Бустинг глубоких нейросетевых ансамблей

Шокоров Вячеслав Александрович

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Научный руководитель к.ф.-м.н. Д. П. Ветров

Москва,
2023 г.

План

План

Проблема генерализации модели в задаче классификации изображений

Проблема

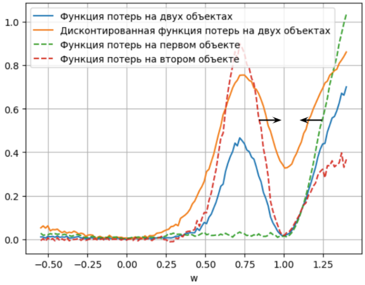
В пространстве весов глубокой, перепараметризованной нейронной сети существует множество точек глобального минимума функции потерь. Минимумы отличаются по генерализации.

Решение

Предлагается использовать дисконтированную функцию потерь, которая дополнительно штрафует модель за попадание в узкий минимум.

Сравнение узких и широких минимумов

(a) Визуализация преимущества широкого минимума: модель, которая сошлась в узкий минимум, будет иметь большую ошибку на тестовой выборке, чем модель, которая сошлась в широкий минимум.



(b) Эффект, который получаем от дисконтированной функции потерь, узкий, глобальный минимум перестает был глобальным.

Дисконтированная функция потерь

В качестве нейронной сети понимается параметрическая функция $f(x, \theta)$, где x - тензор изображения, $\theta \in \mathbb{R}^t$ - вектор параметров модели. $f(\cdot, \theta) : \mathbb{R}^{\mathrm{h} \times \mathrm{w} \times 3} \to \mathbb{R}^{\mathrm{K}}$, где K - число классов классификации.
Причем:

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

W - матрица весов последнего слоя, BN - слой нормировки по батчу.

Дисконтированная функция потерь

Функционал $L(f, g)$ назовем *дисконтированной функцией потерь*, g - значение гэпа:

$$L(f, g) = \frac{1}{N} \sum_{x_i} -\log \frac{e^{f(x_i)_{y_i} - g}}{e^{f(x_i)_{y_i} - g} + \sum_{j \neq y_i} e^{f(x_i)_j}},$$

Архитектура модели

Лемма 1

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

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

При увеличении нормы весов последнего слоя W уменьшается вклад, который достигается дисконтированной функцией потерь при ненулевом значении гэпа. Для компенсации данного эффекта предлагается замораживать веса последнего слоя. Такая модификация нейронной сети используется во всех последующих рассуждениях и экспериментах.

Марджин

Марджином модели $f(\cdot, \theta)$ назовем m:

$$m(\theta, x, y) = f(x, \theta)_y - \max_{j \neq y} f(x, \theta)_j$$

Теорема 1

Пусть дана функция $f(x)$, 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, т.е. $\mathrm{P}_{(x,y) \sim \mathcal{X}}(y = k) = 1/K$. Тогда максимальное значение среднего марджина, которое может получить функция:

$$\bar{m} = \frac{1}{|\mathcal{X}|} \sum_{(x,y) \sim \mathcal{X}} m(\theta, x, y) \leq \frac{2}{K} \|W\|_1.$$

Теорема 1

Для достаточно больших g_1 и g_2, для некоторой функции $f(x)$ с замороженными W - весами последнего слоя и множества объектов $\mathcal{X}$ верно:

$$L(f, g_1) - g_1 \approx L(f, g_2) - g_2$$

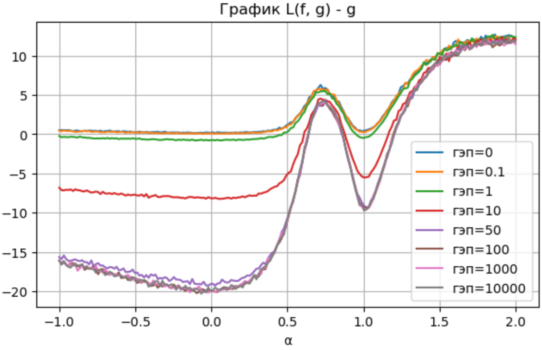
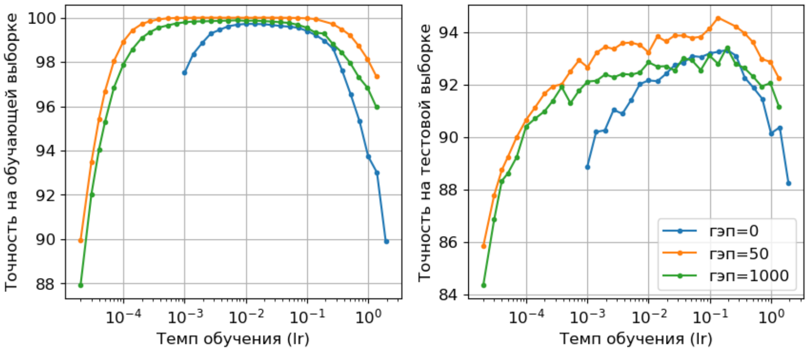


График $L(f, g) - g$ для различных значений гэпа, при движении вдоль некоторого вектора в пространстве весов. Шаг вдоль этого направления описывает коэффициент α.

Результаты обучения с дисконтированной функцией потерь



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

$$L(f(\cdot,\theta), g) + \lambda\|\theta\|_2^2 \to \min_{\theta}, \tag{1}$$

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

План

- **Цель:** получить ансамбль моделей, который будет лучше, чем усреднение предсказаний моделей ансамбля.
- **Идея:** обучать следующую модель компенсировать ошибки предыдущих моделей. В качестве оценки ошибки использовать марджин.

Марджин

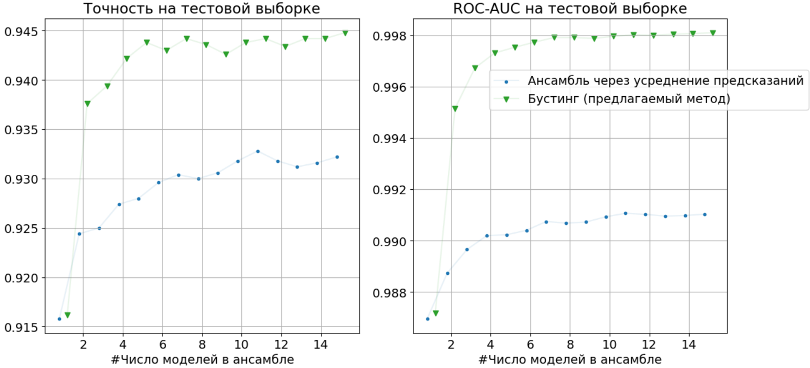
Марджином модели $f(\cdot, \theta)$ назовем m:

$$m(\theta, x, y) = f(x, \theta)_y - \max_{j \neq y} f(x, \theta)_j$$

Algorithm Алгоритм бустинга моделей

1: **Input:** $\mathcal{X}$-множество объектов для обучения, T-размер ансамбля, λ.
Output: $\{\hat{\theta}^1, \hat{\theta}^2, \dots, \hat{\theta}^T\}$.
$\mathbf{g}_i \leftarrow \mathbf{0}$ // Значение гэпа для каждого объекта
$\mathbf{m}_i \leftarrow \mathbf{0}$ // Суммарный марджин ансамбля для каждого объекта
2: **for** $t \leftarrow 1$ to T **do**
3: $\hat{\theta}^t \leftarrow \arg\min_\theta L(f(\cdot, \theta), \mathbf{g}) + \lambda\|\theta\|_2^2$
4: $\mathbf{m}_i \leftarrow \mathbf{m}_i + m(\hat{\theta}^t)_i$
5: $\mathbf{g}_i \leftarrow 2 \cdot \sum_j \mathbf{m}_j - \mathbf{m}_i$
6: **end for**

Вычислительный эксперимент



Эксперимент проводился на датасете CIFAR10, на архитектуре ResNet с замороженным последним слоем.

Анализ ошибки

...

Результаты, выносимые на защиту

- ...
- ...
- ...