

# Оптимизация структур сетей глубокого обучения

Смердов Антон Николаевич

Научный руководитель  
д.ф-м.н. В.В. Стрижов

Московский физико-технический институт  
Факультет управления и прикладной математики  
Кафедра «Интеллектуальные системы»

МФТИ, 13 июня 2018

# Цель исследования

## Проблема




Большое количество параметров в моделях глубокого обучения влечёт сложность оптимизации параметров и переобучение.

## Цель работы

Построение модели глубокого обучения оптимальной структуры.

## Метод решения

Рассматривается вариационный байесовский подход с различными предположениями о распределении вектора параметров модели. Наименее важные параметры удаляются из сети.

-  *Sanborn A., Skryzalin J.* Deep Learning for Semantic Similarity // CS224d: Deep Learning for Natural Language Processing — Stanford, CA, USA: Stanford University, 2015. Unpublished.
-  *Graves A.* Practical variational inference for neural networks // Advances in Neural Information Processing Systems 24 (NIPS 2011). P. 2348–2356.
-  *О.Ю. Бахтеев, В.В. Стрижов.* Выбор моделей глубокого обучения субоптимальной сложности // Автоматика и телемеханика, 2018.

# Постановка задачи

Дано  $\mathcal{D} = \{(\mathbf{a}_i, \mathbf{b}_i, y_i)\}$ ,  $i = \overline{1, N}$ ,  $\mathbf{a}_i, \mathbf{b}_i$  — последовательности векторов слов,  $y_i \in \{0 \dots 5\}$  — экспертная оценка их близости. Для модели  $\mathbf{f} \in \mathfrak{F}$  и вектора параметров  $\mathbf{w}$  определим логарифмическую функцию правдоподобия выборки:

$$L_{\mathcal{D}}(\mathbf{y}, \mathcal{D}, \mathbf{f}, \mathbf{w}) = \log p(\mathbf{y}|\mathcal{D}, \mathbf{w}, \mathbf{f}) = \sum_{(\mathbf{a}, \mathbf{b}, y) \in \mathcal{D}} \log p(y|\mathbf{a}, \mathbf{b}, \mathbf{w}, \mathbf{f})$$

Оптимальная модель  $\mathbf{f}$  находится максимизацией логарифма правдоподобия модели  $L_{\mathbf{f}}(\mathbf{y}, \mathcal{D}, \mathbf{f})$ :

$$L_{\mathbf{f}}(\mathbf{y}, \mathcal{D}, \mathbf{f}) = \log p(\mathbf{y}|\mathcal{D}, \mathbf{f}) = \log \int_{\mathbf{w}} p(\mathbf{y}|\mathcal{D}, \mathbf{w}) p(\mathbf{w}|\mathbf{f}) d\mathbf{w}$$

Априорное и апостериорное распределения параметров будем считать нормальными:

$$p(\mathbf{w}|\mathbf{f}) \sim N(\mu_1, \mathbf{A}_1^{-1}), \quad p(\mathbf{w}|\mathbf{y}, \mathcal{D}, \mathbf{f}) \sim N(\mu_2, \mathbf{A}_2^{-1})$$

# Вариационная нижняя оценка

Рассмотрим вариационную нижнюю оценку  $L_f(y, \mathcal{D}, f)$ , полученную из неравенства Йенсена:

$$\begin{aligned} L_f(y, \mathcal{D}, f) &= \int_{\mathbf{w}} \rho_2(\mathbf{w}) \log p(y|\mathcal{D}, f) d\mathbf{w} \geq \\ &\geq -D_{\text{KL}}(N(\mu_2, \mathbf{A}_2^{-1}) || N(\mu_1, \mathbf{A}_1^{-1})) + \int_{\mathbf{w}} \rho_2(\mathbf{w}) \log p(y|\mathcal{D}, f, \mathbf{w}) d\mathbf{w} \end{aligned}$$

Обозначим за  $L_{\mathbf{w}}(\mathcal{D}, f)$  и  $L_E(y, \mathcal{D}, f, \mathbf{w})$  первое и второе слагаемые со знаком минус:

$$L_{\mathbf{w}}(\mathcal{D}, f) = D_{\text{KL}}(N(\mu_2, \mathbf{A}_2^{-1}) || N(\mu_1, \mathbf{A}_1^{-1}))$$

$$L_E(y, \mathcal{D}, f, \mathbf{w}) = -E_{\mathbf{w} \sim N(\mu_2, \mathbf{A}_2^{-1})} L_{\mathcal{D}}(y, \mathcal{D}, f, \mathbf{w})$$

Искомая модель  $f$  минимизирует суммарную функцию потерь

$$f = \operatorname{argmin}_{f \in \mathcal{F}} L(y, \mathcal{D}, f, \mathbf{w})$$

$$L(y, \mathcal{D}, f, \mathbf{w}) = L_E(y, \mathcal{D}, f, \mathbf{w}) + L_{\mathbf{w}}(\mathcal{D}, f, \mathbf{w})$$

Для оценки  $L_E(\mathbf{y}, \mathcal{D}, \mathbf{f}, \mathbf{w})$  воспользуемся интегрированием Монте-Карло:

$$L_E(\mathbf{y}, \mathcal{D}, \mathbf{f}, \mathbf{w}) \approx \frac{1}{S} \sum_{k=1}^S L_{\mathcal{D}}(\mathbf{y}, \mathcal{D}, \mathbf{f}, \mathbf{w}_k)$$

Сложность модели  $L_{\mathbf{w}}(\mathcal{D}, \mathbf{f}, \mathbf{w})$  может быть найдена аналитически:

$$L_{\mathbf{w}}(\mathcal{D}, \mathbf{f}, \mathbf{w}) = D_{\text{KL}}(N(\mu_1, \mathbf{A}_1^{-1}) || N(\mu_2, \mathbf{A}_2^{-1})) = \frac{1}{2} \left( \log \frac{|\mathbf{A}_2^{-1}|}{|\mathbf{A}_1^{-1}|} - W + \right. \\ \left. + \text{tr}(\mathbf{A}_2 \mathbf{A}_1^{-1}) + (\mu_1 - \mu_2)^T \mathbf{A}_2 (\mu_1 - \mu_2) \right)$$

# Обновление параметров распределений

Скалярные априорная дисперсия и вектор средних, апостериорная матрица ковариаций диагональна:

$$\mathbf{A}_1^{-1} = \sigma \mathbf{I}, \quad \mathbf{A}_2^{-1} = \text{diag}(\sigma), \quad \mu_1 = \mu, \quad \mu_2 = \mathbf{m}.$$

$$\text{Тогда } L_{\mathbf{w}}(\mathfrak{D}, \mathbf{f}, \mathbf{w}) = \sum_{i=1}^d \left( \log \frac{\sigma}{\sigma_i} + \frac{(\mu - m_i)^2 + \sigma_i^2 + \sigma^2}{2\sigma^2} \right)$$

$$\frac{\partial}{\partial \mu} D_{\text{KL}} = \sum_{i=1}^W \frac{\mu - m_i}{\sigma^2} = 0 \Rightarrow \hat{\mu} = \frac{1}{W} \sum_{i=1}^W m_i.$$

$$\frac{\partial}{\partial \sigma^2} D_{\text{KL}} = \sum_{i=1}^W \frac{1}{2\sigma^2} - \frac{(\mu - m_i)^2 + \sigma_i^2}{2\sigma^4} = 0 \Rightarrow \hat{\sigma}^2 = \frac{1}{W} \sum_{i=1}^W (\mu - m_i)^2 + \sigma_i^2.$$

В частности, если апостериорная матрица ковариаций скалярна, т.е.  $\mathbf{A}_2^{-1} = \beta \mathbf{I}$ :

$$\frac{\partial}{\partial \sigma^2} D_{\text{KL}} = \sum_{i=1}^W \frac{1}{2\sigma^2} - \frac{(\mu - m_i)^2 + \beta^2}{2\sigma^4} = 0 \Rightarrow \hat{\sigma}^2 = \frac{1}{W} \sum_{i=1}^W (\mu - m_i)^2 + \beta^2.$$

Оптимизация параметров сводится к следующему алгоритму:

- 1 Инициализация  $\sigma = \mathbf{1}$ ,  $\mathbf{m} = \mathbf{0}$ ,  $\mu = 0$ ,  $\sigma^2 = 1$
- 2 **Повторять**:
- 3 Сделать градиентный шаг  $\sigma := \sigma - \eta \nabla \sigma$ ,  $\mathbf{m} := \mathbf{m} - \eta \nabla \mathbf{m}$
- 4 Обновить параметры априорного распределения  
 $\mu := \hat{\mu}$ ,  $\sigma^2 := \hat{\sigma}^2$ .
- 5 **Пока** значение  $L$  не стабилизируется



## Цели эксперимента

Проверить работоспособность метода. Путём удаления наименее важных весов найти оптимальную структуру сети в задачах поиска парафраз.

## Данные

Вычислительный эксперимент проводился на выборке SemEval 2015. Тренировочная, валидационная и тестовая выборки составили 70%, 15% и 15% соответственно.

Для решения задачи использовалась рекуррентная нейронная сеть с одним скрытым слоем. Векторизация слов проводилась методом GloVe.

Вектор значений скрытого слоя обновляется как:

$$\mathbf{h}_i = \tanh(\mathbf{x}_i^T \mathbf{W} + \mathbf{h}_{i-1}^T \mathbf{U} + \mathbf{b}),$$

где  $\mathbf{x}_i \in R^m$  – входной вектор,  $\mathbf{h}_i \in R^n$ ,  $\mathbf{W} \in R^{n \times m}$ ,  $\mathbf{U} \in R^{n \times n}$ ,  $\mathbf{b} \in R^n$ .

В качестве функционала качества была выбрана F1-мера:

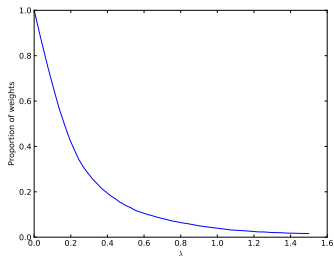
$$F_1 = 2 \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{precision}}$$

Результаты вычислительного эксперимента:

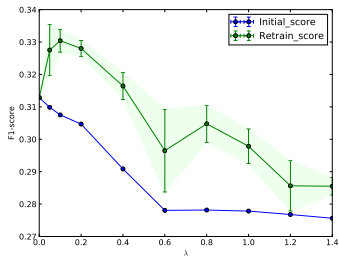
Classifier	F1-measure
Logistic Regression	0.286
SVC	0.290
DecisionTreeClassifier	0.316
KNeighborsClassifier	0.322
RNN	0.362
RNN+variational, I, I	0.311
RNN+variational, D, I	0.330

# Результаты вычислительного эксперимента

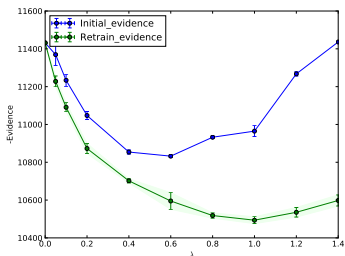
Чем больше плотность вероятности в нуле  $\rho(0) \sim \exp(-\frac{\mu_i^2}{2\sigma_i^2})$  — тем меньше важность параметра.  
 $\lambda = |\frac{\mu_i}{\sigma_i}|$ , тогда  $\rho(0) \sim \exp(-\frac{\lambda^2}{2})$ .  
Веса с большим значением  $\lambda$  имеют высокую плотность в нуле и могут быть удалены.



Количество параметров от  $\lambda$ .



Зависимость F1-меры от  $\lambda$ .



Зависимость правдоподобия от  $\lambda$ .

- Задача выбора оптимальной модели поставлена формально.
- Минимизация правдоподобия модели не приводит к переобучению.
- Алгоритм удаления параметров позволяет упростить структуру модели без существенных потерь качества.