• Введение в машинное обучение •

Эволюция идей машинного обучения

Воронцов Константин Вячеславович
k.v.vorontsov@phystech.edu
http://www.MachineLearning.ru/wiki?title=User:Vokov

Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Введение в машинное обучение (курс лекций, К.В.Воронцов)»

МФТИ.ФПМИ.ИС.ИАД • 5 марта 2026

Содержание

1. **Вектор → вектор → скаляр**
   - Задачи с векторными описаниями объектов
   - Методы преобразования признаков
   - Конструирование признаков
2. **Структура → вектор → скаляр**
   - Свёрточные нейронные сети
   - Векторизация сложно структурированных данных
   - Перенос обучения, самостоятельное обучение
3. **Структура → вектор → структура**
   - Автокодировщики
   - Фундаментальные модели
   - Генеративные модели

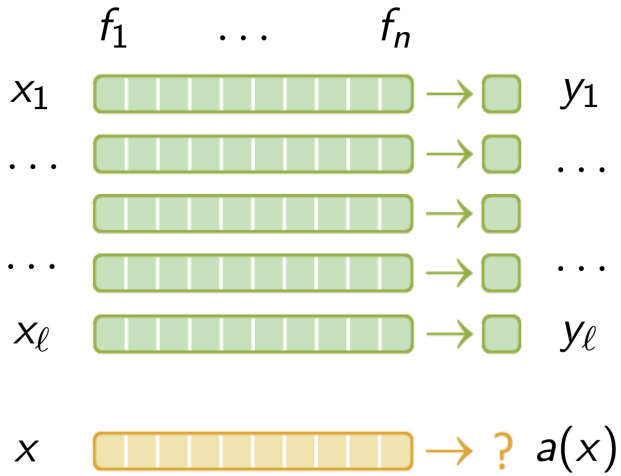
Восстановление зависимостей по эмпирическим данным

Дано: обучающая выборка *объектов*
$x_i = (f_1(x_i), \ldots, f_n(x_i)), \quad i = 1, \ldots, \ell$
$f_j : X \to D_j$ — признаки, $j = 1, \ldots, n$

Найти: параметры w *модели* $a(x, w)$

Критерий: min эмпирического риска

$$\sum_{i=1}^{\ell} \mathscr{L}(w, x_i) + \tau \mathscr{R}(w) \to \min_w,$$



где $\mathscr{L}(w, x)$ — *функция потерь*, $\mathscr{R}(w)$ — *регуляризатор*.

Преобразование признаков (feature transformation/extraction)

$$\sum_{i=1}^{\ell} \mathscr{L}\big(w, f(x_i, w')\big) + \tau \mathscr{R}(w) \to \min_{w, w'}$$

где модель $f(x, w')$ либо задаётся вручную (feature engineering), либо обучается совместно с моделью w (feature generation)

Шкалы измерения

Измерительная шкала — множество D допустимых значений, получаемых в результате измерения признака $f(x)$, $f: X \to D$

Тип шкалы определяется множествами

- допустимых биективных преобразований $\psi: D \to D'$
- допустимых операций над значениями из шкалы D

Классификация типов измерительных шкал по Стивенсу:

шкала	D	$\psi(z)$	операции
логическая (boolean)	$0, 1$	биективные	$\vee \wedge \neg$
номинальная (nominal)	$< \infty$	биективные	$= \neq \in$
порядковая (ordinal)	$< \infty$	монотонные	$= \neq \in < >$
интервальная (interval)	$\mathbb{R}$	$az + b$	$< > + -$
отношений (ratio)	$\mathbb{R}$	az	$< > + - \times \div$
абсолютная (absolute)	$\mathbb{R}$	z	любые

S.S.Stevens. On the Theory of Scales of Measurement // Science, 1946.

Примеры величин, измеряемых в различных шкалах

- **Логическая** (можно переименовать, перенумеровать)
  наличие/отсутствие свойства, истина/ложь, да/нет
- **Номинальная** (можно переименовать, перенумеровать)
  идентификаторы классов, людей, регионов, фирм, товаров
- **Порядковая** (порядок частичный или линейный)
  уровень образования, тяжесть болезни, степень согласия
- **Ранговая** (частный случай порядковой: $1, 2, 3, \ldots, N$)
  оценка в баллах, шкалы Рихтера, Бофорта, Мооса, Бека
- **Интервальная** (можно менять масштаб и сдвигать 0)
  время, географическая широта, температура ($^\circ$C, $^\circ$F)
- **Отношений** (можно менять масштаб, сдвигать 0 нельзя)
  масса, скорость, объём, сила, давление, заряд, яркость, $^\circ$K
- **Абсолютная** (любые преобразования запрещены)
  число предметов, частота события, оценка вероятности

Ослабление шкалы

Номинальный → много бинарных (one-hot-encoding):

- $f_v(x) = \left[f(x) = v\right]$, для всех значений v признака
- $f_A(x) = \left[f(x) \in A\right]$, индикаторный признак подмножества A

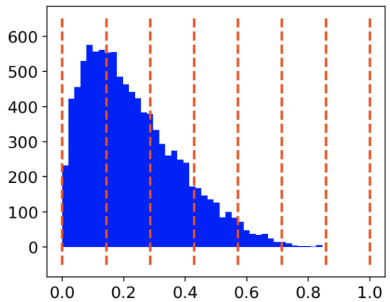
Числовой или порядковый → бинарный:

- $f_{a,b}(x) = \left[a \leqslant f(x) \leqslant b\right]$ для заданного отрезка $[a, b]$

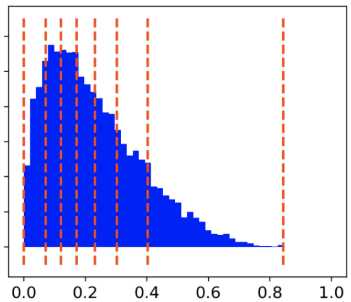
Числовой → ранговый (data binning, quantization):

- $f_a(x) = \sum_{k=1}^{K} \left[f(x) \geqslant a_k\right]$, номер интервала сетки $a_1, \ldots, a_K$

равномерная сетка



квантильная сетка



Ослабление шкалы всегда влечёт потерю информации

Усиление шкалы

Номинальный → числовой:

- категория заменяется частотой:
$$f'(x) = \frac{1}{\ell}\sum_{i=1}^{\ell}\big[f(x_i) = f(x)\big]$$
- условное среднее числового признака $g(x)$:
$$f'(x) = \operatorname{mean}\big(g|f(x)\big) = \frac{\sum_{i=1}^{\ell} g(x_i)\big[f(x_i) = f(x)\big]}{\sum_{i=1}^{\ell}\big[f(x_i) = f(x)\big]},$$
- условное среднее целевой величины $y(x)$:
$f'(x) = \operatorname{mean}\big(y|f(x)\big)$, возможно переобучение!

Порядковый → числовой (монотонное преобразование)

- значение заменяется частотой:
$$f'(x) = \frac{1}{\ell}\sum_{i=1}^{\ell}\big[f(x_i) \leqslant f(x)\big]$$

Нормализация и стандартизация числовых шкал

Многие методы накапливают меньше вычислительных погрешностей, если признаки приведены к одному масштабу

- $f'_j(x) = \dfrac{f_j(x) - f_j^{\min}}{f_j^{\max} - f_j^{\min}}$ — нормализация, приведение к $[0, 1]$
- $f'_j(x) = \dfrac{f_j(x)}{|f_j|^{\max}}$ — масштабирование с сохранением нуля
- $f'_j(x) = \dfrac{f_j(x) - \mu_j}{\sigma_j}$ — стандартизация

$f_j^{\max}$, $|f_j|^{\max}$, $f_j^{\min}$, μ_j, σ_j определяются по обучающей выборке

Для повышения устойчивости к выбросам можно отбрасывать 5% наименьших и наибольших значений признака

Конструирование признаков по «сырым» данным (raw data)

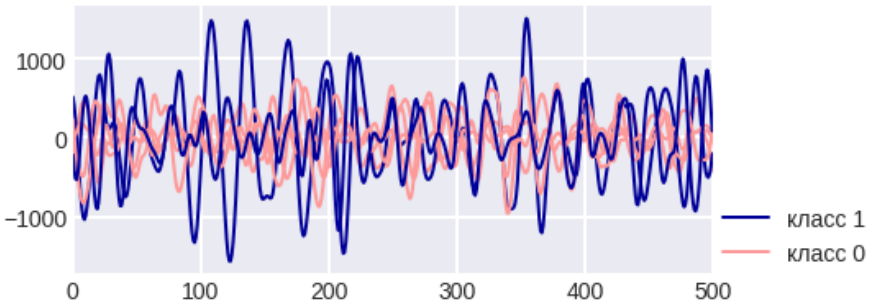
Feature Engineering: признаки вычисляются по формулам, которые зависят от задачи, требуют изобретательности и знаний предметной области. Долго, дорого, неточно.

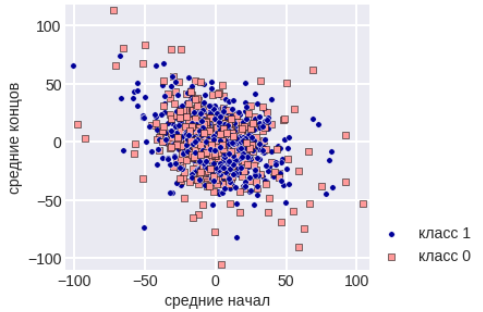
- **Прогнозирование временных рядов:**
  признаки агрегируются по предыстории различной глубины
- **Предсказание оттока клиентов:**
  признаки структуры и объёма услуг, оплаты, тарифов
- **Распознавание лиц:**
  признаки размера и формы черт лица
- **Классификация и поиск текстов:**
  признаки частоты слов, терминов, названий, синонимов
- **Распознавание речи:**
  признаки спектральные, фонетические, лингвистические

Пример. Задача детектирования поломок по сигналу датчика

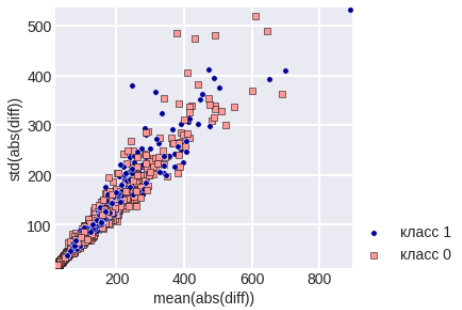
Соревнование «Ford Classification Challenge», 2008
Иногда при удачном выборе признаков задача решается без ML

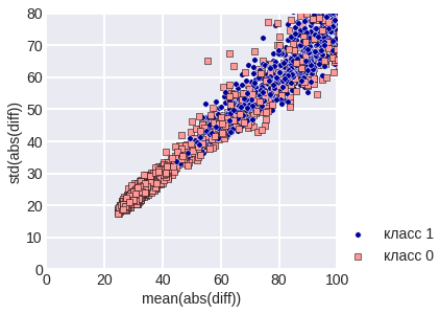
Признаки, генерируемые по исходным временным рядам, слабы:

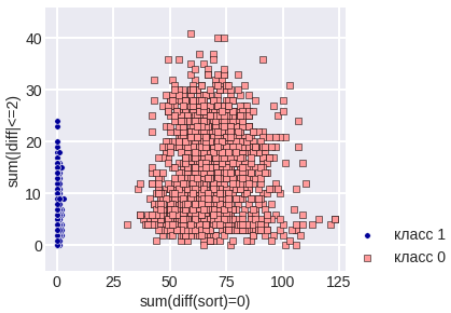




Среди признаков рядов их производных оказывается идеальный:





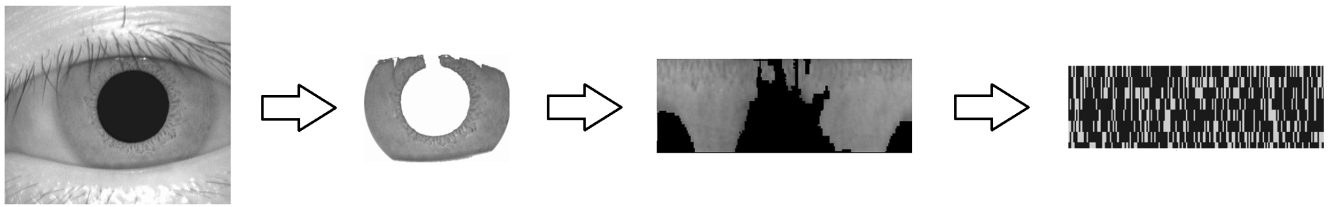


https://dyakonov.org/2018/06/28/простые-методы-анализа-данных

Пример. Задачи биометрической идентификации личности

Идентификация личности по отпечаткам пальцев

Идентификация личности по радужной оболочке глаза



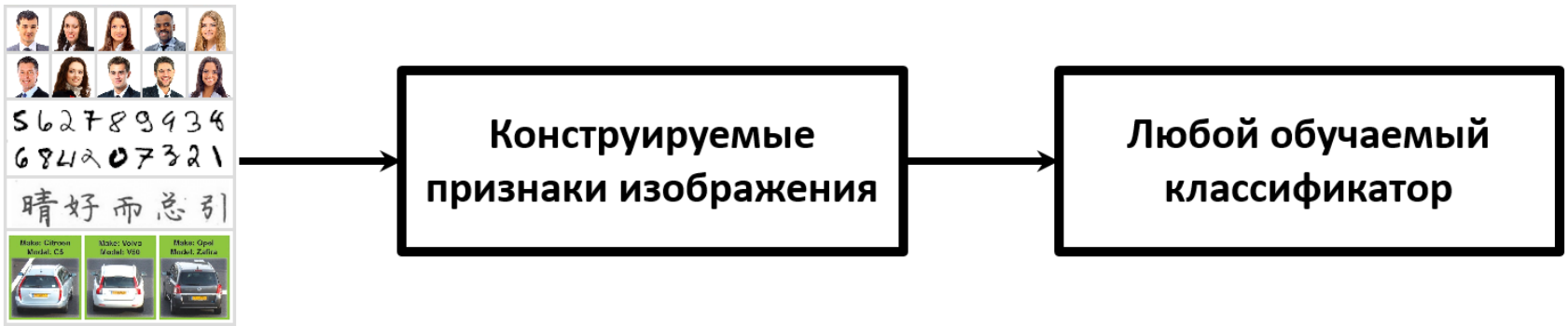
Особенности задач:

- нетривиальная предобработка для извлечения признаков
- высочайшие требования к точности

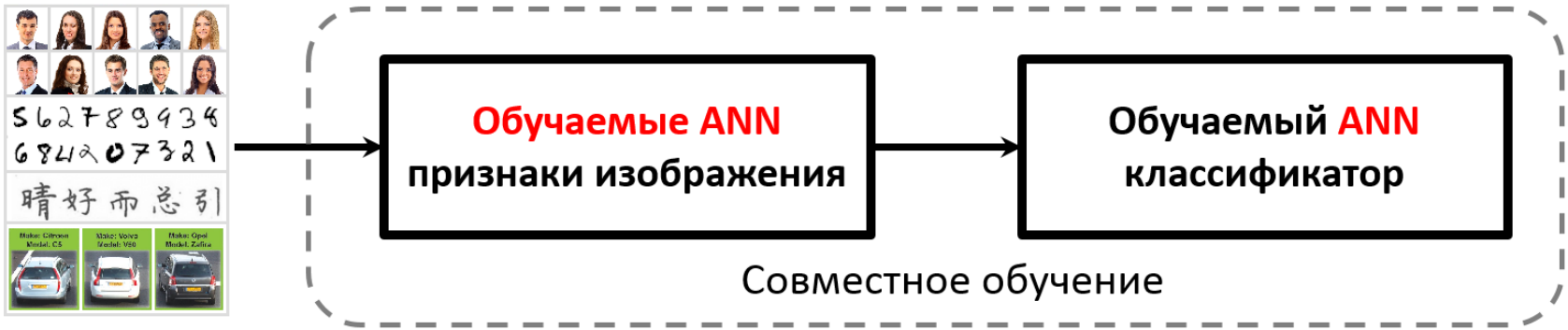
J. Daugman. High confidence visual recognition of persons by a test of statistical independence. 1993

Конструирование признаков для распознавания изображений

Классический подход к распознаванию изображений:



Современный подход — end-to-end deep learning:



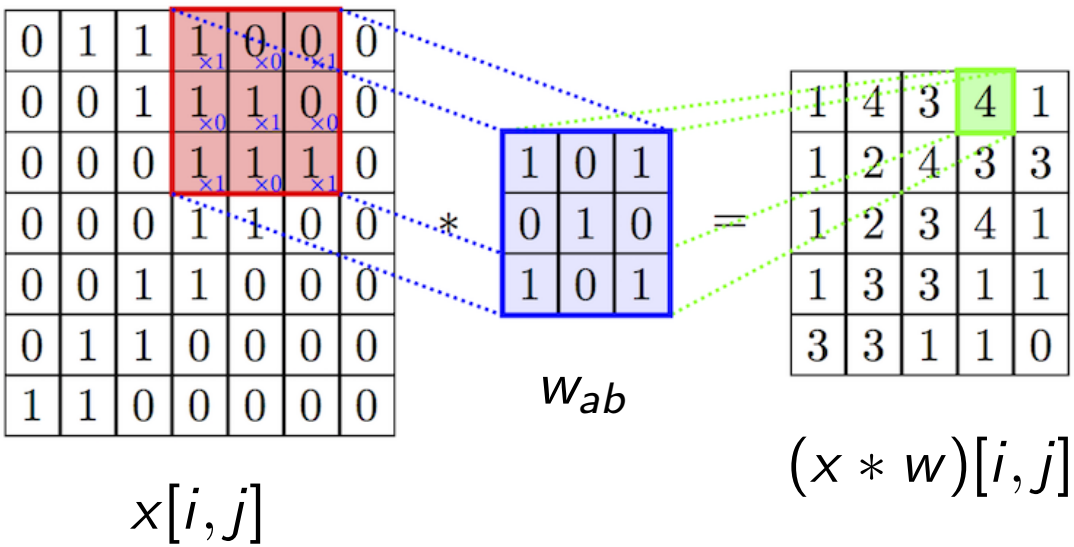
Yann LeCun et al. Learning algorithms for classification: A comparison on handwritten digit recognition. 1995

Свёрточный слой нейронов (convolution layer)

$x[i,j]$ — исходные признаки, пиксели $n \times m$-изображения
w_{ab} — ядро свёртки, $a = -A, \ldots, +A$, $b = -B, \ldots, +B$

Неполносвязный свёрточный нейрон с $(2A+1)(2B+1)$ весами:

$$(x * w)[i,j] = \sum_{a=-A}^{A} \sum_{b=-B}^{B} w_{ab}\, x[i+a, j+b]$$



Объединяющий слой нейронов (pooling layer)

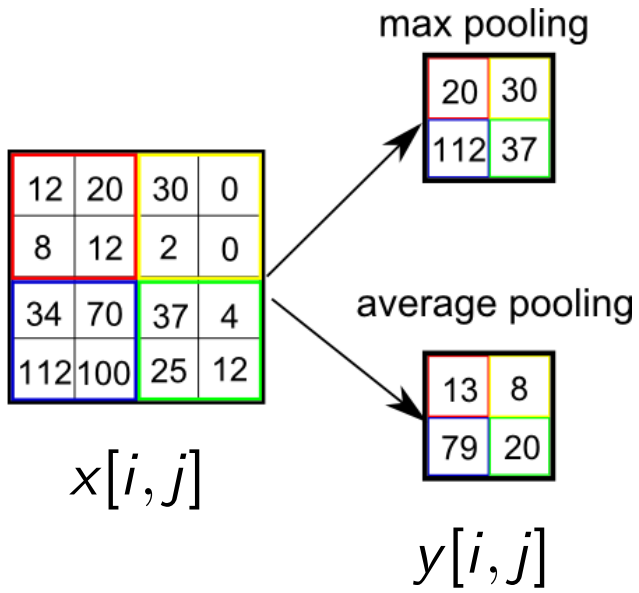
Объединяющий нейрон — необучаемая свёртка с шагом $h > 1$, агрегирующая данные прямоугольной области $h \times h$:

$$y[i,j] = F\big(x[hi,hj],\ldots,x[hi+h-1,hj+h-1]\big),$$

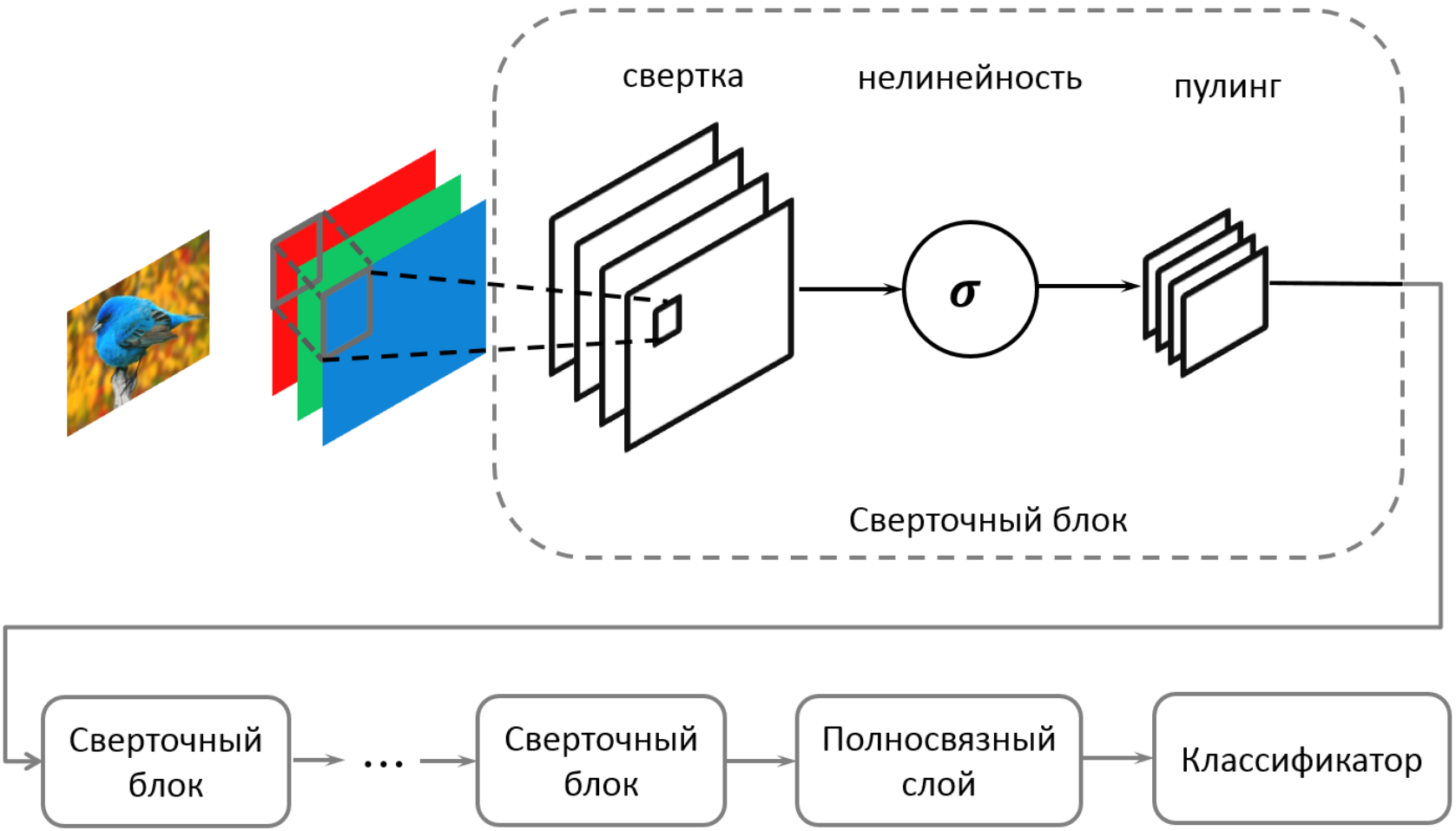
где F — агрегирующая функция: max, average и т.п.

Размер изображения сокращается в h раз по ширине и по высоте

Если нейрон предыдущего слоя отвечал за детектирование некоторого элемента, то max-pooling позволяет обнаружить этот элемент в любом месте из h-окрестности (инвариантность детектирования относительно сдвигов)



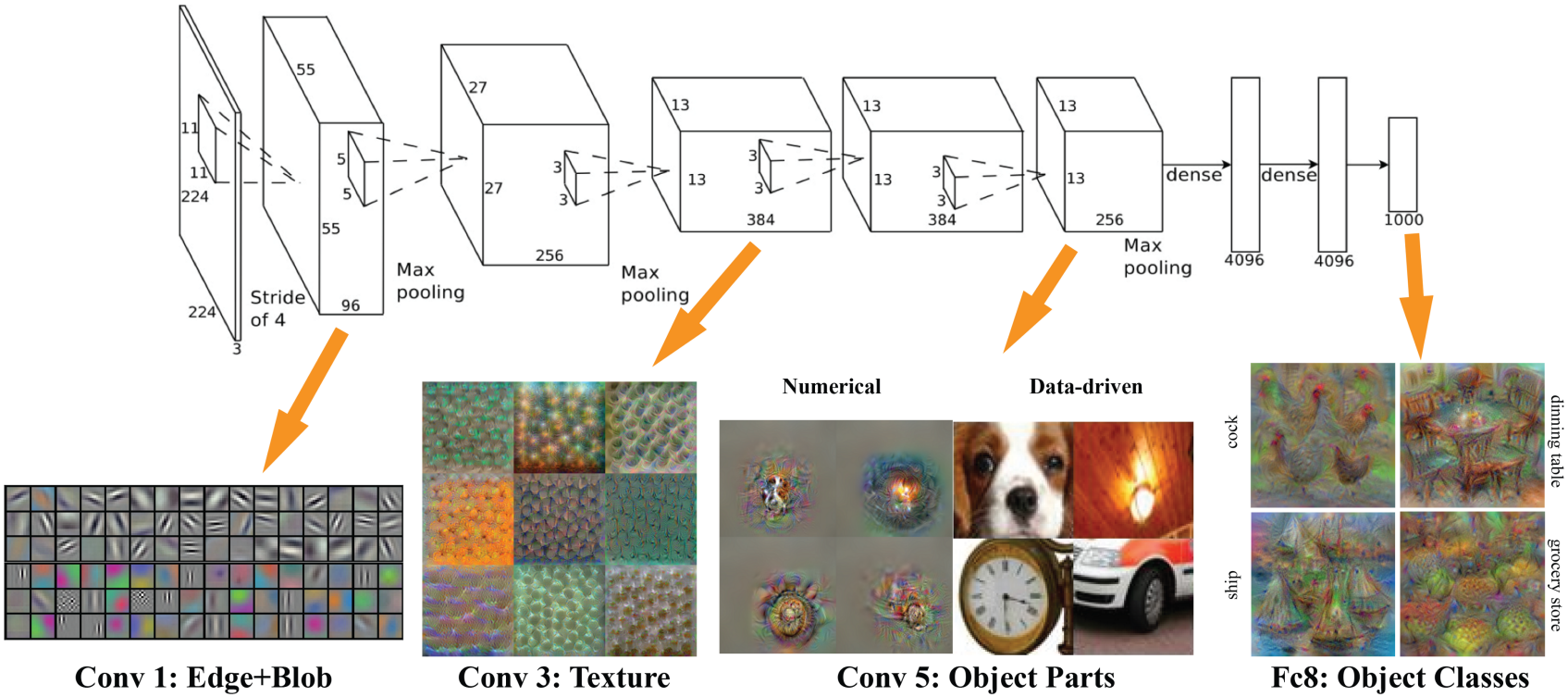
Стандартная схема сверточной сети (Convolutional NN)



Yann LeCun et al. Learning algorithms for classification: A comparison on handwritten digit recognition. 1995

Свёрточная сеть обучается извлечению признаков

Чем выше слой, тем более крупные и сложные элементы изображений он способен распознавать



Krizhevsky A., Sutskever I., Hinton G. ImageNet classification with deep convolutional neural networks. 2012.

ImageNet — большая выборка размеченных изображений

2,5 года на разметку (2008/07–2010/04)

14 197 122 изображений

21 841 классов объектов

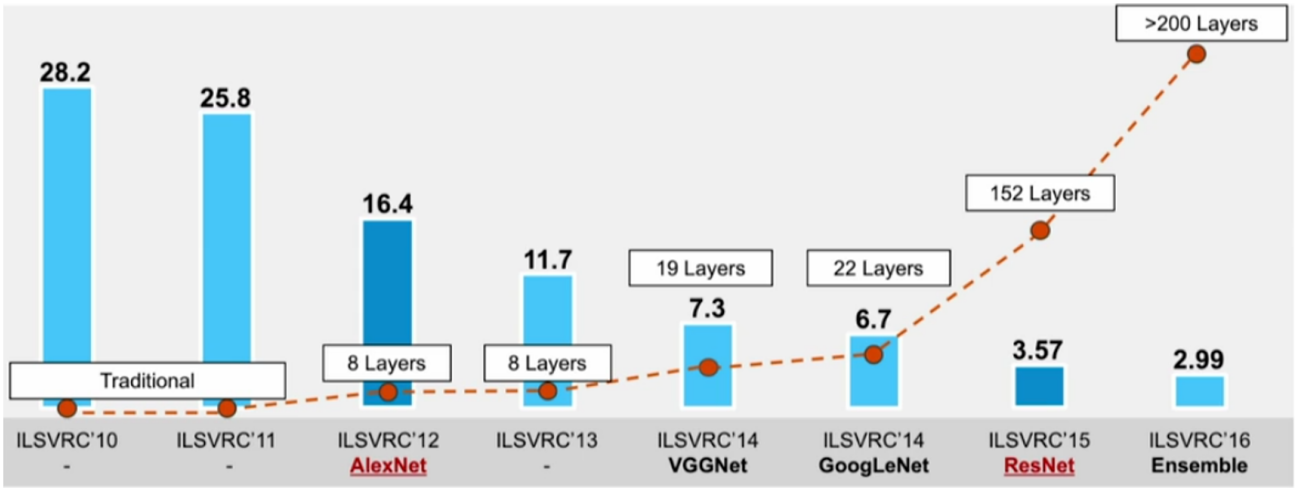
3 разметки каждого изображения

https://image-net.org

Li Fei-Fei et al. ImageNet: A large-scale hierarchical image database. 2009.

Li Fei-Fei et al. Construction and analysis of a large scale image ontology. 2009.

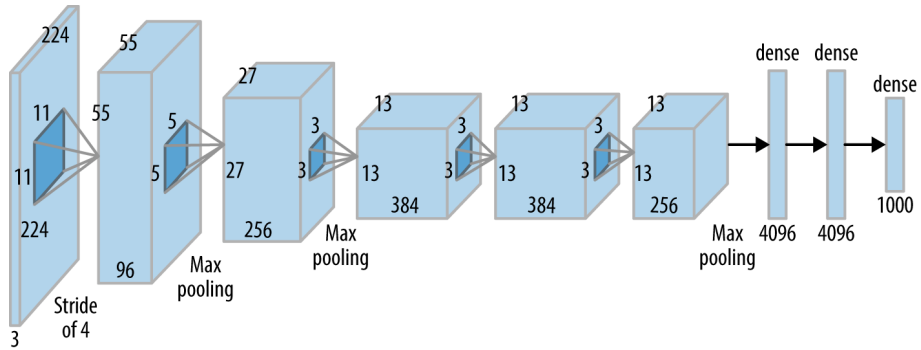
Глубокие свёрточные сети для классификации изображений



Старт в 2009. Человеческий уровень ошибок 5% пройден в 2015

Свёрточная сеть **AlexNet**:

+ ReLU + Dropout
+ 60M параметров
+ пополнение выборки
+ подбор размеров слоёв
+ GPU



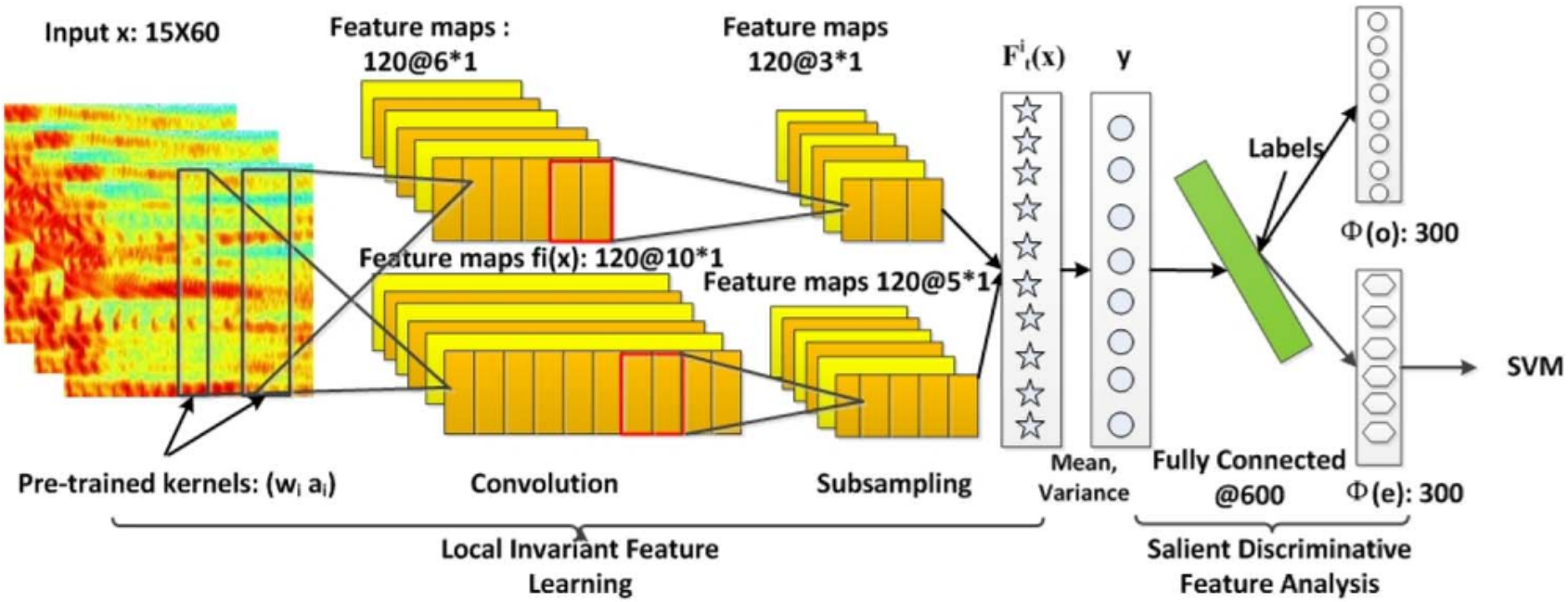
Krizhevsky A. et al. ImageNet classification with deep convolutional neural networks. 2012.

Часто используемые приёмы в CNN

- функции активации без горизонтальных асимптот, типа ReLU
- адаптивные градиентные методы
- остаточные нейронные сети (Residual NN)
- dropout (используется всё реже, считается устаревшим)
- batch normalization
- подбор числа слоёв и их размеров
- dataset augmentation — пополнение выборки с помощью преобразований, сохраняющих класс объекта

Приложение: распознавание речевых сигналов

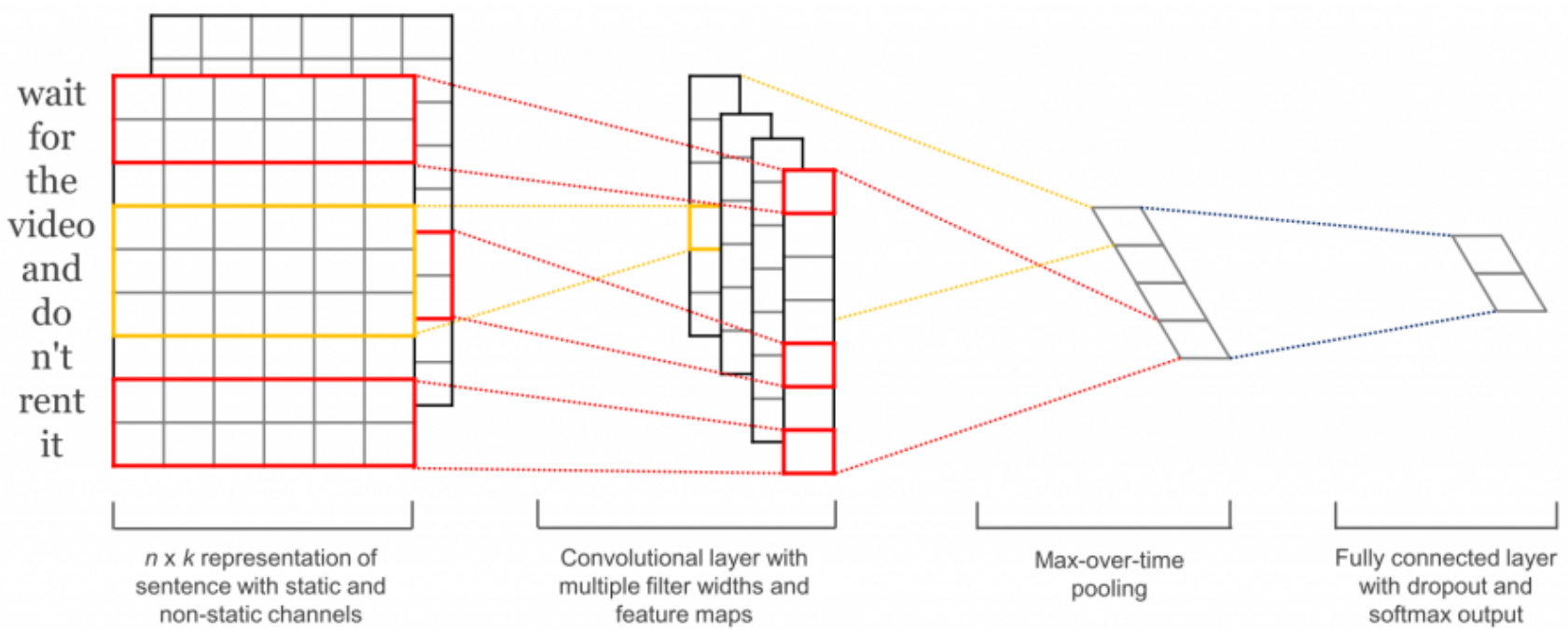
Последовательные фрагменты сигнала представляются векторами спектрального разложения



Qirong Mao, Ming Dong, Zhengwei Huang, Yongzhao Zhan. Learning salient features for speech emotion recognition using convolutional neural networks. 2014.

Приложение: классификация предложений в тексте

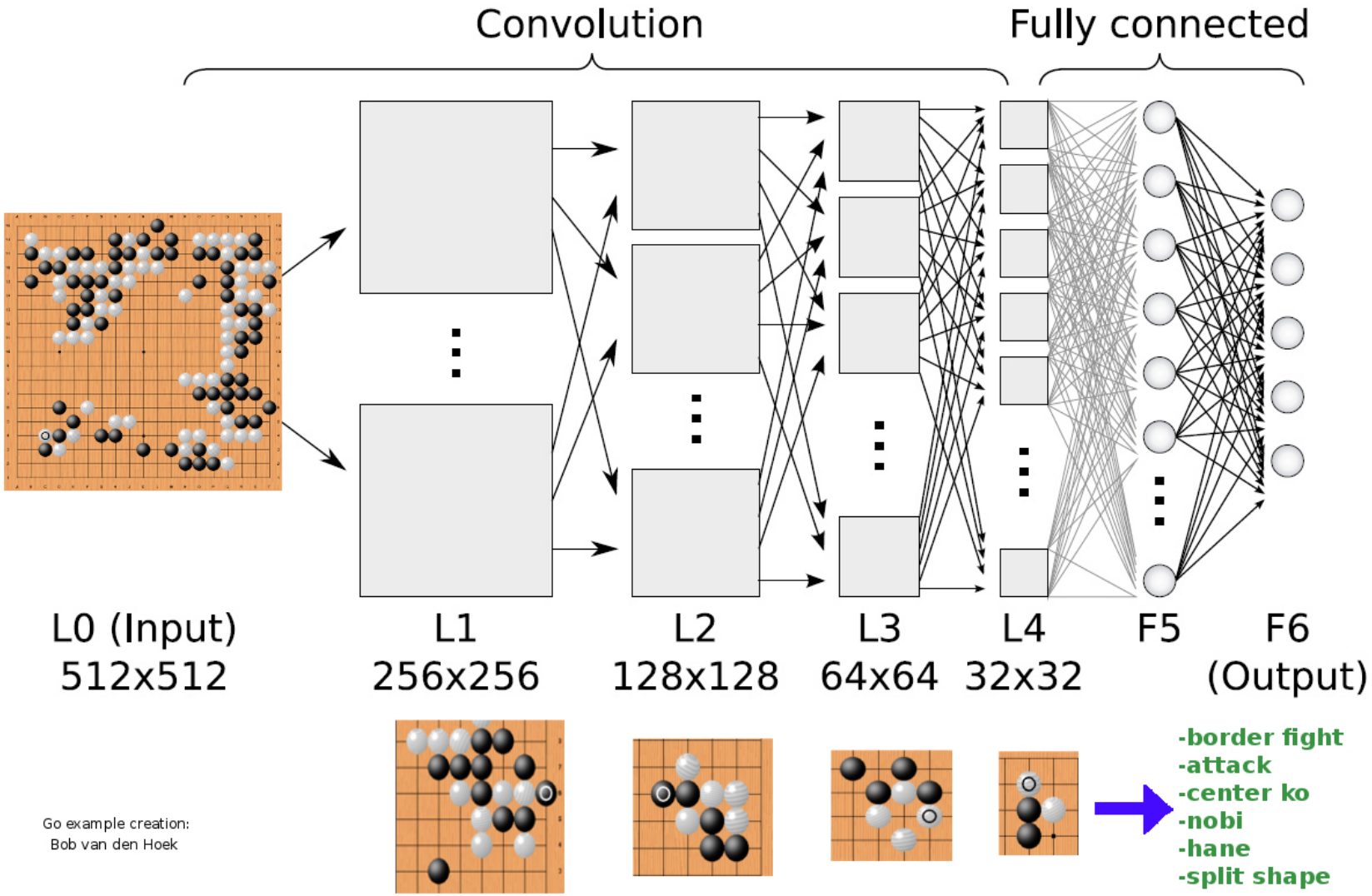
Последовательные слова в тексте представляются векторами с помощью векторных представлений (word2vec и др.)



Yoon Kim. Convolutional neural networks for sentence classification. 2014

Yann N. Dauphin et al. Language modeling with gated convolutional networks, 2017.

Приложение: принятие решений в логических играх



David Silver et al. (DeepMind) Mastering the game of Go without human knowledge. 2017

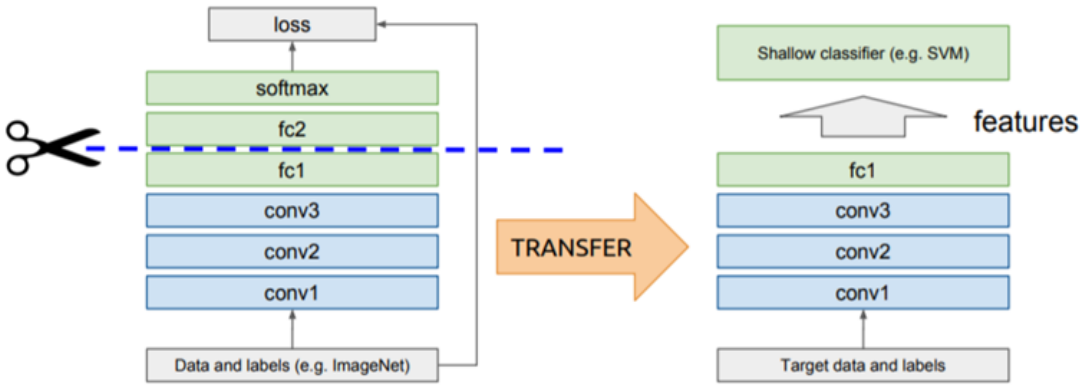
Предобучение (pre-training), перенос обучения (transfer learning)

Обучение модели векторизации $z = f(x, \alpha)$ на выборке $\{x_i\}_{i=1}^{\ell}$:

$$\sum_{i=1}^{\ell} \mathscr{L}_i\big(g(f(x_i, \alpha), \beta)\big) \;\to\; \min_{\alpha,\beta}$$

Обучение целевой модели $y = g(z, \beta)$ на малых данных $\{x'_i\}_{i=1}^{m}$:

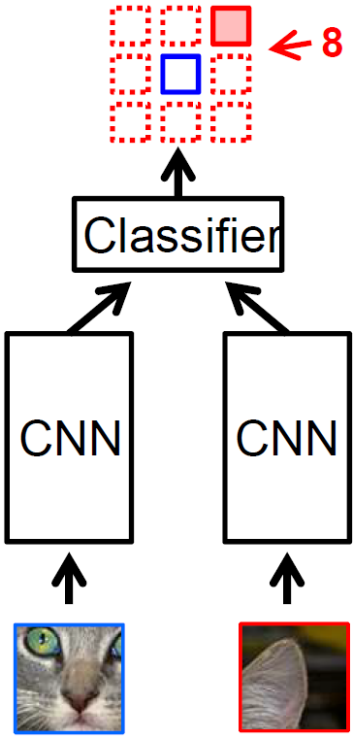
$$\sum_{i=1}^{m} \mathscr{L}'_i\big(g'(f(x'_i, \alpha), \beta')\big) \;\to\; \min_{\beta'}$$

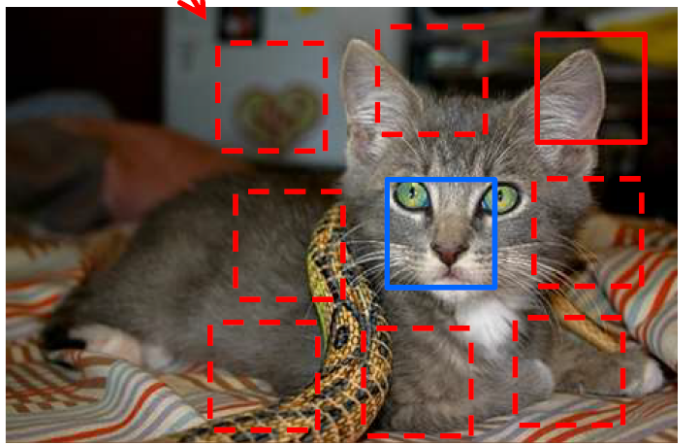


Sinno Jialin Pan, Qiang Yang. A Survey on Transfer Learning. 2009
J.Yosinski et al. How transferable are features in deep neural networks? 2014.

Самостоятельное обучение (self-supervised learning)

Модель векторизации $z = f(x, \alpha)$ обучается предсказывать взаимное расположение пар фрагментов одного изображения





Unsupervised visual representation learning by context prediction,
Carl Doersch, Abhinav Gupta, Alexei A. Efros, ICCV 2015

Преимущество: сеть выучивает векторные представления объектов без размеченной обучающей выборки (без ImageNet).

Автокодировщик (AutoEncoder) — обучение без учителя

Дано: $X^\ell = \{x_1, \ldots, x_\ell\}$ — обучающая выборка

Найти две модели одновременно:

$f: X \to Z$ — кодировщик (encoder), кодовый вектор $z = f(x, \alpha)$

$g: Z \to X$ — декодировщик (decoder), реконструкция $\hat{x} = g(z, \beta)$

Критерий: качество реконструкции исходных объектов x_i:

$$\sum_{i=1}^{\ell} \mathscr{L}\big(g(f(x_i, \alpha), \beta), x_i\big) \to \min_{\alpha,\beta}$$

Квадратичная функция потерь: $\mathscr{L}(\hat{x}, x) = \|\hat{x} - x\|^2$

Пример 1. Линейный автокодировщик: $x \in \mathbb{R}^n$, $z \in \mathbb{R}^m$

$$f(x, A) = \underset{m\times n}{A}\, x, \quad g(z, B) = \underset{n\times m}{B}\, z$$

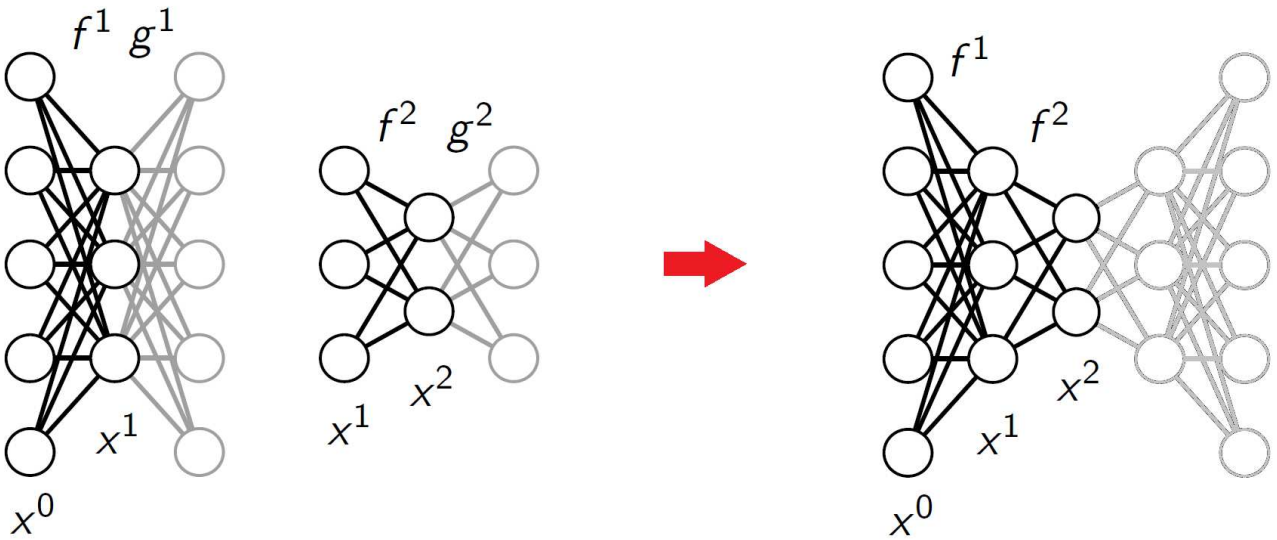
Пример 2. Двухслойная сеть с функциями активации σ_f, σ_g

$$f(x, A) = \sigma_f(Ax + a), \quad g(z, B) = \sigma_g(Bz + b)$$

Многослойный автокодировщик (Stacked AE)

Послойное обучение: $x^h = f^h(x^{h-1}, \alpha^h)$, $x \equiv x^0$, $z \equiv x^H$

- каждая пара f^h, g^h обучается по выборке $\{x_1^{h-1}, \ldots, x_\ell^{h-1}\}$
- декодировщик g^h отбрасывается
- однослойные $f^1, \ldots, f^H$ соединяются в H-слойный



Тонкая настройка (fine tuning): результат послойного обучения используется как начальное приближение для BackProp

---

Y. Bengio et al. Greedy layer-wise training of deep networks. NIPS 2007.

Автокодировщики для векторизации и обучения с учителем

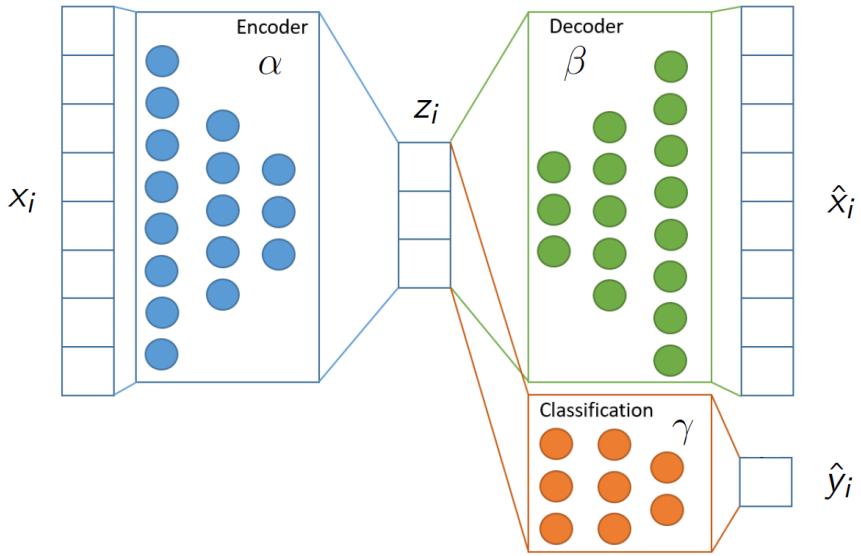
Данные: размеченные $(x_i, y_i)_{i=1}^{k}$, неразмеченные $(x_i)_{i=k+1}^{\ell}$

Найти:

$z_i = f(x_i, \alpha)$ — кодировщик

$\hat{x}_i = g(z_i, \beta)$ — декодировщик

$\hat{y}_i = \hat{y}(z_i, \gamma)$ — предиктор



Задаются функции потерь:

$\mathscr{L}(\hat{x}_i, x_i)$ — реконструкция

$\tilde{\mathscr{L}}(\hat{y}_i, y_i)$ — предсказание

Критерий: совместное обучение автокодировщика и предсказательной модели (классификации, регрессии или др.):

$$\sum_{i=1}^{\ell} \mathscr{L}\big(g(f(x_i, \alpha), \beta), x_i\big) + \lambda \sum_{i=1}^{k} \tilde{\mathscr{L}}(\hat{y}(f(x_i, \alpha), \gamma), y_i) \;\to\; \min_{\alpha,\beta,\gamma}$$

Dor Bank, Noam Koenigstein, Raja Giryes. Autoencoders. 2020

Многозадачное обучение (multi-task learning)

$z = f(x, \alpha)$ — векторизация, универсальная для всех моделей
$g_t(z, \beta)$ — специфичная часть модели для задачи $t \in T$

Одновременное обучение модели f по задачам X_t, $t \in T$:

$$\sum_{t \in T} \sum_{i \in X_t} \mathscr{L}_{ti}\big(g_t(f(x_{ti}, \alpha), \beta_t)\big) \;\to\; \min_{\alpha, \{\beta_t\}}$$

Обучаемость (learnability): качество решения отдельной задачи $\langle X_t, \mathscr{L}_t, g_t \rangle$ улучшается с ростом объёма выборки $\ell_t = |X_t|$.

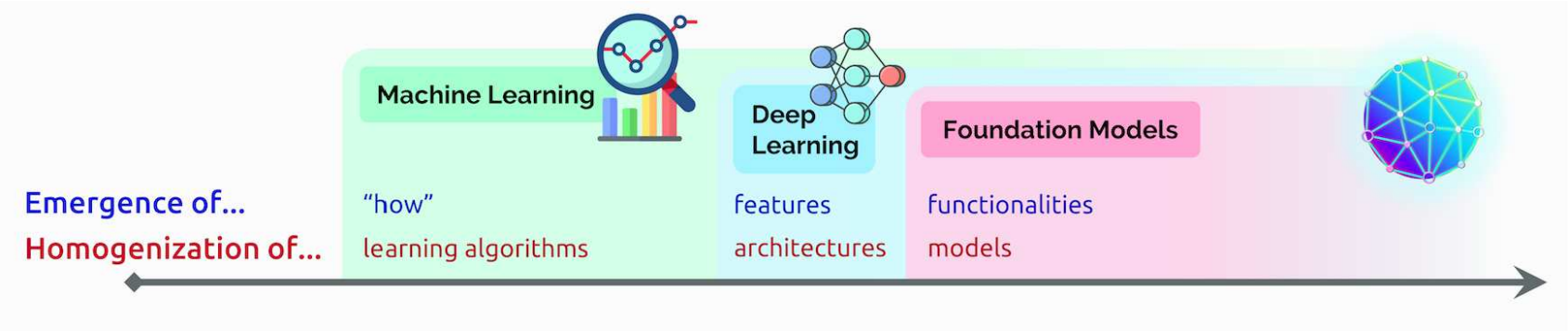
Learning to learn: качество решения каждой из задач $t \in T$ улучшается с ростом как ℓ_t, так и общего числа задач $|T|$.

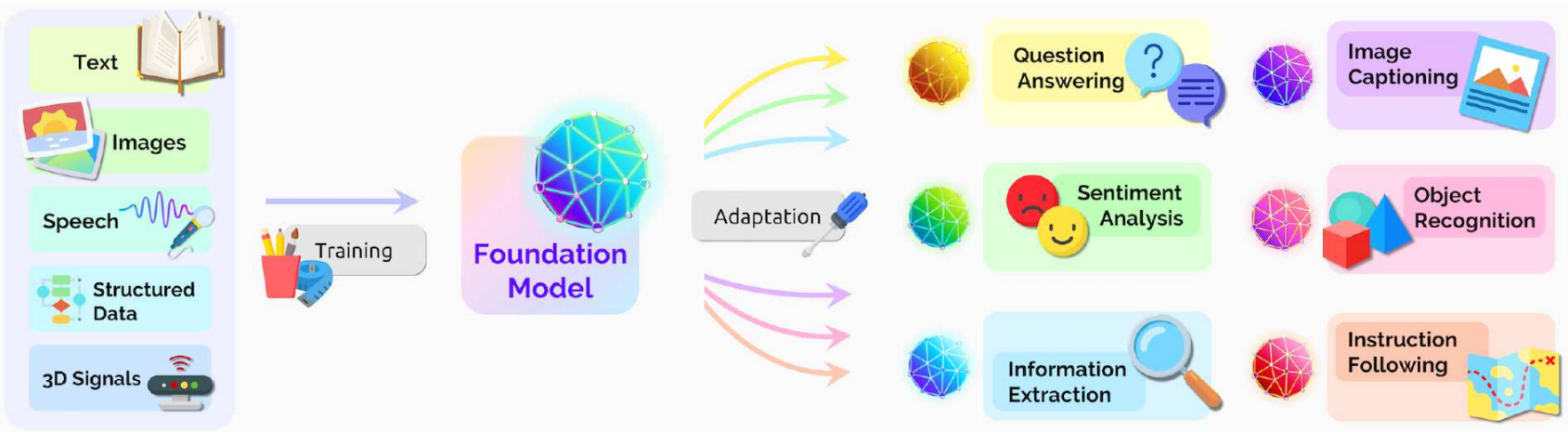
Few-shot learning: для решения новой задачи t достаточно небольшого числа примеров, даже одного (*one-shot learning*).

Yu Zhang, Qiang Yang. A survey on multi-task learning. 2021
M.Crawshaw. Multi-task learning with deep neural networks: a survey. 2020
Y.Wang et al. Generalizing from a few examples: a survey on few-shot learning. 2020

Обучаемая векторизация данных — глобальный тренд AI/ML

Foundation Models — гомогенизация векторных представлений





R.Bommasani et al. (Center for Research on Foundation Models, Stanford University) On the opportunities and risks of foundation models // CoRR, 20 August 2021.

Вариационный автокодировщик (Variational AE)

Дано: выборка объектов $X^\ell = \{x_1, \ldots, x_\ell\}$

Найти: модель для генерации новых объектов, похожих на x_i:

$q_\alpha(z|x)$ — вероятностный кодировщик с параметром α

$p_\beta(\hat{x}|z)$ — вероятностный декодировщик с параметром β

Критерий: максимум log-правдоподобия (его нижней оценки)

$$\sum_{i=1}^{\ell} \ln p(x_i) = \sum_{i=1}^{\ell} \ln \int q_\alpha(z|x_i) \frac{p_\beta(x_i|z)p(z)}{q_\alpha(z|x_i)} dz \geqslant$$

$$\geqslant \sum_{i=1}^{\ell} \int q_\alpha(z|x_i) \ln \frac{p_\beta(x_i|z)p(z)}{q_\alpha(z|x_i)} dz =$$

$$= \sum_{i=1}^{\ell} \int q_\alpha(z|x_i) \ln p_\beta(x_i|z) dz - \mathrm{KL}\big(q_\alpha(z|x_i) \,\big\|\, p(z)\big) \to \max_{\alpha,\beta}$$

D.P.Kingma, M.Welling. Auto-encoding Variational Bayes. 2013.
C.Doersch. Tutorial on variational autoencoders. 2016.

Вариационный автокодировщик (Variational AE)

Оптимизационная задача для вариационного автокодировщика:

$$\sum_{i=1}^{\ell} \underbrace{\mathsf{E}_{z\sim q_\alpha(z|x_i)} \ln p_\beta(x_i|z)}_{\substack{\text{качество реконструкции} \\ \approx \ln p_\beta(x_i|z),\ z\sim q_\alpha(z|x_i)}} - \underbrace{\mathsf{KL}\big(q_\alpha(z|x_i) \,\big\|\, p(z)\big)}_{\text{регуляризатор по } \alpha} \to \max_{\alpha,\beta}$$

где $p(z)$ — априорное распределение, обычно $\mathcal{N}(0, \sigma^2 I)$

Репараметризация $q_\alpha(z|x_i)$: $\quad z = f(x_i, \alpha, \varepsilon),\ \varepsilon \sim \mathcal{N}(0, I)$

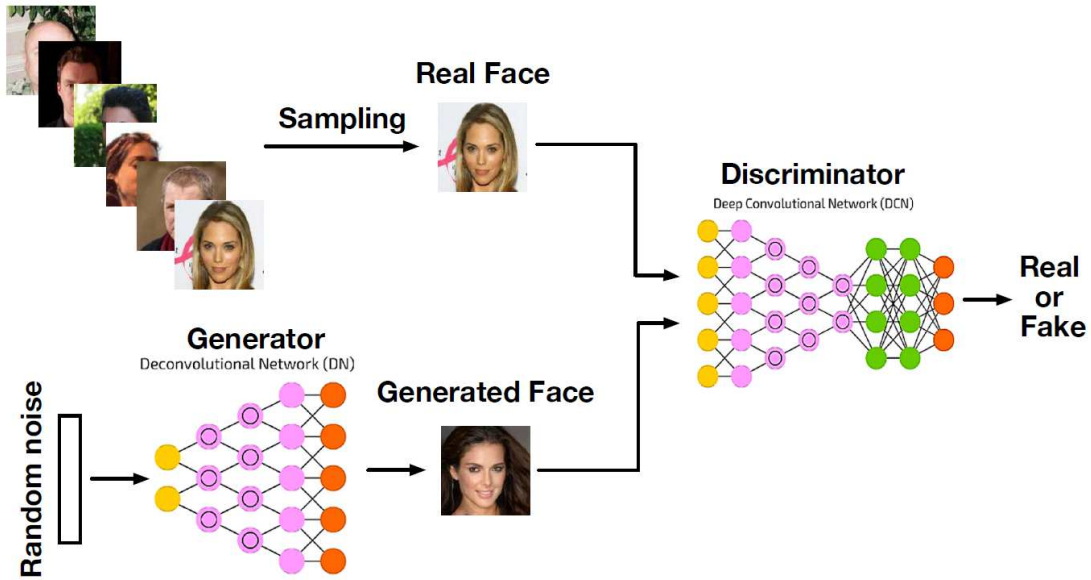
Метод стохастического градиента:

- сэмплировать $x_i \sim X^\ell,\ \ \varepsilon \sim \mathcal{N}(0, I),\ \ z = f(x_i, \alpha, \varepsilon)$
- градиентный шаг:
  $\alpha := \alpha + h\nabla_\alpha \big[\ln p_\beta\big(x_i|f(x_i, \alpha, \varepsilon)\big) - \mathsf{KL}\big(q_\alpha(z|x_i)\|p(z)\big)\big];$
  $\beta := \beta + h\nabla_\beta \big[\ln p_\beta(x_i|f(x_i, \alpha, \varepsilon))\big];$

Генерация похожих объектов: $x \sim p_\beta\big(x|f(x_i, \alpha, \varepsilon)\big),\ \ \varepsilon \sim \mathcal{N}(0, I)$

Генеративная состязательная сеть (Generative Adversarial Net)

Генератор $G(z)$ учится порождать объекты x из шума z
Дискриминатор $D(x)$ учится отличать их от реальных объектов



Antonia Creswell et al. Generative Adversarial Networks: an overview. 2017.
Zhengwei Wang, Qi She, Tomas Ward. Generative Adversarial Networks: a survey and taxonomy. 2019.
Chris Nicholson. A Beginner's Guide to Generative Adversarial Networks. https://pathmind.com/wiki/generative-adversarial-network-gan. 2019.

Постановка задачи GAN

Дано: выборка объектов $\{x_i\}_{i=1}^{\ell}$

Найти:

вероятностную генеративную модель $G(z, \alpha)$: $x \sim p(x|z, \alpha)$

вероятностную дискриминативную модель $D(x, \beta) = p(1|x, \beta)$

Критерий:

max log-правдоподобия для дискриминативной модели D:

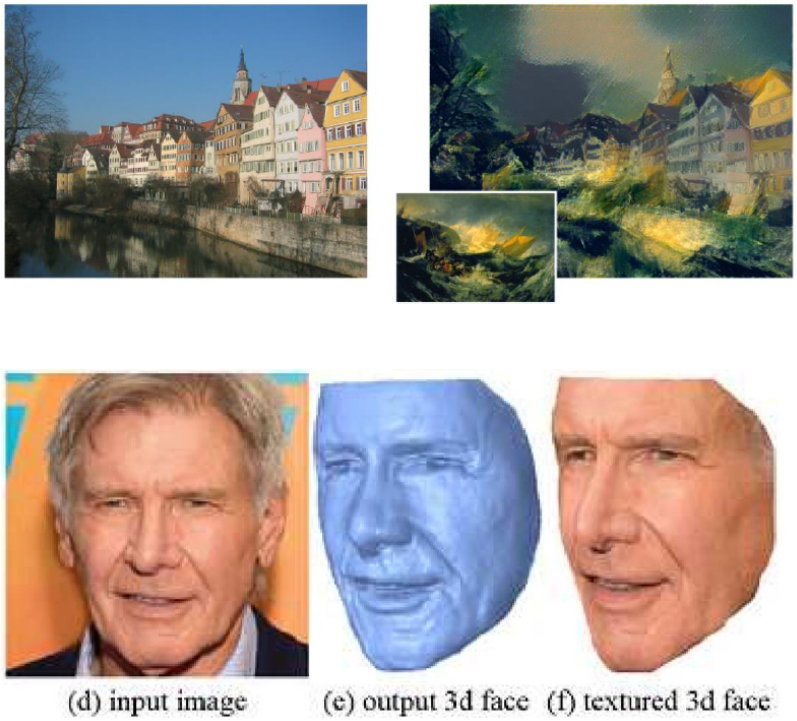
$$\sum_{i=1}^{\ell} \ln D(x_i, \beta) + \ln\big(1 - D(G(z_i, \alpha), \beta)\big) \;\to\; \max_{\beta}$$

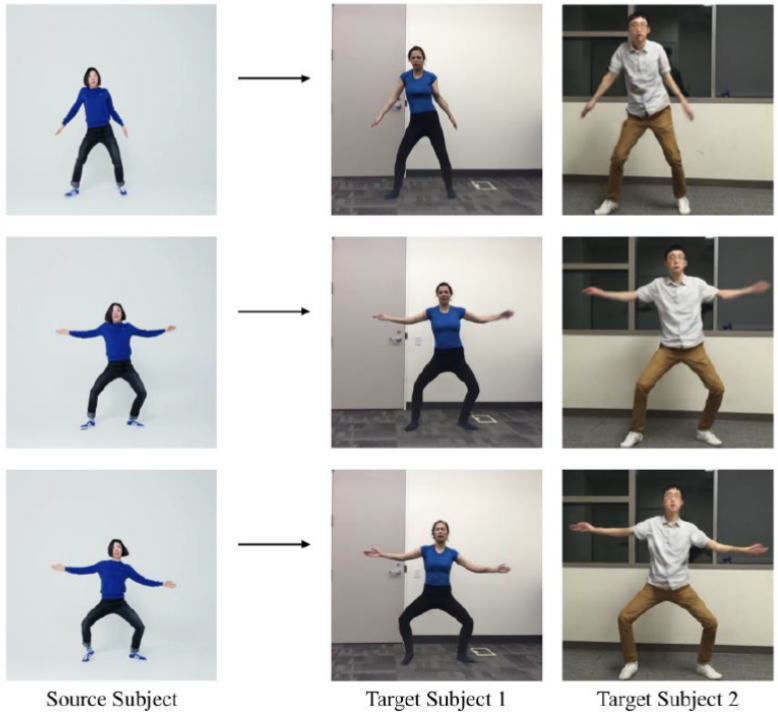
обучение генеративной модели G по случайному шуму $\{z_i\}_{i=1}^{\ell}$:

$$\sum_{i=1}^{\ell} \ln\big(1 - D(G(z_i, \alpha), \beta)\big) \;\to\; \min_{\alpha}$$

Ian Goodfellow et al. Generative Adversarial Nets. 2014

Примеры GAN для синтеза изображений и видео





Chuan Li, Michael Wand. Precomputed Real-Time Texture Synthesis with Markovian Generative Adversarial Networks. 2016.

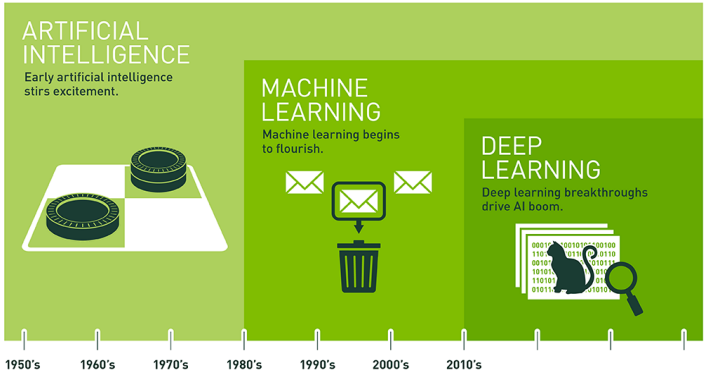
Xiaoxing Zeng, Xiaojiang Peng, Yu Qiao. DF2Net: A Dense Fine Finer Network for Detailed 3D Face Reconstruction. ICCV-2019.

C.Chan, S.Ginosar, T.Zhou, A.Efros. Everybody Dance Now. ICCV-2019.

Резюме

ИИ вне машинного обучения:

— экспертные системы
— системы представления знаний
— системы логического вывода
— автоматическое док-во теорем



Три этапа развития машинного обучения

- **вектор** $\to$ **вектор** $\to$ **скаляр** (shallow learning)
  конструирование информативных признаков
  в задачах предсказательного моделирования
- **структура** $\to$ **вектор** $\to$ **скаляр** (deep learning)
  векторизация сложно структурированных данных,
  обучаемая совместно с предсказательной моделью
- **структура** $\to$ **вектор** $\to$ **структура** (generative models)
  генеративные модели сложно структурированных данных,
  единое пространство эмбедингов (foundation models)