Выбор структуры модели глубокого обучения

Бахтеев Олег

МФТИ

20.11.2019

Резюме прошлых семинаров

Заданы:

- Вариационное распределение $q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \theta_{\mathbf{w}})$ с параметрами θ;
- Априорное распределение $p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})$ с параметрами $\mathbf{h}$;
- Функция потерь L и функция валидации Q.

Требуется: предложить метод выбора структуры модели $\mathbf{\Gamma}$.

Вопросы:

- Как задать структуру модели?
- Как провести ее выбор?
- Какова вероятностная интерпретация структуры?

Automatic relevance determination

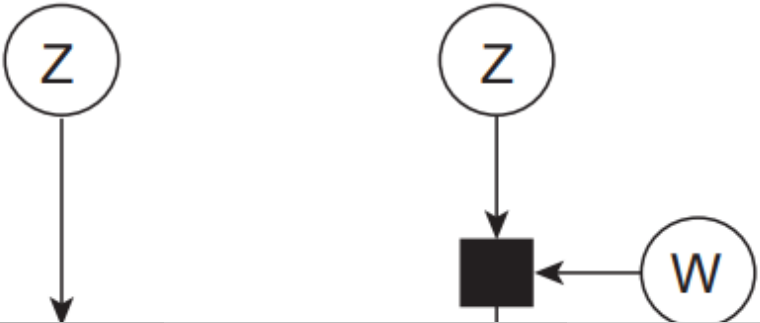
Пример: вариационный автокодировщик + ARD

VAE:

$$L = \int_{\mathbf{z}} p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}.$$

VAE + ARD:

$$L = \iint_{\mathbf{z},\gamma} p(\mathbf{x}|\mathbf{z}\cdot\gamma)p(\mathbf{z})p(\gamma)d\mathbf{z}d\gamma.$$



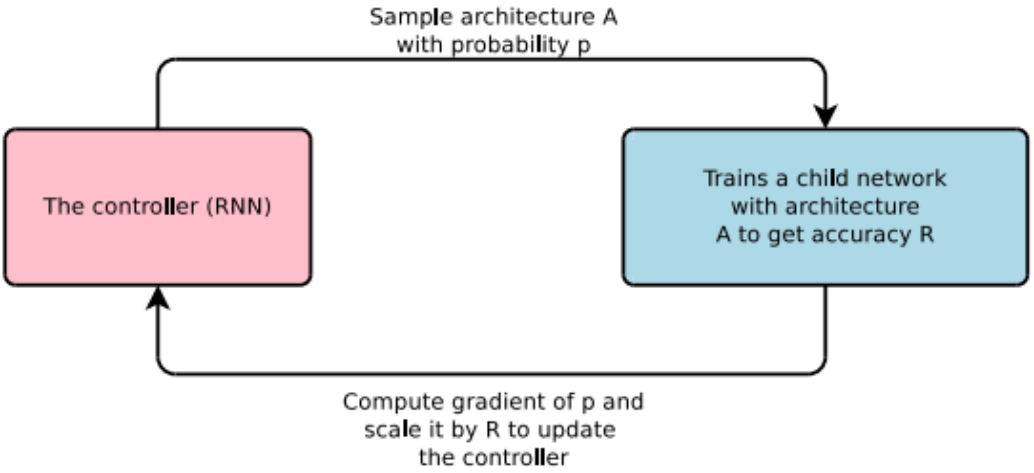
SpikeAndSlab

Гауссовый процесс для выбора структуры модели

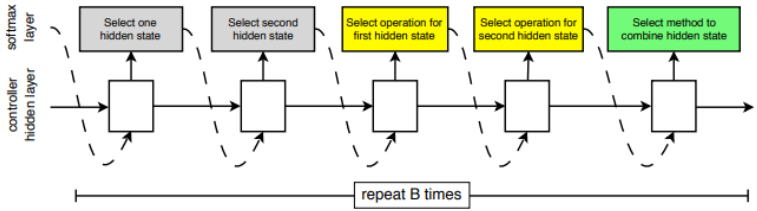
Индийский буфет, еще один пример

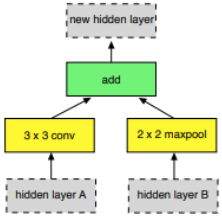
AdaNet

Neural Architecture Search



Neural Architecture Search





Neural Architecture Search: результаты

Model	image size	# parameters	Mult-Adds	Top 1 Acc. (%)	Top 5 Acc. (%)
Inception V2 [29]	224×224	11.2 M	1.94 B	74.8	92.2
NASNet-A (5 @ 1538)	**299×299**	**10.9 M**	**2.35 B**	**78.6**	**94.2**
Inception V3 [59]	299×299	23.8 M	5.72 B	78.0	93.9
Xception [9]	299×299	22.8 M	8.38 B	79.0	94.5
Inception ResNet V2 [57]	299×299	55.8 M	13.2 B	80.4	95.3
NASNet-A (7 @ 1920)	**299×299**	**22.6 M**	**4.93 B**	**80.8**	**95.3**
ResNeXt-101 (64 x 4d) [67]	320×320	83.6 M	31.5 B	80.9	95.6
PolyNet [68]	331×331	92 M	34.7 B	81.3	95.8
DPN-131 [8]	320×320	79.5 M	32.0 B	81.5	95.8
SENet [25]	**320×320**	**145.8 M**	**42.3 B**	**82.7**	**96.2**
NASNet-A (6 @ 4032)	**331×331**	**88.9 M**	**23.8 B**	**82.7**	**96.2**

Zoph et al., 2017. Сложность моделей отличается почти в два раза при одинаковом качестве.

Neural Architecture Search: постановка задачи

TODO

DARTS

DARTS

Графовое представление модели глубокого обучения

Заданы:

1. ациклический граф (V, E);
2. для каждого ребра $(j, k) \in E$: вектор базовых дифференцируемых функций $\mathbf{g}^{j,k} = [\mathbf{g}_0^{j,k}, \ldots, \mathbf{g}_{K^{j,k}}^{j,k}]$ мощности $K^{j,k}$;
3. для каждой вершины $v \in V$: дифференцируемая функция агрегации $\mathbf{agg}_v$.
4. Функция $\mathbf{f} = \mathbf{f}_{|V|-1}$, задаваемая по правилу

$$\mathbf{f}_v(\mathbf{w}, \mathbf{x}) = \mathbf{agg}_v \left(\{ \langle \boldsymbol{\gamma}^{j,k}, \mathbf{g}^{j,k} \rangle \circ \mathbf{f}_j(\mathbf{x}) | j \in \mathrm{Adj}(v_k) \} \right), v \in \{1, \ldots, |V| - 1\}, \quad \mathbf{f}_0(\mathbf{x}) = \mathbf{x} \tag{1}$$

и являющаяся функцией из признакового пространства $\mathbb{X}$ в пространство меток $\mathbb{Y}$ при значениях векторов, $\boldsymbol{\gamma}^{j,k} \in [0,1]^{K^{j,k}}$.

Определение

Граф (V, E) со множестом векторов базовых функций $\{\mathbf{g}^{j,k}, (j,k) \in E\}$ и функций агрегаций $\{\mathbf{agg}_v, v \in V\}$ назовем *параметрическим семейством моделей* $\mathfrak{F}$.

Утверждение

Для любого значения $\boldsymbol{\gamma}^{j,k} \in [0,1]^{K^{j,k}}$ функция $\mathbf{f} \in \mathfrak{F}$ является моделью.

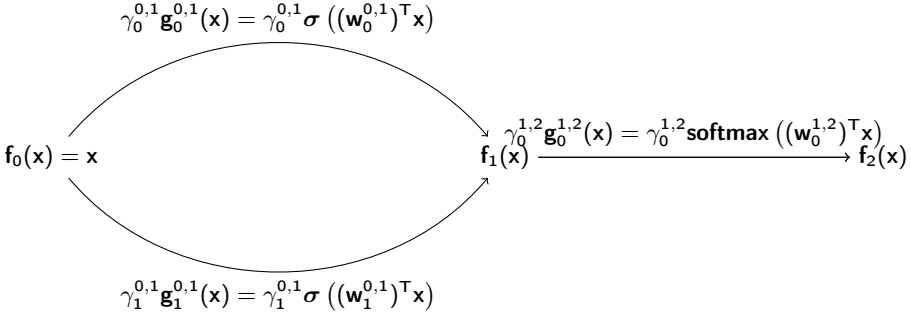
Выбор структуры: двуслойная нейросеть

Модель $\mathbf{f}$ задана **структурой** $\mathbf{\Gamma} = [\gamma^{0,1}, \gamma^{1,2}]$.

$$\text{Модель: } \mathbf{f}(\mathbf{x}) = \mathbf{softmax}\left((\mathbf{w}_0^{1,2})^{\mathsf{T}}\mathbf{f}_1(\mathbf{x})\right), \quad \mathbf{f}(\mathbf{x}) : \mathbb{R}^n \to [0,1]^{|\mathbb{Y}|}, \quad \mathbf{x} \in \mathbb{R}^n.$$

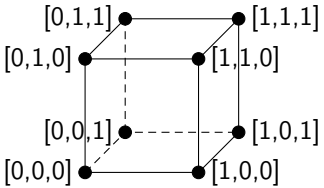
$$\mathbf{f}_1(\mathbf{x}) = \gamma_0^{0,1}\mathbf{g}_0^{0,1}(\mathbf{x}) + \gamma_1^{0,1}\mathbf{g}_1^{0,1}(\mathbf{x}),$$

где $\mathbf{w} = [\mathbf{w}_0^{0,1}, \mathbf{w}_1^{0,1}, \mathbf{w}_0^{1,2}]^{\mathsf{T}}$ — матрицы параметров, $\{\mathbf{g}_{0,1}^0, \mathbf{g}_{0,1}^1, \mathbf{g}_{1,2}^0\}$ — обобщенно-линейные функции скрытых слоев нейросети.

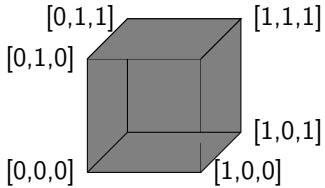


Ограничения на структурные параметры

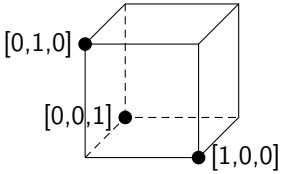
Примеры ограничений для одного структурного параметра $\gamma, |\gamma| = 3$.



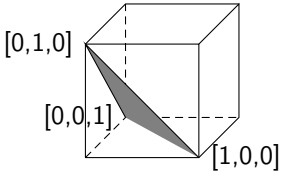
На вершинах куба



Внутри куба



На вершинах симплекса



Внутри симплекса

Репараметризация

Определение

Случайную величину ψ с распределением q с параметрами $\boldsymbol{\theta}_\psi$ назовем репараметризованной через случайную величину ε, чье распределение не зависит от параметров $\boldsymbol{\theta}_\psi$, если:

$$\psi = g(\varepsilon, \boldsymbol{\theta}_\psi)$$

где g — некоторая непрерывная функция.

Пример

$$\mathsf{E}_{q_\mathbf{w}(\mathbf{w}|\boldsymbol{\Gamma},\boldsymbol{\theta}_\mathbf{w})} \log\ p(\mathbf{y}|\mathbf{X},\mathbf{w},\boldsymbol{\Gamma}) = \int_\mathbf{w} \log\ p(\mathbf{y}|\mathbf{X},\mathbf{w},\boldsymbol{\Gamma}) q_\mathbf{w}(\mathbf{w}|\boldsymbol{\Gamma},\boldsymbol{\theta}_\mathbf{w}) d\mathbf{w}.$$

Продифференцируем по параметрам $\boldsymbol{\theta}_\mathbf{w}$:

$$\nabla_{\boldsymbol{\theta}_\mathbf{w}} \mathsf{E}_{q_\mathbf{w}(\mathbf{w}|\boldsymbol{\Gamma},\boldsymbol{\theta}_\mathbf{w})} \log p(\mathbf{y}|\mathbf{X},\mathbf{w},\boldsymbol{\Gamma}) = \int_\mathbf{w} \log p(\mathbf{y}|\mathbf{X},\mathbf{w},\boldsymbol{\Gamma}) \nabla_{\boldsymbol{\theta}_\mathbf{w}} q_\mathbf{w}(\mathbf{w}|\boldsymbol{\Gamma},\boldsymbol{\theta}_\mathbf{w}) d\mathbf{w}.$$

Пусть возможна репараметризация: $\mathbf{w} = \mathbf{g}(\varepsilon, \boldsymbol{\theta}_\mathbf{w})$. Тогда:

$$\nabla_{\boldsymbol{\theta}_\mathbf{w}} \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\theta)} \log p(\mathbf{y}|\mathbf{X},\mathbf{w},\boldsymbol{\Gamma}) = \nabla_{\boldsymbol{\theta}_\mathbf{w}} \mathsf{E}_\varepsilon \log p(\mathbf{y}|\mathbf{X},\mathbf{g}(\varepsilon),\boldsymbol{\Gamma}) =$$

$$= \int_\varepsilon \nabla_{\boldsymbol{\theta}_\mathbf{w}} \log p(\mathbf{y}|\mathbf{X},\mathbf{g}(\varepsilon),\boldsymbol{\Gamma}) p(\varepsilon) d\varepsilon = \mathsf{E}_\varepsilon \nabla_{\boldsymbol{\theta}_\mathbf{w}} \log p(\mathbf{y}|\mathbf{X},\mathbf{g}(\varepsilon),\boldsymbol{\Gamma}).$$

Reparametrization

Logit-Normal

Gumbel-Softmax

Proposed Method

Proposed Method

Properties

Examples

Используемые материалы