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 иследования

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.

Литература

- *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 Unpublished.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 $\mathfrak{D} = \{(\mathbf{a}_i, \mathbf{b}_i, y_i)\}, i = \overline{1, N}$, $\mathbf{a}_i, \mathbf{b}_i$ — 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:

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

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

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

Введём априорное и апостериорные распределения вектора параметров:

$$p(\mathbf{w}|\mathbf{f}) \sim N(\boldsymbol{\mu}_1, \mathbf{A}_1^{-1}),\ \ p(\mathbf{w}|\mathbf{y}, \mathfrak{D}, \mathbf{f}) \sim N(\boldsymbol{\mu}_2, \mathbf{A}_2^{-1})$$

Постановка задачи

Рассмотрим вариационную нижнюю оценку $L_{\mathbf{f}}(\mathbf{y}, \mathfrak{D}, \mathbf{f})$, полученную из неравенства Йенсена:

$$L_{\mathbf{f}}(\mathbf{y}, \mathfrak{D}, \mathbf{f}) = \int_{\mathbf{w}} \rho_2(\mathbf{w}) \log p(\mathbf{y}|\mathfrak{D}, \mathbf{f}) d\mathbf{w} \geq$$
$$\geq -D_{\mathrm{KL}}(N(\boldsymbol{\mu}_2, \mathbf{A}_2^{-1}) || N(\boldsymbol{\mu}_1, \mathbf{A}_1^{-1})) +$$
$$+ \int_{\mathbf{w}} \rho_2(\mathbf{w}) \log p(\mathbf{y}|\mathfrak{D}, \mathbf{f}, \mathbf{w}) d\mathbf{w} = -L(\mathbf{y}, \mathfrak{D}, \mathbf{f}, \mathbf{w})$$

Первое слагаемое назовём сложностью модели $L_{\mathbf{w}}(\mathfrak{D}, \mathbf{f})$:

$$L_{\mathbf{w}}(\mathfrak{D}, \mathbf{f}) = D_{\mathrm{KL}}(N(\boldsymbol{\mu}_2, \mathbf{A}_2^{-1}) || N(\boldsymbol{\mu}_1, \mathbf{A}_1^{-1}))$$

Второе слагаемое формулы является матожиданием правдоподобия выборки:

$$L_E(\mathbf{y}, \mathfrak{D}, \mathbf{f}, \mathbf{w}) = \mathsf{E}_{\mathbf{w} \sim N(\boldsymbol{\mu}_2, \mathbf{A}_2^{-1})} L_{\mathfrak{D}}(\mathbf{y}, \mathfrak{D}, \mathbf{f}, \mathbf{w})$$

Постановка задачи

Искомая модель $\mathbf{f}$ минимизирует суммарную функцию потерь

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

где

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

$L_E(\mathbf{y}, \mathfrak{D}, \mathbf{f}, \mathbf{w})$ — матожидание правдоподобия выборки,
$L_{\mathbf{w}}(\mathfrak{D}, \mathbf{f}, \mathbf{w})$ — сложность модели.

Предлагаемое решение

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

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

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

$$L_{\mathbf{w}}(\mathfrak{D}, \mathbf{f}, \mathbf{w}) = D_{\mathrm{KL}}(N(\boldsymbol{\mu}_1, \mathbf{A}_1^{-1}) || N(\boldsymbol{\mu}_2, \mathbf{A}_2^{-1})) = \frac{1}{2}\Big(\log \frac{|\mathbf{A}_2^{-1}|}{|\mathbf{A}_1^{-1}|} - W +$$

$$+ \mathrm{tr}(\mathbf{A}_2 \mathbf{A}_1^{-1}) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{A}_2 (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\Big)$$

Предлагаемое решение

Скалярные дисперсии и априорный вектор средних:

$\mathbf{A}_1^{-1} = \sigma\mathbf{I},\ \mathbf{A}_2^{-1} = \beta\mathbf{I},\ \boldsymbol{\mu}_1 = \mu,\ \boldsymbol{\mu}_2 = \mathbf{m}.$

Находим $L_{\mathbf{w}}(\mathfrak{D}, \mathbf{f}, \mathbf{w}) = \sum\limits_{i=1}^{W}(\log\frac{\sigma}{\beta} + \frac{(\mu - m_i)^2 + \beta^2 + \sigma^2}{2\sigma^2})$

Необходимые условия экстремума:

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

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

Предлагаемое решение

Скалярные априорная дисперсия и вектор средних, апостериорная матрица диагональна:
$\mathbf{A}_1^{-1} = \sigma\mathbf{I},\ \mathbf{A}_2^{-1} = \mathrm{diag}(\boldsymbol{\sigma}),\ \boldsymbol{\mu}_1 = \mu,\ \boldsymbol{\mu}_2 = \mathbf{m}.$

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

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

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

Алгоритм

Оптимизация параметров сводится к следующему алгоритму:

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

Цели и данные эксперимента

Цели эксперимента

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

Данные

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

Вычислительный эксперимент

Для решения задачи использовалась рекуррентная нейронная сеть с одним скрытым слоем. Векторизация слов проводилась методом GloVe.
Вектор значений скрытого слоя обновляется как:

$$\mathbf{h_i} = \tanh(\mathbf{x_i^T W} + \mathbf{h_{i-1}^T U} + \mathbf{b}),$$

где $\mathbf{x_i} \in R^m$ – входной вектор, $\mathbf{h_i} \in R^n$, $\mathbf{W} \in R^{n*m}$, $\mathbf{U} \in R^{n*n}$, $\mathbf{b} \in R^n$.

Вычислительный эксперимент

В качестве метрики была выбрана F1-мера:

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

Таблица: Результаты вычислительного эксперимента

Classificator	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

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

Результаты вычислительного эксперимента

Результаты удаления параметров для диагональной матрицы ковариаций представлены в таблице, где p — доля оставшихся параметров.

λ	p	Initial_L	Retrain_L	Initial_score	Retrain_score
0	1.000	11431	11431	0.313	0.313
0.05	0.833	11306	11219	0.310	0.336
0.1	0.673	11198	11102	0.308	0.327
0.2	0.420	11071	10851	0.305	0.331
0.4	0.195	10841	10710	0.291	0.321
0.6	0.106	10827	10545	0.278	0.311
0.8	0.064	10925	10502	0.278	0.311
1.0	0.040	10996	10500	0.278	0.299
1.2	0.025	11256	10506	0.277	0.277
1.4	0.018	11439	10569	0.276	0.286

Результаты вычислительного эксперимента

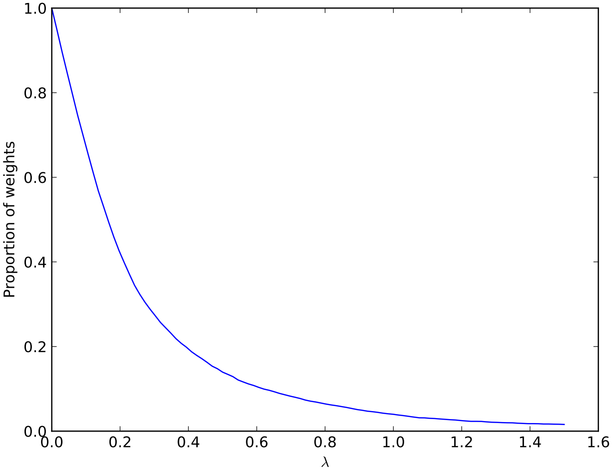

Рис.: Количество оставшихся параметров в зависимости от λ.

Результаты вычислительного эксперимента

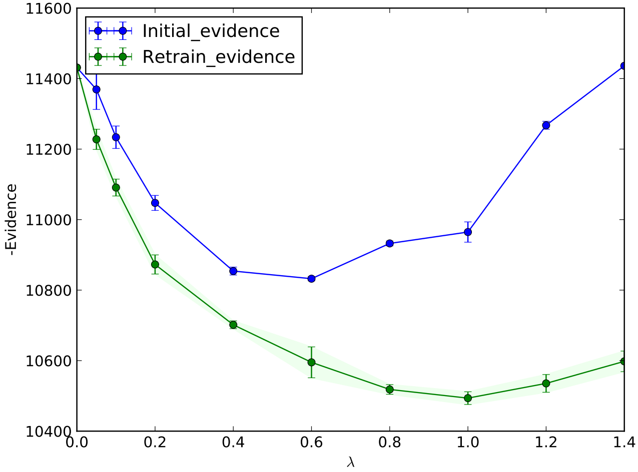

Рис.: Зависимость правдоподобия от λ.

Результаты вычислительного эксперимента

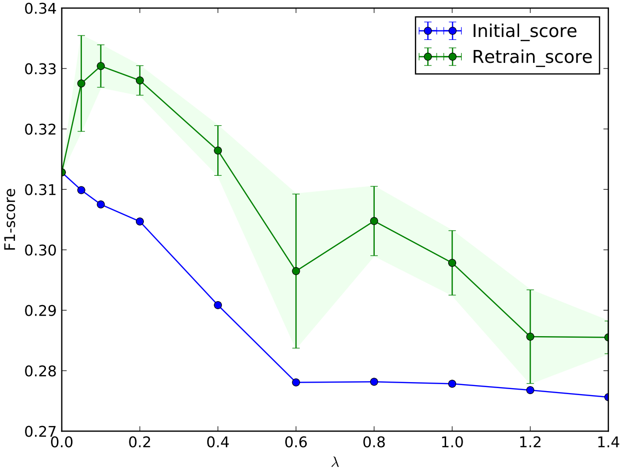

Рис.: Зависимость F1-меры от λ.

Заключение

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