 3.6)

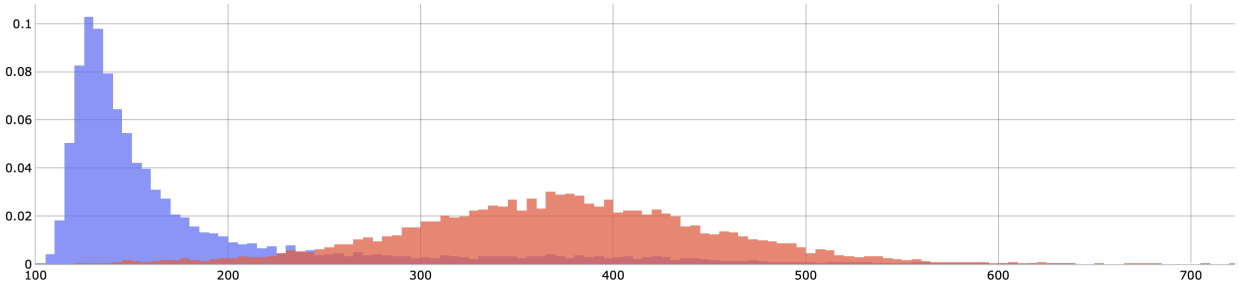

Рис. 5.1. Гистограмма распределения $||G(x, \hat{\theta}, T)||_2$ для inlier (синий цвет) и outlier (красный цвет) прецедентов. CIFAR-10 vs tiny ImageNet.

5.2. Оценка времени работы метода

Предложенный метод достаточно лёгок с вычислительной точки зрения. Внесение ε-пертурбации требует одного прямого прохода по нейронной сети и одного обратного прохода, чтобы посчитать градиент по входам. После этого требуется ещё один прямой и ещё один обратный проход, чтобы посчитать l_2-норму градиента возмущённого прецедента.

Суммарная временная сложность составляет $O(|F| + |B| + |\Theta|)$, где $O(|F|)$ – сложность прямого прохода, $O(|B|)$ – сложность обратного прохода, а $|\Theta|$ – количество параметров в нейронной сети.

На практике время, затраченное на проверку на аномальность одного изображения, приблизительно в два раза дольше чем один шаг обучения (прямой и обратный проходы) и приблизительно в четыре раза дольше чем один шаг применения.

Глава 6

Послесловие

6.1. Заключение

Главным результатом данной работы стал градиентный метод для поиска аномалий в данных, превосходящий прочие рассмотренные решения на датасетах CIFAR-10 и ImageNet-O. Разработанный метод обладает массой интересных свойств и, также, может быть применён к другим задачам, связанным с оценкой неопределённости предсказания, и для интерпретации нейронных сетей.

В рамках данной работы были получены следующие результаты:

- Разработан градиентный метод для поиска аномалий в данных.
- Приведено сравнение результатов работы предложенного метода с прочими актуальными решениями в области поиска аномалий в данных на датасетах: CIFAR-10, ImageNet-O, ImageNet-R.
- Проведены эксперименты по анализу работы метода, в частности исследованы:
  - влияние популярного в исследуемой области метода препроцессинга, основанного на ε-возмущении исходного прецедента;
  - влияние особенностей архитектуры сети ResNet на работу метода;
  - устойчивость гиперпараметров препроцессинга к смене архитектуры нейронной сети;
  - масштабируемость метода на задачи большего размера;
  - оценка надёжности предсказания (робастности) в исходной задаче с помощью нормы градиента;

- Проведён теоретический анализ работы метода, показывающий, в чём принципиальная разница между градиентным методом и ODIN.

6.2. Потенциальные негативные последствия исследования

Важно отметить, что результаты данного исследования могут нести негативные последствия для общества, с которыми необходимо всеми силами бороться, например:

- разработка оружия; Системы поиска аномалий могут повысить точность методов машинного обучения, использующихся в современном вооружении: самонаводящихся ракетах, дронах и прочих автономных системах вооружения.

- системы распознавания лиц, ставящие целью слежку за людьми; Наличие outlier данных на этапе применения затрудняет работу алгоритмов для распознавания лиц. Методы оценки неопределённости предсказания и поиска аномалий могут быть напрямую интегрированы в такие системы с целью улучшения качества.

- adversarial атаки; Изучение неопределённости предсказания и поиска аномалий в данных может быть использовано для улучшения существующих и разработки новых методов adversarial атак на современные системы машинного обучения.

Чтобы минимизировать риски использования научных открытий в злых корыстных целях требуется развивать направление защиты от adversarial атак и систем слежки, а также максимально ограничивать развитие вооружения на всех уровнях.

6.3. Направление для дальнейшей работы

Данное исследование является первым подробным исследованием градиентных подходов к поиску аномалий в данных, поэтому оно имеет большое число потенциальных направлений для дальнейшей работы, например:

- применение предложенных идей к текстовым и речевым данным;
- исследование влияния поворота и нормировки градиента перед взятием нормы в градиентном скоринге; В частности, добавление шага Ньютона-Рафсона.
- замена нормы, как функции градиентного скоринга, на некоторую другую статистику;
- разработка метода поиска аномалий, основанного на проверке гипотез с использованием градиентного скоринга;
- исследование по применимости предложенного метода к бесконечно широким нейронным сетям;
- связь процесса обучения нейронной сети и теории информации с поиском аномалий;
- модификация препроцессинга, вдохновлённая, например, задачей adversarial examples;
- подбор гиперпараметров для препроцессинга без использования outlier данных;
- исследование качества работы метода в защите от adversarial атак;
- исследование других сдвигов распределения и применение метода к другим датасетам с аномалиями;
- более подробное исследование масштабируемости предложенного метода;

- комбинирование градиентного скоринга с ансамблевыми методами.

Список литературы

1. Jacot Arthur, Gabriel Franck, Hongler Clément. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. — 2020. — 1806.07572.
2. Koh Pang Wei, Liang Percy. Understanding Black-box Predictions via Influence Functions. — 2020. — 1703.04730.
3. Hendrycks Dan, Gimpel Kevin. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. — 2018. — 1610.02136.
4. Liang Shiyu, Li Yixuan, Srikant R. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. — 2020. — 1706.02690.
5. Hendrycks Dan, Basart Steven, Mazeika Mantas et al. Scaling Out-of-Distribution Detection for Real-World Settings. — 2020. — 1911.11132.
6. Malinin Andrey, Mlodozeniec Bruno, Gales Mark. Ensemble Distribution Distillation. — 2019. — 1905.00076.
7. Uncertainty Estimation in Autoregressive Structured Prediction. — 2021. — 2002.07650.
8. Lee Jinsol, AlRegib Ghassan. Gradients as a Measure of Uncertainty in Neural Networks. — 2020. — 2008.08030.
9. He Kaiming, Zhang Xiangyu, Ren Shaoqing, Sun Jian. Deep Residual Learning for Image Recognition. — 2015. — 1512.03385.
10. Hochreiter Sepp. The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions // International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. — 1998. — 04. — Vol. 6. — P. 107–116.
11. Krizhevsky A., Hinton G. Learning multiple layers of features from tiny images // Master's thesis, Department of Computer Science, University of Toronto. — 2009.
12. ImageNet Large Scale Visual Recognition Challenge / Olga Russakovsky, Jia Deng, Hao Su et al. // International Journal of Computer Vision (IJCV). — 2015. — Vol. 115, no. 3. — P. 211–252.
13. Hendrycks Dan, Zhao Kevin, Basart Steven et al. Natural Adversarial Examples. — 2021. — 1907.07174.

14. Hendrycks Dan, Basart Steven, Mu Norman et al. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. — 2020. — 2006.16241.
15. Zagoruyko Sergey, Komodakis Nikos. Wide Residual Networks. — 2017. — 1605.07146.
16. Huang Gao, Liu Zhuang, van der Maaten Laurens, Weinberger Kilian Q. Densely Connected Convolutional Networks. — 2018. — 1608.06993.
17. https://pytorch.org/vision/stable/models.html.
18. Lu Zhiyun, Ie Eugene, Sha Fei. Uncertainty Estimation with Infinitesimal Jackknife, Its Distribution and Mean-Field Approximation. — 2020. — 2006.07584.
19. Rumelhart David E., Hinton Geoffrey E., Williams Ronald J. Learning Representations by Back-propagating Errors // Nature. — 1986. — Vol. 323, no. 6088. — P. 533–536. — Access mode: `http://www.nature.com/articles/323533a0`.