МОСКОВСКИЙ ФИЗИКО -ТЕХНИЧЕСКИЙ ИНСТИТУТ
(НАЦИОНАЛЬНО-ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)
ФИЗТЕХ ШКОЛА ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ
ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР ИМ. А. А. ДОРОДНИЦЫНА РАН

ОТЗЫВ РЕЦЕНЗЕНТА
на выпускную квалификационную работу
Шокорова Вячеслава Александровича
«Бустинг глубоких нейросетевых ансамблей»

Вячеслав Александрович в своей работе предложил метод построения бустинга для глубоких нейросетевых моделей, причем решив проблему с тем, что данные модели достигают нулевой ошибки на обучении. Таким образом он предоставил расширенное понимание алгоритма бустинга, которое показало повышение точность по сравнению с базовым решением.

Введенная дисконтированная функция потерь показывает значительное преимущество перед функцией потерь кросс-энтропией и имеет значительный потенциал для дальнейшего развития и изучения.

Из замечаний к работе можно отметить, что требование повышения генерализации, выдвинутое к дисконтированной функции потерь, описано довольно поверхностно. Приведена скудная доказательная база. Также, отмечу, что существуют и другие методы оценки ширины минимума, например, через Гессиан, которые не были упомянуты и описаны в работе.

Проведенные исследования, полнота их описания, теоретические доказательства, подкрепленные экспериментально, показывают, что Вячеслав Александрович выполнил работу на высоком уровне. Актуальность работы высока для всех областей, где используются нейросетевые подходы для задачи классификации, что при текущем уровне популярность данных методов встречается повсеместно. Считаю, что автор достоин присвоения звания магистра, а работа заслуживает оценки 9 (девять) баллов.

Оценка рецензента за ВКР: 9 (девять) баллов.

Рецензент:
к.ф.-м.н. Исаченко Роман Владимирович
12 июня 2023 г.

МОСКОВСКИЙ ФИЗИКО -ТЕХНИЧЕСКИЙ ИНСТИТУТ
(НАЦИОНАЛЬНО-ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)
ФИЗТЕХ ШКОЛА ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ
ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР ИМ. А. А. ДОРОДНИЦЫНА РАН

ОТЗЫВ НАУЧНОГО РУКОВОДИТЕЛЯ
на выпускную квалификационную работу
Шокорова Вячеслава Александровича
«Бустинг глубоких нейросетевых ансамблей»

Направление подготовки / специальность:	03.04.01 Прикладные математика и физика
Направление (профиль) подготовки:	Математическая физика, компьютерные технологии и математическое моделирование в экономике

Задача построения эффективных моделей распознавания изображений с использованием глубоких нейронных сетей будет оставаться актуальной еще долгое время. Вячеслав строит свою работу на использовании такого метода как ассамблирование, который позволяет простым образом повысить итоговое качество алгоритма. Автор, для построения ансамбля, предлагает использовать бустинг на основе марджина нейронной сети, с использованием дисконтированной функции потерь.

Вячеслав провел теоретические исследование, которое доказывает корректность введеной функции потерь, а также доказал это утверждение экспериментально. В работе было рассмотрены различные методы построения бустинга и продемонстрировано их преимущество перед базовым решением.

Стоит отметить, что было потрачено значительное количество времени на то, чтобы сформировать итоговую архитектуру, которая дальше использовалась в экспериментах. Считаю, что автор показал себя с лучше стороны, доведя предложенную концепцию до рабочего алгоритма.

Я рекомендую данную работу к защите, и считаю, что ее автор заслуживает присвоения степени магистра по направлению 03.04.01 «Математическая физика, компьютерные технологии и математическое моделирование в экономике».

Рекомендуемая оценка: 10 баллов (отлично)

Научный руководитель:
к.ф.-м.н. Ветров Дмитрий Петрович

12 июня 2023 г.

1. Введение:

Задача классификации

Рассматривается задача классификации изображений с помощью глубоких нейронных сетей. Модель нейронной сети представляется как последовательность параметрических нелинейных преобразований. Параметры подбираются с помощью решения задачи минимизации функции потерь.

Перепараметризованные сети -> много минимумов. Интуитивное описание почему будет много минимумов (сделать свои картинки)

Проблема такого подхода заключается в том, что размерность входных данных и число объектов обучающей выборки намного меньше, чем размерность пространства параметров нейронной сети. Это значит, что существует множество решений оптимизационной задачи (минимумов). Допустим два скрытых слоя сети представляются, как линейное преобразование с некоторой функцией нелинейности. Тогда можно изменить порядок соответствующих нейронов в этих двух слоях так, чтобы предсказание сети никак не изменилось, но фактически параметры сети получились другие. Таким образом получили вектор параметров, отличный от изначального, но, на котором достигается минимум функционала. Также, если в сети присутствуют нормировки, например, как нормировка по батчу (BatchNorm), то умножение параметров сети на константу не меняет предсказание модели. Более того, можно умножать параметры каждого слоя независимо. Это лишь некоторые примеры, которые показывают неединственность решения оптимизационной задачи. Они не меняют качество модели, потому что сохраняют сеть в том же классе эквивалентности (отношением эквивалентности является совпадение предсказания сети), что и изначальная сеть. И стоит отметить, что таких классов эквивалентности много.

Измерение ширины минимума

Помимо того, что в пространстве параметром модели существует множество глобальных оптимумов, но также, они качественно отличаются друг от друга, например по ширине, по качеству генерализации и тд. Для оценки генерализации модели используется метрика точности модели на тестовой выборке. Для оценки ширины минимума предлагается использовать среднюю норму стохастического градиента по обучающей выборке.

Описание масштаб-инвариантных сетей. Как в этом случае измеряем ширину (норма градиента)

Техника нормировки по батчу (BatchNorm), описанная в статье [https://arxiv.org/pdf/1502.03167.pdf] используется для укорения и стабилизации обучения модели. Предлагается дополнительно добавлять внутрь глубокой нейронной сеть промежуточные слои нормализации, которые приводят разнородные данные к

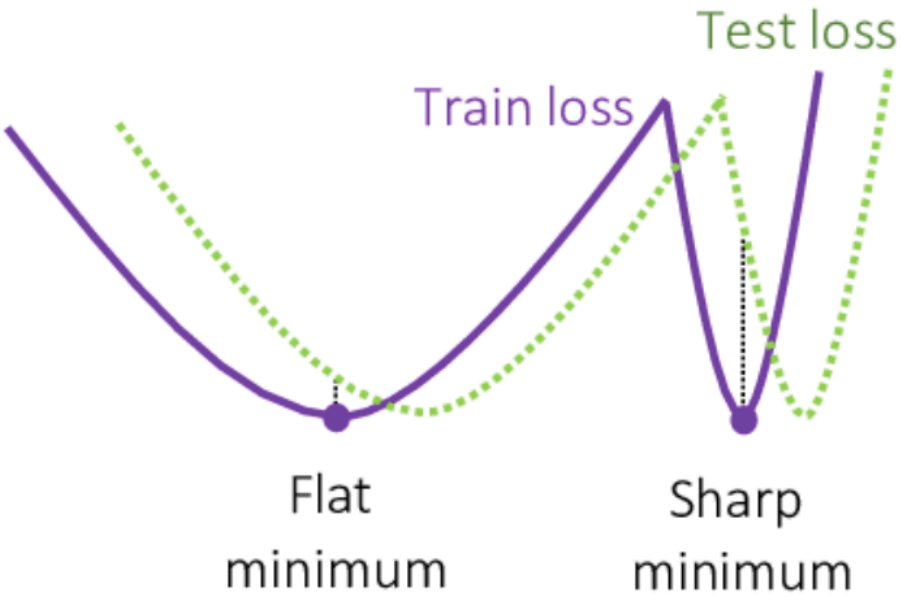
единому виду, то есть к нулевому мат. ожиданию и единичной дисперсии. Таким образом метод позволяет использовать гораздо более высокие темпы обучения (lr) и быть менее аккуратными при начальной инициализации весов.

Но такие промежуточные нормализующие слои порождают дополнительные сложности в анализе поведения модели, в частности они делают нейронные сети масштаб-инвариантными, то есть умножение весов на некоторую константу не меняет предсказание модели. Таким образом две сети, отличающиеся только умножением на константу будут иметь одинаковую точность на тесте, но разную ширину минимума. Чтобы компенсировать этот эффект предлагается перед подсчетом шириной минимума приводить веса сети к единичной норме.

Описание генерализации, картинка из PAC обучаемости говорим, что широкие минимумы лучше.

Для дальнейшего изложения введем понятие зазор генерализации (generalization gap), он показывает разницу качества модели на обучающей и тестовой выборке.

В научном сообществе существует гипотеза [.....], что широкие минимумы обладают большей генерализацией. На рис _ дано интуитивное объяснение этого предположения, на нем изображена кривая функции потерь при линейной интерполяции между двумя минимумами в пространстве весов. причем интерполяция происходит между широким и узким минимумом. Так как поверхность функции потерь на тестовых данных будет немного смещена, это значит, что у широкого минимума зазор генерализации будет меньше, а сама генерализация будет выше, чем у узкого минимума. Таким образом данный эффект основывается на том, что узкие минимумы менее гладкие, поэтому небольшое изменение в данных ведет к сильной просадке качества модели.



Flat minima results in better generalization compared to sharp minima. Pruning neural ODEs flattens the loss around local minima. Figure is reproduced from Keskar et al. (2017).

Например, в [что-то про … https://arxiv.org/pdf/2006.15081.pdf] авторы визуализировали с помощью t-SNE

В [статья про ширину минимумов, например https://arxiv.org/pdf/1906.03291.pdf] авторы исследовали данную проблему, показали, что различные минимумы обладают различной генерализацией (обобщающей способность т.е. качеством на тестовой выборке), а широкие минимумы обладают лучшей генерализацией, чем узкие. Связывают это с тем, что узкие минимумы менее гладкие, поэтому небольшое изменение в данных ведет к сильной просадке качества модели. Например, известно, что оптимизатор SGD [] избегает узких минимумов [].

Что такое марджин, и как он может влиять на генерализацию модели - ???

Глубокие ансамбли:

Глубокие ансамбли, предложенные Lakshminarayanan и др. [[2017], представляют из себя ансамбль глубоких нейронных сетей обученных из различных начальных инициализаций. В [https://arxiv.org/pdf/1912.02757.pdf] авторы эмперически показали, что глубокие ансамбли являются хорошим подходом для повышения точности, за счет того, что модели захватывают различные моды в пространстве весов, в то время как вариационные байесовские методы, как правило, фокусируются на одной моде. Различные, ортогональные веса ансамблей позволяют в совокупности более аккуратно выучивать обучающую выборку.

Бустинг - метод построения ансамбля, основанный на идеи, что каждая следующая обучаемая модели стремится компенсировать ошибку всех предыдущих моделей. Но такой способ применим для слабых алгоритмов, т.е. таких, которые не могут правильно распознать всю обучающую выборку. Глубокие нейронные сети могут обучиться до стопроцентной точности, поэтому классические бустинги как [adaboost, gradboost и тд] для нейронных сетей неприменимы. Мы предлагаем строить бустинг на глубоких нейронных сетях используя марджин в качестве ошибки, которую будут компенсировать следующие модели.

В данной работе предлагается метод регуляризации модели (дисконтированная функция потерь), который сглаживает поверхность функции потерь, тем самым остаются только широкие минимумы. Также рассматривается ансамблирование моделей, обученных с помощью предлагаемого методы и различные эвристики бустинга.

Описать еще план, и плюсы работы, что ввели и рассмотрели, ключевые утверждения

1.1. Обзор литературы

Работы про ширину минимума:

Описание SAM

Схожие методы:

Сравнение Дисконтированного лосса с L2-constrained SoftMax

В задаче верификации лиц популярны так называемые metric Learning методы. Их суть заключается в том, что используется модель, получающая векторное представление изображения. В качестве меры сходства изображений используют косинусное расстояние. Во время обучения модель поощряют изображениям одного класса строить векторные представления более похожие друг на друга (косинусное сходство должно быть большое), а вектора изображений из разных классов должны быть далеки друг от друга. Таким образом строиться функция потерь, которая отражает данную логику.

В статье [https://arxiv.org/pdf/1703.09507.pdf] авторы ввыли дополнительно дополнительные обучаемые параметры, W, которые описывают центроиды соответствующих классов. Таким образом предположение о том, что изображения одного класса должны быть похожи переформулируется в виде того, что вектора соответствующие изображениям одного класса должны быть близки к одному определенному вектору и далеки от всех остальных векторов матрицы W.

Затем исследователи заметили, что можно использовать дополнительный гиперпараметр m, который описывает, насколько вектора соответствующие изображениям одного класса должны быть ближе к одному определенному вектору, чем близость к всем остальным векторам матрицы W. Данная идея была реализована, например, в [CosFace: https://arxiv.org/pdf/1801.09414.pdf]. Итоговая функция потерь описывается как:

$$L_{lmc} = \frac{1}{N}\sum_{i} -\log\frac{e^{s(\cos(\theta_{y_i,i})-m)}}{e^{s(\cos(\theta_{y_i,i})-m)} + \sum_{j\neq y_i} e^{s\cos(\theta_{j,i})}}, \tag{4}$$

subject to

$$\begin{aligned} W &= \frac{W^*}{\|W^*\|}, \\ x &= \frac{x^*}{\|x^*\|}, \\ cos(\theta_j, i) &= {W_j}^T x_i, \end{aligned} \tag{5}$$

Данная процедура имеет очень похожий вид, что и предлагаемый дисконтированная функция потерь, но имеет иную идею.

2. Теоретическая часть

Постановка задачи, введение алгебры, переменных и тд

Алгоритм обучения одной модели

Марджин

Описание дисконтированного лосса

Теорема 1: Пусть модель обучена с дисконтированной функцией потерь, с заданным значением гэпа g, до 100% точности на обучающей выборке. Тогда на всех объектах обучающей выборки зазор (марджин) для данной модели будет не меньше, чем g.

Теорема 2: Существует максимальное значение гэпа, при обучении на котором, модель может показывать 100% точности на обучающей выборке.

- ~~Попробовать связать гэп и марджин как априорную и апостериорную вероятность~~

Измерение ширины минимума (средняя норма стох градиента)

Обучение ансамбля и эвристики построения ансамбля, эвристики построения ансамбля, алгоритм бустинга моделей

- ~~Почему ансамбль это хорошо? - теория (возможно не стоит сюда вставлять)~~

3. Вычислительные эксперименты

Сделать игрушечный тест предназначенный для визуализации того, что дисконтированный лосс работает!

Как себя ведет лосс с гэпом в оптимумах

Связать норму градиента и марджин и другие методы оценки генерализации

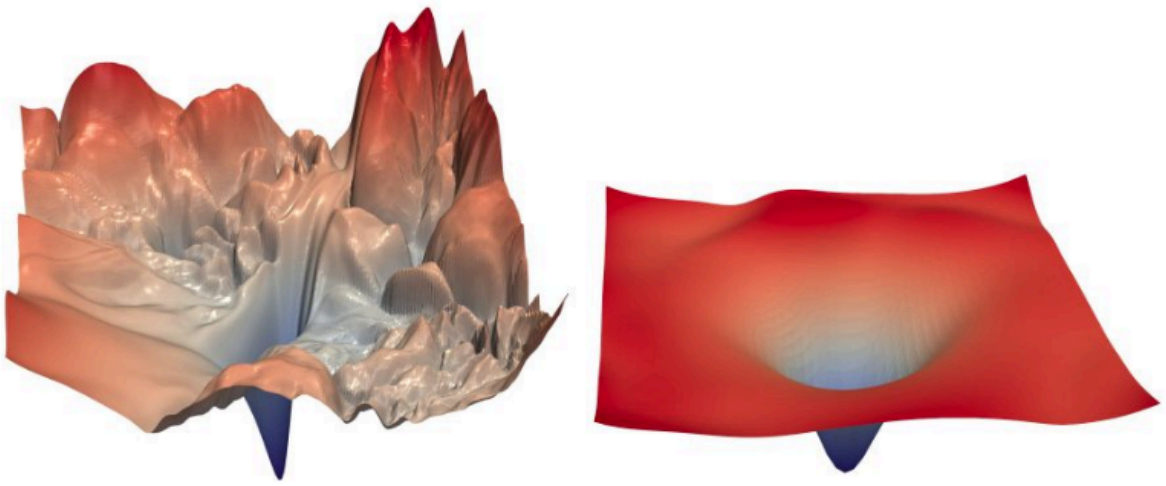
Замораживаем последний слой, обучаем модель, смотрим что происходит

- ~~Верно ли, что задавая гэп мы контролируем марджин. Какое-то теоретической описание придумать. Попробовать предсказывать корреляцию для марджина и гэпа (мб взять простые, линейноразделимые данные и лог. регрессию)~~

Точность дисконтированного лоса при различных гэпах

Сравнение ансамблей

4. Выводы



https://www.researchgate.net/figure/Flat-minima-results-in-better-generalization-compared-to-sharp-minima-Pruning-neural_fig2_353068686