Данные и метрики BatchNorm

Иван Карпухин

Преподаватель

Иван Карпухин

Профессионально занимаюсь машинным обучением более 6 лет

Проекты (Тинькофф, VK, Яндекс):

- Голосовая биометрия
- Распознавание лиц и текстов
- Виртуальный аватар
- Исследования

Задание

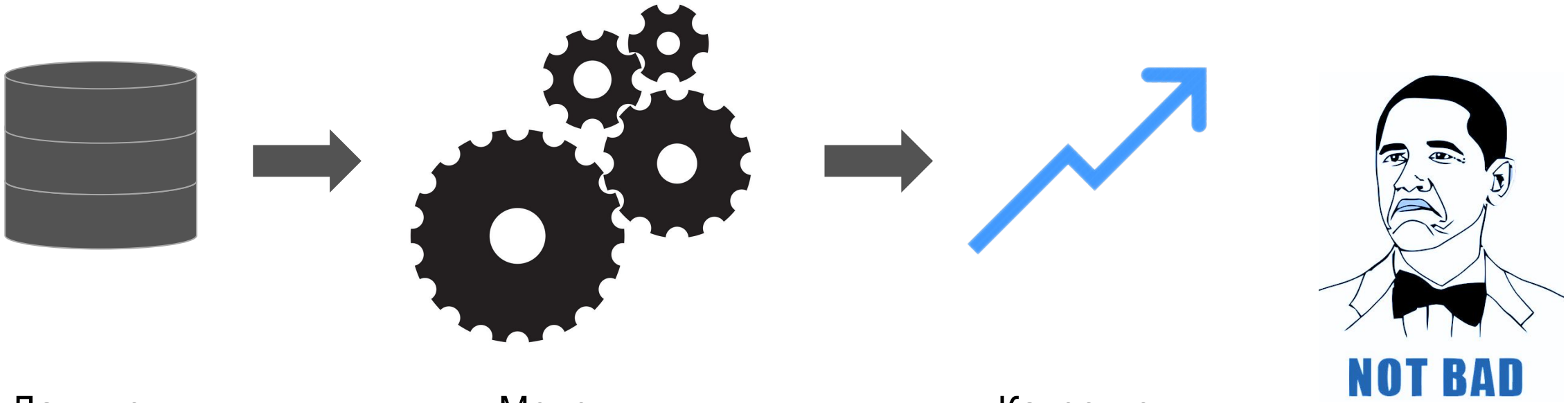
- 3 минуты
- Анонимно
- Ссылка в чате

- Обсудим через несколько слайдов

https://forms.gle/te9K251Zz1sSEy6n6

ML сложнее чем кажется

В теории:



Данные Модель Качество

ML сложнее чем кажется

На практике:

ML сложнее чем кажется

На практике:

Данные
Train/dev/test

Bias / variance recap

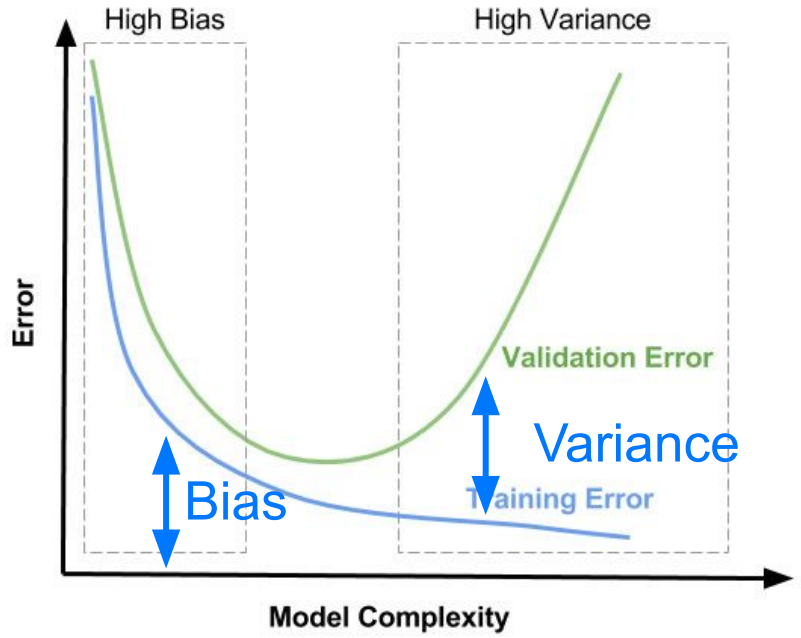
Train set и Validation set из одного распределения

Bias - величина ошибки на Train

Variance - разница ошибок Validation и Train

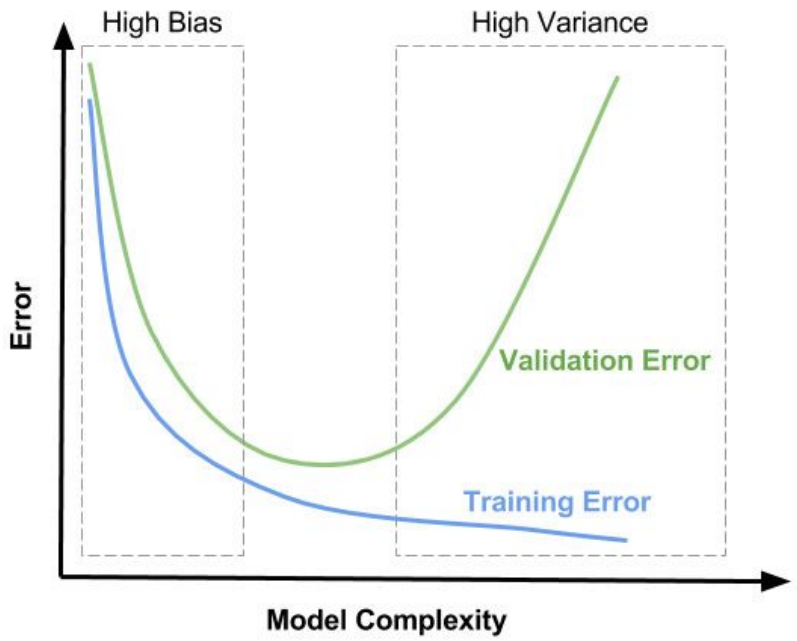
Описывают соответствие модели и данных

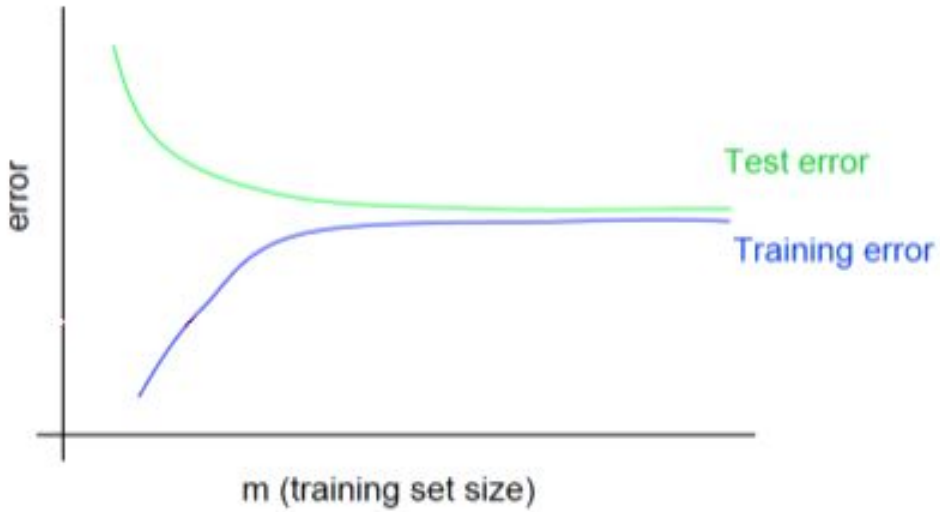
Терминология из анализа MSE*



* https://en.wikipedia.org/wiki/Bias%E2%80%93variance_tradeoff#Bias%E2%80%93variance_decomposition_of_mean_squared_error

Bias / variance recap





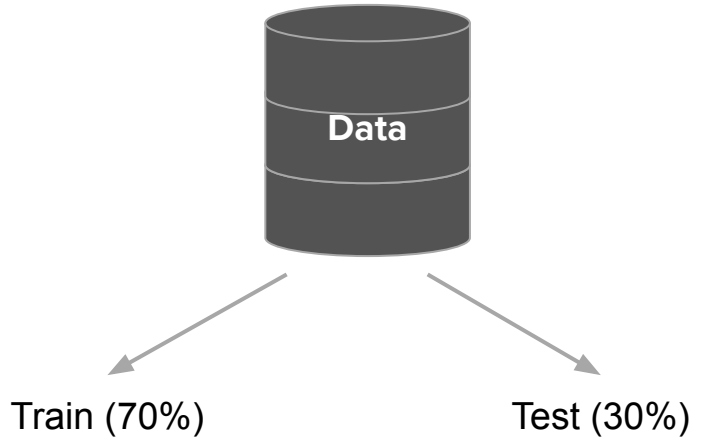
Основные вопросы

- Какие корпусы нужны?
- Какого размера?
- Из какого распределения?

Train / test

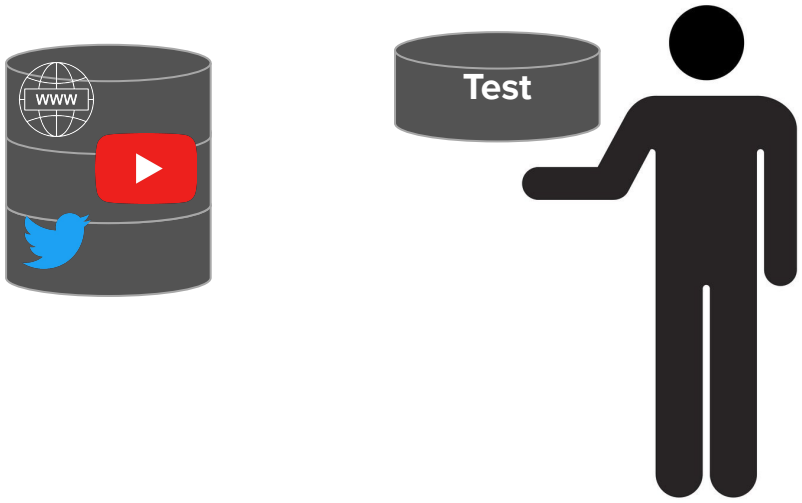
Причина 1

- Алгоритм переобучается под Train
- Нужен независимый Test для оценки



Причина 2

- Train большой, но из другого домена
- Test от заказчика

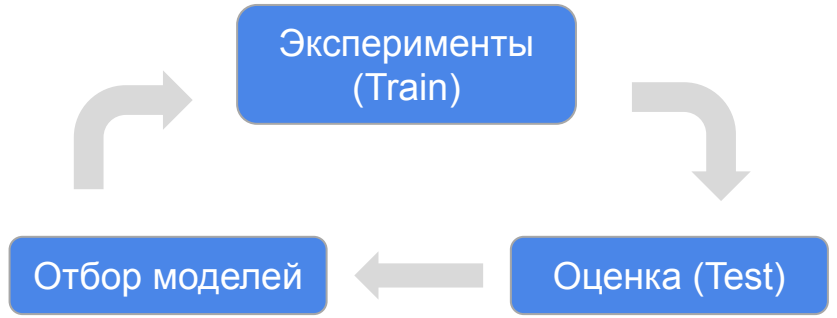


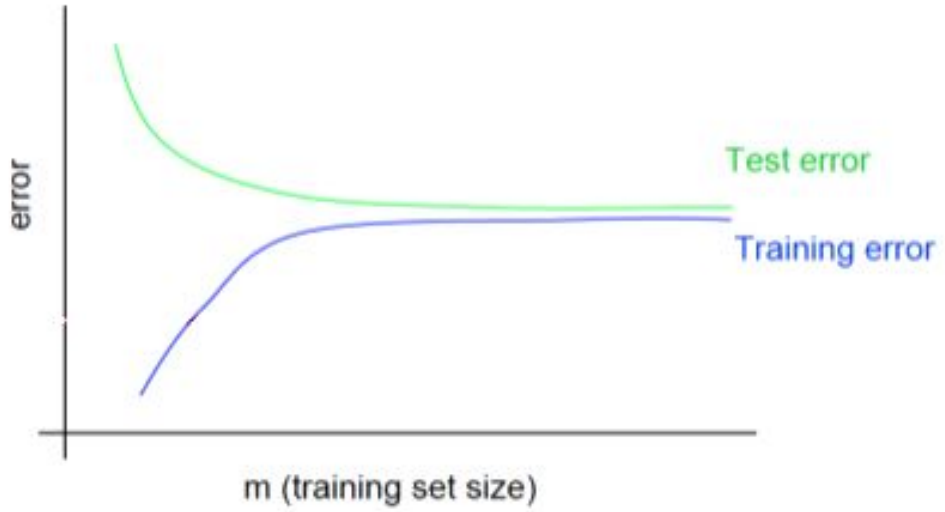
Train / test

Отбираем модели по Test метрикам

=> переобучаемся под Test

Увеличить Test?



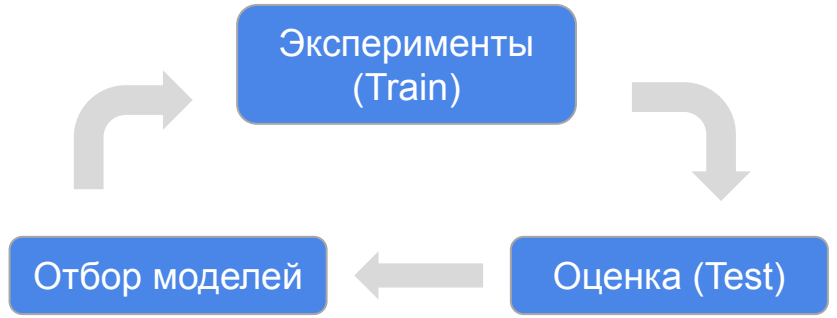


Train / test

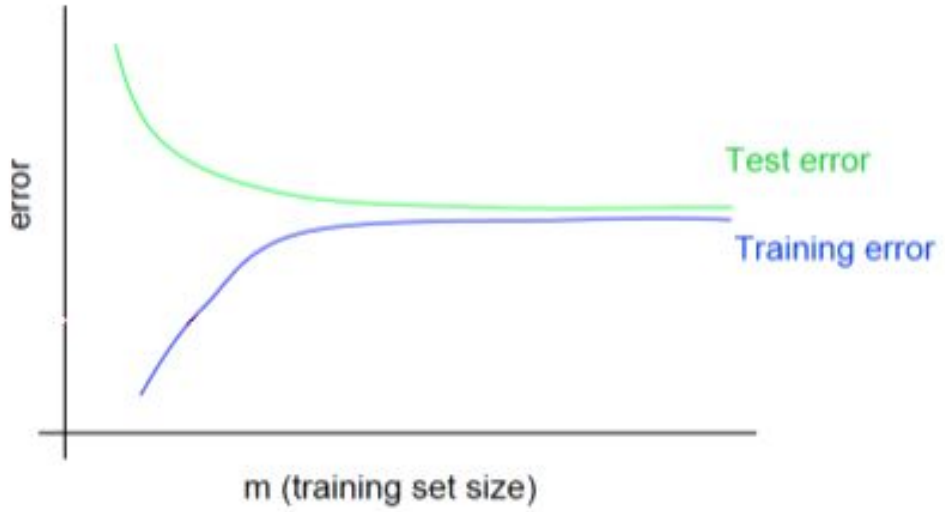
Отбираем модели по Test метрикам

=> переобучаемся под Test

Увеличить Test?



Сперва оценить степень переобучения

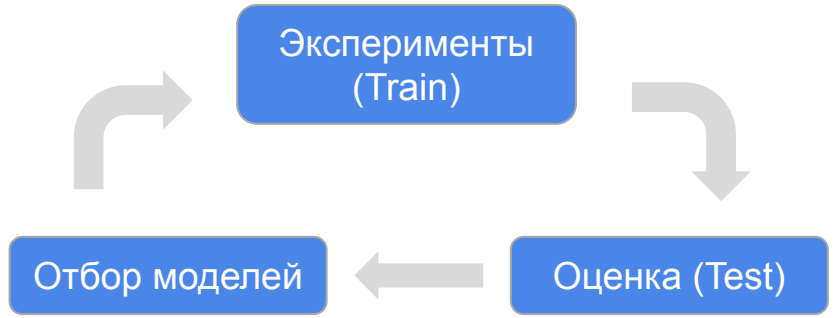


Development set

Отбираем модели по Test метрикам

=> переобучаемся под Test

<u>Решение:</u> Dev корпус



Корпус	Размер*	Распределение	Назначение
Train	10.000 - 10.000.000	М.б. смещенное	Обучение
Dev	1000 - 100.000	Несмещенное	Отбор модели
Test	1000 - 100.000	Несмещенное	Оценка модели

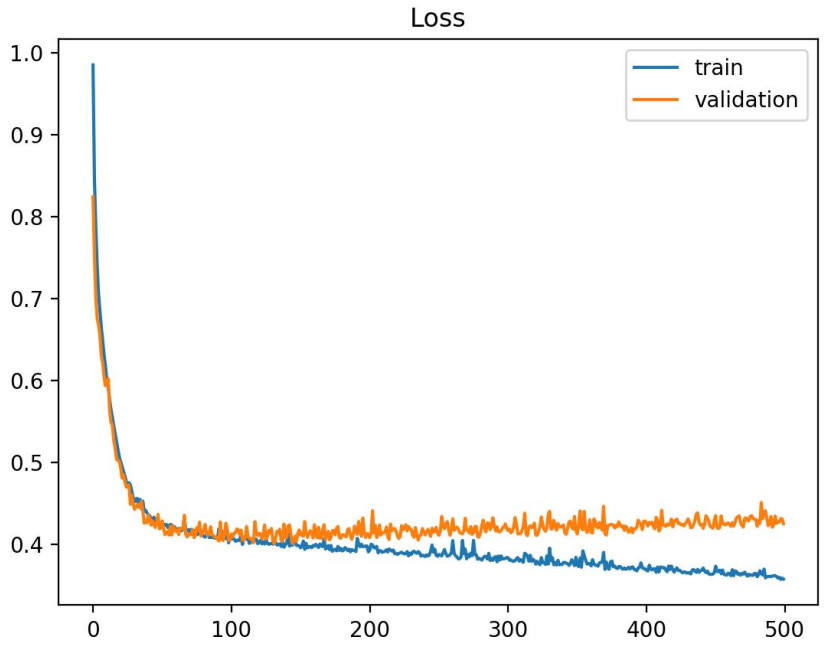
* В некоторых задачах, особенно unsupervised и NLP, размер может заметно отличаться

Проблема

$$Error_{dev} - Error_{train} = 0.1$$

На сколько переобучилась модель?

Как улучшить качество на Dev?



Проблема

$$Error_{dev} - Error_{train} = 0.1$$

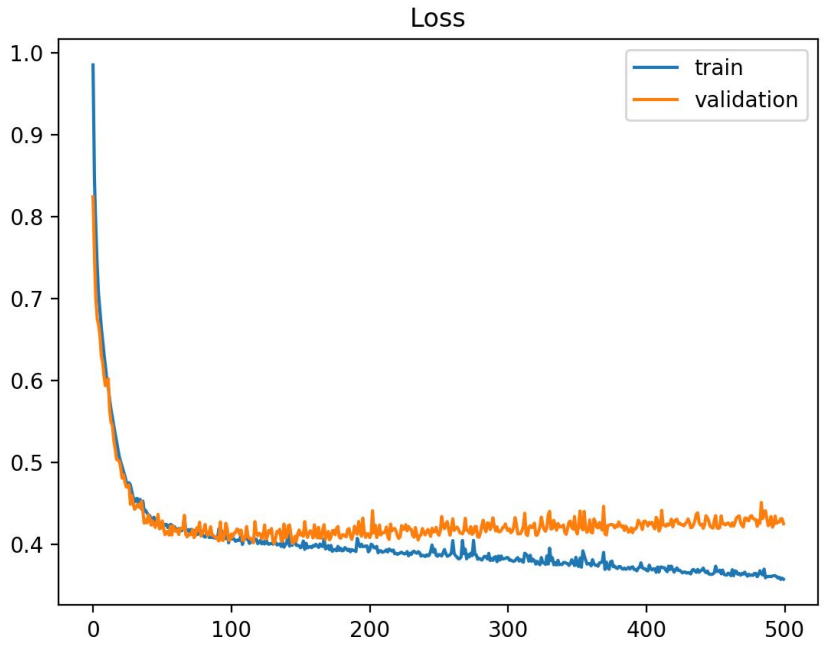
Train - смещённый

Dev - несмещённый

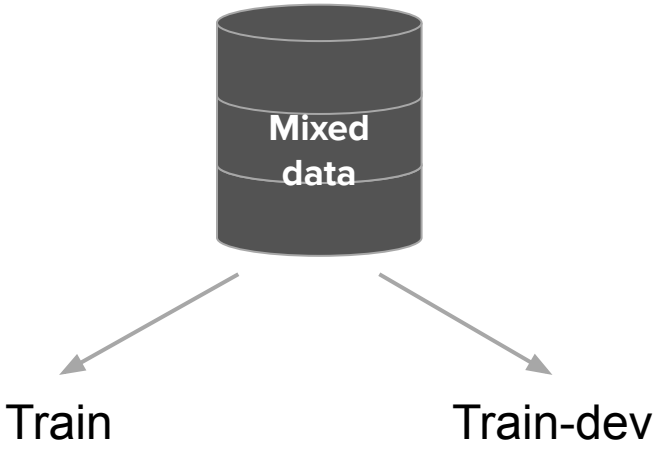
Уменьшать число параметров?

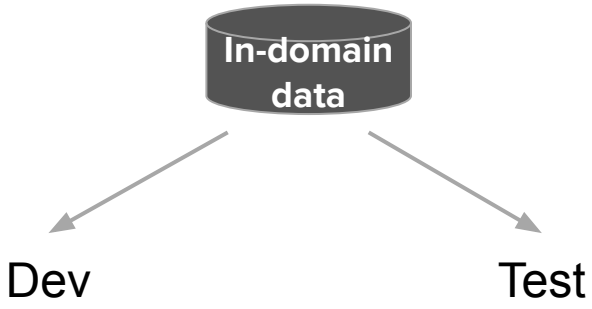
Увеличивать Train?

Искать несмещённые данные для Train?



Train-dev set





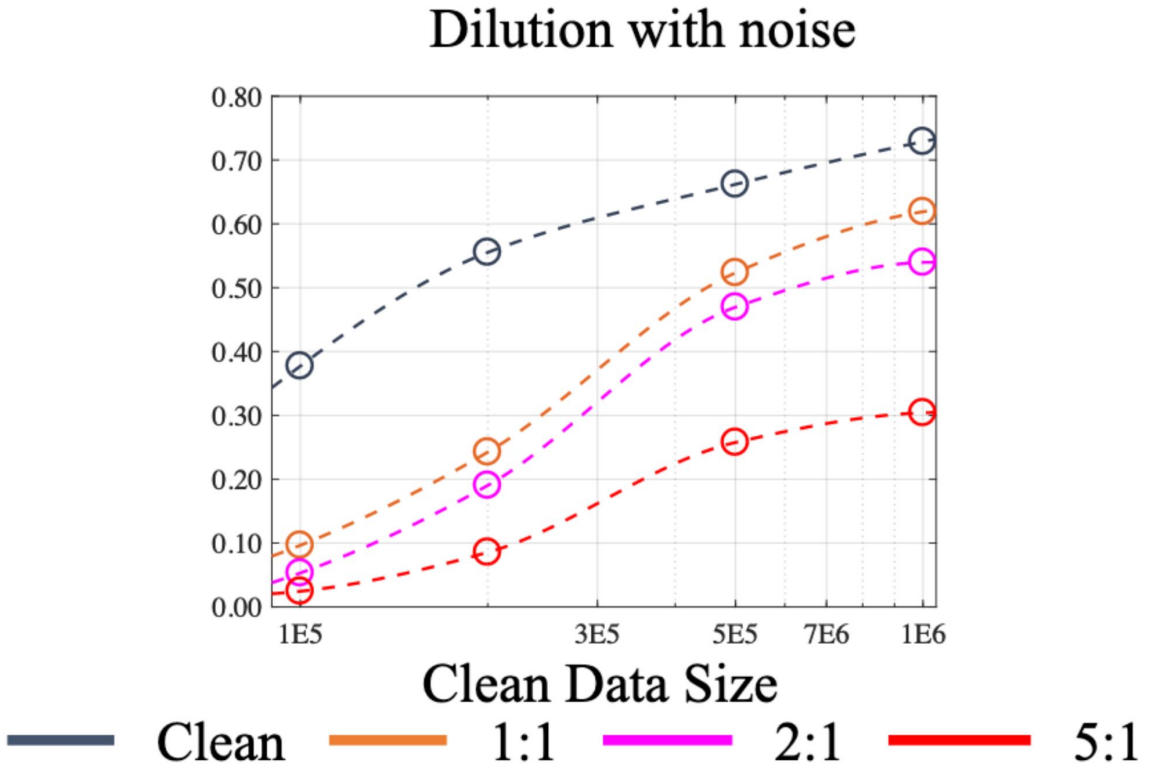
Data quality

Хотим улучшить качество модели за счет данных

- Собрать новых данных?
- Почистить имеющиеся?

Data quality

- Шумных данных нужно больше
- Количество может компенсировать качество



Итоги про данные

- Dataset size is not all you need
- Распределение данных важно
- Quality / dataset size trade-off
- Переобучаемся не только под train, но и под dev (и даже под test)
- Dev и Test стоит иногда менять
- Если train и dev из разных источников, можно выделить train-dev

Вопросы

Размер test set

Процесс оценки

Y* - правильный ответ

$$Error = \begin{cases} 0, Y = Y^* \\ 1, Y \neq Y^* \end{cases}$$

Error - случайная величина Bernoulli(p)

p - вероятность ошибки модели

Задача оценки

N0 - число правильных классификаций

N1 - число ошибок

Размер Test: N = N1 + N0

Error - случайная величина Bernoulli(p)

p - вероятность ошибки модели

$$P(p = x|N_0, N_1)?$$

Оценка

$$P(p = x|N_0, N_1) = \frac{P(N_0, N_1|p = x)P(p = x)}{\int\limits_y P(N_0, N_1|p = y)P(p = y)dy}$$

Оценка

$$P(p = x|N_0, N_1) = \frac{P(N_0, N_1|p = x)P(p = x)}{\int\limits_y P(N_0, N_1|p = y)P(p = y)dy}$$

$$P(p = x) : Uniform(0, 1)$$

Оценка

$$P(p = x|N_0, N_1) = \frac{P(N_0, N_1|p = x)P(p = x)}{\int\limits_y P(N_0, N_1|p = y)P(p = y)dy}$$

$$P(p = x) : Uniform(0, 1)$$

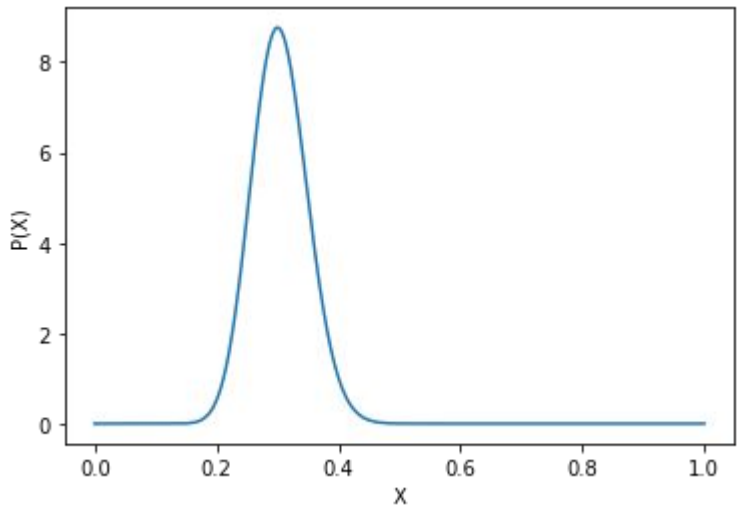
$$P(N_0, N_1|p = x) = x^{N_1}(1 - x)^{N_0}, x \in [0, 1]$$

Оценка

$$P(p = x|N_0, N_1) = \frac{P(N_0, N_1|p = x)P(p = x)}{\int\limits_y P(N_0, N_1|p = y)P(p = y)dy}$$

$$P(p = x) : Uniform(0, 1)$$

$$P(N_0, N_1|p = x) = x^{N_1}(1 - x)^{N_0}, x \in [0, 1]$$



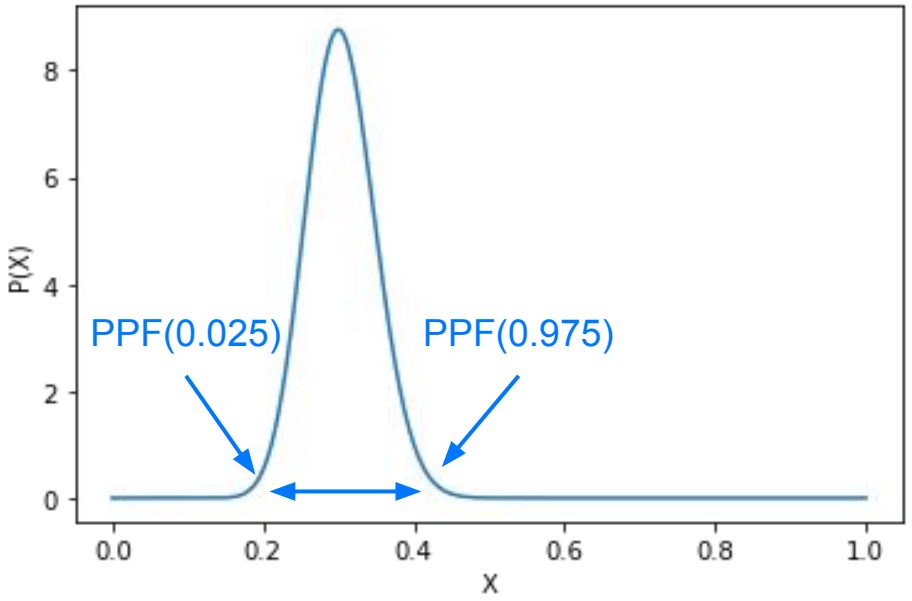
$$P(p = x|N_0, N_1) = Beta(N_1+1, N_0+1) = \frac{x^{N_1}(1 - x)^{N_0}}{B(N_1 + 1, N_0 + 1)}$$

Доверительный интервал

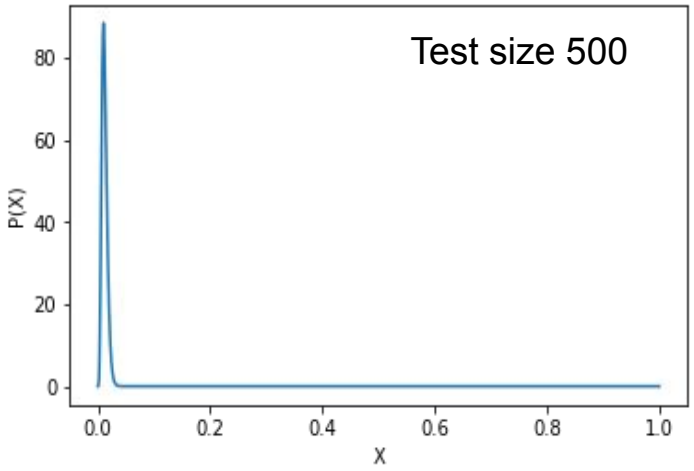
Доверительный интервал с уровнем доверия alpha = 0.95?

$$PPF_p(\alpha) = x : P(p \leq x) = \alpha$$

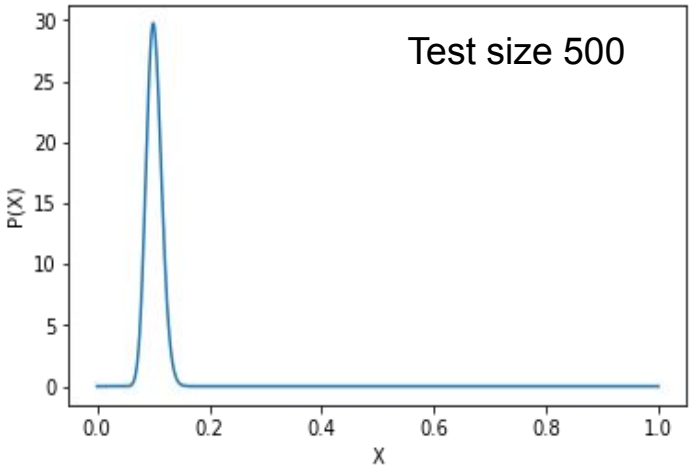
$$\Delta = PPF(\alpha + \frac{1-\alpha}{2}) - PPF(\frac{1-\alpha}{2})$$



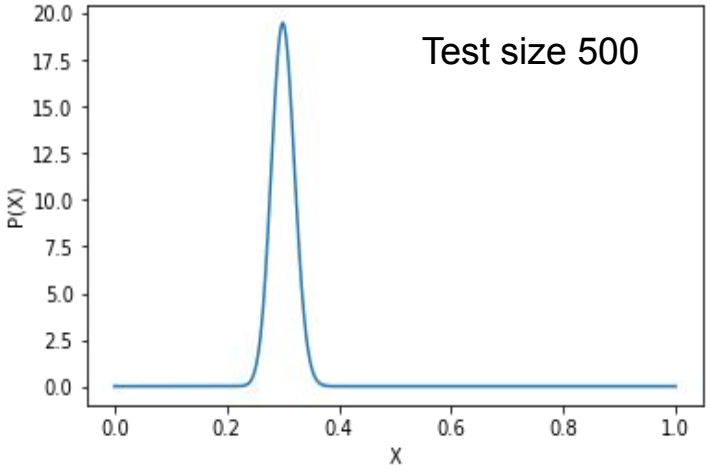
Примеры



Mean Error = 0.01



Mean Error = 0.1



Mean Error = 0.3

Вопросы

Метрики

Виды метрик

Технические

- оценивают подсистемы
- выявляют возможности для улучшений

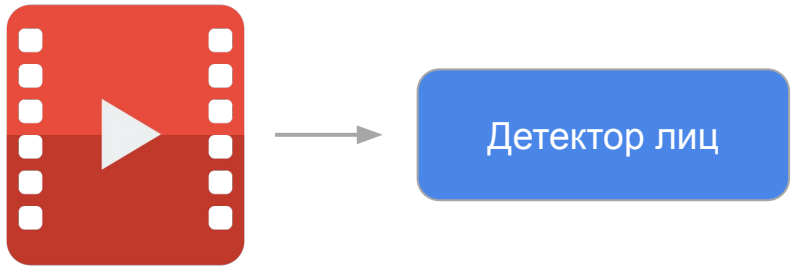
Продуктовые

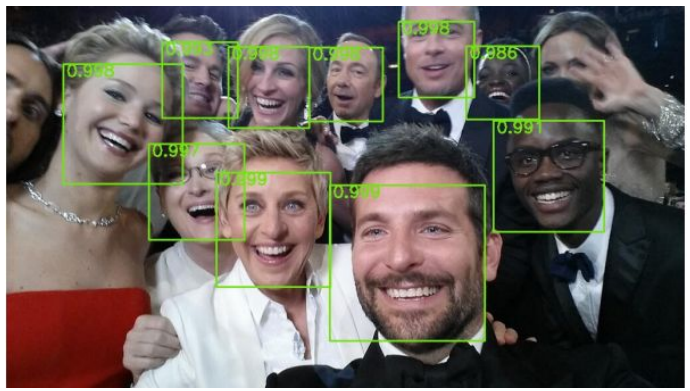
- связаны с бизнесом
- оценивают систему целиком
- одно число

Метрики из статей и стандартов

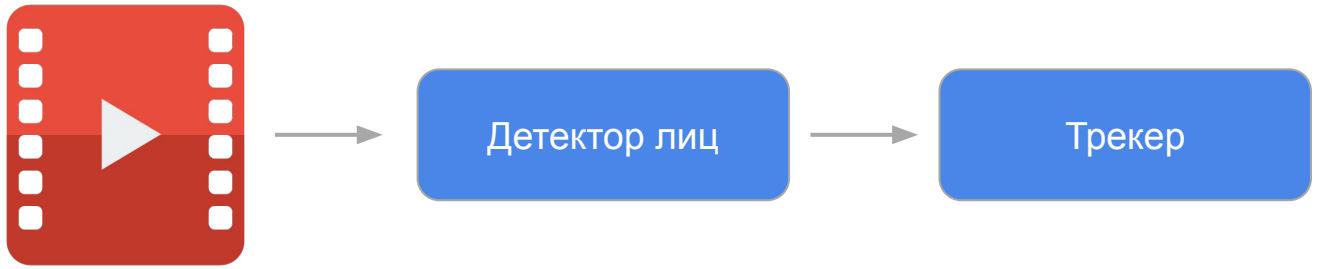
- сравнение с конкурентами

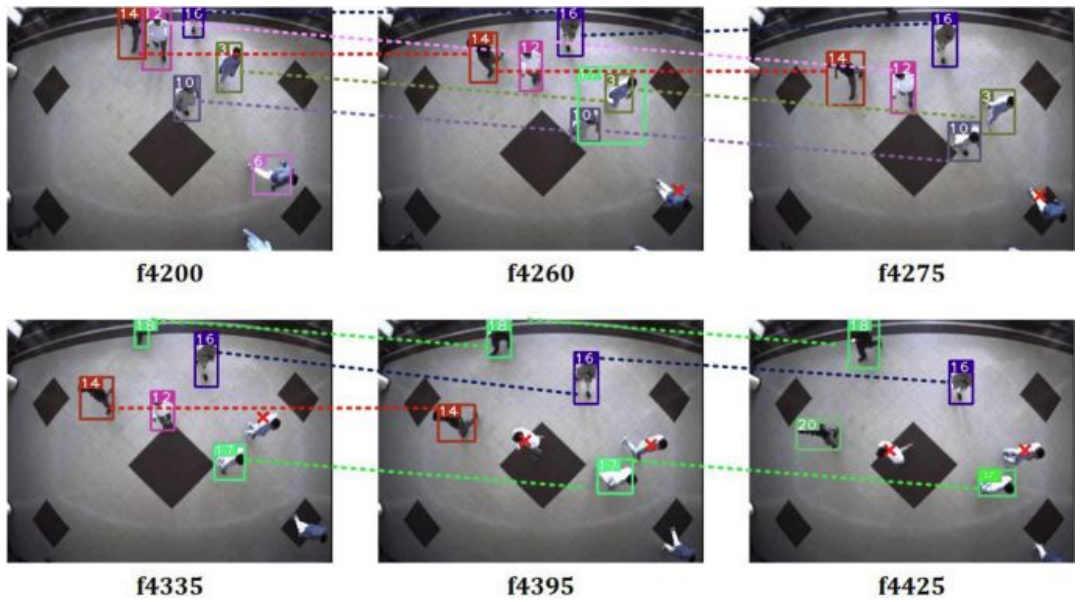
Пример: трекинг людей



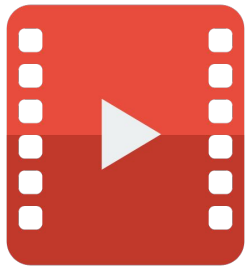


Пример: трекинг людей

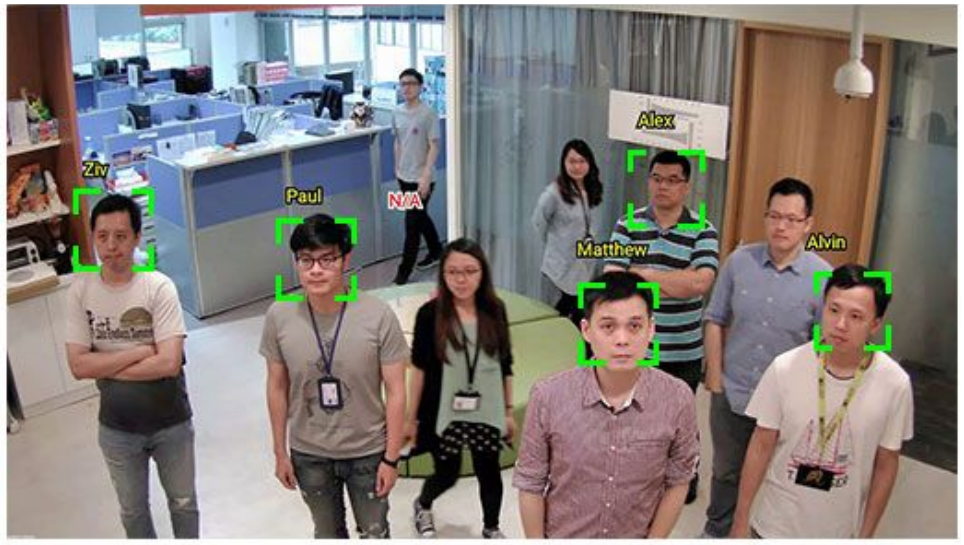




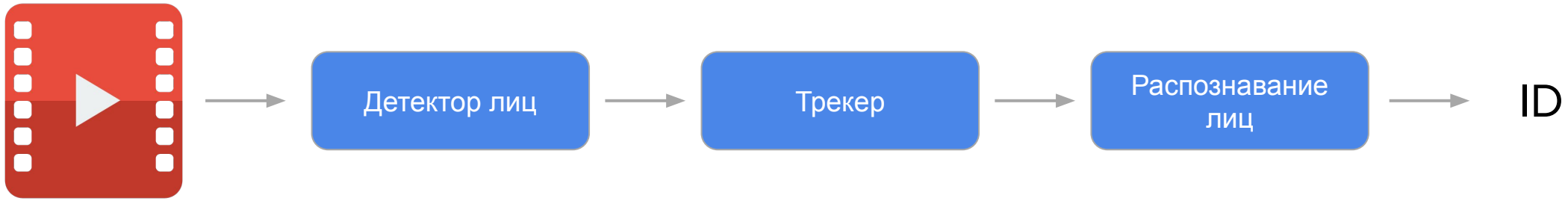
Пример: трекинг людей



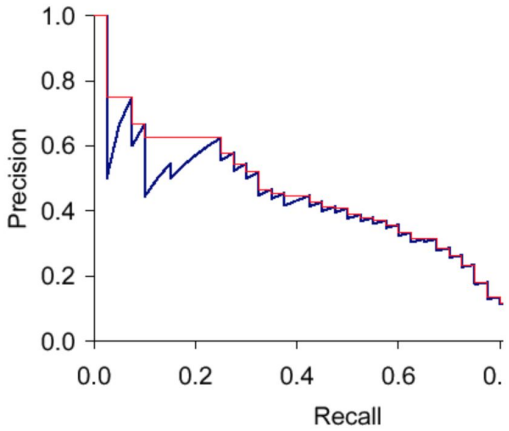
Детектор лиц → Трекер → Распознавание лиц → ID



Технические метрики



Average precision



Mostly tracked
Mostly lost
Identity switches
Fragmentation

Accuracy 1 / 1000

Продуктовые метрики

Задача:

- у заказчика есть база сотрудников с фото
- нужно найти посторонних людей на видео с камеры

Какое число выбрать в качестве продуктовой метрики?

Продуктовые метрики

Какие параметры важны:

- скорость работы пайплайна (ms / frame)
- частота ложных срабатываний (1 / hour)
- вероятность правильной классификации постороннего

Как сделать одно число?

Вариант 1: усреднение

1. Можно связать частоту ложных срабатываний с вероятностью правильной классификации сотрудника
2. Вероятности правильной классификации сотрудника и постороннего можно усреднить (mean, harmonic mean)

Как быть с быстродействием?

Вариант 2: ограничение

Какие параметры важны:

- скорость работы пайплайна (ms / frame)
- частота ложных срабатываний (1 / hour)
- вероятность правильной классификации постороннего

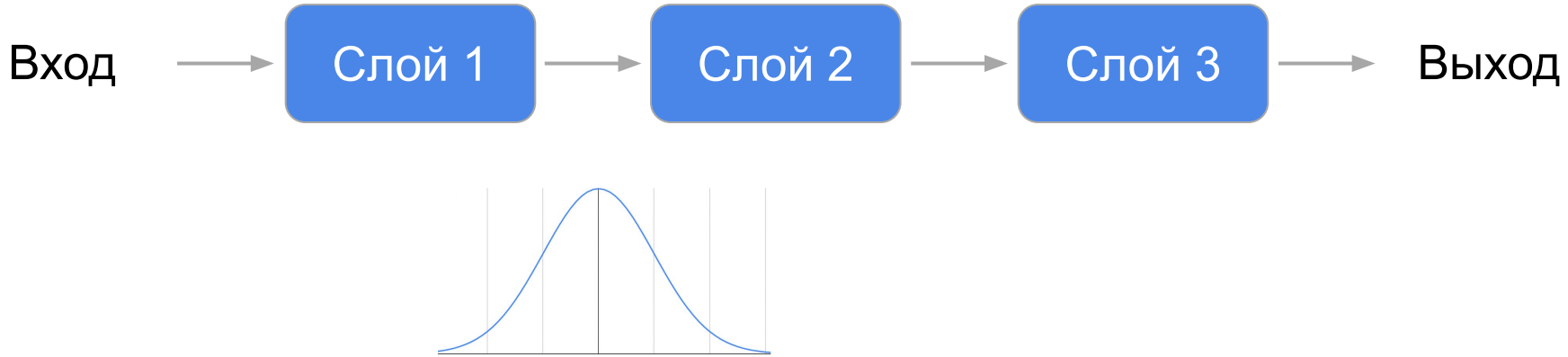
1. Ложные срабатывания допустимы не чаще 1 / час (в среднем)
2. Нужно обрабатывать кадр быстрее 200ms на CPU

=> остается один свободный параметр - вероятность обнаружения постороннего

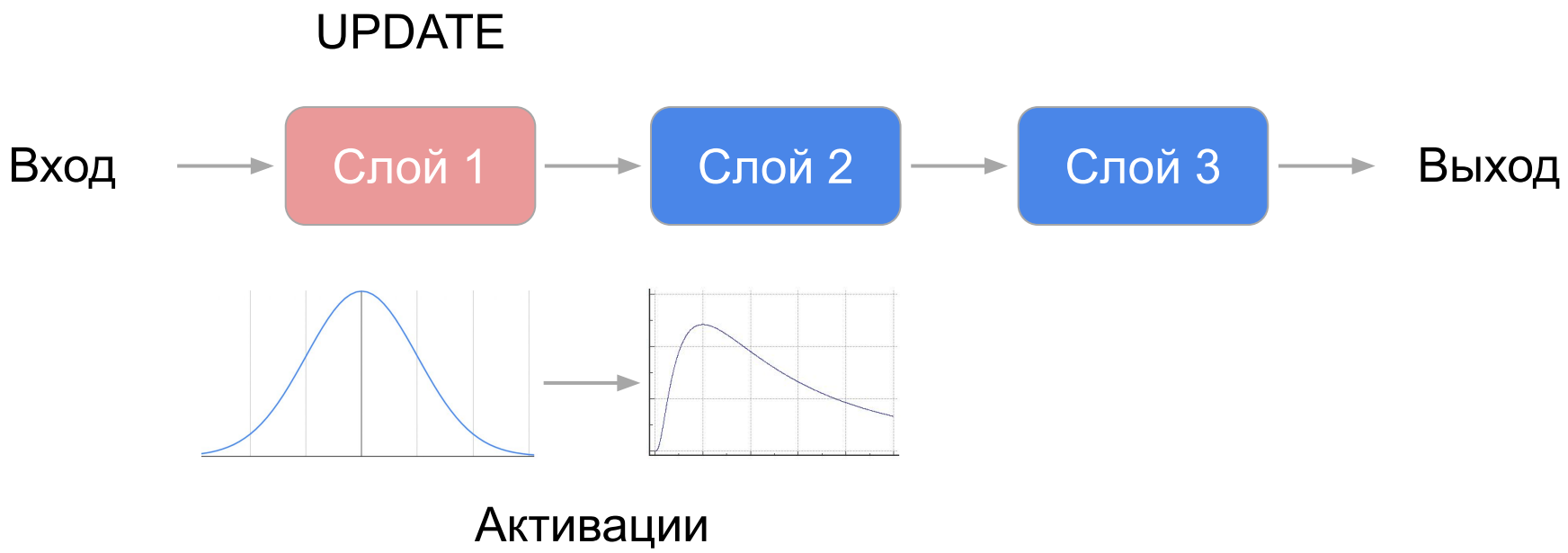
Вопросы

Batch normalization

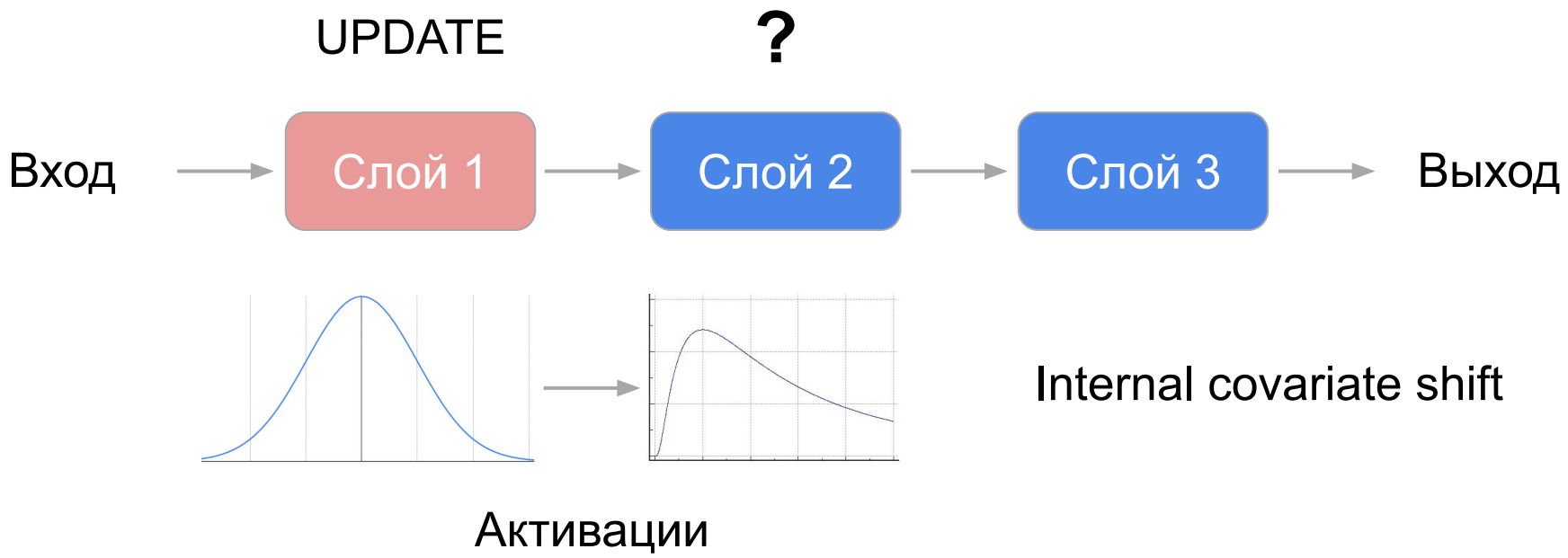
Проблема



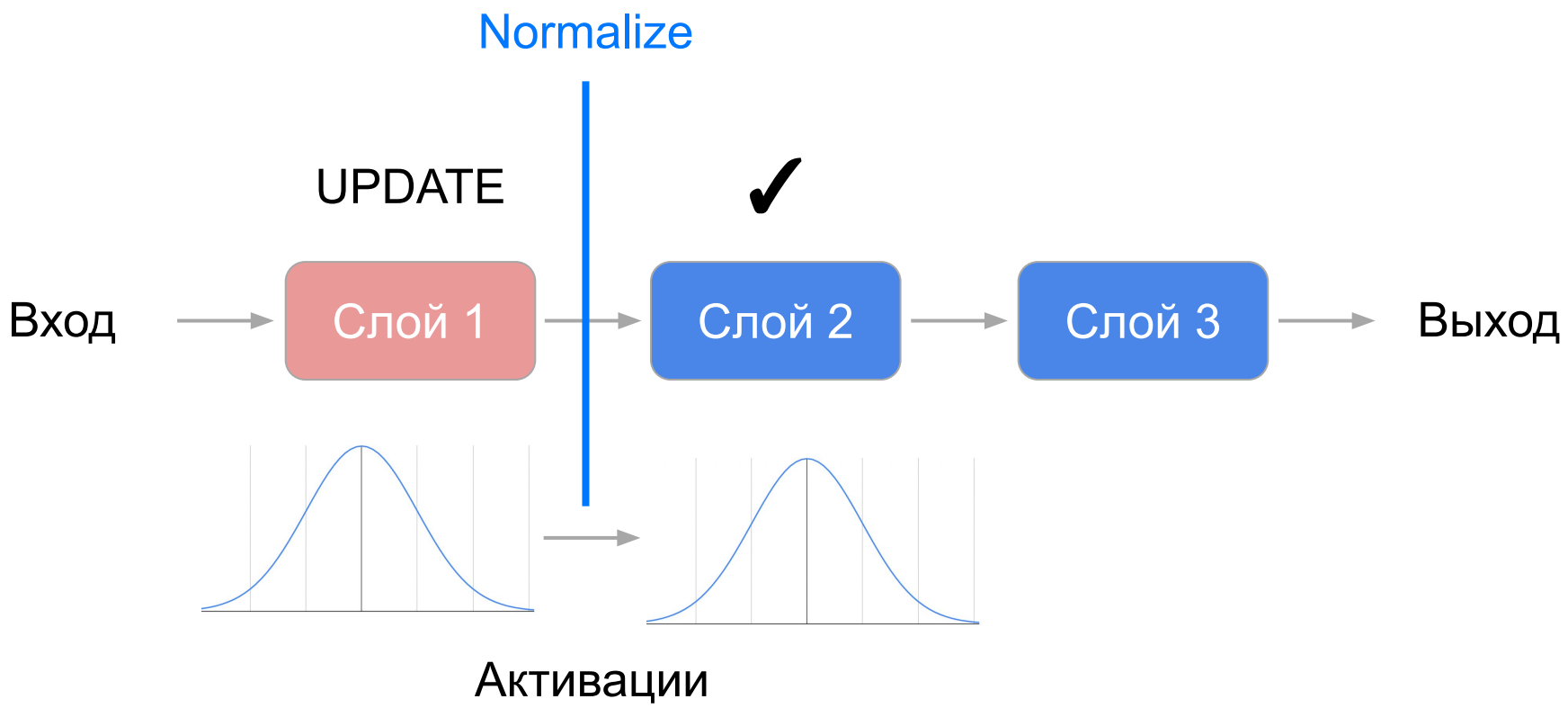
Проблема



Проблема


UPDATE
?
Вход
Слой 1
Слой 2
Слой 3
Выход
Активации
Internal covariate shift

Нормировка



Batch normalization

Во время тренировки:

- Вычесть среднее по батчу
- Разделить на STD батча
- Умножить на обучаемый scale
- Добавить обучаемый bias

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1...m}\}$;
Parameters to be learned: γ, β

Output: $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

$$\sigma^2_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma^2_{\mathcal{B}} + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma\widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

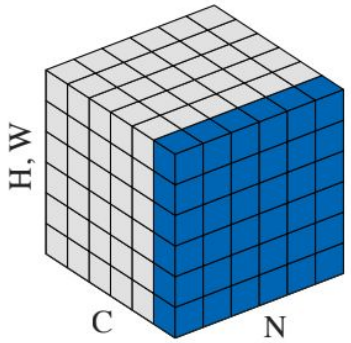
Algorithm 1: Batch Normalizing Transform, applied to activation x over a mini-batch.

Batch normalization

Во время тренировки:

- Вычесть среднее по батчу
- Разделить на STD батча
- Умножить на обучаемый scale
- Добавить обучаемый bias

В свёрточных сетях:



Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1...m}\}$;
Parameters to be learned: γ, β
Output: $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

$$\sigma^2_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma^2_{\mathcal{B}} + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma\widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

Algorithm 1: Batch Normalizing Transform, applied to activation x over a mini-batch.

Batch normalization

Во время тренировки:

- Вычесть среднее по батчу
- Разделить на STD батча
- Умножить на обучаемый scale
- Добавить обучаемый bias

Следствия:

- Поведение зависит от данных батча
- Поведение зависит от размера батча

Что делать в inference если нет батча?

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1...m}\}$;
Parameters to be learned: γ, β
Output: $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

$$\sigma^2_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma^2_{\mathcal{B}} + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma\widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

Algorithm 1: Batch Normalizing Transform, applied to activation x over a mini-batch.

Batch normalization

Во время тренировки:

- Вычесть среднее по батчу
- Разделить на STD батча
- Умножить на обучаемый scale
- Добавить обучаемый bias

При тестировании используются средние и STD усредненные по traininig set

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1...m}\}$;
Parameters to be learned: γ, β
Output: $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

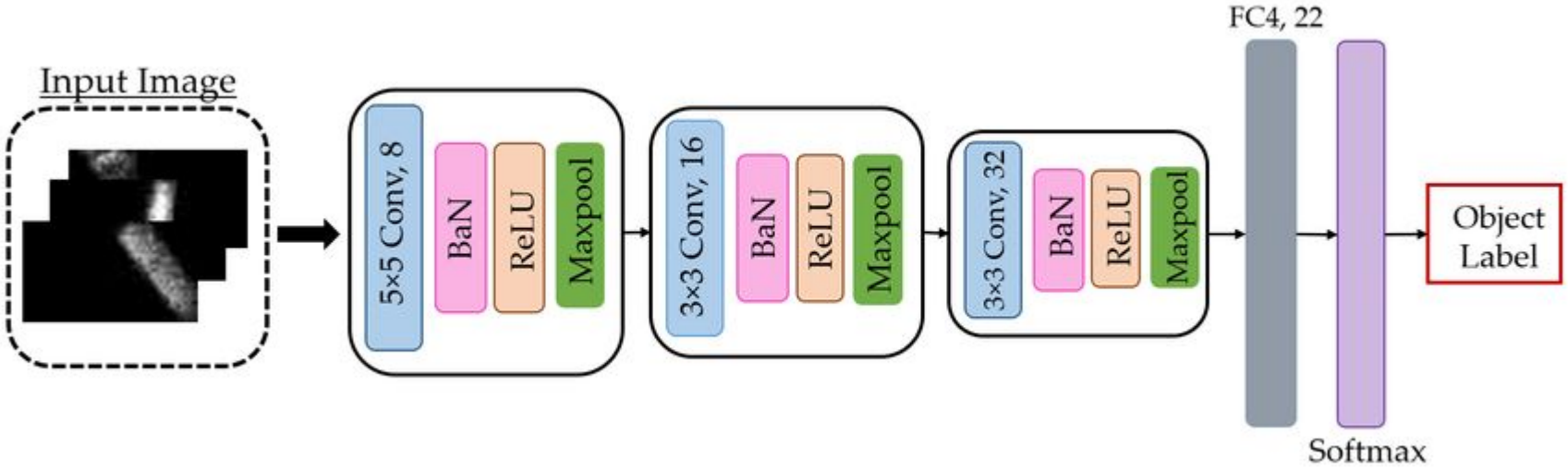
$$\sigma^2_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma^2_{\mathcal{B}} + \epsilon}} \qquad \text{// normalize}$$

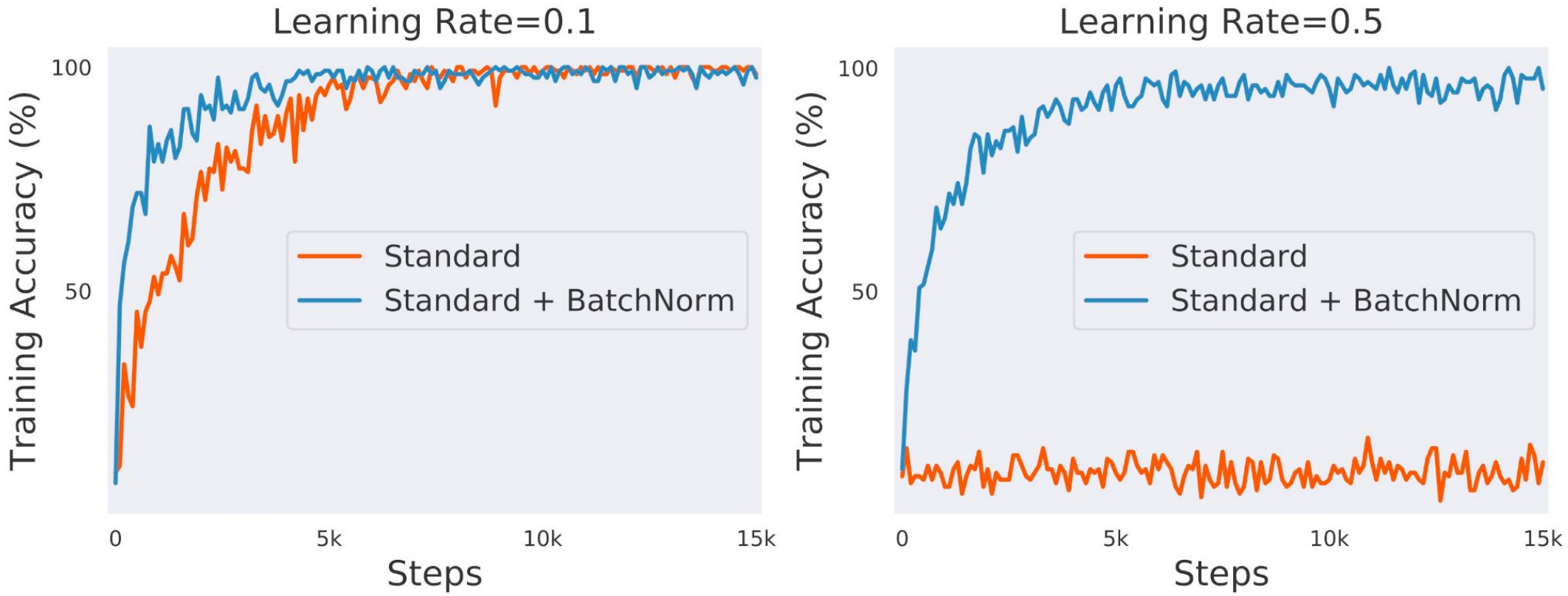
$$y_i \leftarrow \gamma\widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

Algorithm 1: Batch Normalizing Transform, applied to activation x over a mini-batch.

Пример


Input Image
5×5 Conv, 8
BaN
ReLU
Maxpool
3×3 Conv, 16
BaN
ReLU
Maxpool
3×3 Conv, 32
BaN
ReLU
Maxpool
FC4, 22
Softmax
Object Label

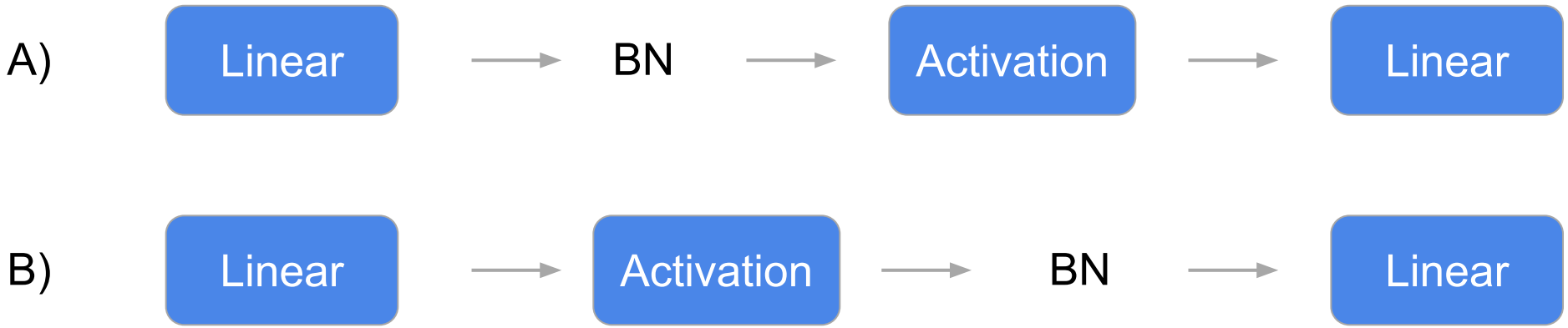
Пример


Learning Rate=0.1
Training Accuracy (%)
100
50
0
5k
10k
15k
Steps
Standard
Standard + BatchNorm
Learning Rate=0.5
Training Accuracy (%)
100
50
0
5k
10k
15k
Steps
Standard
Standard + BatchNorm

Зачем так сложно

- Если использовать скользящие средние в train, обучение может взорваться
- Появляется разница между train и test
- Иногда это благо, т.к. вносит регуляризацию

До или после активации



Batchnorm summary

- BN ускоряет обучение CNN и FC сетей
- BN позволяет использовать больший Learning Rate
- BN вносит разницу между train и test
- При больших batch size разница меньше, а статистики устойчивие

Вопросы

Спасибо
за внимание)