Федеральное государственное автономное образовательное учреждение
высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»
Физтех-школа Прикладной Математики и Информатики
Кафедра интеллектуальных систем

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

БУСТИНГ ГЛУБОКИХ НЕЙРОСЕТЕВЫХ АНСАМБЛЕЙ

(магистерская диссертация)

Студент:
Шокоров Вячеслав Александрович

_____
(подпись студента)

Научный руководитель:
Ветров Дмитрий Петрович,
канд. физ.-мат. наук

_____
(подпись научного руководителя)

Консультант *(при наличии)*:
_____
_____
(подпись консультанта)

Москва 2023

Аннотация

Решается задача классификации, рассматривается метод бустинга глубоких нейросетевых моделей. Классические методы бустинга неприменимы для глубоких нейросетей из-за того, что каждая модель достигает нулевой ошибки на обучающей выборке, следовательно невозможно обучить следующую модель компенсировать ошибки предыдущих моделей, чего требует большинство алгоритмов бустинга.

В работе вводится дисконтированная функция потерь, которая решает две задачи: повышение генерализации каждой модели и добавление гибкости в обучении ансамбля, то есть создание возможности учитывать ошибки моделей. Последнее требование выполняется за счет оценки уверенности модели в предсказании, данная оценка называется марджин. Дисконтированная функция потерь позволяет обучать модель с предсказуемым значением марджина.

Проводится вычислительный эксперимент, на датасетах CIFAR10 и CIFAR100, где предлагаемый метод бустинга превосходит базовое решение, классический глубокий ансамбль через усреднение предсказаний, до 3.2%.

Ключевые слова: *Бустинг, глубокие нейронные сети, повышение генерализации модели.*

Содержание

1 Введение

Рассматривается задача построения ансамбля для классификации изображений с помощью глубоких нейронных сетей. Метод бустинга, предложенный в [16], превосходит по качеству классические подходы построения ансамбля, например когда предсказание ансамбля является усредненным предсказанием всех моделей ансамбля. Идея метода заключается в том, что модели обучаются последовательно, а каждая новая модель учится компенсировать ошибки предыдущих моделей ансамбля. Такая идея хорошо работает со слабыми моделями, то есть с теми, которые не показывают нулевую ошибку на обучении. Современные нейронные сети способны распознавать обучающую выборку с 100%-ой точностью, поэтому методы бустинга не применим для данного типа моделей.

В данной работе вводится дисконтированная функция потерь (1), рассматриваются ее свойства, а также исследуются методы построения бустинга с использованием введенной функцией потерь (в качестве ошибки моделей, используется марджин (4)). Данная функция называется дисконтированной функцией потерь, ключевой задачей которой является обучение модели для получения предсказуемого значения марджина. Данные требование необходимо для применения идеи бустинга.

1.1 Описание дисконтированной функции потерь

Модель нейронной сети представляется как последовательность параметрических нелинейных преобразований, где параметры подбираются с помощью решения задачи минимизации функции потерь.

Проблема данного 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). Например, два скрытых слоя сети представляются, как линейные преобразования 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 тоже достигается минимум функционала. Также у нейронных сетей, которые имеют нормировку по батчу [7] (BatchNorm) в своей архитектуре, наблюдается масштабная инвариантность [9, 13]. Это значит, что умножение параметров сети на ненулевую константу не меняет предсказание модели. Таким образом, в пространстве весов существует луч (вектор), при движении вдоль которого, выход модели никак не изменяется. Это лишь некоторые примеры, которые показывают неединственность решения оптимизационной задачи.

Помимо того, что в пространстве параметром модели существует множество глобальных оптимумов, они также качественно отличаются друг от друга, например

по ширине или 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.

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

В научном сообществе существует гипотеза [8, 1, 14, 12, 19, 6], что широкие минимумы обладают большей генерализацией. На рис. 1 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 (обобщающей способность, то есть качеством на тестовой выборке) 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. Например, в [6] авторы, используя эксперименты и визуализации, исследовали связь между обобщающей способностью модели и геометрией функции потерь. Также известно, что оптимизатор SGD [17] избегает узких минимумов [16, 6], поэтому глубокие нейронные сети показывают хорошую обобщающую способность.

Рис. 1: Широкие минимумы имеют лучшую генерализацию, по сравнению с узкими.

Определение 1 *Функционал $L(f, d)$ назовем дисконтированной функцией потерь, d - значение дисконта:*

$$L(f,d) = \frac{1}{N}\sum_i -\log \frac{e^{X_{i,y_i}-d}}{e^{X_{i,y_i}-d} + \sum_{j\neq y_i} e^{X_{i,j}}}, \tag{1}$$

Причем при $d = 0$, то есть функционал $L(f, 0)$ превращается в функцию потерь кросс-энтропии.

Идея добавления дополнительного гиперпараметра “дисконт” заключается в том, что поверхность функции потерь по всей выборке это среднее значение функций потерь на каждом из объектов. Предложенный функционал “смещает” кривую, соответствующую одному объекта в сторону точки оптимума, на рис. 2 это изображено с помощью стрелок. Поэтому, если в пространстве весов модели существует узкий глобальный минимум для кросс-энтропии, то для дисконтированной функции потерь эта точка будет локальным минимумом.

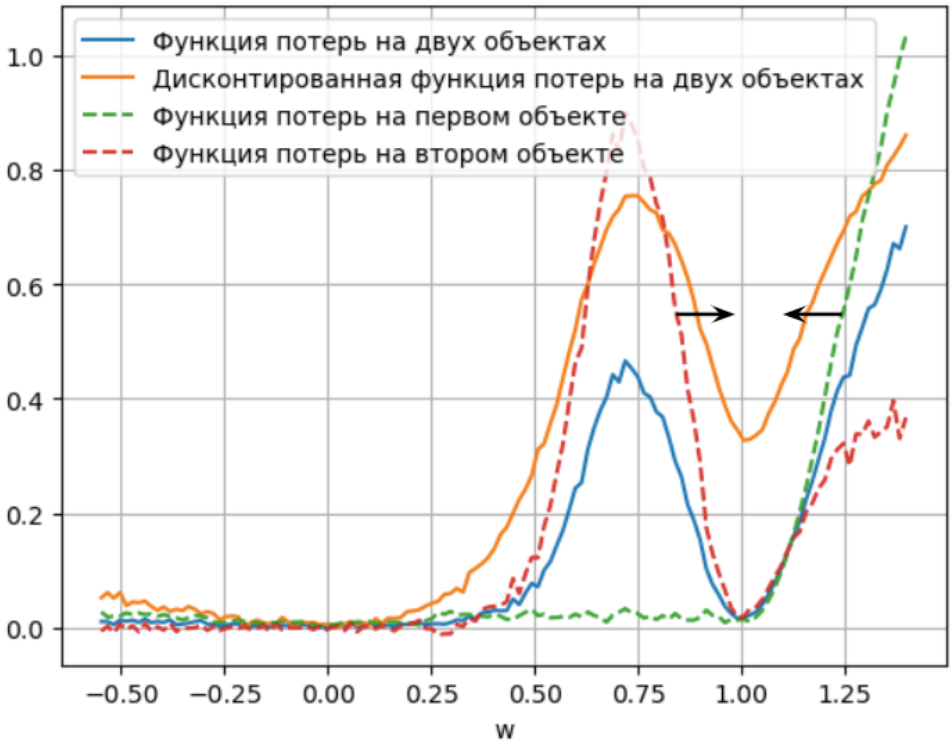

Рис. 2: Схематичное описание поведения дисконтированной функции потерь. Пунктирные линии - кривые функции потерь на определенных объектах, сплошные линии - функция потерь на объектах в совокупности. Стрелками изображен эффект дисконтированной функции потерь, то есть сдвиг, вызванный добавлением дисконта, кривой от каждого объекта в сторону минимума.

1.2 Описание глубоких нейросетевых ансамблей

Глубокие ансамбли, предложенные в [11], представляют собой ансамбль глубоких нейронных сетей обученных из различных начальных инициализаций. В [2] авторы эмпирически показали, что глубокие ансамбли являются хорошим подходом для повышения точности, за счет того, что модели захватывают различные моды в про-

странстве 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 данный способ применим для слабых алгоритмов, то есть таких, которые не могут правильно распознать всю обучающую выборку. Глубокие нейронные сети могут обучиться до стопроцентной точности, поэтому классические бустинги, как AdaBoost [3], GradBoost[4] и подобные, для нейронных сетей неприменимы. Предлагается 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.

1.3 Обзор литературы

Стоит отметить, что существует подход, схожий с дисконтированной функцией потерь по своей структуре. В задаче верификации лиц популярны, так называемые, metric Learning методы. Их суть заключается в том, что используется модель, которая строит векторное представление изображения. В качестве меры сходства изображений используется косинусное расстояние. Во время обучения модель поощряется изображениям одного класса строить более похожие друг на друга векторные представления (косинусное сходство должно быть большое), а вектора изображений из разных классов должны быть далеки друг от друга. Таким образом, строится функция потерь, которая отражает данную логику.

В статье [15] авторы ввели 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 предложили использовать дополнительный гиперпараметр m, который описывает, насколько вектора соответствующие изображениям одного класса должны быть ближе к определенному центроиду, чем близость к всем остальным векторам матрицы W. Данная идея была реализована, например, в [18]. Итоговая функция потерь описывается как:

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

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

где x - векторное представление изображения, W - матрица весов, соответствующая векторам центроидов. Данная процедура имеет очень похожий вид, что и предлагаемый дисконтированная функция потерь, но имеет абсолютно иную идею.

2 Теоретическая часть

Для дальнейшего повествования введем математические описание.

В качестве нейронной сети понимается параметрическая функция $f(x, \theta)$, где x - входное изображение, $\theta \in \mathbb{R}^t$ - вектор параметров модели. Причем используется масштабно-инвариантная сеть, то есть $\forall \alpha > 0 \ \forall x \mapsto f(x, \alpha\theta) = f(x, \theta)$. $f(\cdot, \theta) : \mathbb{R}^{\mathrm{h}\times\mathrm{w}\times 3} \to \mathbb{R}^{\mathrm{K}}$, где K - число классов классификации. $\mathcal{X} = \{(x, y) | x \in \mathbb{R}^{\mathrm{h}\times\mathrm{w}\times 3}, y - \text{меток истинного класса}\}$ - множество объектов (изображений).

Функция $f(x, \theta)$ представляет собой последовательность параметрических преобразований (для экспериментов используется архитектура ResNet [5]), причем последние слои представляют собой слой нормировки по батчу [7] и линейный слой.

Определение 2 *Нормировкой по батчу (BN)* $\mathbb{R}^{n\times m} \to \mathbb{R}^{n\times m}$ *назовется преобразование вида:*

$$\mathrm{BN}(x) = \frac{x - \mathbb{E}[x]}{\mathrm{Var}[x]} \tag{2}$$

$\mathbb{E}$ *и* Var *берется поканально, то есть вдоль второй размерности входной матрицы* x*.*

Таким образом:

$$f(\cdot, \theta) = W \circ \mathrm{BN} \circ \ldots \tag{3}$$

Определение 3 *Марджином модели* $f(\cdot, \theta)$ *для* $(x, y) \sim \mathcal{X}$ *назовем* m*:*

$$m(\theta, x, y) = f(\cdot, \theta)_y - \max_{i \neq y} f(\cdot, \theta)_i \tag{4}$$

Для краткости обозначим $m(\theta, x, y) = m_i$*, если* x *это* i*-ый объект множества* $\mathcal{X}$*.*

Средний марджин модели $f(\cdot, \theta)$ на множестве $\mathcal{X}$ соответственно:

$$\bar{m} = \frac{1}{|\mathcal{X}|} \sum_{(x,y)\sim\mathcal{X}} m(\theta, x, y)$$

Определение 4 *Функцией SoftMax назовем преобразование из* $\mathbb{R}^n$ *в единичный симплекс на* $\mathbb{R}^n$ *вида:*

$$SoftMax(\mathbf{v})_i = \frac{e^{\mathbf{v}_i}}{\sum_j e^{\mathbf{v}_j}}$$

На практике $\mathbf{v}_i$ не может равняться минус бесконечности, поэтому каждая координата вектора $SoftMax(\mathbf{v}) > 0$, сумма всех координат равна единице. Таким образом, без ограничений на область определения логарифма, можем ввести эквивалентное определение дисконтированной функции потерь:

Определение 5 *Функционал $L(f, d)$ назовем дисконтированной функцией потерь, d - значение дисконта:*

$$L(f,d) = \frac{1}{|\mathcal{X}|} \sum_{(x,y)\sim\mathcal{X}} -\log\Big(SoftMax\big(f(x) - \text{onehot}_y \cdot d\big)\Big)_y,$$

где onehot_y *- вектор нулей, с единицей на y-ой координате.*

Для оценки качества модели предлагается использовать две метрики: точность на тестовой выборке и оценку ширины минимума.

Определение 6 *Точностью (Acc.) предсказания модели f на множестве объектов $\mathcal{X}$ назовем:*

$$Acc(f,\mathcal{X}) = \frac{1}{|\mathcal{X}|} \sum_{(x,y)\sim\mathcal{X}} \text{I}\{\arg\max(f(x)) = y\},$$

где I *- индикаторная функция, принимает значение* 1 *при истинном значении аргумента.*

Заметим, что если $m(\theta, x, y) > 0 \Leftrightarrow \text{I}\{\arg\max(f(x)) = y\} = 1$

Как описывалось ранее, для оценки ширины минимума предлагается использовать среднюю норму стохастического градиента, то есть:

$$\frac{1}{|\mathcal{X}|} \sum_{(x,y)\sim\mathcal{X}} \Big\|\nabla_\theta - \log SoftMax\big(f(x,\bar{\theta})\big)_y\Big\|, \tag{5}$$

где $\bar{\theta} = \theta/\|\theta\|$

Лемма 1 *Пусть дана функция $f(x)$ вида (3), и $f \not\equiv 0$ тогда $\forall g > 0$ и $\alpha > 0$ верно:*

$$\lim_{\alpha\to\infty} |L(\alpha f, d) - L(\alpha f, 0)| = 0$$

Доказательство. Обозначим $\mathbf{v} = f(x)$, тогда для доказательства достаточно показать, что

$$\left|\frac{e^{\alpha\mathbf{v}_y - g}}{e^{\alpha\mathbf{v}_y - g} + \sum_{j\neq y} e^{\alpha\mathbf{v}_j}} - \frac{e^{\alpha\mathbf{v}_y}}{e^{\alpha\mathbf{v}_y} + \sum_{j\neq y} e^{\alpha\mathbf{v}_j}}\right| \xrightarrow[\alpha\to\infty]{} 0 \tag{6}$$

С учетом того, что $g/\alpha \to 0$, при $\alpha \to \infty$, перепишем требуемое 6 и докажем утверждение:

$$\left|\frac{e^{\mathbf{v}_y - g/\alpha}}{e^{\mathbf{v}_y - g/\alpha} + \sum_{j\neq y} e^{\mathbf{v}_j}} - \frac{e^{\mathbf{v}_y}}{e^{\mathbf{v}_y} + \sum_{j\neq y} e^{\mathbf{v}_j}}\right| \xrightarrow[\alpha\to\infty]{} 0$$

■

Из описанной выше леммы видно, что при искусственном увеличении нормы весов последнего слоя W из (3) уменьшается вклад, который достигается дисконтированной функцией потерь при ненулевом значении дисконта. Для компенсации данного эффекта предлагается замораживать веса последнего слоя. Такая модификация нейронной сети используется во всех последующих рассуждениях и экспериментах.

2.1 Анализ свойств дисконтированной функции потерь

Для дальнейшего изучения свойств дисконтированной функции потерь, докажем несколько теорем. Покажем, что существует максимальное значение марджина, который может показывать модель, и как следствие получим, что существует предельное значение дисконта в дисконтированной функции потерь, после которого все функции потерь ведут себя эквивалентно, то есть существует предельное значение функции.

Для случая бинарной классификации существует достижимая верхняя оценка на максимальное значение марджина.

Лемма 2 *Пусть дана функция $f(x)$ вида (3), с замороженными W - весами последнего слоя. $\mathcal{X} = \{(x, y)|x \in \mathbb{R}^{\mathrm{h}\times\mathrm{w}\times 3}, y \in \{-1, 1\}\}$ - множество объектов бинарной классификации, причем выборка равновесна, то есть $\mathrm{P}_{(x,y)\sim\mathcal{X}}(y = 1) = 1/2$. Тогда максимальное значение среднего марджина, которое может получить функция:*

$$\bar{m} = \|W_{[:,0]} - W_{[:,1]}\|_1 \leq \|W\|_1$$

Доказательство.

Так как решается задача бинарной классификации, следовательно $W \in \mathbb{R}^{d\times 2}$. Введем вектор $\mathbf{w} \in \mathbb{R}^d$, $\mathbf{w}_j = W_{j,0} - W_{j,1}$. Также верно, что маржин из выражения (4) переписывается как $m_i = f(x_i)_{y_i} - f(x_i)_{-y_i} \Rightarrow m_i = y_i \cdot X_i^T\mathbf{w}$.

Пусть X это матрица предсказаний функции $f(x)$ до применения последнего слоя на множестве объектов $\mathcal{X}$. Для доказательства леммы сформулируем оптимизационную задачу:

$$\frac{1}{N}\sum_i y_i \cdot X_i^T\mathbf{w} \to \max \tag{7}$$

$$\forall j \mapsto \ \frac{1}{N}\sum_i X_{ij} = 0, \tag{8}$$

$$\forall j \mapsto \ \frac{1}{N}\sum_i X_{ij}^2 = 1, \tag{9}$$

Где $N = |\mathcal{X}|$ - размер выборки, выражение (7) описывает максимальное среднее значение марджина, (8) добавляет ограничение на нулевое математическое ожидание, а (9) на единичную дисперсию. Такие ограничения появляются из-за того, что X является выходом слоя BN.

Для аналитического решения воспользуемся условием Каруша-Куна-Таккера, введем двойственные переменные $\mu, \nu \in \mathbb{R}^d$:

$$\nabla_X \Big[\sum_i \big(- y_i \cdot X_i^T \mathbf{w} \big) + \sum_j \mu_j \sum_i X_{ij} + \sum_j \nu_j \big(\sum_i X_{ij}^2 - c \big) \Big] = 0$$

Упростим выражение и сгруппируем слагаемые:

$$\nabla_X \sum_{i,j} \big(\nu_j X_{ij}^2 + X_{ij}(\mu_j - y_i \mathbf{w}_j) - \nu_j c \big) = 0$$

Поэтому:

$$X_{ij}^* = \frac{y_i \mathbf{w}_j - \mu_j}{2\nu_j} \tag{10}$$

Подставим полученное значение X^* из (10) в (8):

$\forall j \mapsto \frac{1}{N} \sum_i \frac{y_i \mathbf{w}_j - \mu_j}{2\nu_J} = 0 \Rightarrow \mathbf{w}_j \sum_i y_i = N \cdot \mu_j \Rightarrow \mu_j = \mathbf{w}_j \cdot \frac{\sum_i y_i}{N} = 0$, так как выборка $\mathcal{X}$ равновесна.

Подставим полученное значение X^* из (10) в (9):

$\forall j \mapsto \frac{1}{N} \sum_i \big(\frac{y_i \mathbf{w}_j - \mu_j}{2\nu_J} \big)^2 = 1 \Rightarrow \nu_j = \frac{|\mathbf{w}_j|}{2} \sqrt{\frac{\sum_i y_i^2}{N}} = \frac{|\mathbf{w}_j|}{2}$.

Таким образом, максимальное значение среднего марджина получается после подстановки всех найденных значений в (7):

$$\frac{1}{N} \sum_i y_i \cdot X_i^T \mathbf{w} = \frac{1}{N} \sum_{ij} y_i X_{ij} \mathbf{w}_j = \frac{1}{N} \sum_{ij} y_i \frac{y_i \mathbf{w}_j}{|\mathbf{w}_j|} \mathbf{w}_j = \sum_j |\mathbf{w}_j| = \|\mathbf{w}\|_1$$

■

Теорема 3 *(О максимальном значении среднего марджина)*
Пусть дана функция $f(x)$ вида (3), с замороженными W - весами последнего слоя. $\mathcal{X} = \{(x, y) | x \in \mathbb{R}^{\mathrm{h} \times \mathrm{w} \times 3}, y \in \{1, 2 \dots K\}\}$ - множество объектов многоклассовой классификации, причем выборка равновесна, то есть $\mathrm{P}_{(x,y) \sim \mathcal{X}}(y = k) = 1/K$. Тогда максимальное значение среднего марджина, которое может получить функция:

$$\bar{m} \leq \frac{2}{K} \|W\|_1.$$

Доказательство.

Для каждого класса зададим пару, например, для класса k парным классом будет $k' = k + 1 \mod K$. Таким образом:

$$m_i = f(x)_k - \max_{j \neq k} f(x)_j \leq f(x)_k - f(x)_{k'},$$

Из доказательства леммы 2 известно, что $f(x)_k - f(x)_{k'} \leq \|W_{[:,k]} - W_{[:,k']}\|_1$. Таким образом:

$$\bar{m} = \frac{1}{N}\sum_i f(x_i)_{y_i} - \max_{j \neq y_i} f(x_i)_j \leq$$

$$\frac{1}{N}\sum_i f(x_i)_{y_i} - f(x_i)_{y_i+1 \mod K} \leq$$

$$\frac{1}{N}\sum_i \|W_{[:,y_i]} - W_{[:,y_i+1 \mod K]}\|_1 =$$

$$\frac{1}{N \cdot K}\sum_i \|W_{[:,1]} - W_{[:,2]}\|_1 + \cdots + \|W_{[:,K-1]} - W_{[:,K]}\|_1 + \|W_{[:,K]} - W_{[:,1]}\|_1 \leq$$

$$\frac{2}{N \cdot K}\sum_{i,k} \|W_{[:,k]}\|_1 = \frac{2}{K}\|W\|_1$$

При последнем переходе используется неравенство треугольника.

■

Также отметим, что из-за того, что если модель обучилась до стопроцентной точности на обучающей выборке, значит марджин на каждом объекте больше нуля. Так как среднее значения марджина ограничено (теорема 3), следовательно и максимальный марджин также ограничен.

Из теоремы 3 следует, что средний марджин, который получает модель ограничен, поэтому следующая теорема имеет смысл:

Теорема 4 *(О существовании предельной функции дисконтированной функции потерь)*
Для достаточно больших d_1 и d_2, для некоторой функции $f(x)$ вида (3), с замороженными W - весами последнего слоя и множества объектов $\mathcal{X}$ верно, что существует предел:

$$\exists \lim_{d \to \infty} L(f, d) - d < \infty \tag{11}$$

Доказательство.

Пусть X это матрица предсказаний функции $f(x)$ на множестве объектов $\mathcal{X}$, то есть $X_i = f(x_i)$ для $(x_i, y_i) \sim \mathcal{X}$, $N = |\mathcal{X}|$, тогда выражение $L(f, d)$ переписывается как:

$$L(f,d) = \frac{1}{N}\sum_i -\log\frac{e^{X_{y_i,i}-d}}{e^{X_{y_i,i}-d} + \sum_{j\neq y_i} e^{X_{j,i}}}$$

Для упрощения обозначим $\sum_{j\neq y_i} e^{X_{j,i}} = c_i$, тогда:

$$L(f,d) - d = \frac{1}{N}\sum_i -\log\frac{e^{X_{y_i,i}-d}}{e^{X_{y_i,i}-d} + c_i} - d = \\ = \frac{1}{N}\sum_i [d - X_{y_i,i} + \log(e^{X_{y_i,i}-d} + c_i) - d] \quad (12)$$

Рассмотрим отдельное слагаемое из всей суммы:

$$\log(e^{X_{y_i,i}-d} + c_i) - X_{y_i,i} = \log(c_i) - \log\left(\frac{e^{X_{y_i,i}}}{c_i}e^{-d} + 1\right) - X_{y_i,i}$$

Очевидно, что если $d_1 > d_2 \Rightarrow 0 \leq \log\left(\frac{e^{X_{y_i,i}}}{c_i}e^{-d_1} + 1\right) < \log\left(\frac{e^{X_{y_i,i}}}{c_i}e^{-d_2} + 1\right) \to 0$ для $d_1 \to \infty$.

Таким образом предельное значение:

$$\lim_{d\to\infty} L(f,d) - d = \frac{1}{N}\sum_i \left[\log\left(\sum_{j\neq y_i} e^{X_{j,i}}\right) - X_{y_i,i}\right]$$

■

Также отметим, что значение дисконта, после которого выражение 12 перестает значительно меняться (меньше чем машинный эпсилон) сопоставимо по максимальным значением марджина. А оно известно из теоремы 3. Из теоремы 4 следует, что при достаточно больших значениях дисконта дисконтированные функции потерь совпадают с точностью до константы. Отсюда важным следствием является, что в данном случае градиенты у таких функций совпадают, а значит и алгоритм стохастического градиентного спуска [17] будет обновлять веса эквивалентным образом, независимо от значения дисконта.

2.2 Бустинг

Определение 7 *Ансамблем моделей $f(x)$ вида (3) называется множество весов моделей $\Theta = \{\theta^i\}_{i=1}^T$, где T - размер ансамбля.*

Предсказание ансамбля $\{\theta^i\}_{i=1}^T$ - усреднённое предсказание моделей.

$$y_{pred} = \frac{1}{M}\sum_{i=1}^{M} SoftMax\big(f(x,\theta^i)\big)$$

Как было описано ранее, предлагается использовать дисконтированную функцию потерь для построения ансамбля. Для оценки ошибки модели на каждом объекте из обучающей выборке используется значение марджина. Таким образом, даже при правильной классификации данного объекта, сохраняется возможность оценить качество распознавания объекта, потому что чем больше марджин, тем больше правдоподобие данного объекта с точки зрения модели и соответственно уверенность модели в предсказании.

В задаче обучения ансамбля марджин и дисконт берутся по каждому объекту, то есть $\mathbf{m}, \mathbf{d} \in \mathbb{R}^{|\mathcal{X}|}$. Общая схема построения ансамбля с помощью бустинга описана в алгоритме 1. На вход алгоритм принимает обучающую выборку $\mathcal{X}$, требуемый размер ансамбля T, функцию, которая по истории обученных моделей генерирует значения дисконта для следующей обучаемой модели get_next_discount$(\cdot)$ и гиперпараметр λ. Детальное описание функции get_next_discount$(\cdot)$ будет дано дальше.

Algorithm 1 Алгоритм бустинга для построения ансамбля

Input: $\mathcal{X}$-множество объектов для обучения, T-размер ансамбля, get_next_discount$(\cdot)$-функция генерирующая следующее значение дисконта, λ.
Output: $\{\hat{\theta}^1, \hat{\theta}^2, \ldots, \hat{\theta}^T\}$.
$\mathbf{d} \leftarrow$ get_next_discount$(\mathbf{0})$ // Начальное значение дисконта для каждого объекта
$\mathcal{M} \leftarrow \{\}$ // Множество значений марджинов для моделей ансамбля
for $t \leftarrow 1$ to T **do**
    $\hat{\theta}^t \leftarrow \arg\min_\theta L(f(\cdot,\theta), \mathbf{d}) + \lambda\|\theta\|_2^2$ // см. выражение (13)
    $\mathcal{M} \leftarrow \mathcal{M} \cup \{m(\hat{\theta}^t)\}$
    $\mathbf{d} \leftarrow$ get_next_discount$(\mathcal{M})$
end for

Основная задача функции get_next_discount$(\cdot)$ заключается в построении значения дисконта для следующей модели. На вход подается множество значений марджина моделей, которые уже есть в ансамбле. Так как ключевая задача бустинга это обучить следующую модель компенсировать ошибки предыдущих. Поэтому используется эвристика, что объекты, которые имеют маленький марджин, то есть плохо распознаются моделью, должны иметь большой дисконт, чтобы следующая модель училась распознавать данные объекты лучше.

Методы, рассматриваемые в работе, для построения данной функции описаны в таблице 1.

Название	Формула	Текстовое описание
д-т.0	0	Дисконт равен 0 для любого объекта для каждой модели.
д-т.34	34	Дисконт равен 34 для любого объекта для каждой модели.
д-т.34_по_послед.	$34 + \frac{1}{N}\sum_{j=1}(\mathcal{M}_{T,j}) - \mathcal{M}_{T,i}$	Средний дисконт по всем объектам равен 34, дисконт строиться по последней модели.
д-т.34_ср.кум.	$34 + \frac{1}{T}\sum_t \left(\frac{1}{N}\sum_j \mathcal{M}_{t,j} - \mathcal{M}_{t,i}\right)$	Средний дисконт по всем объектам равен 34, дисконт строиться по среднему марджину всех моделей.
д-т.ср.мар.	$\frac{1}{N}\sum_j \mathcal{M}_{T,j}$	Дисконт по всем объектам равен среднему марджину последней модели.
д-т.сохр.ср.	$\frac{2}{N}\sum_{j=1}(\mathcal{M}_{T,j}) - \mathcal{M}_{T,i}$	Средний дисконт по всем объектам равен среднему марджину, дисконт строиться по последней модели.

Таблица 1: Различные методы построения функции get_next_discount$(\cdot)$. Для краткости, в столбце формула указывается только правая часть равенства, левая соответственно: get_next_discount$(\mathcal{M})_i$. Значение дисконта 34 выбрано эмпирически, как значение на котором достигается максимальная точность при обучении с дисконтированной функцией потерь.

2.3 Постановка задачи обучения модели

Оптимальные параметры нейронной сети f подбираются методом стохастического градиентного спуска [17] при минимизации:

$$L(f(\cdot,\theta),d) + \lambda\|\theta\|_2^2 \to \min_\theta, \tag{13}$$

где λ - коэффициент L_2 регуляризации весов ($\|\theta\|_2^2$).

3 Вычислительный эксперимент

Вычислительный эксперимент проводился на базе данных CIFAR10 [10] - набор изображений разбитых на 10 классов.

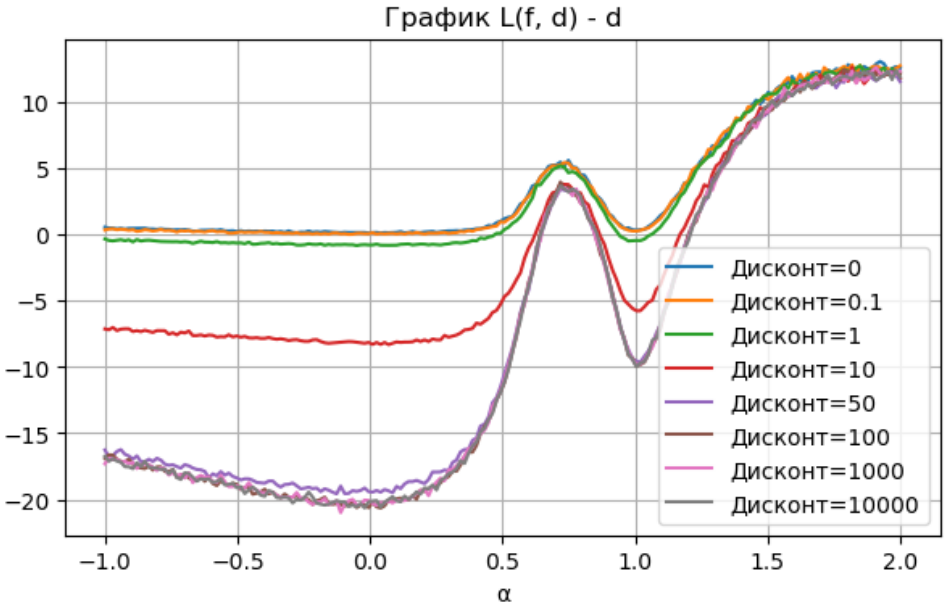

Рис. 3: График $L(f, d) - d$ для различных значений дисконта, при движении вдоль некоторого направления в пространстве весов. Данное направлено проходит через два глобальных минимума, каждый получен с через (13). Шаг вдоль этого направления описывает коэффициент α.

На рис. 3 экспериментально подтверждается теорема 4. Видно, что при значении дисконта больше 100 все кривые дисконтированных функций потерь совпадают. Отметим, что узкий минимум больше не является глобальным минимумом, таким образом, подтверждаются требования выдвинутые для дисконтированной функции потерь. А также видно, что значение производной предельной дисконтированной функции потерь больше, чем у функции при нулевом дисконте. Это значит, что оптимальный темп обучения, который используется в методе стохастического градиентного спуска для минимизации (13) будут отличаться для различных значений дисконта.

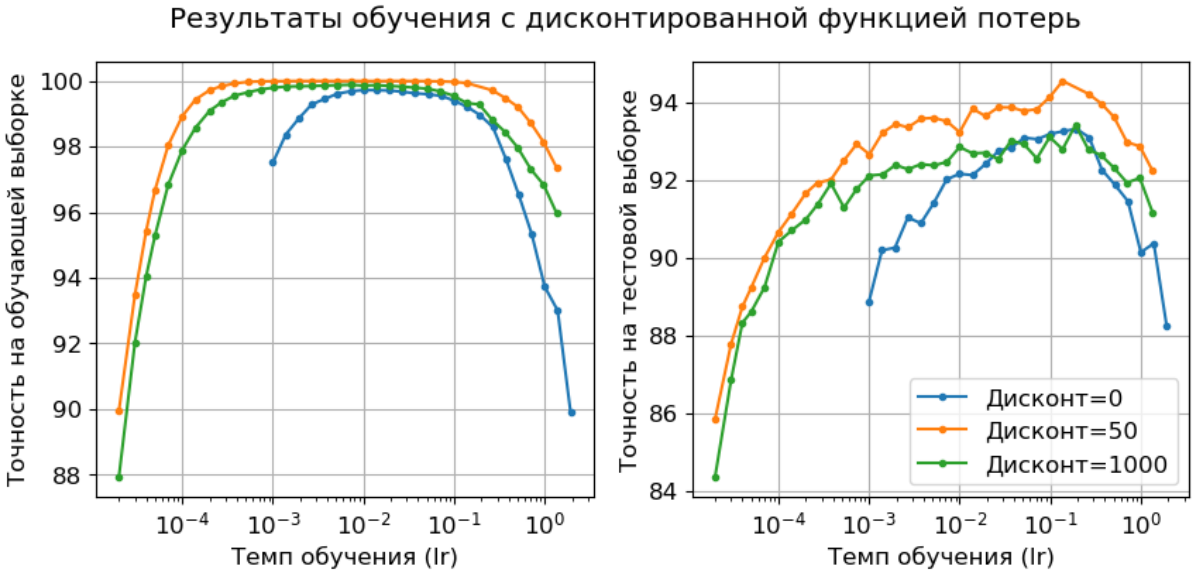

Рис. 4: Качество модели полученной при минимизации (13) для различных темпов обучения и различных значений дисконта.

На рис. 4 изображено качество моделей на тестовой выборке, обученных с различными темпами обучения. Видно, что модели, полученные при ненулевом дисконте, имеют преимущество перед моделями, обученными с нулевым дисконтом.

Дисконт	Темп обучения	Ср. норма градиента
1000	0.00721	116.3
50	0.01924	11.9
0	0.01	22.9

Таблица 2: Оценка ширина минимума на обучающей выборке (5).

Для оценки ширины минимума были взять значения темпов обучения на которых достигалось максимальная точность на обучающей выборке (для каждого значения дисконта свое значение темпа обучения). Результаты показаны в таблице 2. Ширина минимума измерялась как средняя норма градиента по объектам выборки для нормированных весов модели, см. (5).

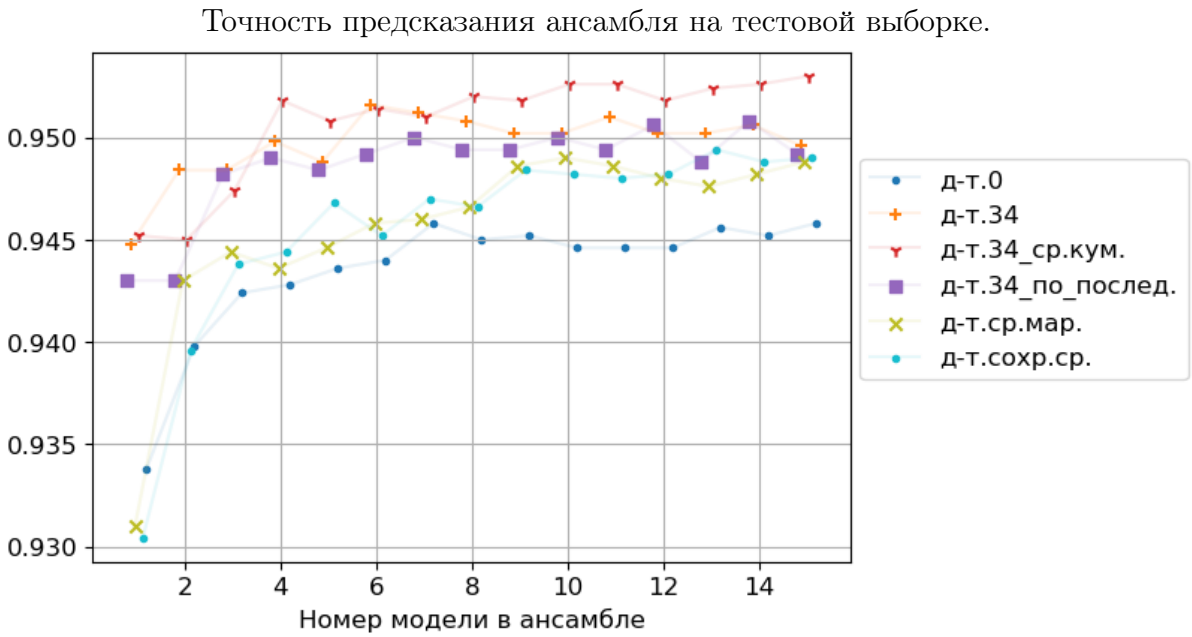

Рис. 5: Сравнение различных вариантов бустинга и ансамблей через усреднение. Детальное описание методов дано в таблице 1. Эксперимент проводился на датасете CIFAR10 [10].

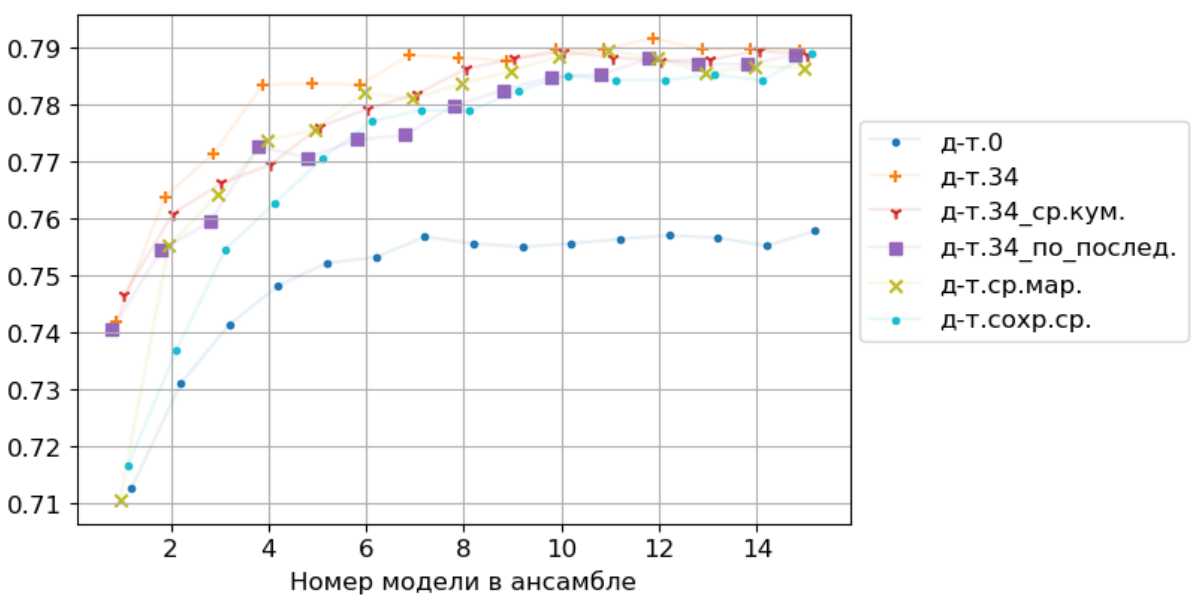

Рис. 6: Сравнение различных вариантов бустинга и ансамблей через усреднение. Детальное описание методов дано в таблице 1. Эксперимент проводился на датасете CIFAR100 [10].

4 Заключение

В работе предложен метод бустинга ансамбля глубоких нейросетевых моделей, с помощью дисконтированной функции потерь. Теоретически показаны свойства дисконтированной функции потерь, а также экспериментально продемонстрировано ее преимущество перед классической функцией потерь, кросс-энтропией. Также показано, что добавление дисконта повышает генерализацию модели, позволяет модели сходиться в более широкий минимум, ускоряет обучение и балансирует выборку, то есть дисперсия марджина на обучающей выборке становится меньше, даже при увеличении среднего значения марджина.

Рассмотрены различные эвристики бустинга. Показано, что предлагаемый алгоритм превосходит построение ансамбля через усреднение по качеству на тестовой выборке.

На рис. 4 продемонстрировано преимущество дисконтированной функции потерь, однако неочевидно, почему дисконт 1000, показывает себя хуже, чем дисконт 50. Данный эффект планируется исследовать в дальнейшем.

Список литературы

[1] *Dinh L. et al.* Sharp minima can generalize for deep nets, International Conference on Machine Learning. – PMLR, 2017. – С. 1019-1028.

[2] *Fort S., Hu H., Lakshminarayanan B.* Deep ensembles: A loss landscape perspective, arXiv preprint arXiv:1912.02757. – 2019.

[3] *Freund Y., Schapire R. E.* A decision-theoretic generalization of on-line learning and an application to boosting, Journal of computer and system sciences. – 1997. – Т. 55. – №. 1. – С. 119-139.

[4] *Friedman J. H.* Greedy function approximation: a gradient boosting machine, Annals of statistics. – 2001. – С. 1189-1232.

[5] *He K. et al.* Deep residual learning for image recognition, Proceedings of the IEEE conference on computer vision and pattern recognition. – 2016. – С. 770-778.

[6] *Huang W. R. et al.*

[7] *Ioffe S., Szegedy C.* Batch normalization: Accelerating deep network training by reducing internal covariate shift, International conference on machine learning. – pmlr, 2015. – С. 448-456.

[8] *Kaddour J. et al.* Questions for flat-minima optimization of modern neural networks, arXiv e-prints. – 2022. – С. arXiv: 2202.00661.

[9] *Kodryan M. et al.* Training Scale-Invariant Neural Networks on the Sphere Can Happen in Three Regimes, Advances in Neural Information Processing Systems. – 2022. – Т. 35. – С. 14058-14070.

[10] *Krizhevsky A. et al.* Learning multiple layers of features from tiny images. – 2009.

[11] *Lakshminarayanan B., Pritzel A., Blundell C.* Simple and scalable predictive uncertainty estimation using deep ensembles, Advances in neural information processing systems. – 2017. – Т. 30.

[12] *Li H. et al.* Visualizing the loss landscape of neural nets, Advances in neural information processing systems. – 2018. – Т. 31.

[13] *Lobacheva E. et al.* On the periodic behavior of neural network training with batch normalization and weight decay, Advances in Neural Information Processing Systems. – 2021. – Т. 34. – С. 21545-21556.

[14] *Neyshabur B. et al.* Exploring generalization in deep learning, Advances in neural information processing systems. – 2017. – Т. 30.

[15] *Ranjan R., Castillo C. D., Chellappa R.* L2-constrained softmax loss for discriminative face verification, arXiv preprint arXiv:1703.09507. – 2017.

[16] *Schapire R. E.* The boosting approach to machine learning: An overview, Nonlinear estimation and classification. – 2003. – C. 149-171.

[17] *Smith S., Elsen E., De S.* On the generalization benefit of noise in stochastic gradient descent, International Conference on Machine Learning. – PMLR, 2020. – C. 9058-9067.

[18] *Sutskever I. et al.* On the importance of initialization and momentum in deep learning, International conference on machine learning. – PMLR, 2013. – C. 1139-1147.

[19] *Wang H. et al.* Cosface: Large margin cosine loss for deep face recognition, Proceedings of the IEEE conference on computer vision and pattern recognition. – 2018. – C. 5265-5274.

[20] *Yang Y. et al.* Taxonomizing local versus global structure in neural network loss landscapes, Advances in Neural Information Processing Systems. – 2021. – T. 34. – C. 18722-18733.