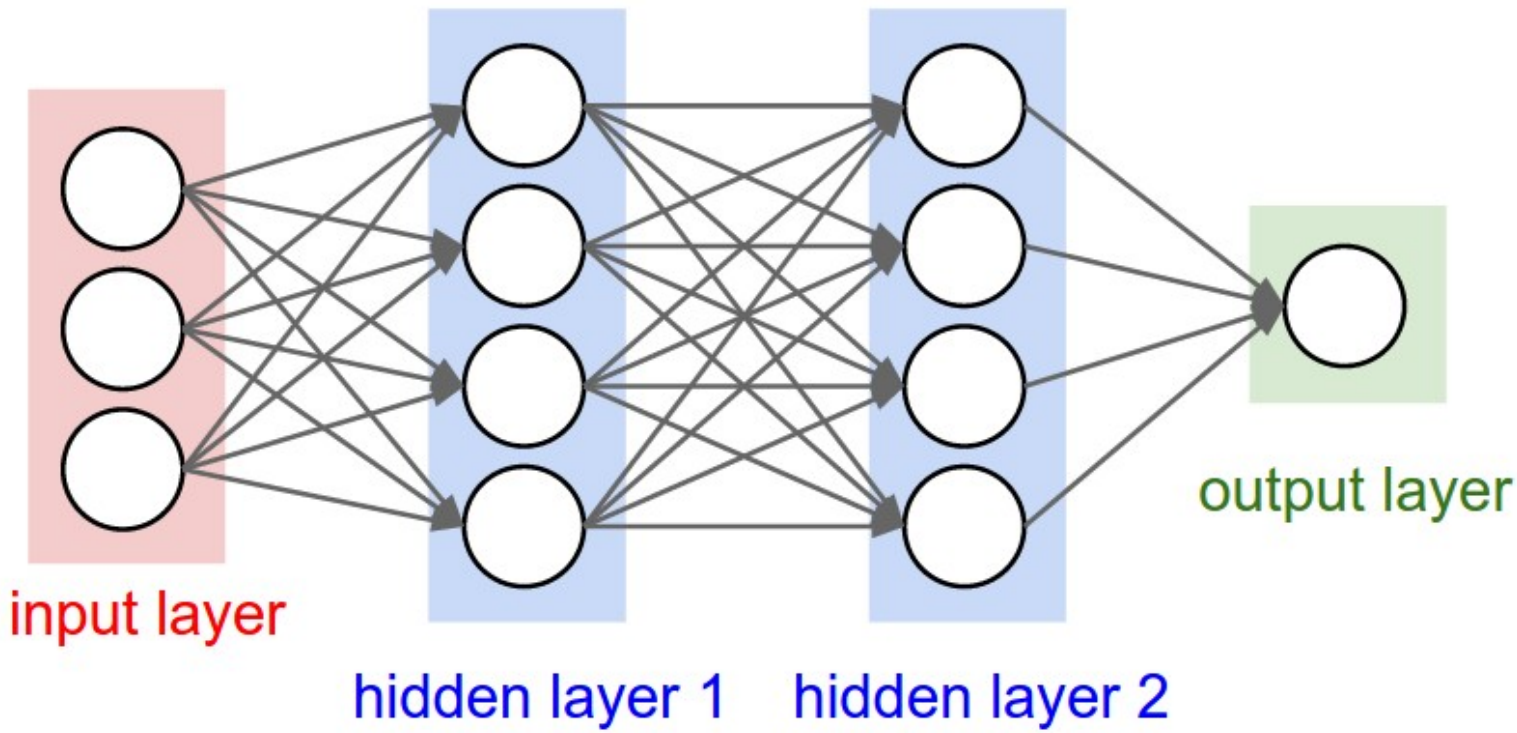
Модификации метода стохастического градиентного спуска для задач машинного обучения с большими объемами данных

Владислав Чабаненко

Научный руководитель: научный сотрудник
Кропотов Дмитрий Александрович

2016

Нейронные сети



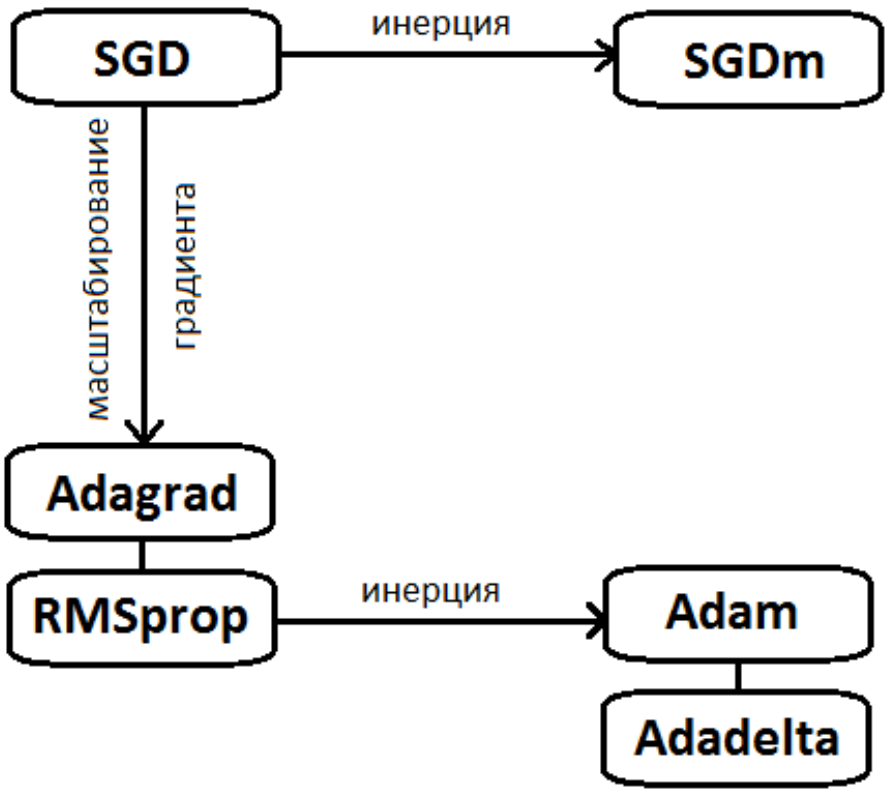
Стохастический градиентный спуск (SGD) и его модификации

$$\theta_{t+1} = \theta_t - \eta \nabla f(\theta_t) \tag{1}$$

Наиболее популярные модификации:

- Стохастический градиентный спуск с инерцией (SGDm)
- Метод адаптивного градиента (Adagrad)
- Метод адаптивного скользящего среднего градиентов (RMSprop)
- Метод адаптивного шага обучения (Adadelta)
- Метод адаптивной инерции (Adam)

Интуиция методов



Ковариационный сдвиг и батч-нормализация

- Проблема *ковариационного сдвига* (Shimodaira, 2000)
- Нормализация входных данных (LeCun et al., 1998b; Wiesler & Ney, 2011)
- Батч-нормализация (Ioffe & Szegedy, 2015)

$$\hat{x}^k = \frac{x^k - \mathbb{E}[x^k]}{\sqrt{\mathrm{Var}[x^k]}} \tag{2}$$

$$y^k = \gamma^k \hat{x}^k + \beta^k \tag{3}$$

Батч-нормализация

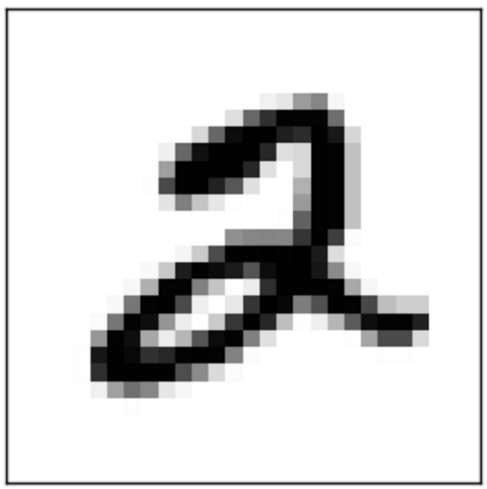
- Уменьшает ковариационный сдвиг и ускоряет обучение.
- Является дифференцируемым преобразованием.
- Использует мини-батчи, тем самым подходит для работы с большими данными.

Авторы идеи (Ioffe & Szegedy, 2015) обучали нейронную сеть с батч-нормализацией с помощью метода SGD. Мы исследуем совместимость батч-нормализации и модификаций SGD, расмотренных выше.

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

Для экспериментов использовались следующие архитектуры:

- полносвязная сеть (MLP): 3 полносвязных скрытых слоя по 100 нейронов;
- сверточная сеть (CNN): 2 сверточных слоя с макс-пулингом, затем один полносвязный слой с 256 нейронами;
- полносвязная глубокая сеть (deep MLP): 20 полносвязных скрытых слоёв по 30 нейронов;
- сверточная глубокая сеть (deep CNN): 5 сверточных подсетей (каждая состоит из 3-х сверточных слоёв с макс-пулингом), затем один полносвязный слой с 256 нейронами.

Гипотезы

- Добавление батч-нормализации в сеть увеличивает скорость сходимости обучения сети для всех методов.
- Чем метод сложнее, тем батч-нормализация слабее ускоряет его сходимость.
- Батч-нормализация сильнее проявляет ускорение обучения на глубоких сетях.

Результаты

- Предварительно для каждого метода на различных датасетах и архитектурах был подобран оптимальный шаг обучения.
- Для измерения повышения качества работы методов был выбран показатель относительного улучшения:

$$\mathrm{rel} = \frac{y - x}{100 - x}, \ rel \leq 1 \tag{4}$$

Влияние БН на сверточные сети, топ-3 наименьшего улучшения

Номер эпохи	**2**	**5**	**20**	**35**	**50**
SGD	0.64	0.57	0.48	0.49	0.37
SGDm	0.7	0.52	0.47	0.34	0.3
Adam	**0.44**	**0.17**	**0.29**	**0.15**	**0.11**
Adagrad	0.46	0.38	0.31	**0.3**	0.33
Adadelta	**0.19**	**0.07**	**0.23**	**0.15**	**0.18**
RMSprop	**0.34**	**0.22**	**0.11**	0.36	**0.27**

Улучшения качества по эпохам MNIST, CNN

Номер эпохи	**2**	**5**	**20**	**35**	**50**
SGD	0.32	0.4	0.3	0.34	0.32
SGDm	0.34	0.34	0.35	0.27	0.24
Adam	**0.19**	**0.22**	**0.23**	**0.2**	**0.15**
Adagrad	**0.26**	**0.26**	0.35	0.37	0.35
Adadelta	0.27	0.34	**0.17**	**0.22**	**0.22**
RMSprop	**0.23**	**0.22**	**0.26**	**0.21**	**0.19**

Улучшения качества по эпохам CIFAR-10, CNN

Влияние БН на полносвязные сети, топ-3 наименьшего улучшения

Номер эпохи	**2**	**5**	**10**	**25**	**50**
SGD	0.67	0.63	0.42	0.25	0.22
SGDm	0.47	0.34	**0.13**	**-0.08**	**0.05**
Adam	**0.21**	**0.2**	**-0.05**	**-0.15**	**0.07**
Adagrad	0.34	0.27	0.17	0.12	0.14
Adadelta	**0.3**	**0.08**	**0.08**	**-0.0**	**0.03**
RMSprop	**0.28**	**0.27**	0.35	0.21	0.25

Улучшения качества по эпохам MNIST, MLP

Номер эпохи	**2**	**5**	**10**	**25**	**50**
SGD	0.23	0.18	0.22	0.21	0.17
SGDm	**0.17**	**0.13**	**0.15**	**0.13**	**0.1**
Adam	**0.13**	**0.12**	**0.11**	**0.13**	**0.12**
Adagrad	0.25	0.23	0.23	0.18	0.17
Adadelta	**0.18**	**0.14**	**0.19**	**0.17**	**0.15**
RMSprop	0.2	0.18	0.21	0.22	0.16

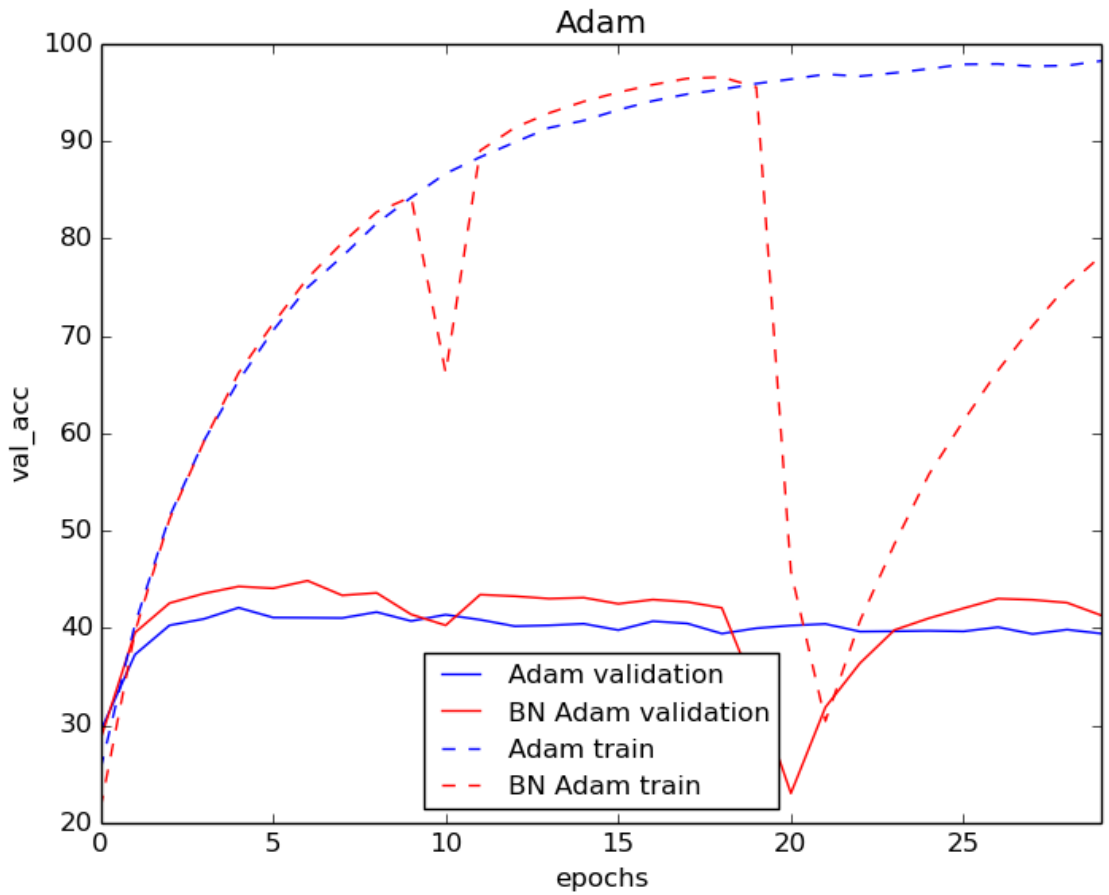
Улучшения качества по эпохам CIFAR-10, MLP

БН и глубокая сверточная сеть

Номер эпохи	2	5	20	35	50
SGD	0.1	0.23	0.41	0.44	0.34
SGDm	0.1	0.31	0.36	0.42	0.18
Adam	0.17	0.35	0.37	0.41	0.37
Adagrad	-0.05	0.21	0.24	0.28	0.28
Adadelta	0.14	0.25	0.51	0.45	0.43
RMSprop	-0.1	0.27	0.37	0.46	0.42

Разница в улучшении качества deep CNN и CNN для CIFAR-10

Комбинация БН с методом Adam



Рекомендации

- Для полносвязной неглубокой архитектуры сети батч-нормализацию стоит применять к более простым методам.
- Для глубоких сетей важно использовать батч-нормализацию.
- Если время или количество эпох ограничено и очень мало, то однозначно стоит добавить в сеть батч-нормализацию.
- Для метода Adam с батч-нормализацией нужно быть аккуратным: чтобы не возникло проблем при обучении, требуется аккуратно подобрать параметры метода.

На защиту выносится:

1. Экспериментальное исследование гипотезы о зависимости величины улучшения качества работы метода при добавлении в сеть батч-нормализации от сложности метода.
2. Экспериментальное исследование влияния глубины нейронной сети на совместимость батч-нормализации и модификаций метода стохастического градиентного спуска.
3. Рекомендации по применению батч-нормализации для рассмотренных архитектур и методов.

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

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

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

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

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

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