

Курс «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 • 6 марта 2025

1 Вектор \rightarrow вектор \rightarrow скаляр

- Задачи с векторными описаниями объектов
- Методы преобразования признаков
- Конструирование признаков

2 Структура \rightarrow вектор \rightarrow скаляр

- Свёрточные нейронные сети
- Векторизация сложно структурированных данных
- Перенос обучения, самостоятельное обучение

3 Структура \rightarrow вектор \rightarrow структура

- Автокодировщики
- Фундаментальные модели
- Генеративные модели

Восстановление зависимостей по эмпирическим данным

Дано: обучающая выборка объектов

$$x_i = (f_1(x_i), \dots, f_n(x_i)), \quad i = 1, \dots, \ell$$

$f_j: X \rightarrow D_j$ — признаки, $j = 1, \dots, n$

Найти: параметры w модели $a(x, w)$

Критерий: \min эмпирического риска

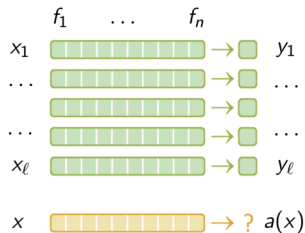
$$\sum_{i=1}^{\ell} \mathcal{L}(w, x_i) + \tau \mathcal{R}(w) \rightarrow \min_w$$

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

Преобразование признаков (feature transformation/extraction)

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

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 $f(x)$, $f: X \rightarrow D$

Тип шкалы определяется множествами

- допустимых биективных преобразований $\psi: D \rightarrow D'$
- 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: $1, 2, 3, \dots, N$)
оценка в баллах, шкалы Рихтера, Бофорта, Мооса, Бека
- **Интервальная** (можно сдвигать и менять масштаб)
время, географическая широта, температура ($^{\circ}\text{C}$, $^{\circ}\text{F}$)
- **Отношений** (можно менять масштаб)
масса, скорость, объём, сила, давление, заряд, яркость, $^{\circ}\text{K}$
- **Абсолютная**
число предметов, частота события, оценка вероятности

Ослабление шкалы

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

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

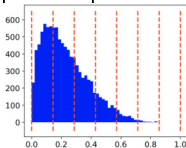
Числовой или порядковый → бинарный:

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

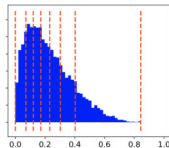
Числовой → ранговый (data binning, quantization):

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

равномерная сетка



квантильная сетка



Ослабление шкалы всегда влечёт потерю информации

Усиление шкалы

Номинальный → числовой:

- категория заменяется частотой:

$$f'(x) = \frac{1}{\ell} \sum_{i=1}^{\ell} [f(x_i) = f(x)]$$

- условное среднее числового признака $g(x)$:

$$f'(x) = \text{mean}(g|f(x)) = \frac{\sum_{i=1}^{\ell} g(x_i) [f(x_i) = f(x)]}{\sum_{i=1}^{\ell} [f(x_i) = f(x)]},$$

- условное среднее целевой величины $y(x)$:

$$f'(x) = \text{mean}(y|f(x)), \text{ возможно переобучение!}$$

Порядковый → числовой (монотонное преобразование)

- значение заменяется частотой:

$$f'(x) = \frac{1}{\ell} \sum_{i=1}^{\ell} [f(x_i) \leq f(x)]$$

Нормализация и стандартизация числовых шкал

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

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

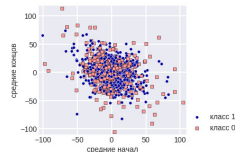
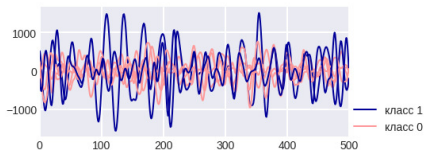
Feature Engineering: признаки вычисляются по формулам, которые зависят от задачи, требуют изобретательности и знаний предметной области. Долго, дорого, неточно.

- **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

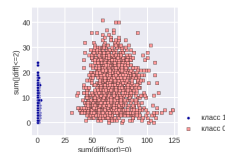
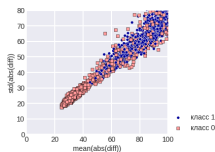
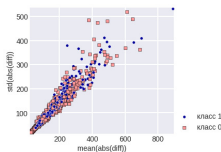
Пример. Задача детектирования поломок по сигналу датчика (соревнование «Ford Classification Challenge», 2008)

Иногда при удачном выборе признаков задача решается без ML

Признаки, генерируемые по исходным временным рядам, слабы:



Среди признаков рядов их производных оказывается идеальный:



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

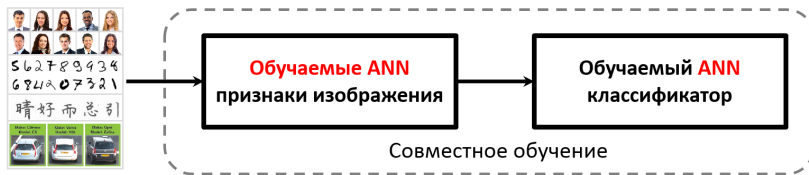
J. Daugman. High confidence visual recognition of persons by a test of statistical independence. 1993

Конструирование признаков для распознавания изображений

Классический подход к распознаванию изображений:



Современный подход — end-to-end deep learning:



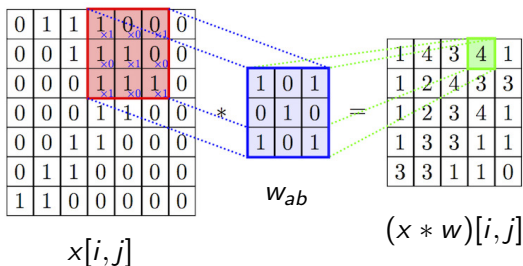
Yann LeCun et al. Learning algorithms for classification: A comparison on handwritten digit recognition. 1995

Свёрточный слой нейронов (convolution layer)

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

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)

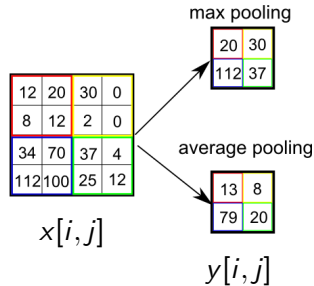
Объединяющий нейрон — необучаемая свёртка с шагом $h > 1$, агрегирующая данные прямоугольной области $h \times h$:

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

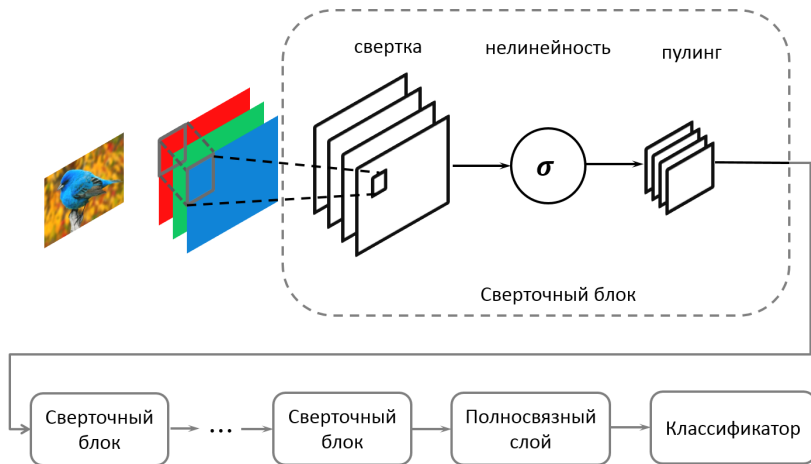
где F — агрегирующая функция: max, average и т.п.

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)



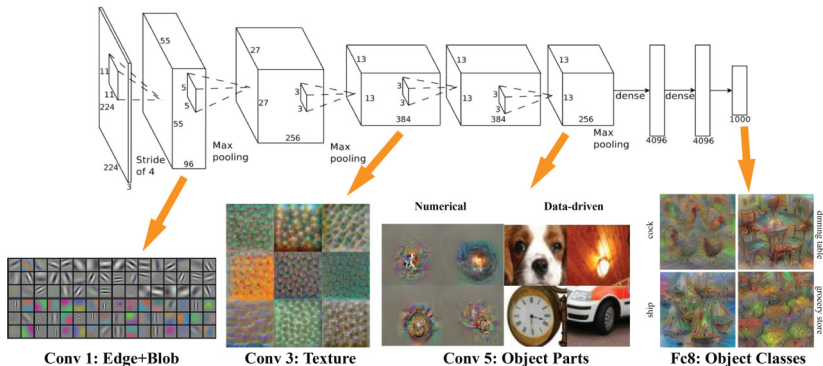
Стандартная схема сверточной сети (Convolutional NN)



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

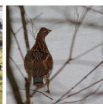
3 разметки
каждого
изображения



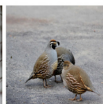
flamingo



cock



ruffed grouse



quail



partridge



Egyptian cat



Persian cat



Siamese cat



tabby



lynx



dalmatian



keeshond



miniature schnauzer



standard schnauzer



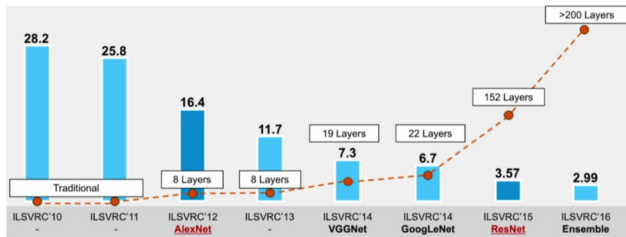
giant schnauzer

<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**:

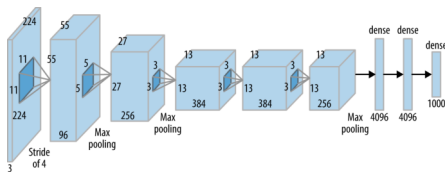
+ ReLU + Dropout

+ 60M параметров

+ пополнение выборки

+ подбор размеров слоёв

+ GPU



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

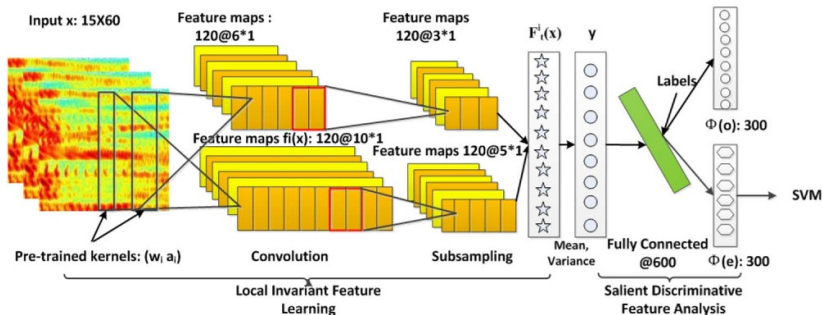
Часто используемые приёмы в CNN

- функции активации без горизонтальных асимптот, типа ReLU
- адаптивные градиентные методы
- dropout
- batch normalization
- сквозные связи (skip connection, Residual NN)
- подбор числа слоёв и их размеров
- dataset augmentation — пополнение выборки с помощью преобразований, сохраняющих класс объекта



Приложение: распознавание речевых сигналов

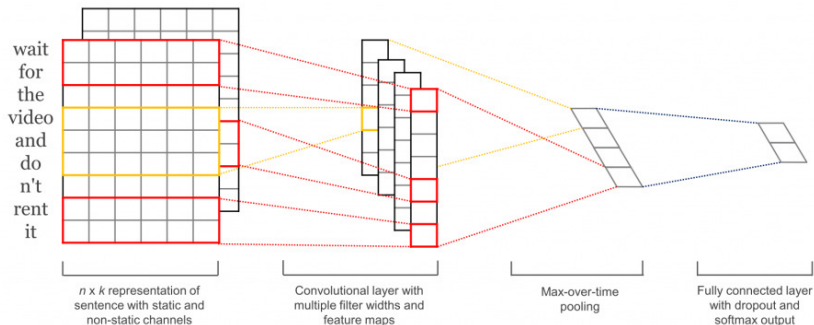
Последовательные фрагменты сигнала представляются векторами спектрального разложения



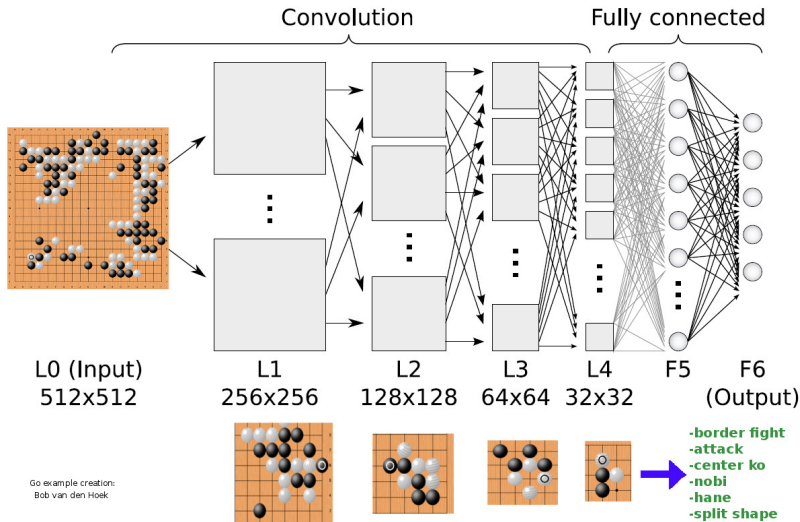
Qirong Mao, Ming Dong, Zhengwei Huang, Yongzhao Zhan. Learning salient features for speech emotion recognition using convolutional neural networks. 2014.

Приложение: классификация предложений в тексте

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



Приложение: принятие решений в логических играх



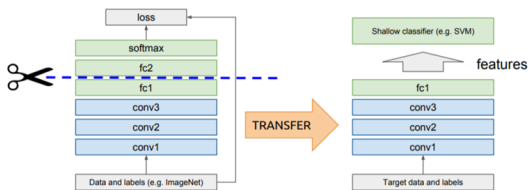
Предобучение (pre-training), перенос обучения (transfer learning)

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

$$\sum_{i=1}^{\ell} \mathcal{L}_i(g(f(x_i, \alpha), \beta)) \rightarrow \min_{\alpha, \beta}$$

Обучение целевой модели $y = g(z, \beta)$ на малых данных:

$$\sum_{i=1}^m \mathcal{L}'_i(g'(f(x'_i, \alpha), \beta')) \rightarrow \min_{\beta'}$$

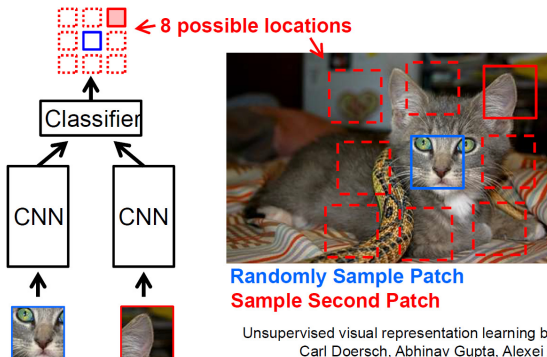


Sinno Jialin Pan, Qiang Yang. A Survey on Transfer Learning. 2009

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



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

Построение автокодировщика — задача обучения без учителя

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

Найти две модели одновременно:

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

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

Критерий: качество реконструкции исходных объектов x_i :

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

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

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

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

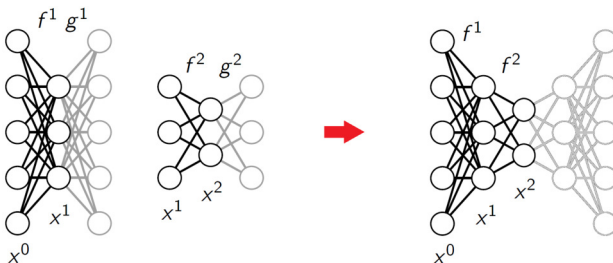
Пример 2. Двухслойная сеть с функциями активации σ_f, σ_g

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

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 $\{x_1^{h-1}, \dots, x_\ell^{h-1}\}$
- декодировщик g^h отбрасывается
- однослойные f^1, \dots, 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

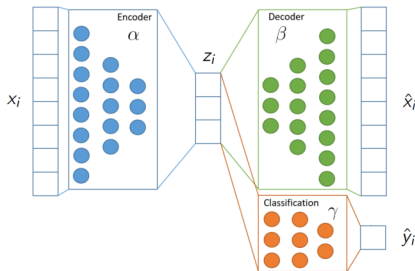
$\hat{x}_i = g(z_i, \beta)$ — декодировщик

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

Задаются функции потерь:

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

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



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

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

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 $X_t, t \in T$:

$$\sum_{t \in T} \sum_{i \in X_t} \mathcal{L}_{ti}(g_t(f(x_{ti}, \alpha), \beta_t)) \rightarrow \min_{\alpha, \{\beta_t\}}$$

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

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

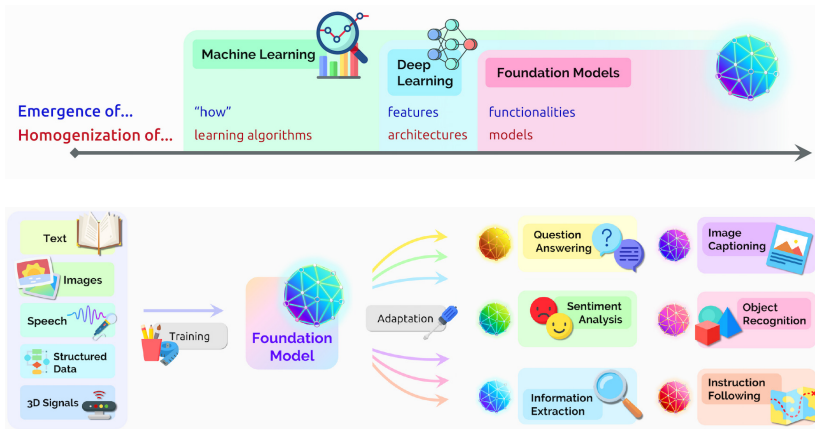
Few-shot learning: для решения новой задачи t достаточно небольшого числа примеров, даже одного (*one-shot learning*).

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 $X^\ell = \{x_1, \dots, x_\ell\}$

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: max нижней оценки log-правдоподобия

$$\begin{aligned} \sum_{i=1}^{\ell} \ln p(x_i) &= \sum_{i=1}^{\ell} \ln \int q_\alpha(z|x_i) \frac{p_\beta(x_i|z)p(z)}{q_\alpha(z|x_i)} dz \geq \\ &\geq \sum_{i=1}^{\ell} \int q_\alpha(z|x_i) \ln \frac{p_\beta(x_i|z)p(z)}{q_\alpha(z|x_i)} dz = \\ &= \sum_{i=1}^{\ell} \int q_\alpha(z|x_i) \ln p_\beta(x_i|z) dz - \text{KL}(q_\alpha(z|x_i) \parallel p(z)) \rightarrow \max_{\alpha, \beta} \end{aligned}$$

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:

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

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

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

Метод стохастического градиента:

- сэмплировать $x_i \sim X^{\ell}$, $\varepsilon \sim \mathcal{N}(0, I)$, $z = f(x_i, \alpha, \varepsilon)$
- градиентный шаг:

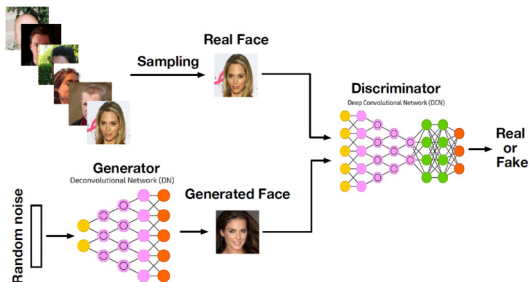
$$\alpha := \alpha + h \nabla_{\alpha} [\ln p_{\beta}(x_i | f(x_i, \alpha, \varepsilon)) - \text{KL}(q_{\alpha}(z|x_i) \parallel p(z))];$$

$$\beta := \beta + h \nabla_{\beta} [\ln p_{\beta}(x_i | z)];$$

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

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 $\{x_i\}_{i=1}^m$ из X

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 :

$$\sum_{i=1}^m \ln D(x_i, \beta) + \ln(1 - D(G(z_i, \alpha), \beta)) \rightarrow \max_{\beta}$$

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

$$\sum_{i=1}^m \ln(1 - D(G(z_i, \alpha), \beta)) \rightarrow \min_{\alpha}$$

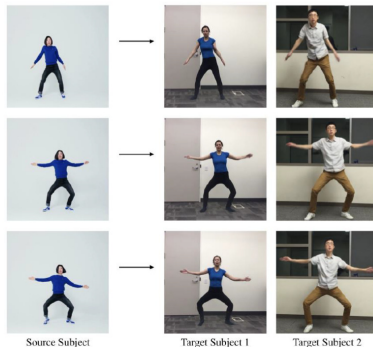
Примеры GAN для синтеза изображений и видео



(d) input image

(e) output 3d face

(f) textured 3d face



Source Subject

Target Subject 1

Target Subject 2

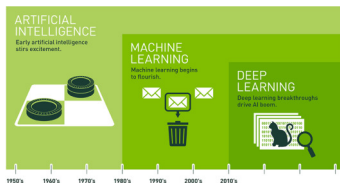
Chuan Li, Michael Wand. Precomputed Real-Time Texture Synthesis with Markovian Generative Adversarial Networks. 2016.

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.

ИИ до машинного обучения:

- экспертные системы
- системы представления знаний
- системы логического вывода



Три этапа развития машинного обучения

- вектор \rightarrow вектор \rightarrow скаляр (shallow learning)
конструирование более информативных признаков
в задачах предсказательного моделирования
- структура \rightarrow вектор \rightarrow скаляр (deep learning)
векторизация сложно структурированных данных,
обучаемая совместно с предсказательной моделью
- структура \rightarrow вектор \rightarrow структура (deep learning)
генеративные модели сложно структурированных данных