

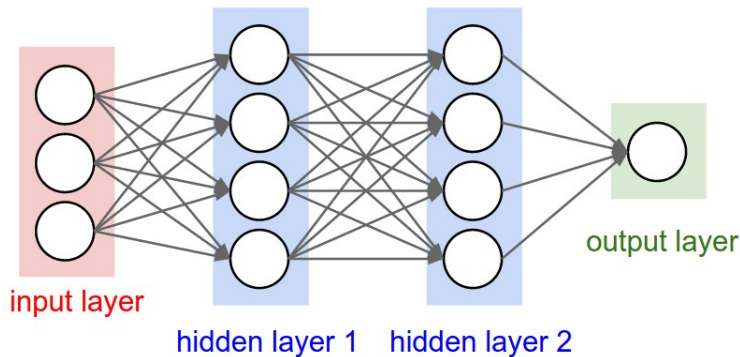
Модификации метода стохастического градиентного спуска для задач машинного обучения с большими объемами данных

Владислав Чабаненко

Научный руководитель: научный сотрудник
Кропотов Дмитрий Александрович

2016

Нейронные сети



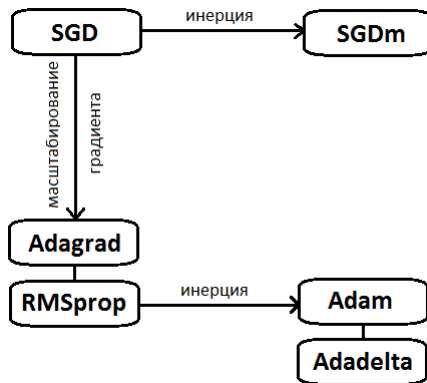
Стохастический градиентный спуск (SGD) и его модификации

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

Наиболее популярные модификации:

- Стохастический градиентный спуск с инерцией (SGDm)
- Метод адаптивного градиента (Adagrad)
- Метод адаптивного скользящего среднего градиентов (RMSprop)
- Метод адаптивного шага обучения (Adadelata)
- 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)

$$\hat{x}^k = \frac{x^k - \mathbb{E}[x^k]}{\sqrt{\text{Var}[x^k]}} \quad (2)$$

$$y^k = \gamma^k \hat{x}^k + \beta^k \quad (3)$$

Батч-нормализация

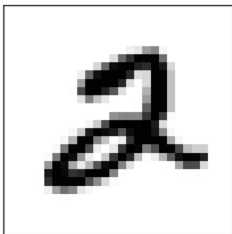
- Является дифференцируемым преобразованием.
- Используем мини-батчи, тем самым подходит для работы с большими данными.
- Уменьшает ковариационный сдвиг и ускоряет обучение.

Авторы идеи (Ioffe & Szegedy, 2015) применили батч-нормализацию для метода SGD. Мы исследуем применение батч-нормализации к рассмотренным выше модификациям SGD.

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:

$$\text{rel} = \frac{y - x}{100 - x}, \text{ rel} \leq 1 \quad (4)$$

Влияние БН на все методы

Методы	MNIST + MLP	CIFAR-10 + MLP	MNIST + CNN	CIFAR-10 + CNN
SGD	0.24	0.2	0.35	0.36
SGDm	0.06	0.1	0.29	0.25
Adam	0.05	0.09	0.13	0.14
Adagrad	0.13	0.17	0.3	0.33
Adadelta	0.03	0.11	0.19	0.2
RMSprop	0.22	0.16	0.24	0.19

Улучшение качества всех методов при добавлении БН

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	0.23	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

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

Зависимость влияния БН от первоначального качества

Номер эпохи	1	2	25	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. Жирным выделены топ-3 наименьшего улучшения

Номер эпохи	1	2	25	35	50
SGD	24.26	22.14	58.95	63.23	65.45
SGDm	30.03	37.98	64.37	68.3	71.54
Adam	43.28	47.3	68.02	70.79	72.54
Adagrad	29.82	37.25	60.86	62.35	64.63
Adadelta	31.31	32.96	67.12	70.56	70.81
RMSprop	25.6	39.89	64.24	68.96	70.31

Качество методов по эпохам на CIFAR-10, CNN. Выделены топ-3 лучшего качества

БН и глубокая полносвязная сеть

Номер эпохи	2	5	10	25	50
SGD	0.84	0.9	0.94	0.93	0.7
SGDm	0.8	0.83	0.32	0.19	0.03
Adam	0.72	0.66	0.58	0.4	0.28
Adagrad	0.85	0.84	0.45	0.37	0.37
Adadelta	0.58	0.69	0.27	0.18	0.14
RMSprop	0.72	0.82	0.45	0.39	0.37

Улучшения качества для MNIST, deep MLP

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 для MNIST, MLP

БН и глубокая сверточная сеть

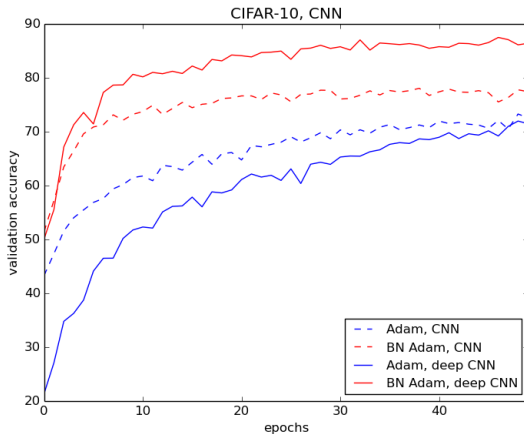
Номер эпохи	2	5	20	35	50
SGD	0.42	0.63	0.71	0.78	0.68
SGDm	0.44	0.65	0.71	0.69	0.42
Adam	0.36	0.57	0.6	0.61	0.52
Adagrad	0.21	0.47	0.59	0.65	0.63
Adadelta	0.41	0.59	0.68	0.67	0.65
RMSprop	0.13	0.49	0.63	0.67	0.61

Улучшение качества для CIFAR-10, deep 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

Улучшение качества для CIFAR-10, CNN

БН и глубокая архитектура сети на примере метода Adam



Рекомендации

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

На защиту выносятся:

- 1 Экспериментальные исследования гипотезы об ускорении сходимости модификаций метода стохастического градиентного спуска при добавлении в сеть батч-нормализации.
- 2 Экспериментальные исследования гипотезы о зависимости величины улучшения качества работы метода при добавлении в сеть батч-нормализации от сложности метода.
- 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

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

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

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

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

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