Анализ сходимости оптимизационной поверхности сверточных нейросетевых моделей на основе Гессиана функции потерь

Владислав Сергеевич Мешков
Научный руководитель: к.ф.-м.н. А. В. Грабовой
Научный консультант: Н. С. Киселев

Кафедра интеллектуальных систем ФПМИ МФТИ
Специализация: Интеллектуальный анализ данных
Направление: 01.03.02 Прикладные математика и информатика

2025

Анализ сходимости оптимизационной поверхности сверточных нейросетевых моделей на основе Гессиана функции потерь

Проблема

Поверхность функции потерь сложным образом зависит от архитектуры нейронной сети.

Цель

Предложить оценку изменения функции потерь при изменении размера обучающей выборки.

Решение

Предлагается провести исследование

1. Рассмотреть абсолютное изменение функции потерь при добалении в выборку нового элемента;
2. Аппроксимировать функцию потерь с помощью аппроксимации Тейлора второго порядка.

Постановка задачи

Выборка

$$\mathfrak{D}_m = \{(\mathbf{x}_i, y_i)\}, i \in \mathcal{I} = \{1, \ldots, m\}$$

- $\mathbf{x} \in \mathbb{X} \subseteq \mathbb{R}^n$ — вектор признакового описания объекта;
- $y \in \mathbb{Y}$ — значение целевой переменной.

Модель

$f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathcal{Y}, \boldsymbol{\theta} \in \mathbb{R}^P$ - нейронная сеть, которая аппроксимирует условное распределение данных $p(\mathbf{y}|\mathbf{x})$

Функция потерь

$$\mathcal{L}_m(\boldsymbol{\theta}) = \frac{1}{m} \sum_{i=1}^{m} \ell(f_{\boldsymbol{\theta}}(\mathbf{x}_i), \mathbf{y}_i) \approx \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p(\mathbf{x},\mathbf{y})} \left[\ell(f_{\boldsymbol{\theta}}(\mathbf{x}), \mathbf{y})\right]$$

Изменение значения функции потерь при добавлении объекта

$$\mathcal{L}_{k+1} - \mathcal{L}_k = \frac{1}{k+1} \left(\ell(f_{\boldsymbol{\theta}}(\mathbf{x}_{k+1}), \mathbf{y}_{k+1}) - \mathcal{L}_k(\boldsymbol{\theta})\right)$$

Предположение о точке минимума

Предположение 1

Пусть $\boldsymbol{\theta}^*$ является точкой минимума обеих функций $\mathcal{L}_k(\boldsymbol{\theta})$ и $\mathcal{L}_{k+1}(\boldsymbol{\theta})$, то есть $\nabla\mathcal{L}_k(\boldsymbol{\theta}^*) = \nabla\mathcal{L}_{k+1}(\boldsymbol{\theta}^*) = \mathbf{0}$.

Аппроксимация второго порядка

$$\mathbf{H}^{(k)}(\boldsymbol{\theta}) = \nabla^2_{\boldsymbol{\theta}}\mathcal{L}_k(\boldsymbol{\theta}) = \frac{1}{k}\sum_{i=1}^{k}\nabla^2_{\boldsymbol{\theta}}\ell(f_{\boldsymbol{\theta}}(\mathbf{x}_i), \mathbf{y}_i) = \frac{1}{k}\sum_{i=1}^{k}\mathbf{H}_i(\boldsymbol{\theta})$$

$$\mathcal{L}_k(\boldsymbol{\theta}) \approx \mathcal{L}_k(\boldsymbol{\theta}^*) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\mathsf{T}\mathbf{H}^{(k)}(\boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$$

Абсолютное изменение функции потерь

$$|\mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta})| \leqslant \frac{1}{k+1}\left|\ell(f_{\boldsymbol{\theta}^*}(\mathbf{x}_{k+1}), \mathbf{y}_{k+1}) - \frac{1}{k}\sum_{i=1}^{k}\ell(f_{\boldsymbol{\theta}^*}(\mathbf{x}_i), \mathbf{y}_i)\right| +$$

$$+\frac{1}{k+1}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2\left\|\mathbf{H}_{k+1}(\boldsymbol{\theta}^*) - \frac{1}{k}\sum_{i=1}^{k}\mathbf{H}_i(\boldsymbol{\theta}^*)\right\|_2$$

Связь изменения функции потерь с матрицей Гессе

Абсолютное изменение функции потерь

$$|\mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta})| \leqslant \frac{2}{k+1} \max_{i=\overline{1,k+1}} |\ell(f_{\boldsymbol{\theta}^*}(\mathbf{x}_i, \mathbf{y}_i))| + \\ + \frac{2}{k+1} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2 \max_{i=\overline{1,k+1}} \left\|\mathbf{H}_i(\boldsymbol{\theta}^*)\right\|_2$$

Декомпозиция Гессиана

$$\mathbf{H}_i(\boldsymbol{\theta}) = \underbrace{\nabla_{\boldsymbol{\theta}} \mathbf{z}_i \frac{\partial^2 \ell(\mathbf{z}_i, \mathbf{y}_i)}{\partial \mathbf{z}_i^2} \nabla_{\boldsymbol{\theta}} \mathbf{z}_i^\intercal}_{\mathbf{H}_O} + \underbrace{\sum_{k=1}^{K} \frac{\partial \ell(\mathbf{z}_i, \mathbf{y}_i)}{\partial z_{ik}} \nabla_{\boldsymbol{\theta}}^2 z_{ik}}_{\mathbf{H}_F}$$

Аппроксимация Гессиана

- Аппроксимируем Гессиан, пренебрегая $\mathbf{H}_F$. В задаче K-классовой классификации $\|\mathbf{H}_F\| \ll \|\mathbf{H}_O\|$
- Тогда можно оценить $\|\mathbf{H}\| \approx \|\mathbf{H}_O\|$.

Структура $\mathbf{H}_O$ компоненты матрицы Гессе

Пусть нейросеть $f_{\boldsymbol{\theta}}(\mathbf{x})$ представляется в виде :
$f_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbf{T}^{(L+1)}\boldsymbol{\Lambda}^{(L)}\dots\boldsymbol{\Lambda}^{(1)}\mathbf{T}^{(1)}\mathbf{x}$, где $\mathbf{T}^{(l)}$-матрицы с параметрами из $\boldsymbol{\theta}$, $\boldsymbol{\Lambda}^{(l)}$-матрицы активации не содержащие параметров.
Рассмотрим матрицы:

- $\mathbf{F}$ — матрица состоящая из кронекеровских произведений частей нейросети $\mathbf{T}^{(i)}$ и $\boldsymbol{\Lambda}^{(i)}$.
- $\mathbf{A} = \operatorname{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\intercal$, где $\mathbf{p}$ — вектор вероятностей классов для $\mathbf{x}$
- $\mathbf{Q}^{(p)} := \frac{\partial \mathbf{T}^{(p)}}{\partial \mathbf{W}^{(p)}}$, где $\mathbf{W}^{(p)}$ — параметры p-го слоя.

Лемма 1
$\mathbf{H}_O(\boldsymbol{\theta}) = \mathbf{Q}^\intercal\mathbf{F}^\intercal\mathbf{A}\mathbf{F}\mathbf{Q}$.

Данная Лемма позволяет представить норму $\mathbf{H}_O$ компоненты Гессиана как произведение норм более простых блоков.

Оценка нормы матрицы Гессе

Лемма 2
Пусть $\left\|\mathbf{Q}^{(p)}\right\|_2 \leqslant q, \left\|\mathbf{T}^{(p)}\right\|^2 \leqslant w_{\mathbf{T}}^2 \ \ \forall p$, тогда
$\|\mathbf{H}_O\| \leqslant \sqrt{2} q^2 \|\mathbf{x}\|^2 (L+1) w_{\mathbf{T}}^{2L}$.

Оценка нормы $\mathbf{H}_O$ как функция весов является степенной, а как функция числа слоев — показательной.

Лемма 2 является основой для оценки нормы гессиана в дальнейшем.

Из Леммы, для того, чтобы оценить Гессиан, достаточно оценить $\left\|\mathbf{Q}^{(p)}\right\|$ и $\left\|\mathbf{T}^{(p)}\right\|$ одновременно для всех слоев, что и будет проделано в будущих результатах.

Норма матрицы Гессе сверточных нейронных сетей

Теорема 2

(о верхней границе нормы Гессиана сверточных сетей) Пусть $|\mathbf{W}^{(p)}_{i,j,k,t}|^2 \leqslant w^2$, где $\mathbf{W}^{(p)}$ — веса p-го слоя свртки $C_l, k_l, (m_l, n_l)$ — число каналов, размер ядра, пространственные размеры карты признаков соответственно на l-м слое нейросети. Пусть $C_l \leqslant C,\ \ k_i \leqslant k,\ m_i \leqslant m,\ \ n_i \leqslant n$, тогда

$$\|\mathbf{H}_O\| \leqslant \sqrt{2}\,\|x\|^2\, C^2k^2mn(L+1)(C^2k^2w^2mn)^L.$$

Данный результат показывает как именно различные параметры сверточных сетей влияют на оценку нормы матрицы Гессе, в частности оценка является показательной функцией числа слоев, в то же время степенной функцией от числа каналов.

Пулинги и полносвязная голова

Лемма 3

Пусть на месте l-й нелинейности находится max/avg пулинг, причем пусть ядро: $k_{\text{pool}} \times k_{\text{pool}}$, $\text{stride} = k_{\text{pool}}, \text{padding} = 0$, при этом верны все ограничения предыдущей теоремы, тогда:

$$\|\mathbf{H}_O\| \leqslant \sqrt{2}\,\|x\|^2\, q^2 \frac{1}{k_{\text{pool}}^{2(L-l+2)}} (L+1)(C^2k^2w^2mn)^L$$

Лемма 4

Пусть после сверточных слоев нахожится полносвязная голова размера P с не более чем h нейронами в скрытом слое:

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbf{T}^{(L+P+1)}\boldsymbol{\Lambda}^{(L+P)} \dots \boldsymbol{\Lambda}^{(L+1)}\mathbf{T}^{(L+1)}\boldsymbol{\Lambda}^{(L)} \dots \boldsymbol{\Lambda}^{(1)}\mathbf{T}^{(1)}\mathbf{x},$$

А также пусть имеет место ограничения Теоремы 2, и все веса в полносвязной голове не превосходят $\tilde{w}$, тогда имеет место оценка:

$$\|\mathbf{H}_O\| \leqslant \sqrt{2}\,\|\mathbf{x}\|^2\, C^2k^2mn\left(h^2\tilde{w}^2\right)^P \left(k^2C^2w^2mn\right)^L \times$$
$$\left(L+1+P\frac{h^2\tilde{w}^2}{k^2C^2w^2mn}\right).$$

Выносится на защиту

1. Теоремы об оценке нормы гессиана произвольный сетей, представимых в виде произведения матриц.
2. Теоремы об оценке нормы Гессиана сверточных нейронных сетей общего вида.
3. Теоремы об оценке нормы Гессиана для сверточных нейронных сетей с полносвязной головой или слоем max pooling'a.

Публикации

1. V. Meshkov, N. Kiselev, A. Grabovoy. ConvNets Landscape Convergence: Hessian-Based Analysis of Matricized Networks // 2024 Ivannikov Ispras Open Conference (ISPRAS)