Анализ сходимости поверхности функции потерь сверхточных нейросетевых моделей на основе Гессиана

Владислав Мешков, Никита Киселев, Андрей Грабовой

Московский Физико-Технический Институт

Мотивация

Проблема

Число данных для обучения нейросетей, а также число параметров растет. Необходимо изучать связь между сложностью моделей и необходимым числом объектов в обучающей выборке.

Цель

Изучить сходимость поверхности функции потерь в пространстве параметров при изменении размера выборки для сетей, представленных в виде произведения зависимых от входа матриц.

Решение

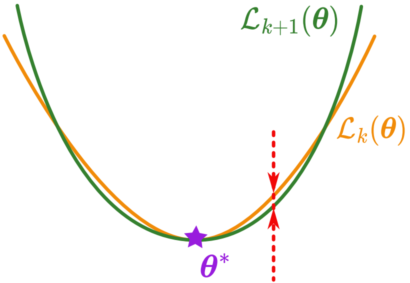
1. Рассмотреть взаимосвязь гессиана со сходимостью функции потерь.
2. Посчитать норму гессиана для сетей, представленных в виде произведения матриц, и применить данные результат к сверточным сетям.

Постановка задачи

Пусть $f_{\boldsymbol{\theta}}$ — нейросеть, $\mathbf{x}_k$ — входы, $\mathbf{y}_k$ — one-hot метки классов, $\mathcal{L}_k$ — функция потерь на k первых объектах.

Изменение значения при добавлении одного объекта

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



Предположение 1

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

Аппроксимация второго порядка

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

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

Связь изменения функции потерь с Гессианом

Абсолютное изменение функции потерь

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

Декомпозиция Гессиана

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

Аппроксимация Гессиана

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

Нейросети представимые в виде произведения матриц

Рассмотрим нейросети специального вида, а именно представимые в виде произведения матриц, возможно зависимых от входа.
Пусть нейросеть $f_{\boldsymbol{\theta}}(\mathbf{x}) := \mathbf{T}^{(L+1)}\boldsymbol{\Lambda}^{(L)} \dots \boldsymbol{\Lambda}^{(1)}\mathbf{T}^{(1)}\mathbf{x}$, где

- $\mathbf{x}$ — вход
- $\mathbf{T}^{(l)}$ — матрица линейного слоя l
- $\boldsymbol{\Lambda}^{(l)}$ — матрица $ReLU$-активации l-го слоя, зависимая от входа $\mathbf{x}$

Структура $\mathbf{H}_O$ компоненты Гессиана

Рассмотрим матрицы:

- $\mathbf{F}$ — матрица сложной структуры, зависящая только от всех $\mathbf{T}^{(i)}, \mathbf{\Lambda}^{(i)}$ и от $\mathbf{x}$
- $\mathbf{A} = \mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^{\mathsf{T}}$, 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

$\mathbf{H}_O(\boldsymbol{\theta}) = \mathbf{Q}^{\mathsf{T}}\mathbf{F}^{\mathsf{T}}\mathbf{A}\mathbf{F}\mathbf{Q}$.

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

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.

Свертки как линейная операция

Действие свертки на $\mathbf{x} \in \mathbb{R}^{C \times d}$ представим линейным оператором, действующим на $\mathbf{x} \in \mathbb{R}^{Cd}$, причем с сохранением обозначений : $\mathbf{T}^{(l)} * \mathbf{x} \to \mathbf{T}^{(l)}\mathbf{x}$

$$\begin{pmatrix} t_1 & t_2 & t_3 \end{pmatrix} * \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} \to \begin{pmatrix} t_1 & t_2 & t_3 & 0 & 0 \\ 0 & t_1 & t_2 & t_3 & 0 \\ 0 & 0 & t_1 & t_2 & t_3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix}$$

Результаты для 1-D сверток

Теорема 1

Пусть $C_l \leqslant C,\ k_i \leqslant k,\ d_i \leqslant d,\ |W_{i,j,k}^{(p)}|^2 \leqslant w^2$, где C_l, k_l, d_l — число каналов, размер ядра и пространственный размер соответственно, тогда

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

Из Теоремы видно, что оценка нормы является степенной функцией от числа каналов, весов, размера ядра и размера последовательности

Результаты для 2-D сверток

Теорема 2

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

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

Результат полученный в **Теореме 1** отличается от данного лучшей оценкой $\left\|\mathbf{Q}^{(p)}\right\|$, это связано с различием в структуре Теплицевых матриц 1D и 2D сверток.

Результаты для пулингов и полносвязной головы

Лемма 3

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

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

Лемма 4

Пусть после сверточных слоев нахожится полносвязная голова размера P:

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

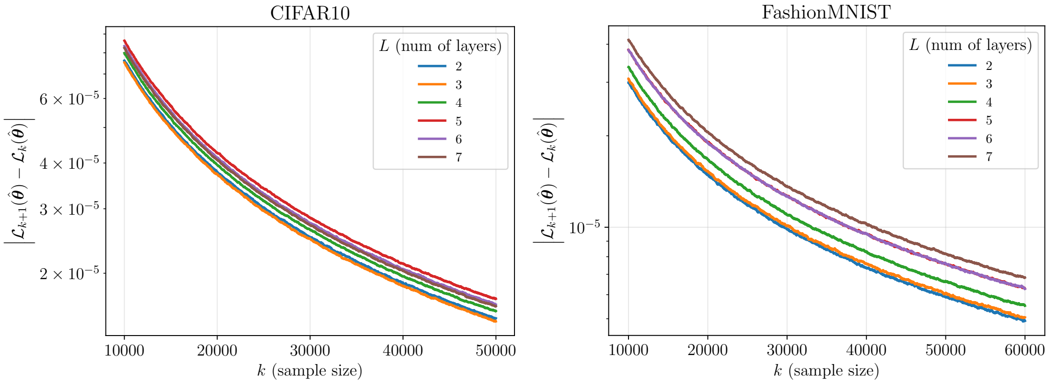
где $\mathbf{T}^{(L+1+i)}$ — Линейный слой с параметрами h_i, h_{i+1}, $\mathbf{T}^{(r)}$-2D-свертки. Пусть имеют место оценки: $\left\|\mathbf{T}_{ij}^{(L+1+i)}\right\| \leqslant \tilde{w}$ и $h_p \leqslant h$.Тогда в условиях Теоремы 2 имеем

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

Постановка эксперимента

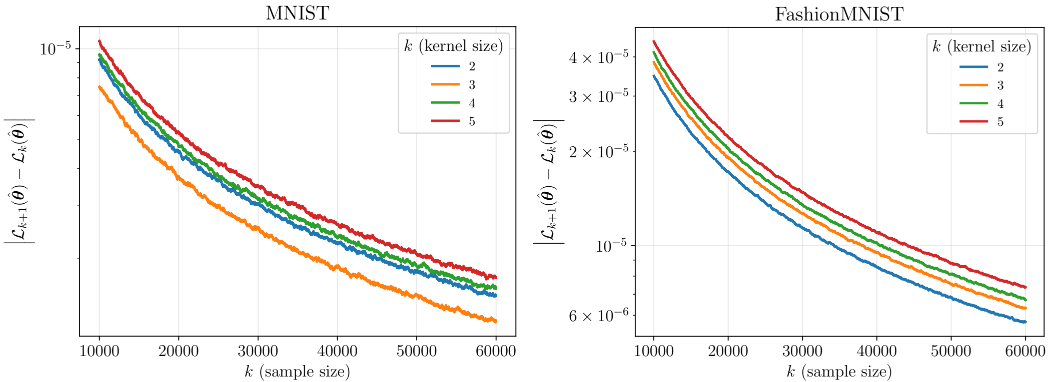
- **Задача:** классификация изображений.
- **Выборка:** MNIST, FashionMNIST, CIFAR10.
- **Архитектура:** L-слойная сверточная нейросеть с ReLU активациями.
- **Проверяеые гипотезы:** Абсолютная разница функции потерь растет в зависимости от
  1. числа сверточных слоев,
  2. размера ядра,
  3. числа фильтров.
- **Постановка эксперимента**
  1. Обучаем модель на полном наборе данных и получаем близкое к оптимальному $\hat{\boldsymbol{\theta}}$,
  2. Находим для всех k: $\left|\mathcal{L}_{k+1}(\hat{\boldsymbol{\theta}}) - \mathcal{L}_k(\hat{\boldsymbol{\theta}})\right|$ для всех $k = 1, \ldots, m$,
  3. Предыдущий пункт повторяем меняя порядок элементов в выборке.

Число слоев



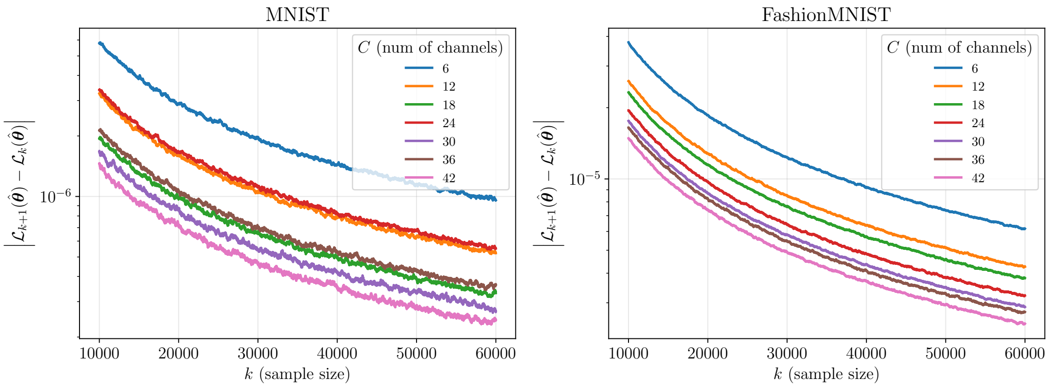
Графики демонстрируют шумное поведение, но видна тенденция: при увеличении числа слоев абсолютная разница функции потерь также увеличивается.

Размер ядра



При увеличении размера ядра, абсолютная разница функции потерь также растет.

Число фильтров



Графики, как видно, демонстрируют обратный результат: при большем числе каналов разница явно меньше. Предположительно, результат таков, так как сети при большем числе фильтров лучше обучились и компонента с Гессианом влияла значимо меньше.

Заключение

Основные результаты

1. Предложен способ оценки нормы гессиана для сетей, представленных в виде произведения матриц.
2. Данный результат применен для 2D/1D сверток, а также для пулингов и fc-головы.
3. Предложен способ оценки абсолютной разницы функции потерь для сверточных сетей, основываясь на гессиане.

Будущее работы

1. Улучшить теоретические оценки, пользуясь разреженностью матриц $\mathbf{\Lambda}^{(i)}$.
2. Применить данные оценки к другим видам нейронных сетей.
3. Проанализировать другие способы оценки изменения ландшафта функции потерь.