Изменение поверхности функции потерь в нейронных сетях при увеличении размера выборки

Отчет о научно-исследовательской работе

Киселев Никита Сергеевич
Научный руководитель: к.ф.-м.н. А. В. Грабовой

Московский физико-технический институт
(национальный исследовательский университет)
Физтех-школа прикладной математики и информатики
Кафедра интеллектуальных систем

Изменение поверхности функции потерь

Исследуется ландшафт функции потерь в нейронных сетях.

Проблема

Рассматриваемая поверхность нетривиально зависит от выбранной архитектуры, функции потерь и обучающей выборки.

Цель

Оценить изменение поверхности функции потерь при изменении обучающего набора данных.

Решение

1. Рассмотреть изменение функции потерь в окрестности минимума при добавлении одного нового объекта;
2. Использовать аппроксимацию второго порядка в окрестности экстремума.

Обзор полученных результатов

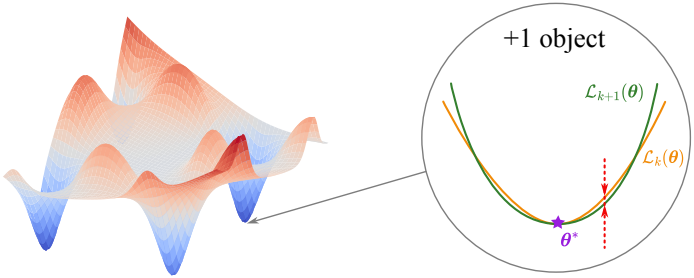

Рис. 1: Обзор результатов исследования. Часть (a) показывает ландшафт функции потерь, который является поверхностью в пространстве параметров. Часть (b) показывает разность значений функции ошибки. Она возникает, когда в выборку добавляется один новый объект. Здесь мы демонстрируем поведение для двумерного пространства параметров. Рядом с точкой минимума $\boldsymbol{\theta}^*$ среднее значение функции ошибки на $k+1$ объектах $\mathcal{L}_{k+1}(\boldsymbol{\theta})$ стремится к тому же, но на k объектах $\mathcal{L}_k(\boldsymbol{\theta})$.

Обозначения

Выборка

$$\mathfrak{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}, \quad i = 1, \ldots, m, \quad \mathbf{x} \in \mathcal{X},\ \mathbf{y} \in \mathcal{Y}$$

Модель

Условное распределение $p(\mathbf{y}|\mathbf{x})$ аппроксимируется $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathcal{Y}$, $\boldsymbol{\theta} \in \mathbb{R}^P$.

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 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 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

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

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

Абсолютное изменение функции потерь

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

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

Декомпозиция Гессиана

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

Аппроксимация Гессиана

В задаче K-классовой классификации спектр Гессиана состоит из двух частей[1]: "bulk" (большое количество около нуля) и "outliers" (ровно K ненулевых). Поэтому часто для изучения нормы Гессиана пренебрегают H-членом, то есть

$$\mathbf{H}_i(\boldsymbol{\theta}) \approx \nabla_{\boldsymbol{\theta}}\mathbf{z}_i \frac{\partial^2 \ell(\mathbf{z}_i, \mathbf{y}_i)}{\partial \mathbf{z}_i^2} \nabla_{\boldsymbol{\theta}}\mathbf{z}_i^\mathsf{T}.$$

[1] Sagun, L. et al. (2017). Empirical analysis of the hessian of over-parametrized neural networks.
Papyan, V. (2018). The full spectrum of deepnet hessians at scale: Dynamics with sgd training and sample size.
Ghorbani, B. et al. (2019). An Investigation into Neural Net Optimization via Hessian Eigenvalue Density.

Полносвязная нейронная сеть

Пусть $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^K$. Рассмотрим L-слойную сеть для задачи классификации, то есть $f_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbf{z}$ и $\ell(\mathbf{z}, \mathbf{y}) = \mathrm{CE}(\mathbf{p}, \mathbf{y}) = -\sum_{k=1}^{K} y_k \log p_k$:

$$\begin{aligned} \mathbf{z}^{(p)} &= \mathbf{W}^{(p)}\mathbf{x}^{(p)} + \mathbf{b}^{(p)}, \\ \mathbf{x}^{(p+1)} &= \sigma(\mathbf{z}^{(p)}), \end{aligned}$$

где $\mathrm{softmax}(\mathbf{z}) = \mathbf{p}$, а $\sigma(\mathbf{x}) = [\mathbf{x} \geqslant \mathbf{0}]\,\mathbf{x}$ есть функция активации ReLU.

Аппроксимация Гессиана полносвязной сети на одном объекте

$$\boxed{\mathbf{H}(\boldsymbol{\theta}) \overset{2}{\approx} \mathbf{F}^\intercal \mathbf{A} \mathbf{F}}$$

- $\mathbf{D}^{(p)} = \operatorname{diag}([\mathbf{z}^{(p)} \geqslant \mathbf{0}])$
- $\mathbf{G}^{(p)} = \dfrac{\partial \mathbf{z}}{\partial \mathbf{z}^{(p)}} = \mathbf{W}^{(L)} \mathbf{D}^{(L-1)} \mathbf{W}^{(L-1)} \mathbf{D}^{(L-2)} \cdot \ldots \cdot \mathbf{D}^{(p)}$
- $$\mathbf{F}^\intercal = \begin{pmatrix} (\mathbf{G}^{(1)})^\intercal \otimes \mathbf{x}^{(1)} \\ (\mathbf{G}^{(1)})^\intercal \\ \vdots \\ (\mathbf{G}^{(L)})^\intercal \otimes \mathbf{x}^{(L)} \\ (\mathbf{G}^{(L)})^\intercal \end{pmatrix}$$
- $\mathbf{A} = \nabla^2_{\mathbf{z}} \ell(\mathbf{z}, \mathbf{y}) = \operatorname{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\intercal$

[2] Wu, Y. et al. (2020). Dissecting hessian: Understanding common structure of hessian in neural networks.

Ограниченность и пропорциональность нормы Гессиана

Теорема 1 (об ограниченности нормы Гессиана)

Пусть задана L-слойная полносвязная нейронная сеть с функцией активацией ReLU и без параметров смещения, применямая для решения задачи K-классовой классификации. Если выполнено следующее: $\|\mathbf{W}^{(p)}\|_2 \leqslant M_{\mathbf{W}}$ и $\|\mathbf{x}_i\|_2 \leqslant M_{\mathbf{x}}$ для всех слоев $p = 1, \ldots, L$ в сети и для всех объектов $i = 1, \ldots, m$ в наборе данных, то для любого объекта $i = 1, \ldots, m$ верно следующее неравенство:

$$\|\mathbf{H}_i(\boldsymbol{\theta})\|_2 \leqslant L\sqrt{2}M_{\mathbf{x}}^2 M_{\mathbf{W}}^{2L} + \sqrt{2}\frac{M_{\mathbf{W}}^2(M_{\mathbf{W}}^{2L} - 1)}{M_{\mathbf{W}}^2 - 1}.$$

Лемма 1 (о пропорциональности нормы Гессиана)

Если каждый параметр модели ограничен константой $M > 0$, то есть $|w_{ij}^{(p)}| \leqslant M$ для всех $i, j = 1, \ldots, h$ и для всех слоев $p = 1, \ldots, L$, тогда в условиях Теоремы 1 верно следующее:

$$\|\mathbf{H}_i(\boldsymbol{\theta})\|_2 \leqslant L\sqrt{2}M_{\mathbf{x}}^2 (hM)^{2L} + \sqrt{2}\frac{(hM)^2((hM)^{2L} - 1)}{(hM)^2 - 1}.$$

Таким образом, верна следующая пропорция:

$$\|\mathbf{H}_i(\boldsymbol{\theta})\|_2 \propto L(hM)^{2L}.$$

Лемма 2 (о сходимости абсолютной разности функции потерь)

Пусть точка $\boldsymbol{\theta}$ выбрана так, что $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2 \leqslant R^2$ для некоторого $R > 0$. Если существует неотрицательная константа M_ℓ, что $|\ell(f_{\boldsymbol{\theta}^*}(\mathbf{x}_i), \mathbf{y}_i)| \leqslant M_\ell$ для всех объектов $i = 1, \ldots, m$ в наборе данных, то в условиях Теоремы 1 верно следующее:

$$|\mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta})| \leqslant \frac{2}{k+1}\left(M_\ell + \left(L\sqrt{2}M_{\mathbf{x}}^2 M_{\mathbf{W}}^{2L} + \sqrt{2}\frac{M_{\mathbf{W}}^2(M_{\mathbf{W}}^{2L} - 1)}{M_{\mathbf{W}}^2 - 1}\right)R^2\right) \to 0 \text{ при } k \to \infty.$$

Таким образом, верна следующая пропорция:

$$|\mathcal{L}_{k+1}(\boldsymbol{\theta}) - \mathcal{L}_k(\boldsymbol{\theta})| \propto \frac{L(hM)^{2L}R^2}{k}.$$

Вычислительный эксперимент

Постановка эксперимента

- **Задача:** классификация изображений:
  1. Прямая, то есть пиксели есть признаки,
  2. С предварительной векторизацией при помощи предобученной модели ViT;
- **Выборка:** MNIST, FashionMNIST, CIFAR10, CIFAR100;
- **Архитектура:** полносвязная L-слойная сеть со скрытым размером h на каждом слое и ReLU после каждого из них;
- **Варьирование архитектуры:**
  1. Фиксируем число слоев L и варьируем размер скрытого слоя h,
  2. Наоборот, фиксируем размер скрытого слоя h и варьируем число слоев L;
- **Постановка эксперимента:**
  1. Обучаем модель на полном наборе данных $\rightarrow$ получаем $\hat{\boldsymbol{\theta}}$,
  2. Подсчитываем $\left|\mathcal{L}_{k+1}(\hat{\boldsymbol{\theta}}) - \mathcal{L}_k(\hat{\boldsymbol{\theta}})\right|$ для всех $k = 1, \ldots, m$,
  3. Повторяем пункт 2 несколько раз, используя разный порядок добавления.

Прямая классификация изображений

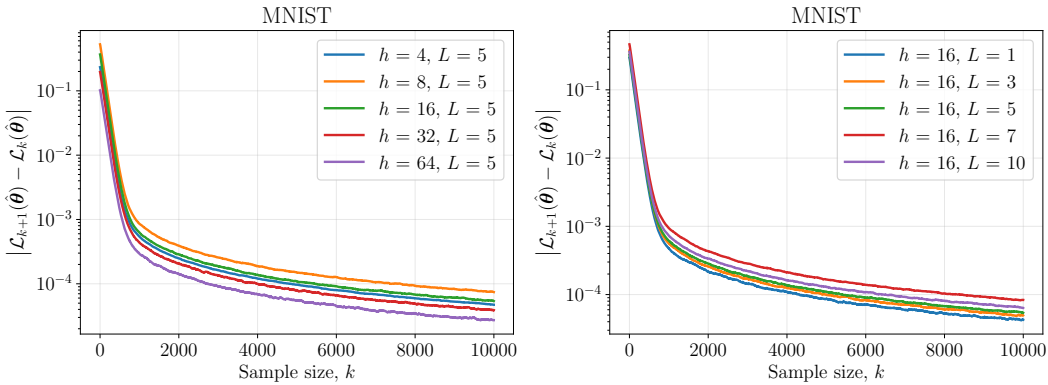

Рис. 2: Зависимость абсолютного значения разности значений функции ошибки от доступного размера выборки, на задаче прямой классификации изображений. Графики слева демонстрируют уменьшение значений при увеличении размера скрытого слоя. Графики справа демонстрируют увеличение значений при увеличении числа слоев.

Предварительная векторизация изображений

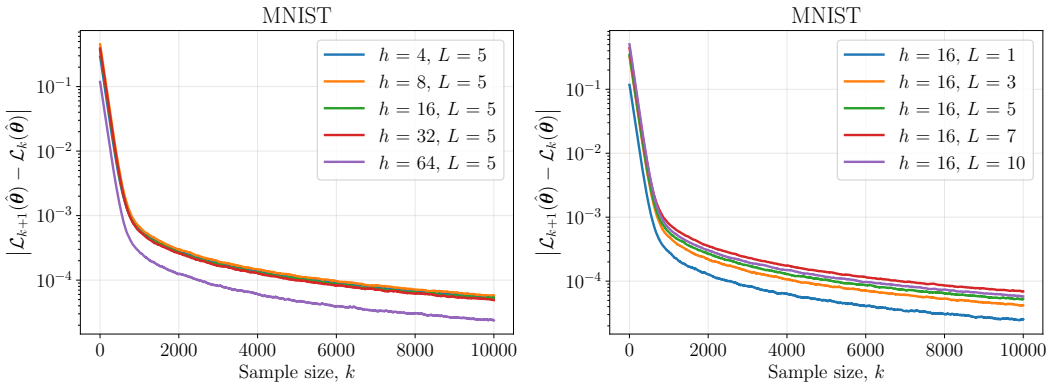

Рис. 3: Зависимость абсолютного значения разности значений функции ошибки от доступного размера выборки, на задаче с предварительной векторизацией изображений. Графики слева демонстрируют уменьшение значений при увеличении размера скрытого слоя. Графики справа демонстрируют увеличение значений при увеличении числа слоев.

Заключение

1. Доказаны теоретические результаты о сходимости поверхности функции потерь в полносвязной нейронной сети при увеличении размера выборки;
2. Вычислительный эксперимент демонстрирует справедливость полученных результатов на практике;
3. Дальнейшее исследование лежит в направлении 1) рассмотрения других архитектур[2], 2) определения достаточного размера выборки.

Публикации

1. N. Kiselev, A. Grabovoy. Unraveling the Hessian: A Key to Smooth Convergence in Loss Function Landscapes // arXiv preprint arXiv:2409.11995. – 2024.
   Отобрана для публикации в рамках конференции AI Journey 2024. Будет опубликована в журнале Doklady Mathematics (**Q2**).

[2]Ожидается выступление студента 4-го курса Владислава Мешкова на конференции ИСП РАН по совместному исследованию для сверточных архитектур