Методы типа градиентного клиппинга для задач на больших данных

Жолобов Владимир Александрович

Московский физико-технический институт

Группа 774
Научный руководитель: д.ф.-м.н. Гасников А.В.

Цель и задача

Цель

Проверить экспериментально сходимость методов на больших данных

Задача

Для исследования рассматриваются задачи классификации изображений, семантической сегментации и super resolution.

Список литературы

Ключевые работы

1. Gorbunov E., Danilova M., Gasnikov A. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping //arXiv preprint arXiv:2005.10785. – 2020.
2. Aleksandr Viktorovich Nazin, AS Nemirovsky, Aleksandr Borisovich Tsybakov, and AB Juditsky. Algorithms of robust stochastic optimization based on mirror descent method. Automation and Remote Control, 80(9):1607–1627, 2019.
3. Simsekli U., Sagun L., Gurbuzbalaban M. A tail-index analysis of stochastic gradient noise in deep neural networks //International Conference on Machine Learning. – PMLR, 2019. – С. 5827-5837.

Постановка задачи

Целевая функция

$$\min_{x \in \mathrm{R}^n} f(x), \quad f(x) = \mathrm{E}_\xi[f(x, \xi)],$$

Предположения

$$\mathrm{E}_\xi[\nabla f(x, \xi)] = \nabla f(x), \quad \mathrm{E}_\xi[||\nabla f(x, \xi) - \nabla f(x)||_2^2] \leq \sigma^2$$

Определение

Случайный вектор с узким хвостом (light-tailed). Будем говорить, что случайный вектор η имеет узкий хвост распределения, если существует $\mathrm{E}[\eta]$ и $\mathrm{P}\{||\eta - \mathrm{E}[\eta]||_2 > b\} \leq 2 \exp\left(-\frac{b^2}{2\sigma^2}\right)$ для всех $b > 0$.

Задача классификации изображений

Постановка

Для набора изображений и классов строится отображение каждого изображения к классу.

Критерий качества

$$Top1 = \sum_{i=1}^{m} |\arg\max a(x_i) = y_i|$$

Задача семантической сегментации изображений

Постановка задачи

Для набора изображений для каждого изображения по пиксельно предсказывается отношение к классу

Критерий качества

$$IoU = \frac{TP}{TP + FN + FP}$$

Задача Super resolution

Постановка задачи

Заданы два набора изображений высокого разрешения $\{I_{y_i}\}_{i=1}^N$ и низкого разрешения $\{I_{x_i}\}_{i=1}^N$. Требуется построить модель аппроксимации изображения высокого разрешения $\{\hat{I}_{y_i}\}_{i=1}^N$

$$\hat{I}_y = \mathcal{F}(I_x, \Theta)$$

Функция ошибки

$$MSE = \frac{1}{mnk} \sum_{d=1}^{k} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I_y(i,j,d) - \hat{I}_y(i,j,k)]^2$$

Критерий качества

$$PSNR = 10 \cdot \log_{10} \left(\frac{MAX_I^2}{MSE} \right)$$

Градиентный клиппинг

$$clip(\nabla f(x,\xi),\lambda) = \min\{1, \frac{\lambda}{||\nabla f(x,\xi)||_2}\}\nabla f(x,\xi),$$

где λ определяет порог клиппинга.

Метод подобных треугольников с клиппингом

Algorithm 1: Clipped Stochastic Similar Triangles Method (clipped-SSTM)

Input: начальная точка x^0, число итераций N, размеры батчей $\{m_k\}_{k=1}^N$, параметр шага a, параметр клиппинга B

Обозначим $A_0 = \alpha_0 = 0$, $y^0 = z^0 = x^0$

for $k = 0, \ldots, N-1$ **do**

  Вычисляем $\alpha_{k+1} = \frac{k+2}{2aL}, A_{k+1} = A_k + \alpha_{k+1}, \lambda_{k+1} = \frac{B}{\alpha_{k+1}}$

  $x^{k+1} = \frac{A_k y^k + \alpha_{k+1} z^k}{A_{k+1}}$

  Снова получаем $\xi_1^k, \ldots, \xi_{m_k}^k$ и вычисляем $\nabla f(x^{k+1}, \xi^k) = \frac{1}{m_k} \sum_{i=1}^{m_k} \nabla f(x^{k+1}, \xi_i^k)$

  Вычисляем $\widetilde{\nabla} f(x^{k+1}, \xi^k) = clip(\nabla f(x^{k+1}, \xi^k), \lambda_{k+1})$

  $z^{k+1} = z^k - \alpha_{k+1} \widetilde{\nabla} f(x^{k+1}, \xi^k)$

  $y^{k+1} = \frac{A_k y^k + \alpha_{k+1} z^{k+1})}{A_{k+1}}$

Output: y^N

Скорость сходимости

Теорема Горбунова (2020)

Пусть функция f выпукла и L-гладка. Тогда для всех $\beta \in (0, 1)$ и $N \geq 1$ такой что $\ln(\frac{4N}{\beta} \geq 2$ имеем, что после N итераций clipped-SSTM с $m_k = \Theta(\max\{1, \frac{\sigma^2 \alpha_{k+1}^2 N \ln(N/\beta)}{R_0^2}\}), B = \Theta(\frac{R_0}{\ln(N/\beta)})$ и $a = \Theta(\ln^2(N/\beta))$ что $f(y^N) - f(x^*) = O(\frac{aLR_0^2}{N^2})$ справедливо с вероятностью по крайней мере $1 - \beta$, где $R_0 = ||x^0 - x^*||_2$.

Метод стохастического градиентного спуска с клиппингом

Algorithm 2: Clipped Stochastic Gradient Descent (clipped-SGD)

Input: начальная точка x^0, число итераций N, размеры батчей $\{m_k\}_{k=0}^{N-1}$, шаг $\gamma > 0$, порог клиппинга $\lambda > 0$

for $k = 0, \ldots, N-1$ **do**

  Получаем $\xi_1^k, \ldots, \xi_{m_k}^k$ и вычисляем $\nabla f(x^{k+1}, \xi^k) = \frac{1}{m_k} \sum_{i=1}^{m_k} \nabla f(x^{k+1}, \xi_i^k)$

  Вычисляем $\widetilde{\nabla} f(x^{k+1}, \xi^k) = clip(\nabla f(x^{k+1}, \xi^k), \lambda_{k+1})$

  $x^{k+1} = x^k - \gamma \widetilde{\nabla} f(x^k, \xi^k)$

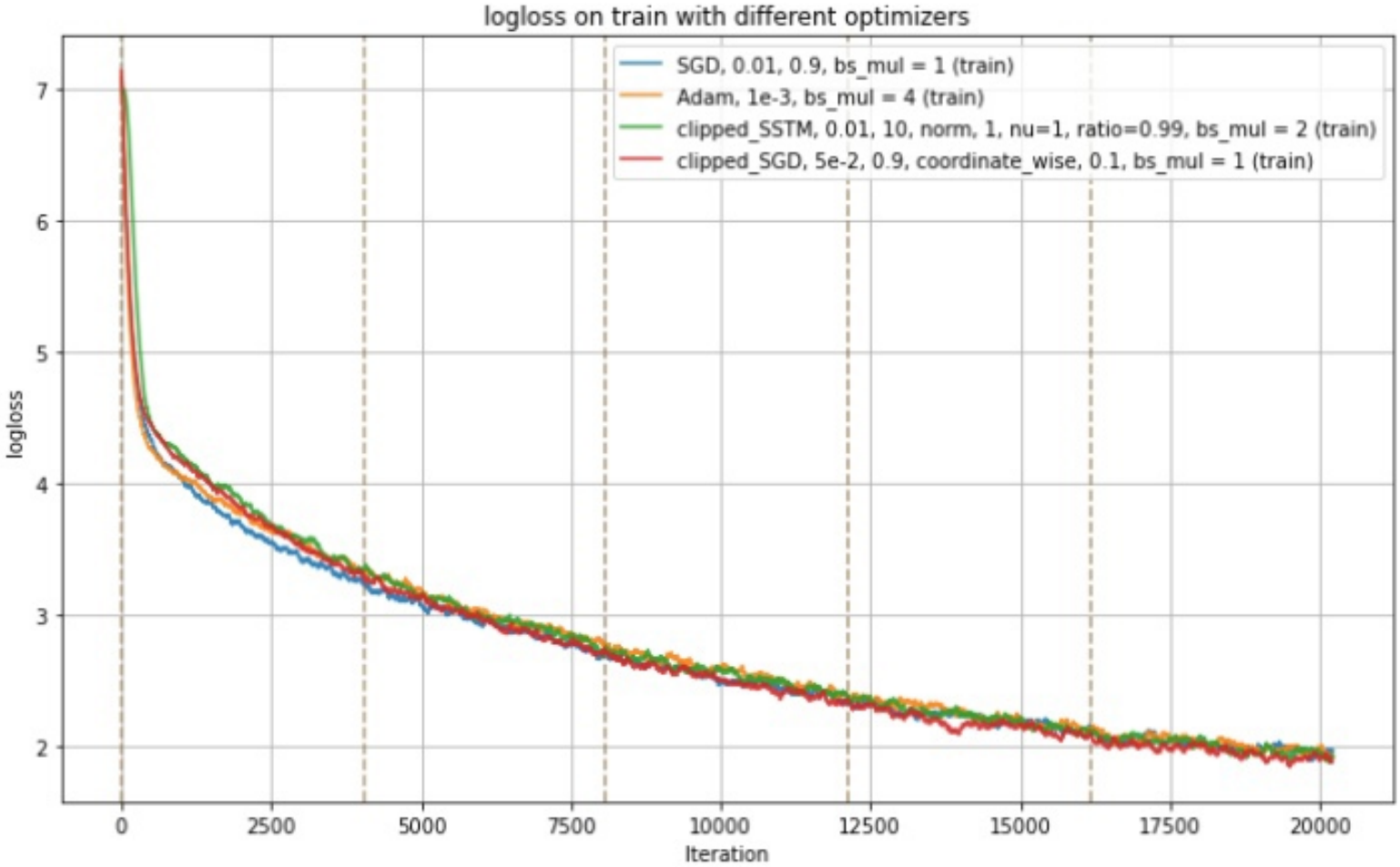
Output: $\overline{x}^N = \frac{1}{N} \sum_{k=0}^{N-1} x^k$

Скорость сходимости

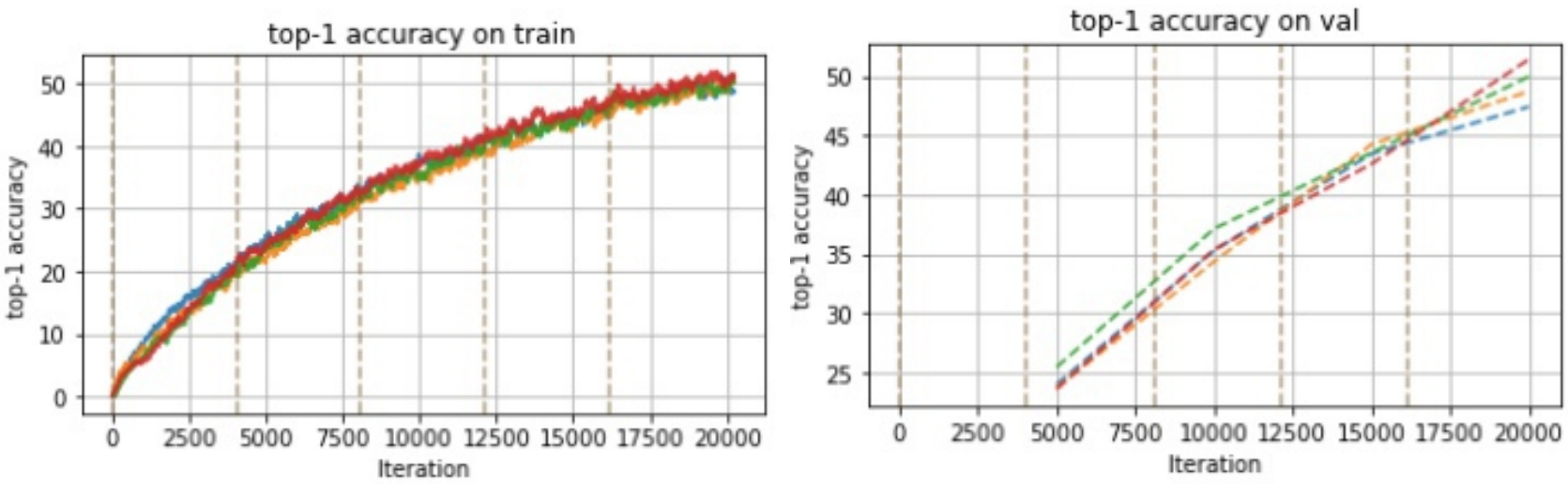
Теорема Горбунова (2020)

Пусть функция f выпукла и L-гладка. Тогда для всех $\beta \in (0, 1)$ и $N \geq 1$ такой, что $\ln(\frac{4N}{\beta} \geq 2$ имеем, что после N итераций clipped-SGD с $\lambda = \Theta(LR_0)$ и $m_k = m = \Theta(\max\{1, \frac{\sigma^2 N}{R_0^2 L^2 \ln(N/\beta)}\})$, где $R_0 = ||x^0 - x^*||_2$ и шаг $\gamma = \frac{1}{80L\ln(4N/\beta)}$, имеем $f(\overline{x}^N) - f(x^*) = O(\frac{LR_0^2 \ln(4N/\beta)}{N})$ с вероятностью не менее $1 - \beta$, где $\overline{x}^N = \frac{1}{N}\sum_{k=0}^{N-1} x^k$.

ImageNet-100k: сходимость



ImageNet-100k:

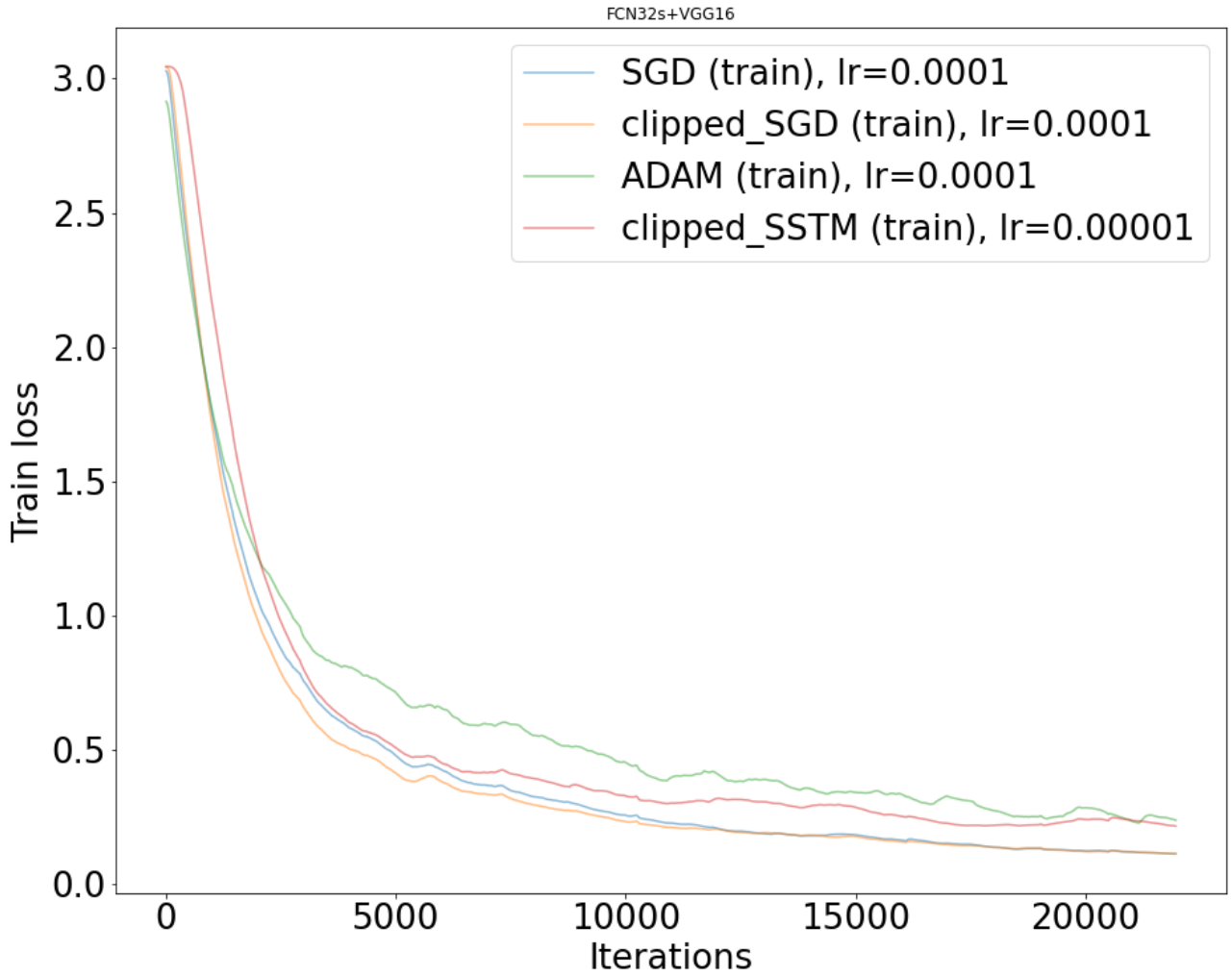


(a) Top-1 на Train

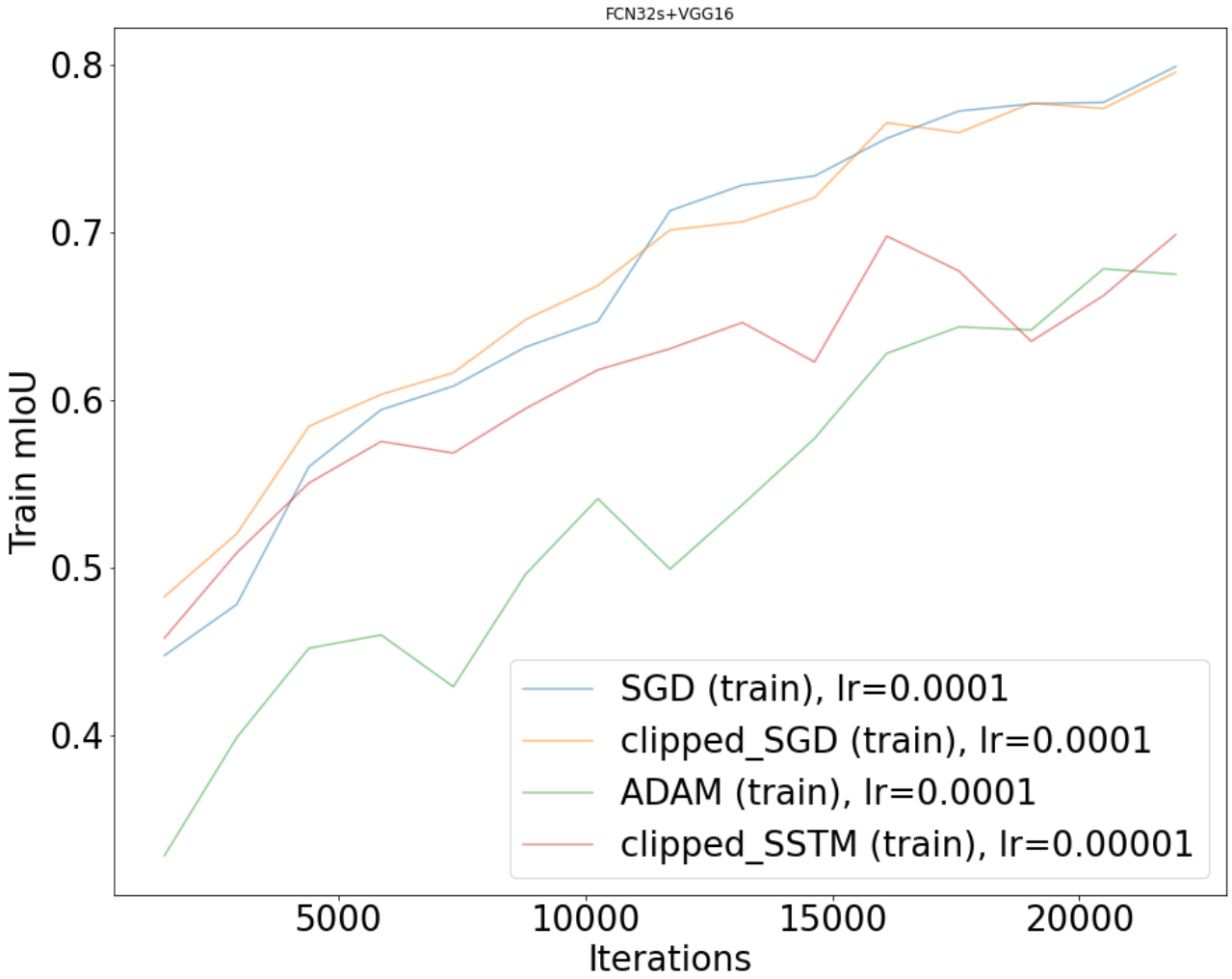
(b) Top-1 на Validation

Рис.: Сравнение точностей моделей классификации изображений на обучении и валидации

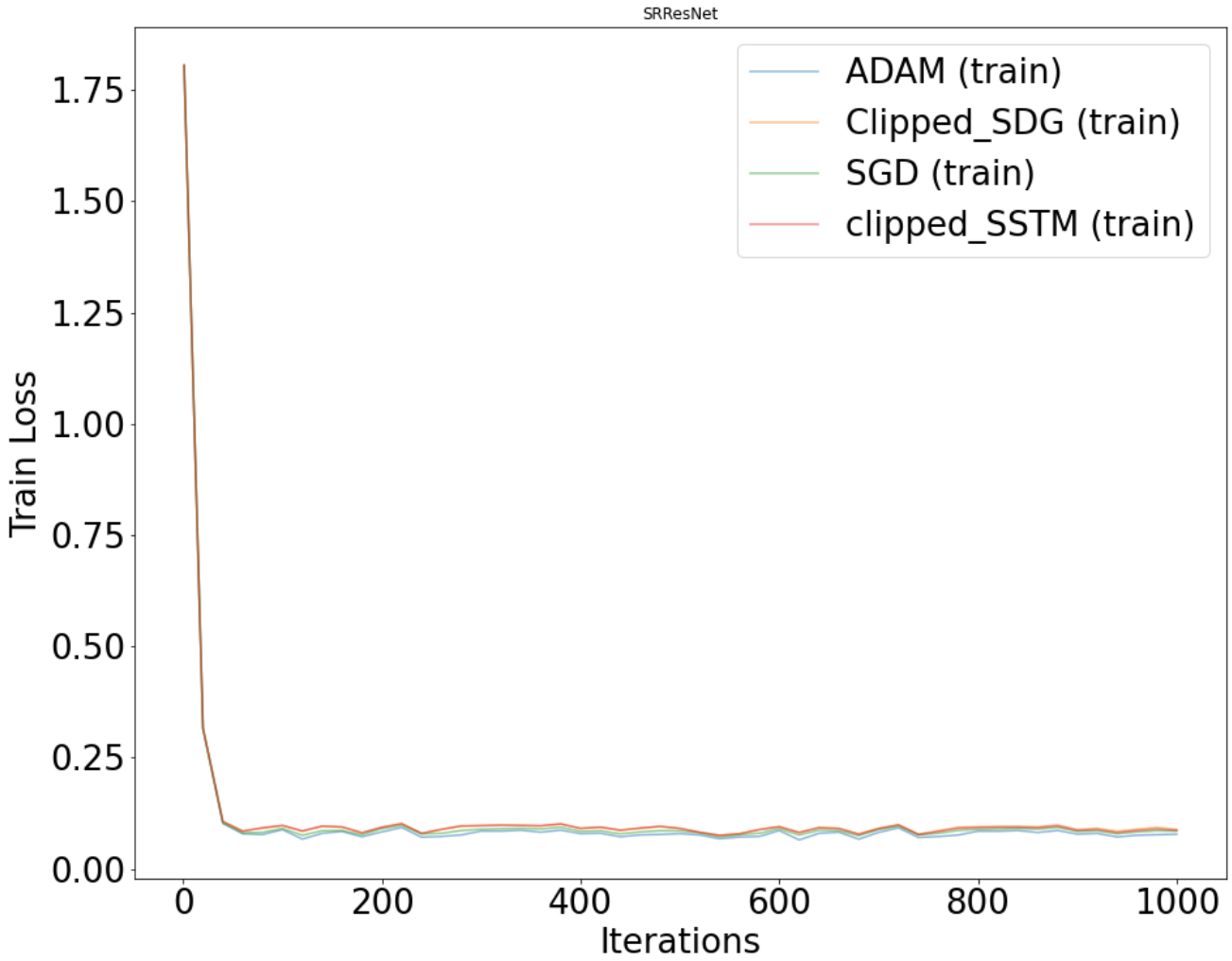
PascalVOC2012



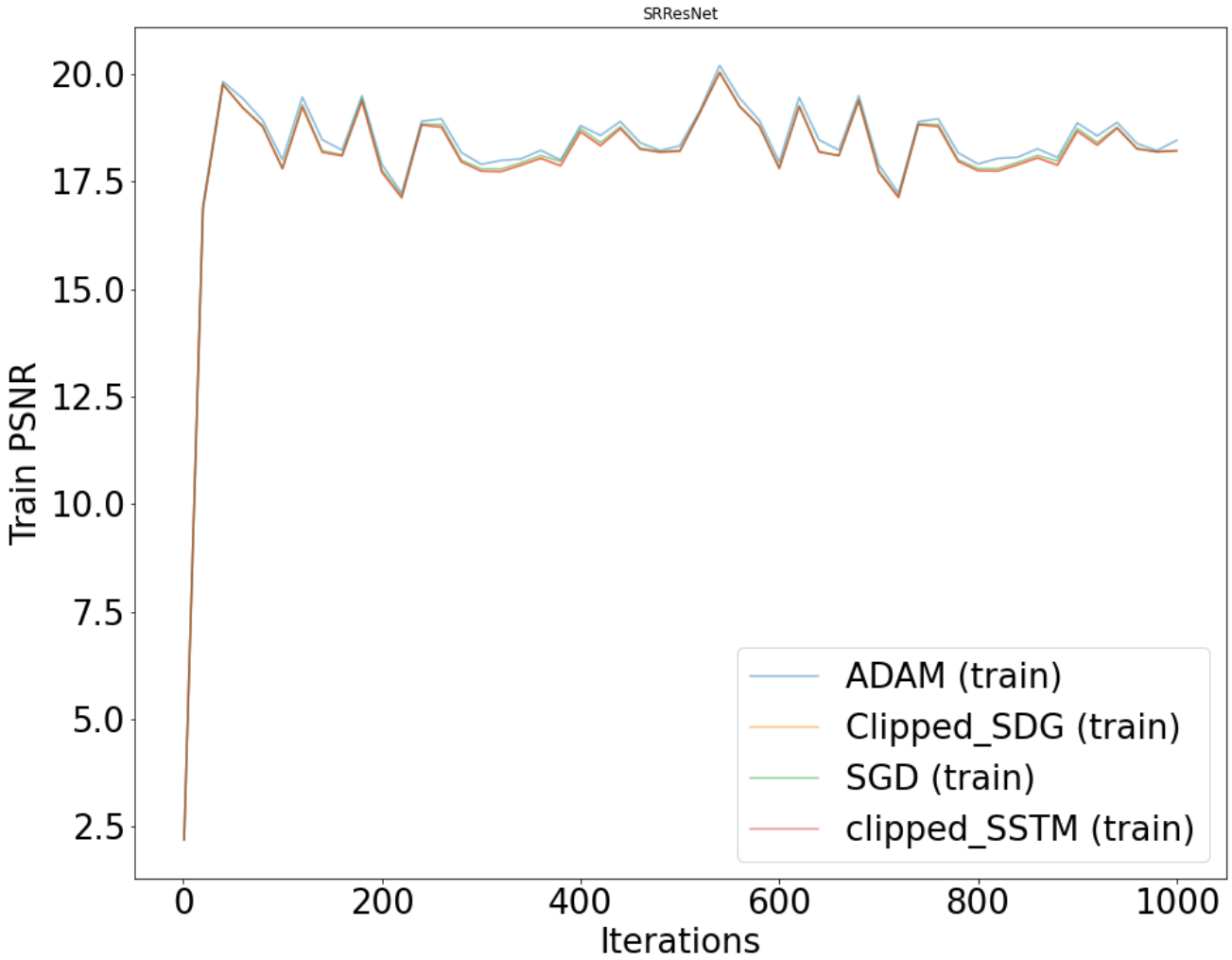
PascalVOC2012: mIoU на валидации



DIV2K: сходимость



DIV2K: mPSNR на обучении



Результаты экспериментов

Задача	ImCl	SemSeg	SupRes
Модель	ResNet-18	FCN32-s+VGG16	SRResNet
Качество	Top-1	val mIoU	val mPSNR
ADAM	47.21	0.466	18.046
Clipped-SGD	52.32	0.576	17.855
SGD	45.8	0.578	17.873
clipped-SSTM	47.2	0.549	17.869

Таблица: Качество моделей

Заключение

Выносятся на защиту:

1. Реализация методов типа градиентного клиппинга
2. Подтверждение работаспособности методов и хорошее качество на задачах heavy-tailed
3. Практическое решение задачи на выборке ImageNet-100k, PascalVOC2012, DIV2K

Публикации

Список

1. Dvurechensky P., Gasnikov A., Tiurin A., Zholobov V. Unifying Framework for Accelerated Randomized Methods in Convex Optimization //arXiv preprint arXiv:1707.08486. – 2020.
2. Potanin M. S. et al. Deep learning neural network structure optimization //Informatika i Ee Primeneniya [Informatics and its Applications]. – 2020. – Т. 14. – №. 4. – С. 55-62.