

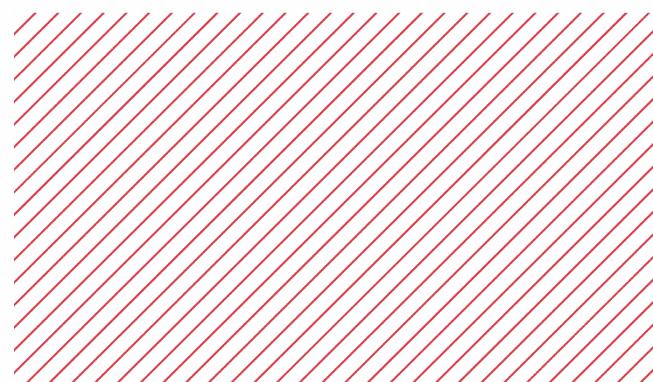
академия
больших
данных

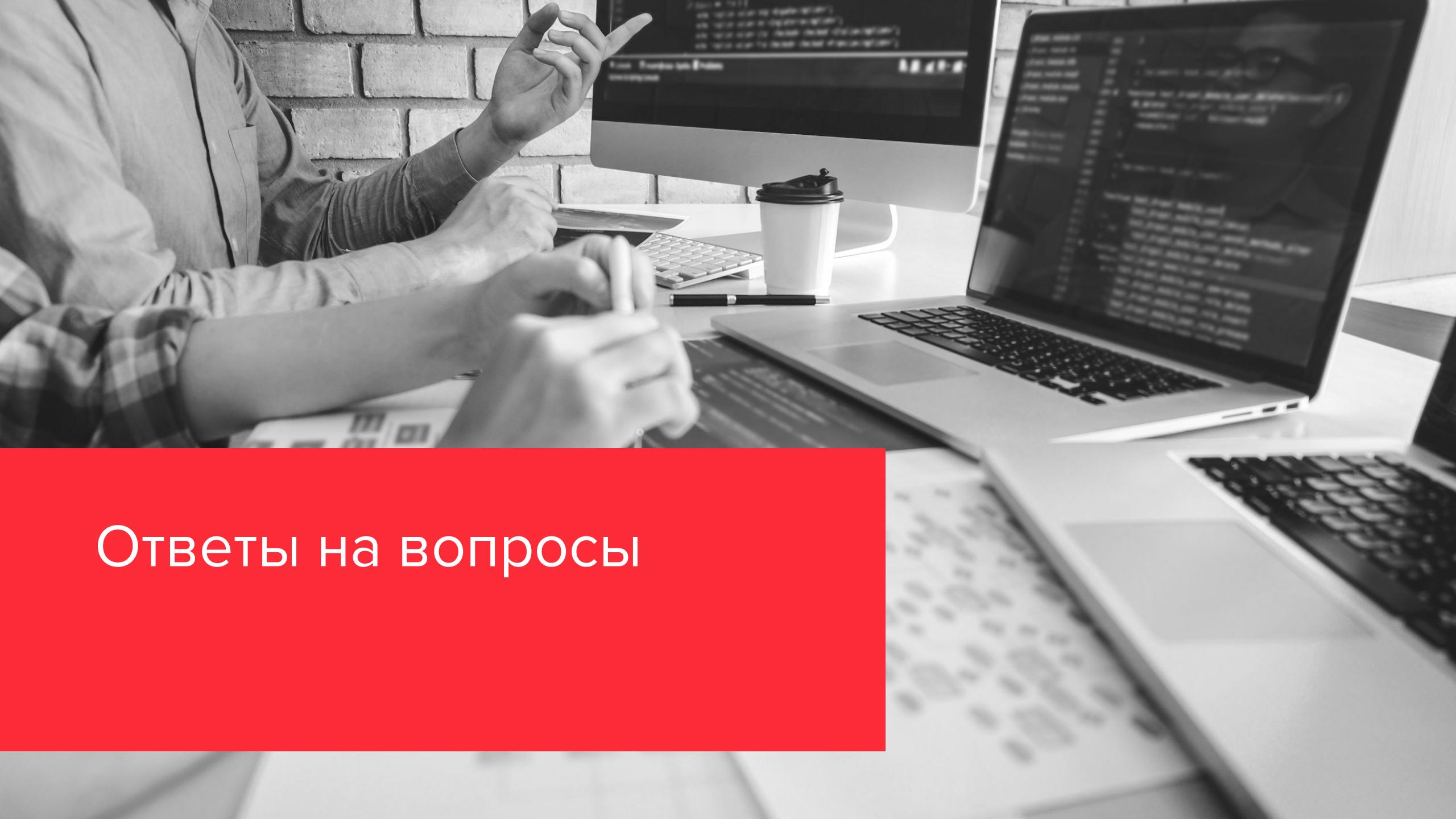


Пайpline ML Обучение FC сетей

Иван Карпухин

Ведущий программист-исследователь в команде
машинного зрения



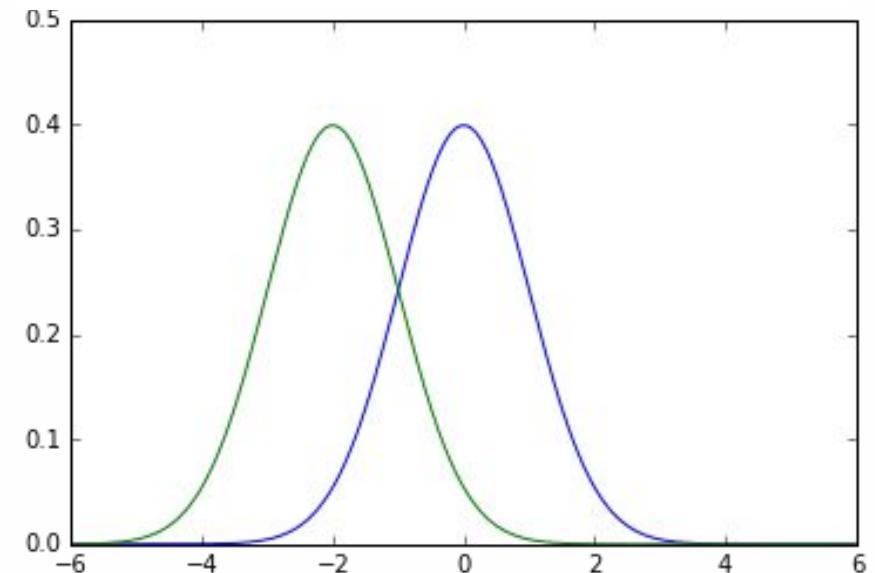
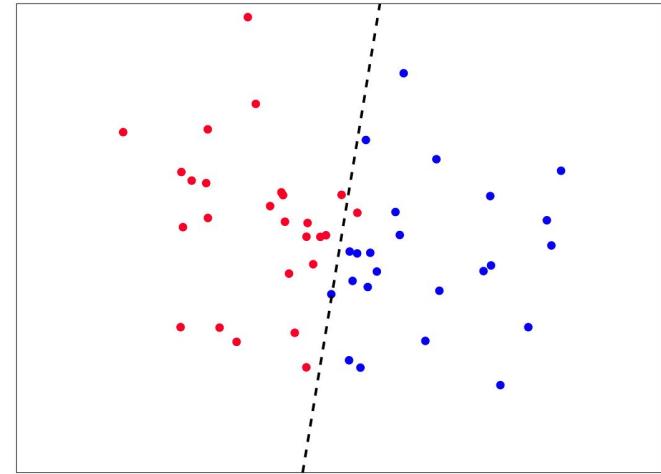


Ответы на вопросы

Перекрестная энтропия

Почему на выходе сети вероятности?

$$H(p, q) = - \sum_x p(x) \log q(x)$$

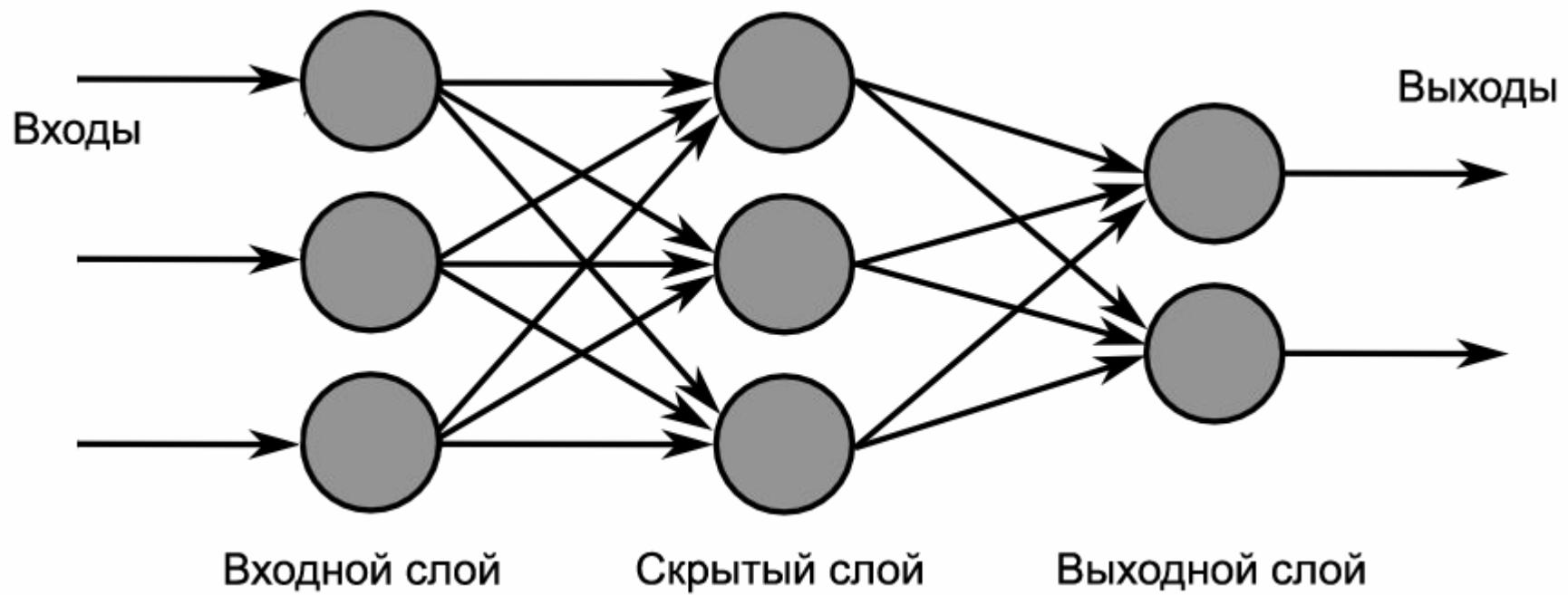




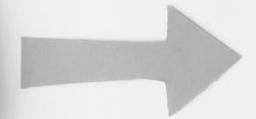
План курса

- ❖ Base (3 лекции, до 5 марта)
 - CV
 - backprop, FC, CNN
- ❖ Classics (5 лекций, до 9 апреля)
 - аугментация, dropout, batchnorm
 - классификация, сегментация, детекция, регрессия
- ❖ Challenging (4 лекции, до 7 мая)
 - рекуррентные сети, metric learning, GAN
- ❖ Advanced (3 лекции, до 28 мая)
 - Semi-supervised, few-shot learning, domain adaptation, video processing
- ❖ Прод (2 лекции, до 11 июня)

Универсальный аппроксиматор



Пайпайн машинного обучения



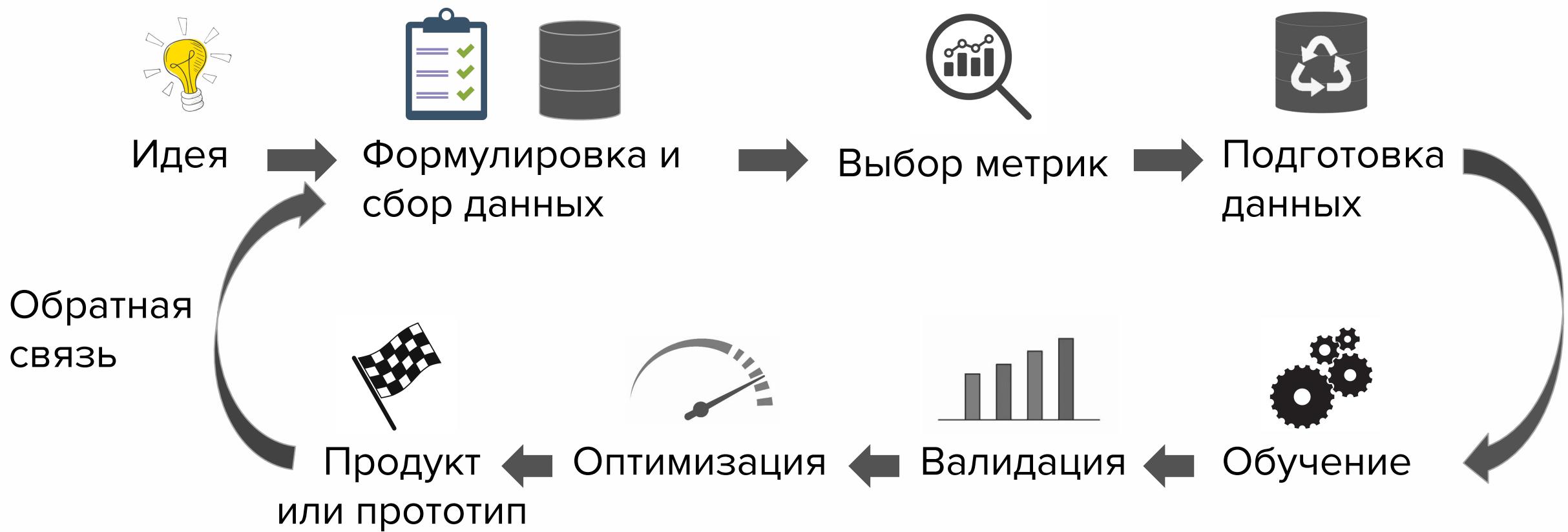
ML сложнее, чем кажется

В теории:



ML сложнее, чем кажется

На практике:



Пайплайн



Формулировка задачи

ML решает трудно формализуемые задачи.

Задача ставится на данных:

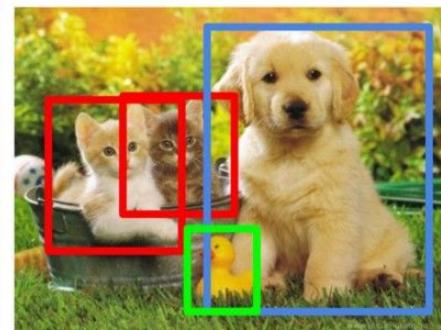
- ❖ входные данные
- ❖ разметка



CAT



CAT



CAT, DOG, DUCK



CAT, DOG, DUCK



Сбор данных

Три корпуса данных:

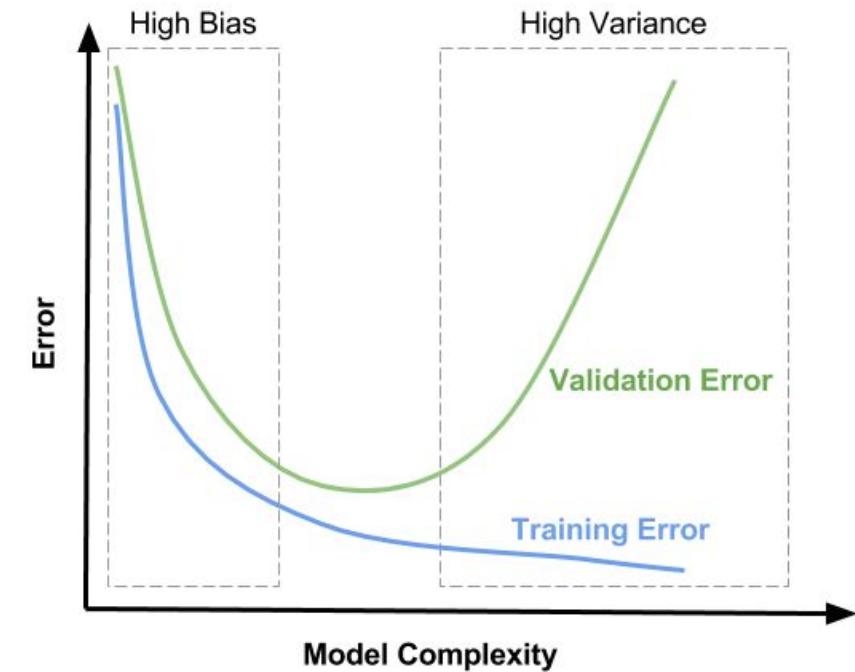
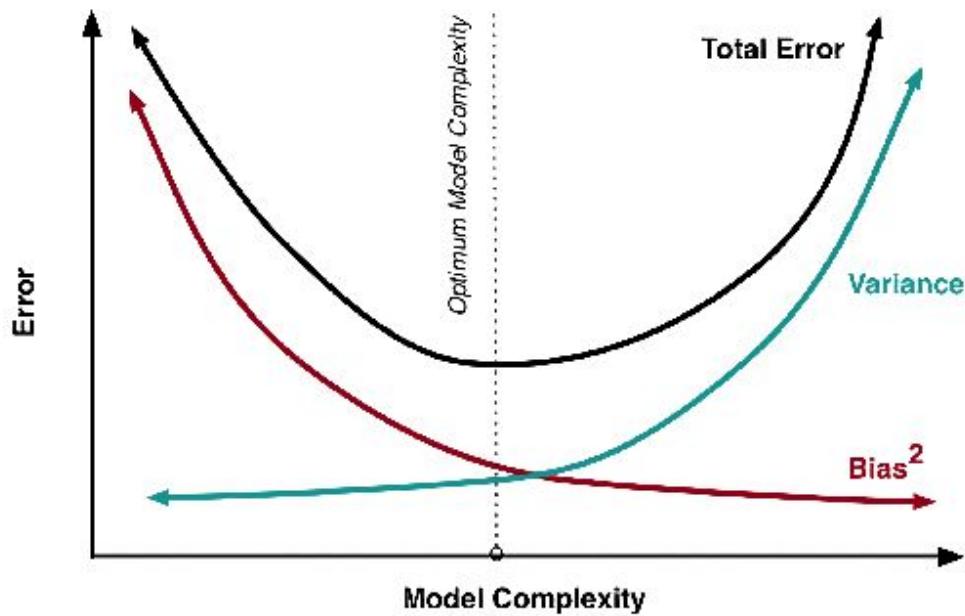
- ❖ тренировочный (train) для построения модели
- ❖ валидационный (dev) для выбора модели
- ❖ тестовый (test) для оценки качества

Источники разметки:

- ❖ имеется по построению (аватарки одного человека)
- ❖ аксессоры
- ❖ краудсорсинг
- ❖ другая система (knowledge distillation)

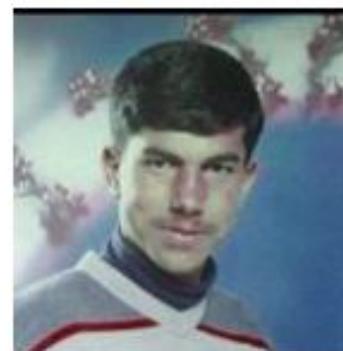
Дilemma смещения–дисперсии

Bias / variance



Сбор данных: распознавание лиц

Тренировочные данные из публичного корпуса



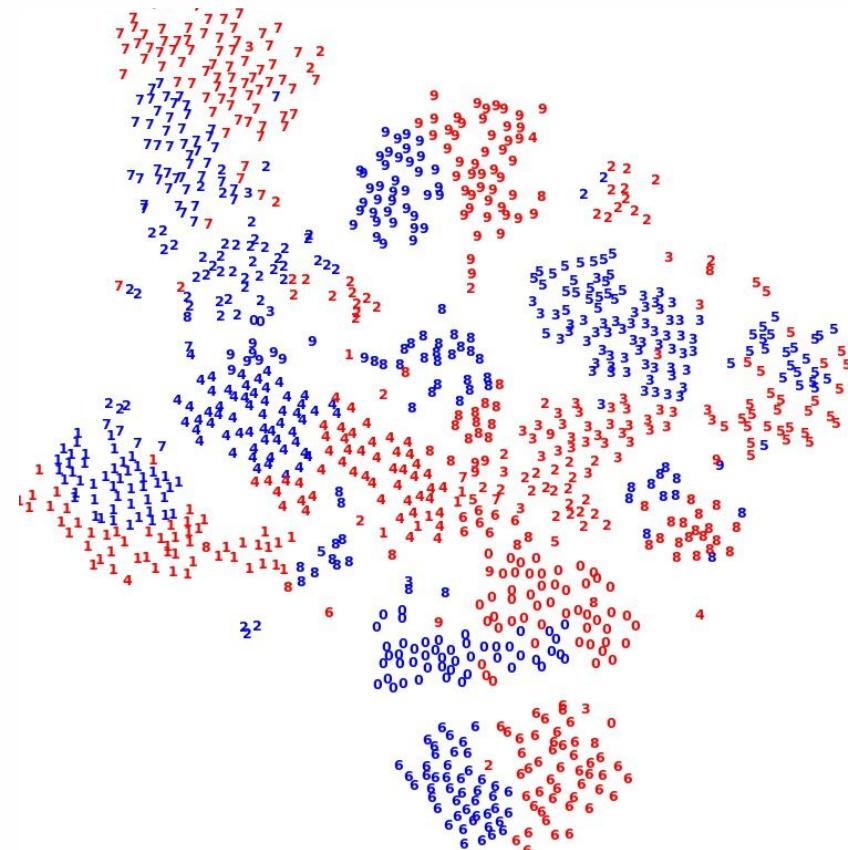
Сбор данных: распознавание лиц

Тестовые данные от заказчика



Сбор данных: распознавание лиц

Разница доменов (domain mismatch)



Сбор данных

- ❖ Корпусы максимально независимы
- ❖ Валидационный и тестовый одинаково распределены

Корпус	Размер	Распределение	Назначение
Тренировочный	10.000-1.000.000	Смещеное	Обучение
Валидационный	1000 - 10.000	Несмещеное	Отбор модели
Тестовый	1000 - 10.000	Несмещеное	Тестирование

Пайплайн



Метрики

Технические

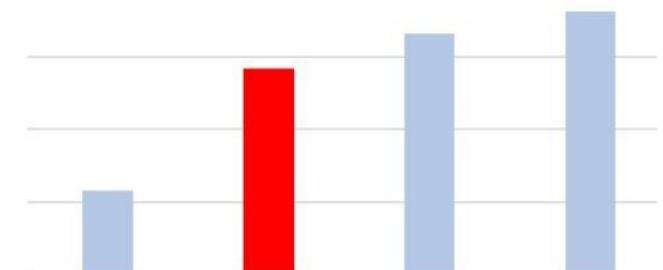
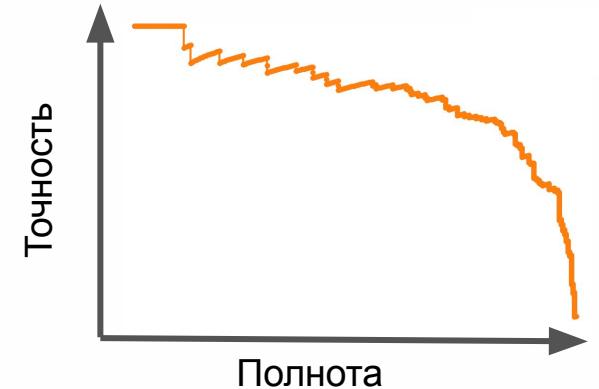
- ❖ Оценивают части системы
- ❖ Влияют на выбор функции потерь

Продуктовые

- ❖ Оценивают мнение пользователей
- ❖ Оценивают систему целиком

Метрики из статей и стандартов

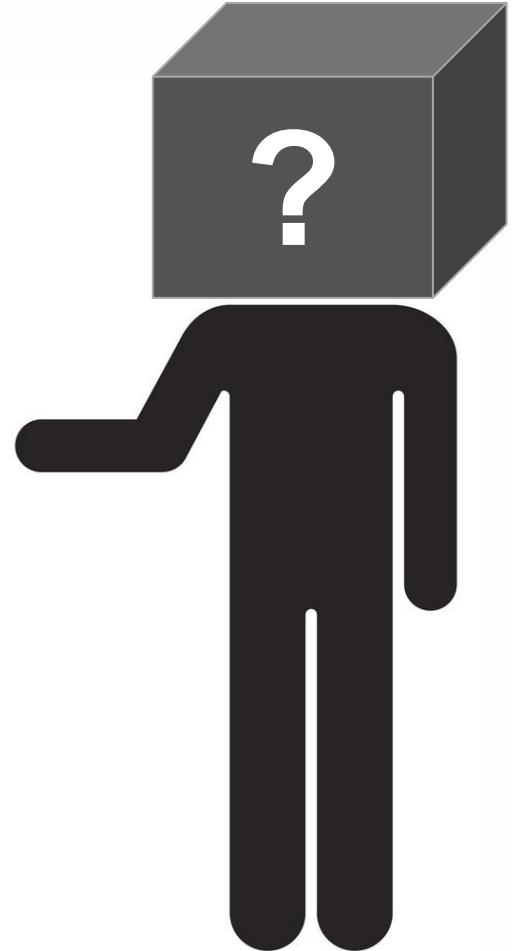
- ❖ Позволяют сравниваться с аналогами



Метрики: трудности

Заказчик не имеет четких ожиданий

- ❖ первые версии продукта
- ❖ отсутствие аналогов



Продуктовые метрики - модель мнения заказчика

- ❖ если метрик нет, их нужно придумать
- ❖ метрики уточняются на основе мнения заказчика

Метрики: трудности

Нежелание принимать риски

- ❖ ошибки есть всегда

Заказчик оценивает “глазами”

- ❖ может ускорить разработку на старте



Стремимся провести бета-тестирование как можно раньше

Пайплайн



Подготовка данных

Поиск и устранение ошибок



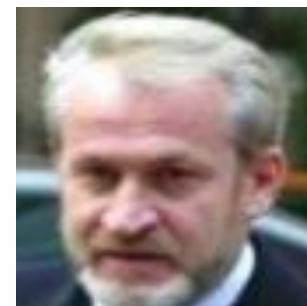
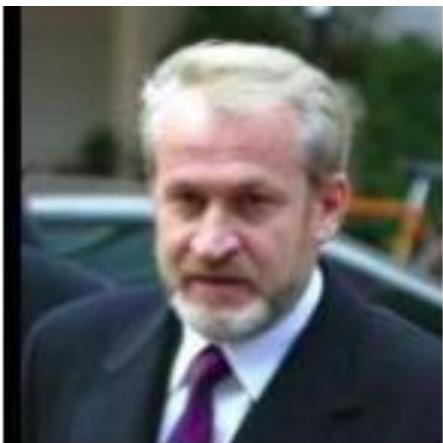
Подготовка данных

Аугментация

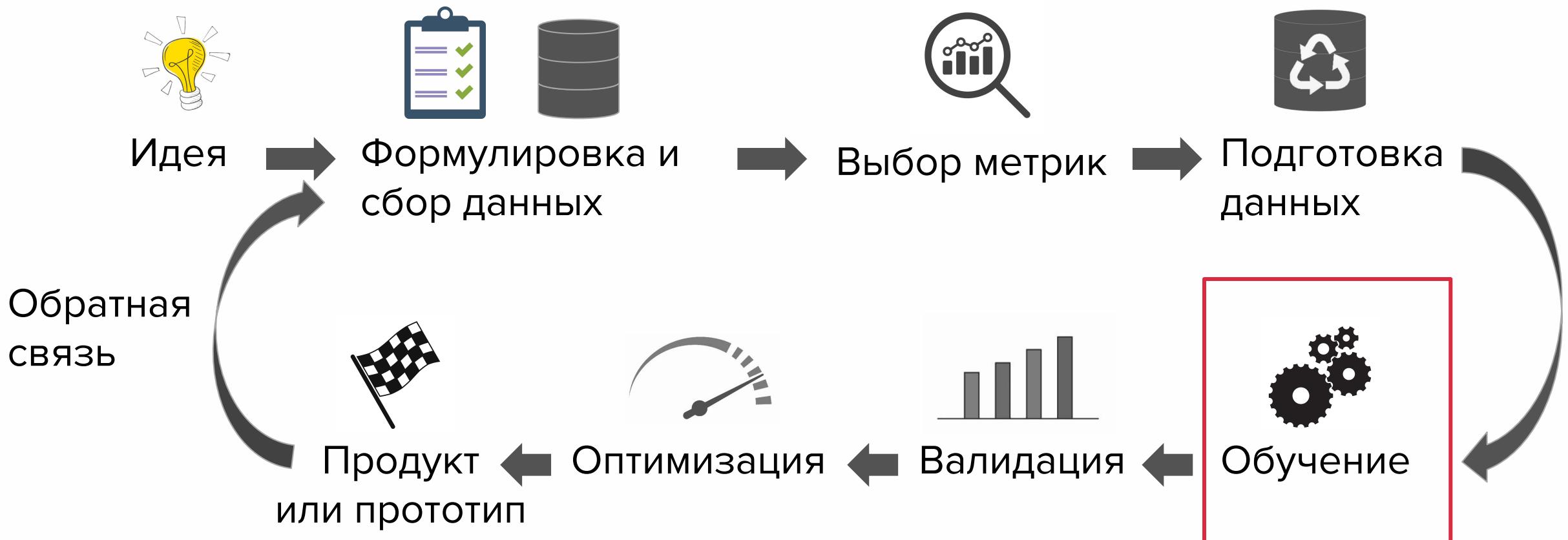


Подготовка данных

Приведение к стандартному виду



Пайплайн



Обучение

- ❖ Выбор модели
- ❖ Выбор функции потерь
- ❖ Оптимизация
- ❖ Подбор параметров модели и обучения



Пайплайн

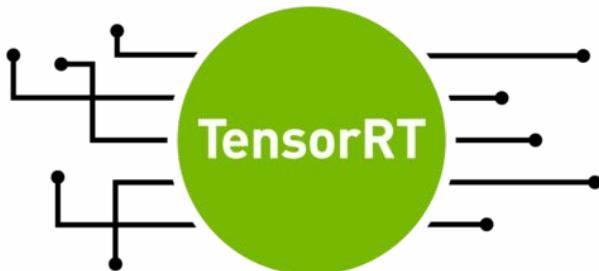


Пайплайн



Оптимизация быстродействия

- ❖ Упрощение модели (дистилляция)
- ❖ Оптимизация под железо
- ❖ Квантование (использование пониженной точности)
 - FP 16
 - INT 8



Пайплайн



Современные методы оптимизации

Градиентный спуск (GD)

$$Error(W) = \frac{1}{N} \sum_{i=1}^N Error(X_i, W)$$

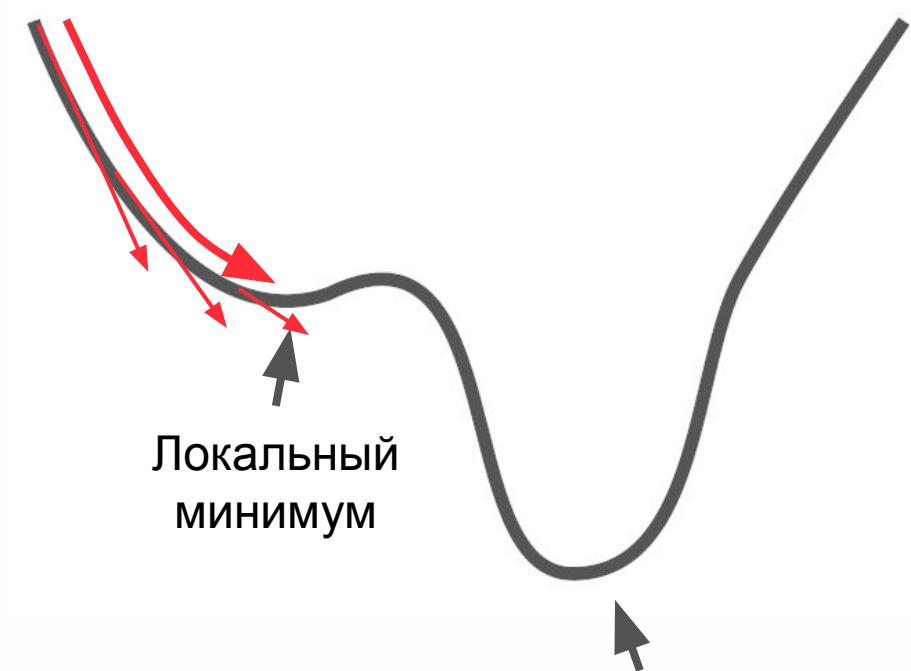
$$g_{n+1} = \nabla Error(W_n)$$

$$W_{n+1} = W_n - \lambda g_{n+1}$$

λ - скорость обучения (learning rate)

Проблемы:

- ❖ Один шаг - просмотр всей обучающей выборки
- ❖ Медленная сходимость



Глобальный
минимум

Стохастический градиентный спуск (SGD)

$$Error_n(W) = \frac{1}{B} \sum_{i=Bn}^{Bn+B} Error(X_i, W)$$

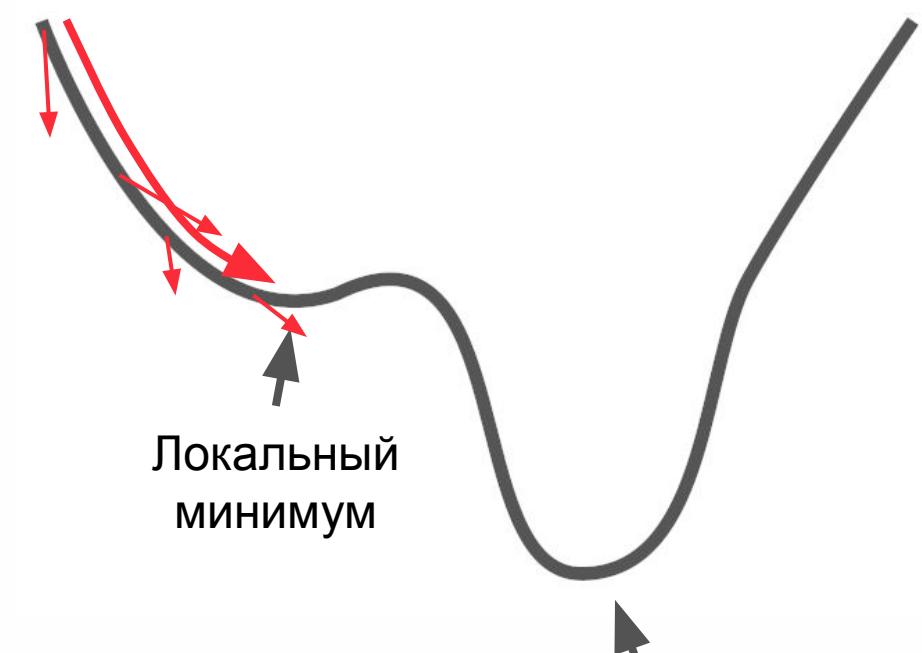
$$g_{n+1} = \nabla Error_n(W_n)$$

$$W_{n+1} = W_n - \lambda g_{n+1}$$

B - размер батча (batch)

Проблемы:

- ❖ Сходимость к локальному оптимуму



Локальный
минимум

Глобальный
минимум

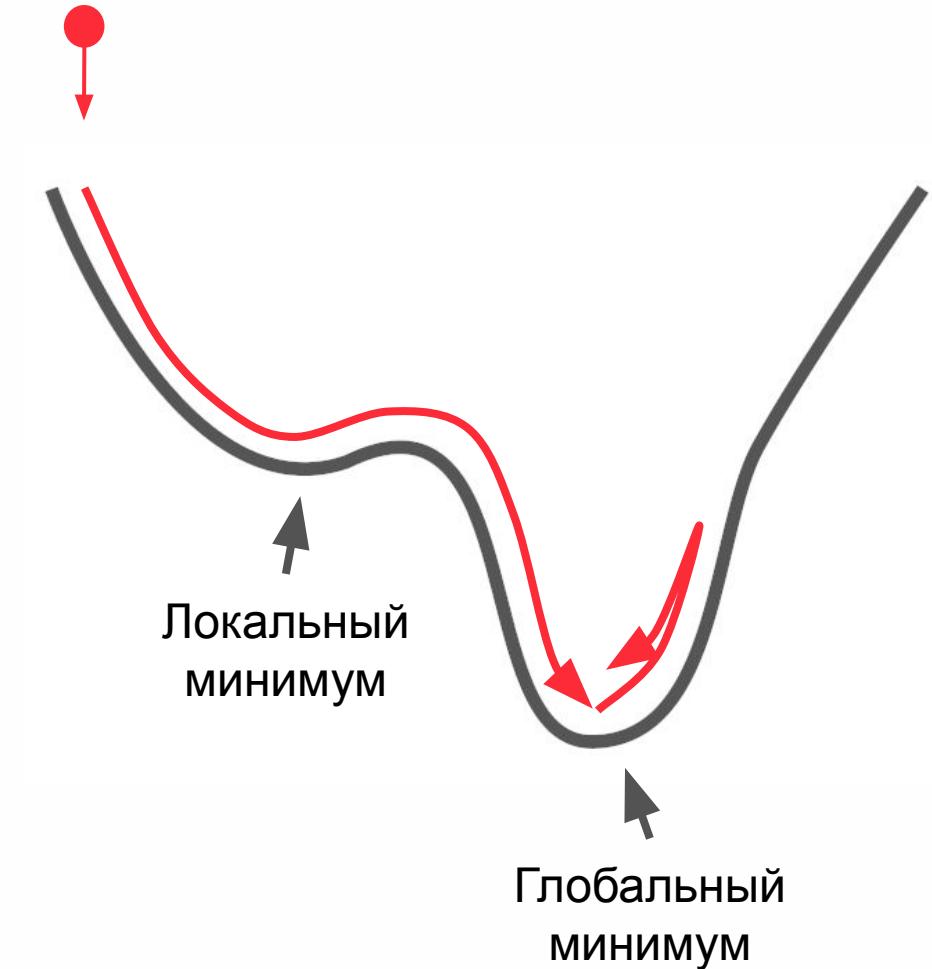
SGD с моментом

$$V_{n+1} = \mu V_n - \lambda g_{n+1}$$

$$W_{n+1} = W_n + V_{n+1}$$

Проблемы:

- ❖ Смещение в направлении скорости и градиент не учитывают друг друга



* Поляк Б. Т. О некоторых способах ускорения сходимости итерационных методов. 1964

Момент Нестерова

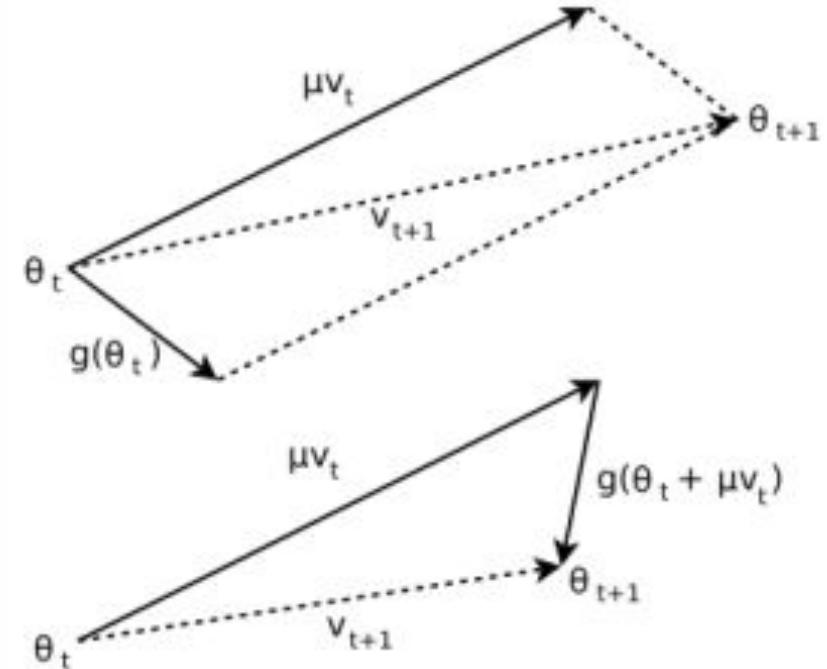
$$\tilde{W}_{n+1} = W_n + \mu V_n$$

$$g_{n+1} = -\lambda \nabla Error_{n+1}(\tilde{W}_{n+1})$$

$$V_{n+1} = \mu V_n + g_{n+1}$$

$$W_{n+1} = \tilde{W}_{n+1} + g_{n+1}$$

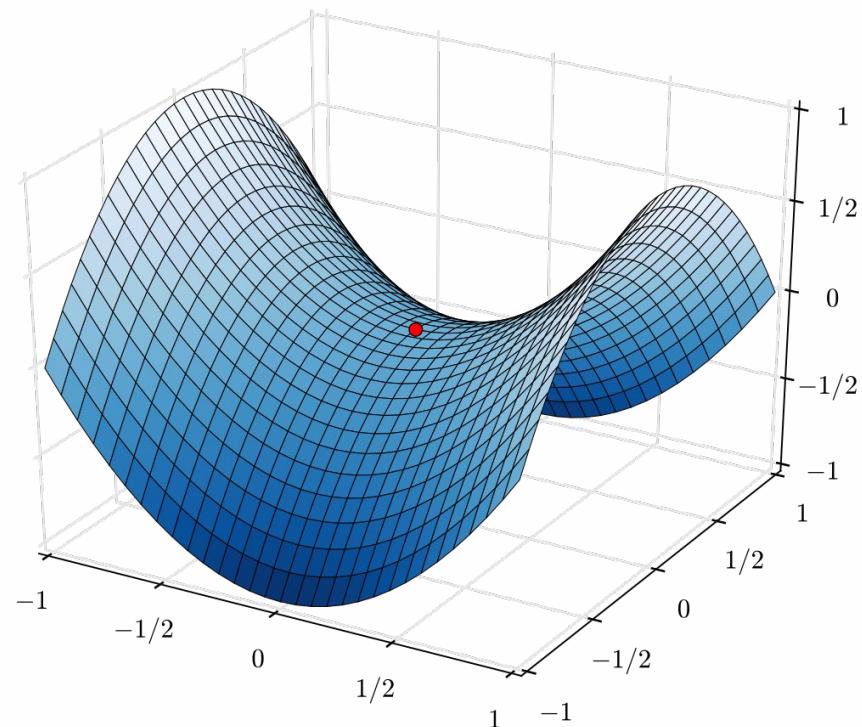
- ❖ Градиент корректирует движение
- ❖ Градиент становится стабильнее



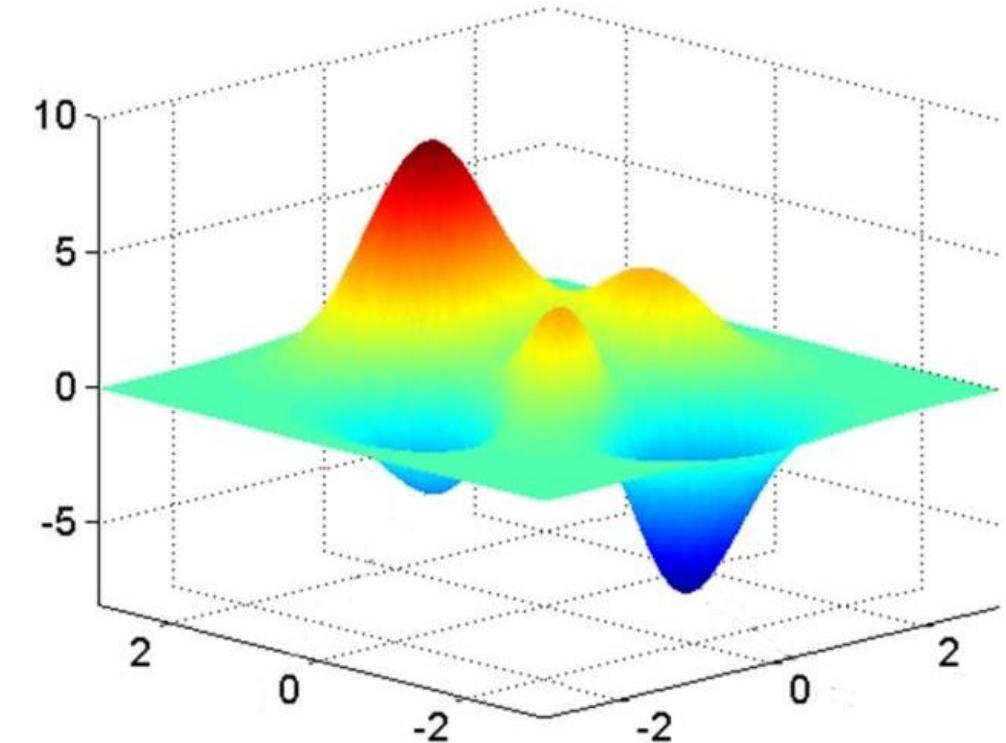
* Метод решения задачи выпуклого программирования со скоростью сходимости $O(1/k^2)$. 1983

* Sutskever I. et al. On the importance of initialization and momentum in deep learning. 2013

RMSProp



Седловая точка



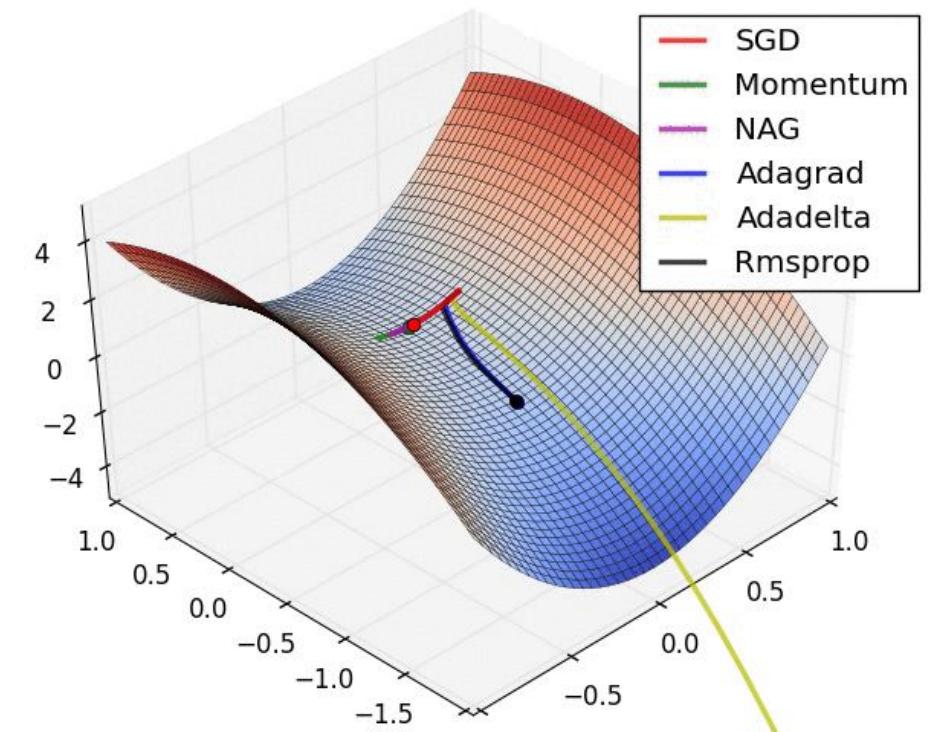
Плато

RMSProp

Идея: уменьшать скорость обучения для часто изменяющихся параметров

$$E[g^2]_{n+1} = \gamma E[g^2]_n + (1 - \gamma) g_{n+1}^2$$

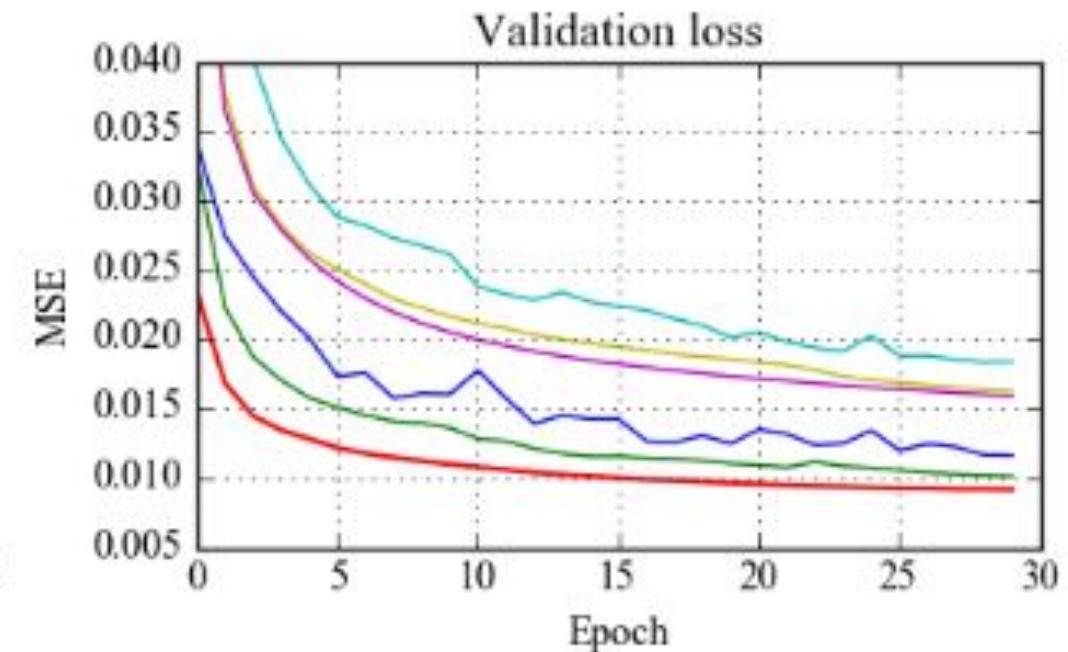
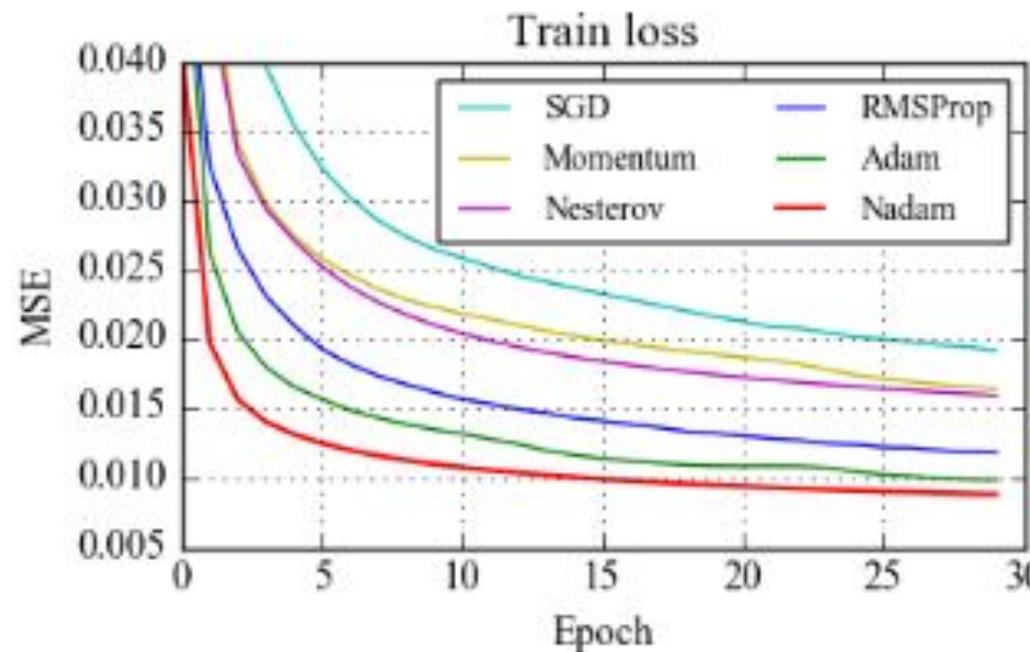
$$W_{n+1} = W_n - \frac{\lambda}{\sqrt{E[g^2]_{n+1} + \epsilon}} g_{n+1}$$



Adam

Adam = RMSProp + Momentum

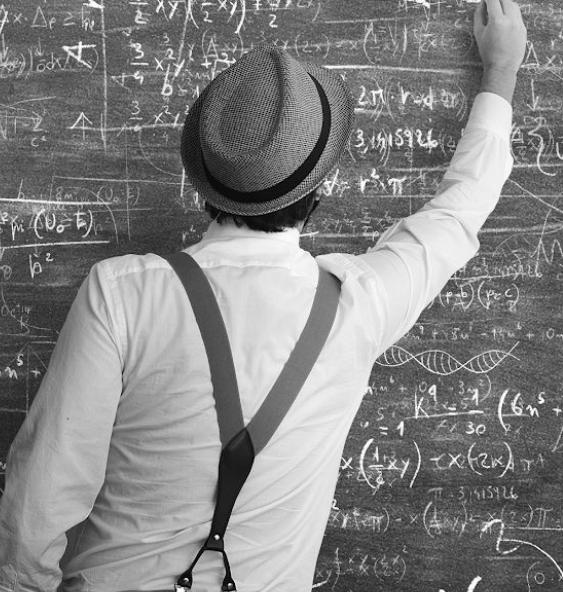
Nadam = RMSProp + Nesterov momentum



* Diederik P. Kingma and Jimmy Lei Ba. Adam : A method for stochastic optimization. 2014

* Dozat T. Incorporating nesterov momentum into adam. 2016.

На следующей лекции





На следующей лекции

- ❖ Сверточные сети
 - свертки
 - пулинг
 - функции активации