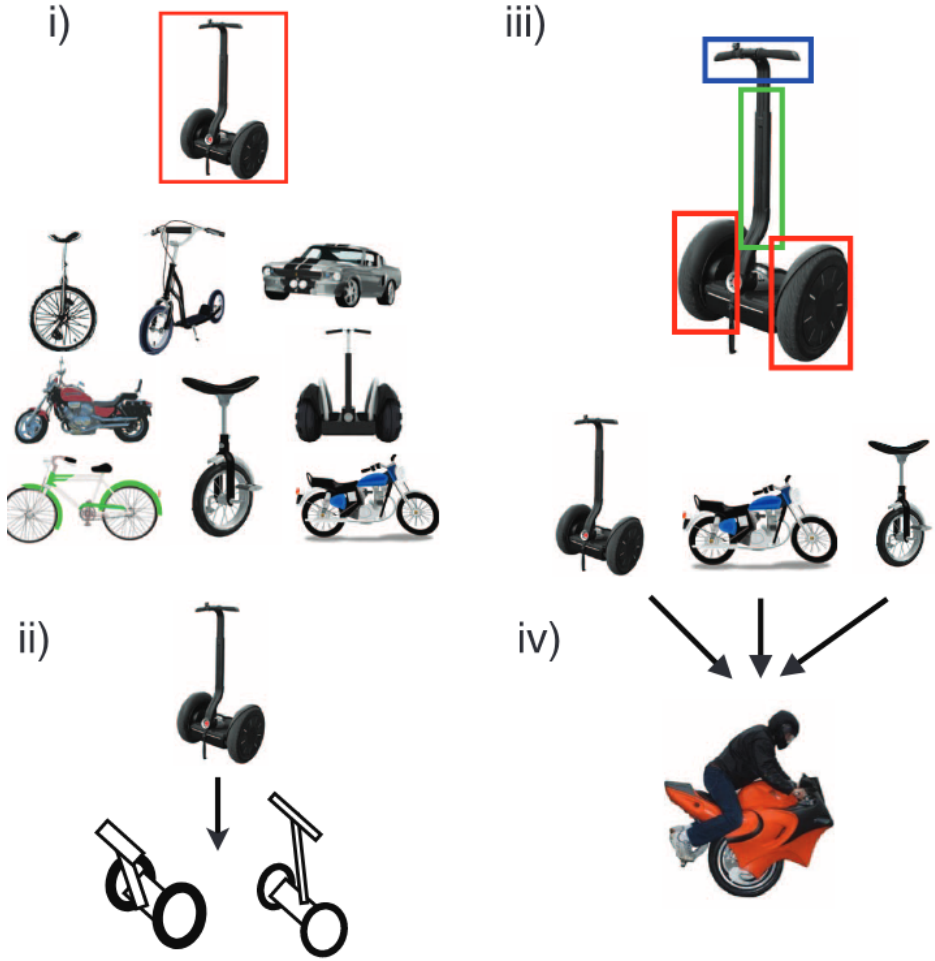
Bayesian programming

Бахтеев Олег

11.11.2020

Мотивация

Любой предмет представляется человеку как совокупность связанных между собой понятий.



Основные работы

- Lake B. M., Salakhutdinov R., Tenenbaum J. B. Human-level concept learning through probabilistic program induction //Science. – 2015. – Т. 350. – №. 6266. – С. 1332-1338.
- Lake B. M. et al. Building machines that learn and think like people //Behavioral and brain sciences. – 2017. – Т. 40.
- Lake B. M., Salakhutdinov R. R., Tenenbaum J. One-shot learning by inverting a compositional causal process //Advances in neural information processing systems. – 2013. – С. 2526-2534.

Постановка

One-shot classification

The tasks tested within-alphabet classification on 10 alphabets. Each trial (of 400 total) consists of a single test image ofa new character compared to 20 new characters from the same alphabet, given just one image eachproduced by a typical drawer of that alphabet.

Идея подхода

Три ключевые концепции подхода

- Композиционность
  - Объект разбивается на несколько элементов
  - Каждый элемент характеризуется своей генеративной моделью
- Причинность
  - Над элементами вводится вероятностная иерархия
  - Элемент сам по себе представляется в иерархическом виде (токен и тип токена)
- Обучение обучению
  - Настройка гиперпараметров на обучающем датасете

Модель

A character type $\psi = \{\kappa, S, R\}$ is defined by a set of κ strokes $S = \{S_1, ..., S_\kappa\}$ and spatial relations $R = \{R_1, ..., R_\kappa\}$ between strokes. The joint distribution can be written as

$$P(\psi) = P(\kappa) \prod_{i=1}^{\kappa} P(S_i) P(R_i | S_1, ..., S_{i-1}). \tag{2}$$

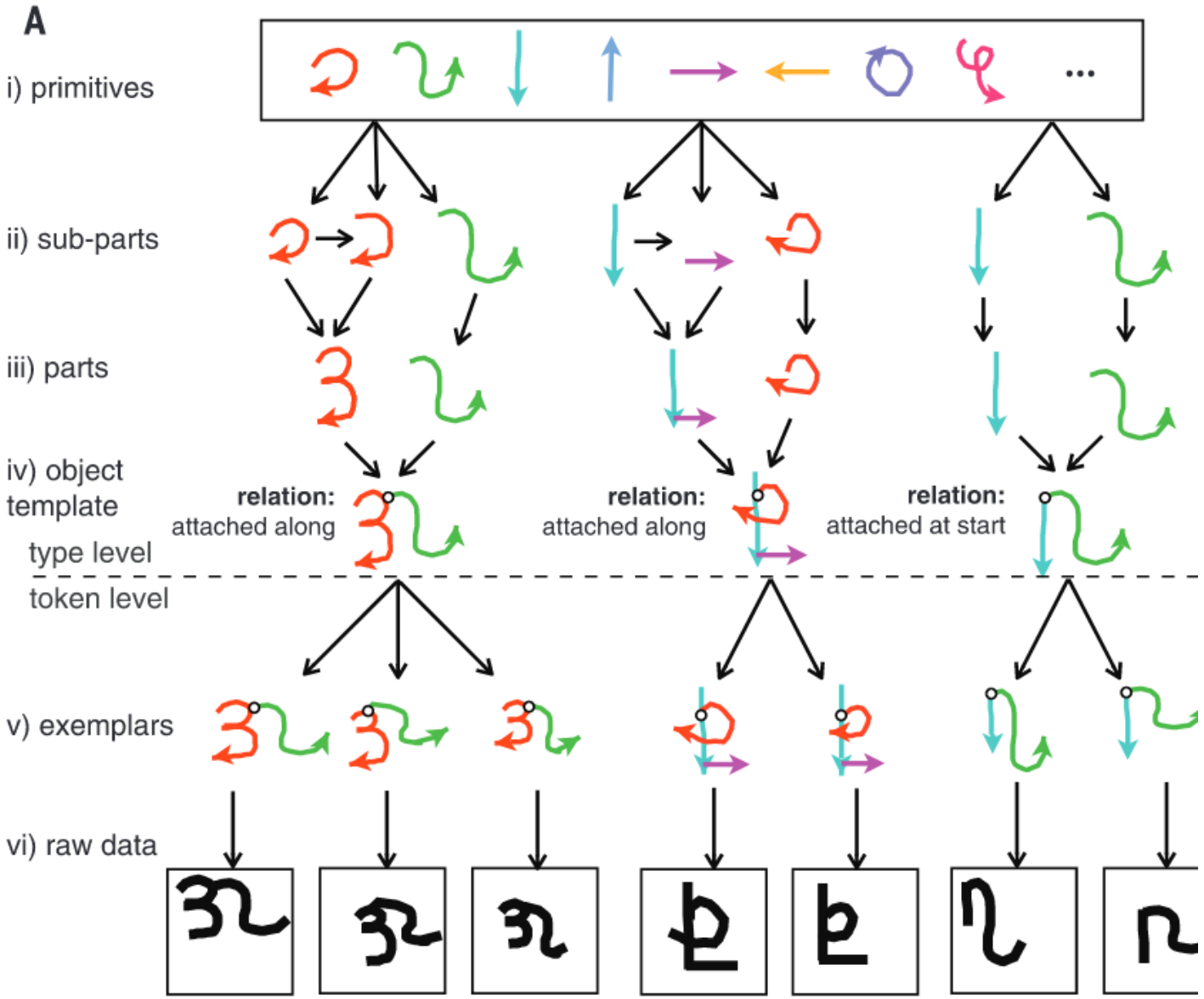
Relations. The spatial relation R_i specifies how the beginning of stroke S_i connects to the previous strokes $\{S_1, ..., S_{i-1}\}$. The distribution $P(R_i | S_1, ..., S_{i-1}) = P(R_i | z_1, ..., z_{i-1})$, since it only depends on the number of sub-strokes in each stroke. Relations can come in four types with probabilities θ_R, and each type has different sub-variables and dimensionalities:

- *Independent* relations, $R_i = \{J_i, L_i\}$, where the position of stroke i does not depend on previous strokes. The variable $J_i \in \mathbb{N}$ is drawn from $P(J_i)$, a multinomial over a 2D image grid that depends on index i (Figure 4c). Since the position $L_i \in \mathbb{R}^2$ has to be real-valued, $P(L_i | J_i)$ is then sampled uniformly at random from within the image cell J_i.
- *Start* or *End* relations, $R_i = \{u_i\}$, where stroke i starts at either the beginning or end of a previous stroke u_i, sampled uniformly at random from $u_i \in \{1, ..., i-1\}$.
- *Along* relations, $R_i = \{u_i, v_i, \tau_i\}$, where stroke i begins along previous stroke $u_i \in \{1, ..., i-1\}$ at sub-stroke $v_i \in \{1, ..., n_{u_i}\}$ at type-level spline coordinate $\tau_i \in \mathbb{R}$, each sampled uniformly at random.

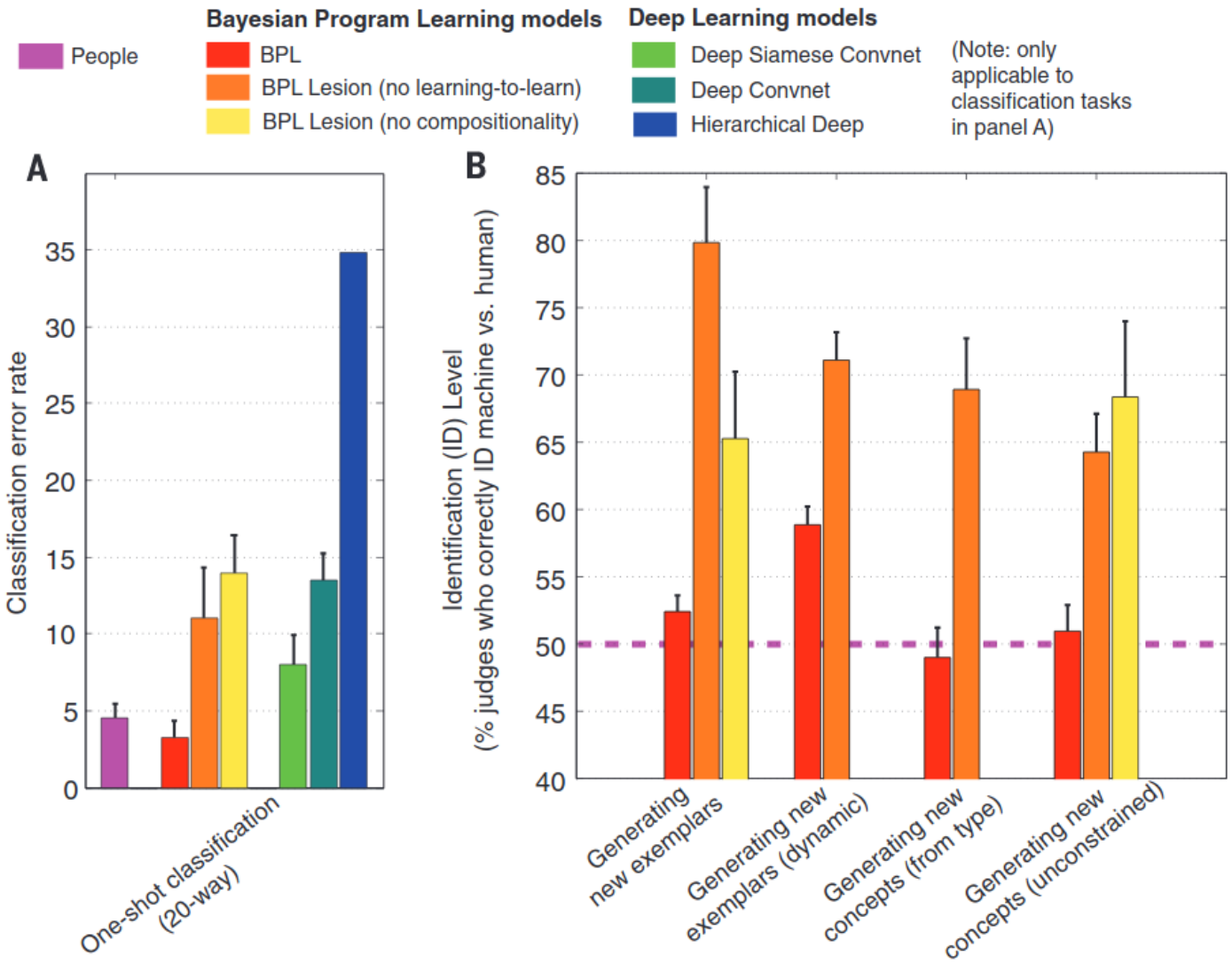
The token-level variables, $\theta^{(m)} = \{L^{(m)}, x^{(m)}, y^{(m)}, R^{(m)}, A^{(m)}, \sigma_b^{(m)}, \epsilon^{(m)}\}$, are distributed as

$$P(\theta^{(m)} | \psi) = P(L^{(m)} | \theta^{(m)}_{\setminus L^{(m)}}, \psi) \prod_i P(R_i^{(m)} | R_i) P(y_i^{(m)} | y_i) P(x_i^{(m)} | x_i) P(A^{(m)}, \sigma_b^{(m)}, \epsilon^{(m)})$$

Модель


A
i) primitives
ii) sub-parts
iii) parts
iv) object template
type level
token level
v) exemplars
vi) raw data
relation: attached along
relation: attached along
relation: attached at start

Результат


Bayesian Program Learning models
People
BPL
BPL Lesion (no learning-to-learn)
BPL Lesion (no compositionality)
Deep Learning models
Deep Siamese Convnet
Deep Convnet
Hierarchical Deep
(Note: only applicable to classification tasks in panel A)
A
Classification error rate
0
5
10
15
20
25
30
35
One-shot classification (20-way)
B
Identification (ID) Level (% judges who correctly ID machine vs. human)
40
45
50
55
60
65
70
75
80
85
Generating new exemplars
Generating new exemplars (dynamic)
Generating new concepts (from type)
Generating new concepts (unconstrained)

Идея подхода

Где нашли продолжение описанные концепции?

- Композиционность
- Причинность
- Обучение обучению

Идея подхода

Где нашли продолжение описанные концепции?

- Композиционность
- Причинность
  - Attention и Self-attention
  - Иерархические генеративные модели и смеси
- Обучение обучению

Работа с разными модальностями

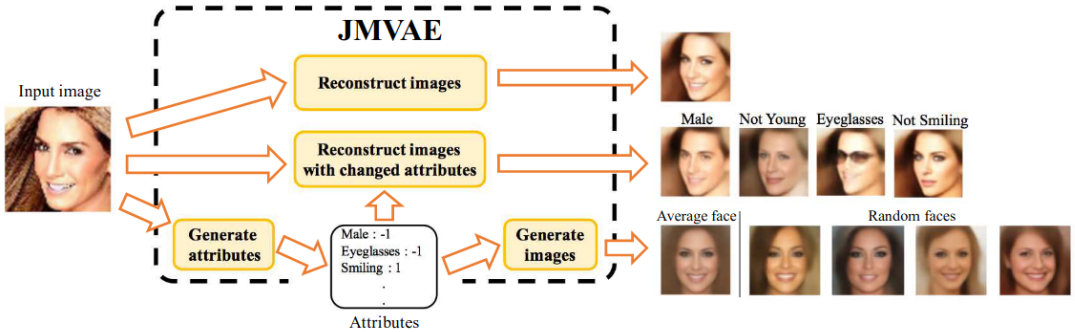


Figure 1: Suzuki et al., 2017

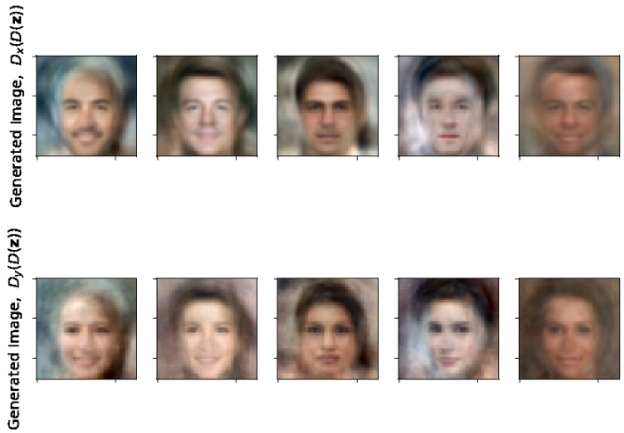


Figure 2: Kuznetsova et al., 2018

Метапараметры

Wikipedia

A parameter that controls the value of one or more others.

Определение

Метапараметрами $\boldsymbol{\lambda}$ модели назовем параметры оптимизации.

Чаще всего метапараметры назначаются экспертно и не подлежат оптимизации в ходе решения задачи выбора модели.
Что можно считать метапараметрами:

- параметры оператора оптимизации;
- параметры задачи оптимизации;
- структуру модели;
- функции активации слоев сети;
- вид априорного распределения и функции правдоподобия.

A neural network that embeds its own meta-levels

Предлагается разделить подмодели внутри модели сети по назначениям:

- “Normal” model: обучение и вывод.
- Evaluation model: оценка качества Q.
- Analyzing model: анализ параметров модели.
- Modifiyng model: модификация параметров.

Представлен градиентный алгоритм оптимизации нейронной сети.

L2L by gradient descent by gradient descent

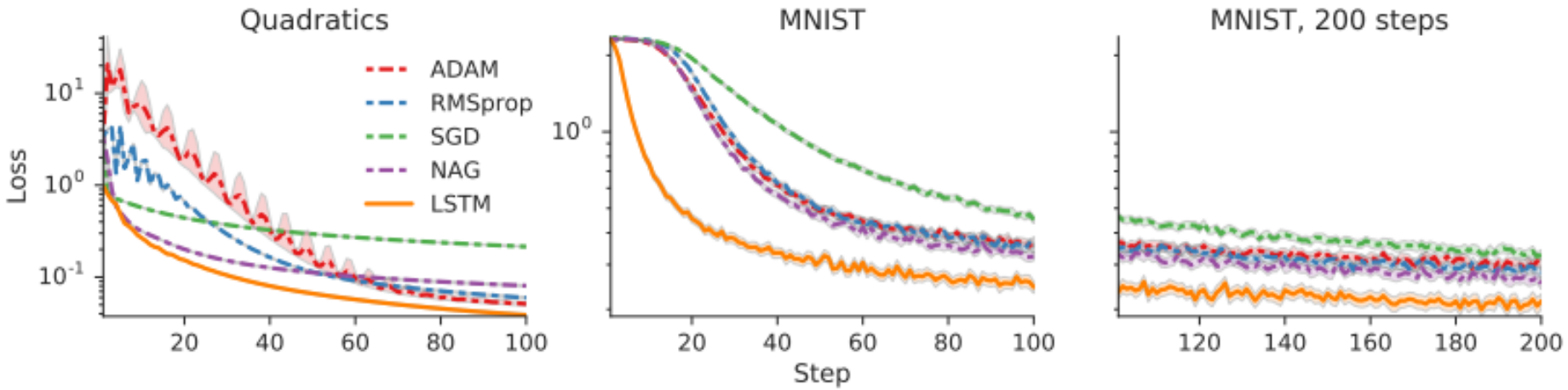
Идея: рассматривать результат применения градиентного спуска T как дифференцируемую функцию:

$$T(\theta) = \text{LSTM}(\theta).$$

Оптимизационная задача:

$$\sum_{t=t_0}^{t_\eta} L\left(T^t(\theta_{t_0})\right) \to \min.$$

LSTM имеет небольшое число параметров и делит параметры между всеми параметрами, подлежащими оптимизации.



Список источников

- Lake B. M., Salakhutdinov R., Tenenbaum J. B. Human-level concept learning through probabilistic program induction //Science. – 2015. – Т. 350. – №. 6266. – С. 1332-1338.
- Lake B. M. et al. Building machines that learn and think like people //Behavioral and brain sciences. – 2017. – Т. 40.
- Lake B. M., Salakhutdinov R. R., Tenenbaum J. One-shot learning by inverting a compositional causal process //Advances in neural information processing systems. – 2013. – С. 2526-2534.
- Suzuki M., Nakayama K., Matsuo Y. Joint multimodal learning with deep generative models //arXiv preprint arXiv:1611.01891. – 2016.
- Kuznetsova R., Bakhteev O., Ogaltsov A. Variational learning across domains with triplet information //arXiv preprint arXiv:1806.08672. – 2018.
- Schmidhuber J. A neural network that embeds its own meta-levels //IEEE International Conference on Neural Networks. – IEEE, 1993. – С. 407-412.
- Andrychowicz M. et al. Learning to learn by gradient descent by gradient descent //Advances in neural information processing systems. – 2016. – С. 3981-3989.