# Модификации метода стохастического градиентного спуска для задач машинного обучения с большими объемами данных

Владислав Чабаненко

Научный руководитель: к.ф-м.н., доцент
Ветров Дмитрий Петрович

2016

# Обучение нейронных сетей

Проблема:

- Медленная сходимость из-за изменения распределений на слоях в процессе обучения

Решение — батч-нормализация (БН) [1]:

- Подсчет необходимых статистик в слое по батчам (эмпирические средние и дисперсии)
- Нормализация распределения на каждом слое нейронной сети с помощью подсчитанных статистик

[1] http://arxiv.org/abs/1502.03167

# Актуальность батч-нормализации

- Батч-нормализация сейчас очень популярна, так как многие state-of-the-art архитектуры сетей без нее совсем плохо обучаются
- Работает при больших объемах данных, что в сейчас очень актуально

# Применение батч-нормализации

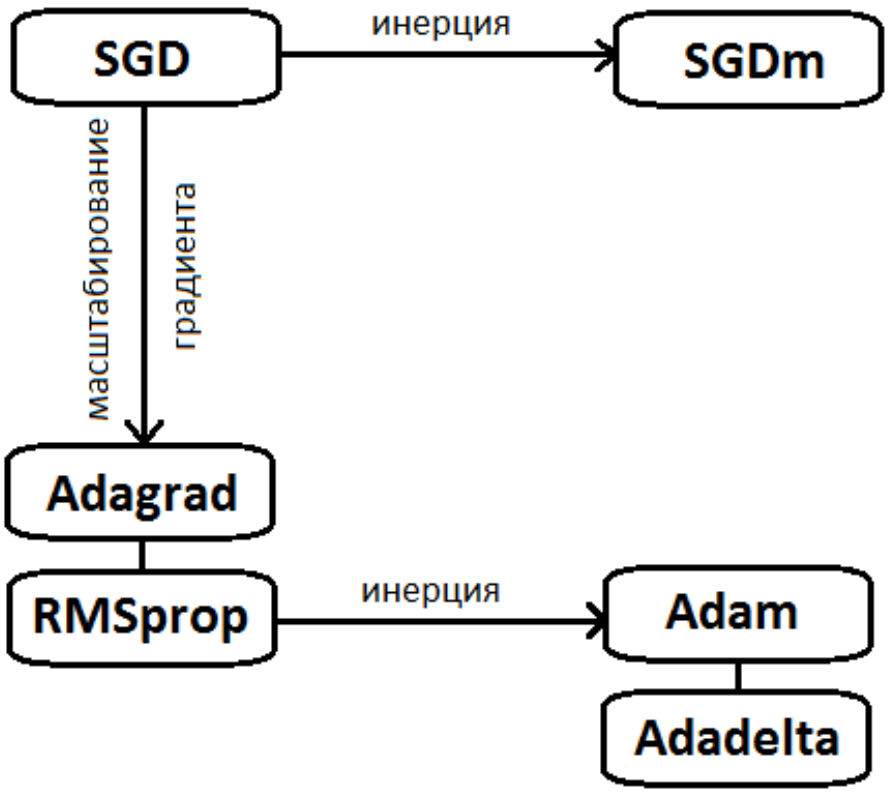
- Авторы идеи применили БН для метода стохастического градиентного спуска (SGD)
- Мы исследовали применение БН к различным популярным модификациям SGD, используемых при обучении нейронных сетей

# Модификации SGD

В работе рассматривались следующие модификации SGD:

- SGD с инерцией (SGDm)
- Adagrad
- RMSprop
- Adadelta
- Adam

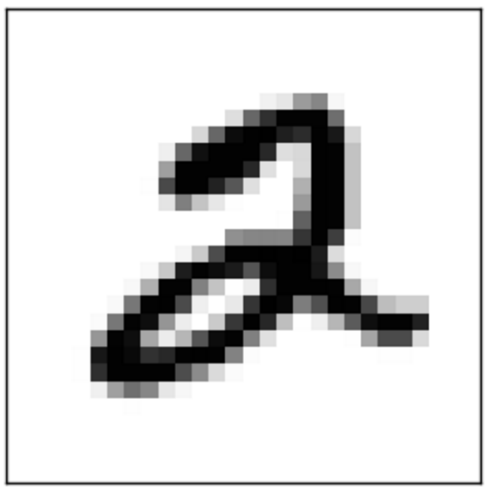
# Интуиция методов



# Датасеты

В рамках работы мы проводили экспериментальные исследования на следующих датасетах:

- MNIST (70 тыс. рукописных цифр — 10 классов)
- CIFAR-10 (60 тыс. изображений — 10 классов)



# Архитектуры

Использовались следующие архитектуры:

- Полносвязная сеть с 3 скрытыми слоями по 100 нейронов (MLP)
- Сеть с двумя сверточными слоями (32 фильтра $5 \times 5$ + макс-пулинг с окном $2 \times 2$) и один полносвзяный слой с 256 нейронами (CNN)

Для обеих архитектур использовалась функция активации ReLU.

# Результаты

Для каждого метода в комбинации с различными архитектурами, датасетами и батч-нормализацией мы подобрали лучший шаг обучения

- Вычислили повышение качества методов на валидационной выборке при добавлении батч-нормализации
- Абсолютное улучшение не показательно, поэтому использовали относительное:

$$\text{rel} = \frac{y - x}{100 - x}$$

## Качество на MNIST + MLP, топ-3 наименьшего улучшения

| Номер эпохи | 1 | 3 | 15 | 35 | 50 |
|---|---|---|---|---|---|
| **SGD** | 89.15 | 92.84 | 96.81 | 97.58 | 97.73 |
| BN SGD | 96.42 | 97.38 | 98.14 | 98.18 | 98.23 |
| Улучшение | 0.67 | 0.63 | 0.42 | 0.25 | 0.22 |
| **SGDm** | 93.39 | 96.4 | 97.83 | 98.03 | 98.03 |
| BN SGDm | 96.49 | 97.62 | 98.11 | 97.87 | 98.12 |
| Улучшение | 0.47 | 0.34 | **0.13** | **-0.08** | **0.05** |
| **Adam** | 93.47 | 96.03 | 97.62 | 98.04 | 98.02 |
| BN Adam | 94.85 | 96.81 | 97.51 | 97.75 | 98.16 |
| Улучшение | **0.21** | **0.2** | **-0.05** | **-0.15** | **0.07** |
| **Adagrad** | 93.92 | 96.1 | 97.37 | 97.55 | 97.64 |
| BN Adagrad | 96.01 | 97.16 | 97.81 | 97.84 | 97.97 |
| Улучшение | 0.34 | 0.27 | 0.17 | 0.12 | 0.14 |
| **Adadelta** | 94.28 | 96.7 | 97.51 | 98.04 | 98.2 |
| BN Adadelta | 95.98 | 96.97 | 97.72 | 98.04 | 98.26 |
| Улучшение | **0.3** | **0.08** | **0.08** | **-0.0** | **0.03** |
| **RMSprop** | 94.58 | 96.14 | 96.75 | 97.34 | 97.27 |
| BN RMSprop | 96.12 | 97.18 | 97.88 | 97.91 | 97.95 |
| Улучшение | **0.28** | **0.27** | 0.35 | 0.21 | 0.25 |

## Качество на MNIST + MLP, топ-3 лучшего качества

| Номер эпохи | 1 | 3 | 15 | 35 | 50 |
|---|---|---|---|---|---|
| **SGD** | 89.15 | 92.84 | 96.81 | 97.58 | 97.73 |
| BN SGD | 96.42 | 97.38 | 98.14 | 98.18 | 98.23 |
| Улучшение | 0.67 | 0.63 | 0.42 | 0.25 | 0.22 |
| **SGDm** | 93.39 | **96.4** | **97.83** | **98.03** | **98.03** |
| BN SGDm | 96.49 | 97.62 | 98.11 | 97.87 | 98.12 |
| Улучшение | 0.47 | 0.34 | 0.13 | -0.08 | 0.05 |
| **Adam** | 93.47 | 96.03 | **97.62** | **98.04** | **98.02** |
| BN Adam | 94.85 | 96.81 | 97.51 | 97.75 | 98.16 |
| Улучшение | 0.21 | 0.2 | -0.05 | -0.15 | 0.07 |
| **Adagrad** | **93.92** | 96.1 | 97.37 | 97.55 | 97.64 |
| BN Adagrad | 96.01 | 97.16 | 97.81 | 97.84 | 97.97 |
| Улучшение | 0.34 | 0.27 | 0.17 | 0.12 | 0.14 |
| **Adadelta** | **94.28** | **96.7** | **97.51** | **98.04** | **98.2** |
| BN Adadelta | 95.98 | 96.97 | 97.72 | 98.04 | 98.26 |
| Улучшение | 0.3 | 0.08 | 0.08 | -0.0 | 0.03 |
| **RMSprop** | **94.58** | **96.14** | 96.75 | 97.34 | 97.27 |
| BN RMSprop | 96.12 | 97.18 | 97.88 | 97.91 | 97.95 |
| Улучшение | 0.28 | 0.27 | 0.35 | 0.21 | 0.25 |

## Качество на CIFAR-10 + CNN, топ-3 наименьшего улучшения

| Номер эпохи | 1 | 2 | 25 | 35 | 50 |
|---|---|---|---|---|---|
| **SGD** | 24.26 | 22.14 | 58.95 | 63.23 | 65.45 |
| BN SGD | 48.47 | 52.95 | 71.31 | 75.89 | 76.58 |
| Улучшение | 0.32 | 0.4 | 0.3 | 0.34 | 0.32 |
| **SGDm** | 30.03 | 37.98 | 64.37 | 68.3 | 71.54 |
| BN SGDm | 53.72 | 58.76 | 76.85 | 76.89 | 78.39 |
| Улучшение | 0.34 | 0.34 | 0.35 | 0.27 | 0.24 |
| **Adam** | 43.28 | 47.3 | 68.02 | 70.79 | 72.54 |
| BN Adam | 53.94 | 59.13 | 75.4 | 76.63 | 76.66 |
| Улучшение | **0.19** | **0.22** | **0.23** | **0.2** | **0.15** |
| **Adagrad** | 29.82 | 37.25 | 60.86 | 62.35 | 64.63 |
| BN Adagrad | 48.24 | 53.83 | 74.66 | 76.11 | 76.96 |
| Улучшение | **0.26** | **0.26** | 0.35 | 0.37 | 0.35 |
| **Adadelta** | 31.31 | 32.96 | 67.12 | 70.56 | 70.81 |
| BN Adadelta | 49.62 | 55.96 | 72.75 | 76.99 | 77.2 |
| Улучшение | 0.27 | 0.34 | **0.17** | **0.22** | **0.22** |
| **RMSprop** | 25.6 | 39.89 | 64.24 | 68.96 | 70.31 |
| BN RMSprop | 42.61 | 52.98 | 73.7 | 75.47 | 75.81 |
| Улучшение | **0.23** | **0.22** | **0.26** | **0.21** | **0.19** |

## Качество на CIFAR-10 + CNN, топ-3 лучшего качества

| Номер эпохи | 1 | 2 | 25 | 35 | 50 |
|---|---|---|---|---|---|
| **SGD** | 24.26 | 22.14 | 58.95 | 63.23 | 65.45 |
| BN SGD | 48.47 | 52.95 | 71.31 | 75.89 | 76.58 |
| Улучшение | 0.32 | 0.4 | 0.3 | 0.34 | 0.32 |
| **SGDm** | **30.03** | **37.98** | **64.37** | 68.3 | **71.54** |
| BN SGDm | 53.72 | 58.76 | 76.85 | 76.89 | 78.39 |
| Улучшение | 0.34 | 0.34 | 0.35 | 0.27 | 0.24 |
| **Adam** | **43.28** | **47.3** | **68.02** | **70.79** | **72.54** |
| BN Adam | 53.94 | 59.13 | 75.4 | 76.63 | 76.66 |
| Улучшение | 0.19 | 0.22 | 0.23 | 0.2 | 0.15 |
| **Adagrad** | 29.82 | 37.25 | 60.86 | 62.35 | 64.63 |
| BN Adagrad | 48.24 | 53.83 | 74.66 | 76.11 | 76.96 |
| Улучшение | 0.26 | 0.26 | 0.35 | 0.37 | 0.35 |
| **Adadelta** | **31.31** | 32.96 | **67.12** | **70.56** | **70.81** |
| BN Adadelta | 49.62 | 55.96 | 72.75 | 76.99 | 77.2 |
| Улучшение | 0.27 | 0.34 | 0.17 | 0.22 | 0.22 |
| **RMSprop** | 25.6 | **39.89** | 64.24 | **68.96** | 70.31 |
| BN RMSprop | 42.61 | 52.98 | 73.7 | 75.47 | 75.81 |
| Улучшение | 0.23 | 0.22 | 0.26 | 0.21 | 0.19 |

# Выводы

- Батч-нормализация тем слабее улучшает метод, чем он сам по себе качественнее работает
- Возможно, батч-нормализация плохо совместима с инерцией для полносвязных сетей
- Возможно, батч-нормализация плохо совместима с методами, наследуемыми от RMSprop, для сверточных сетей

# На защиту выносится:

1. Реализация двух архитектур нейронной сети с батч нормализацией и пяти модификаций стохастического градиентного спуска
2. Экспериментальные исследования модификаций стохастического градиентного спуска для обучения нейронных сетей с и без батч-нормализации
3. ?Рекомендации по использованию батч-нормализации и модификаций стохастического градиентного спуска при обучении нейронных сетей

# Конец

Конец!

# SGD

$$\theta_{t+1} = \theta_t - \eta \nabla F(\theta)$$

# SGD momentum

$$v_{t+1} = \mu v_t - \eta \nabla F(\theta)$$

$$\theta_{t+1} = \theta_t + v_{t+1}$$

# Adagrad

$$g_{t+1} = g_t + \nabla F(\theta)^2$$

$$\theta_{t+1} = \theta_t - \frac{\eta \nabla F(\theta)}{\sqrt{g_{t+1}} + \epsilon}$$

# RMSprop

$$g_{t+1} = \gamma g_t + (1 - \gamma)\nabla F(\theta)^2$$

$$\theta_{t+1} = \theta_t - \frac{\eta \nabla F(\theta)}{\sqrt{g_{t+1} + \epsilon}}$$

# Adadelta

$$g_{t+1} = \gamma g_t + (1 - \gamma)\nabla F(\theta)^2$$

$$v_{t+1} = -\frac{\sqrt{x_t + \epsilon}\nabla F(\theta)}{\sqrt{g_{t+1} + \epsilon}}$$

$$x_{t+1} = \gamma x_t + (1 - \gamma)v_{t+1}^2$$

$$\theta_{t+1} = \theta_t + v_{t+1}$$

# Adam

$$m_{t+1} = \gamma_1 m_t + (1 - \gamma_1)\nabla F(\theta)$$

$$g_{t+1} = \gamma_2 g_t + (1 - \gamma_2)\nabla F(\theta)^2$$

$$\hat{m}_{t+1} = \frac{m_{t+1}}{1 - \gamma_1^{t+1}}$$

$$\hat{g}_{t+1} = \frac{m_{t+1}}{1 - \gamma_2^{t+1}}$$

$$\theta_{t+1} = \theta_t - \frac{\eta \hat{m}_{t+1}}{\sqrt{\hat{g}_{t+1}} + \epsilon}$$