

Выбор оптимальной модели рекуррентной сети в задачах поиска парафразы

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, 28 июня 2018

Цель работы




Оптимизация структур моделей глубокого обучения и выбор модели с наибольшим правдоподобием.

Проблема

Избыточное число параметров в моделях глубокого обучения влечёт переобучение и сложность оптимизации параметров.

Метод решения

Используется вариационный байесовский подход с предположением о нормальном распределении вектора параметров модели. Для оптимизации структуры предлагается удалять параметры с наибольшей плотностью распределения в нуле.

-  *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 Unpublished.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.

Задача нахождения оптимальной модели

Дано $\mathcal{D} = \{(\mathbf{a}_i, \mathbf{b}_i, y_i)\}$, $i = \overline{1, N}$, $\mathbf{a}_i, \mathbf{b}_i$ — последовательности векторов слов, $y_i \in \mathbb{Y}$ — экспертная оценка их близости. Оптимальная модель \mathbf{f} находится максимизацией логарифма правдоподобия модели:

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

Априорное и апостериорное распределения параметров будем считать нормальными:

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

Решение предлагается искать в классе \mathfrak{F} рекуррентных нейронных сетей с одним скрытым слоем.

Вектор значений скрытого слоя:

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

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

Вариационная нижняя оценка $L_f(\mathbf{y}, \mathcal{D}, \mathbf{f})$

Из неравенства Йенсена:

$$\begin{aligned} L_f(\mathbf{y}, \mathcal{D}, \mathbf{f}) &= \int_{\mathbf{w}} \rho_2(\mathbf{w}) \log p(\mathbf{y}|\mathcal{D}, \mathbf{f}) d\mathbf{w} \geq \\ &\geq \underbrace{-D_{\text{KL}}(\mathcal{N}(\boldsymbol{\mu}_2, \mathbf{A}_2^{-1}) \parallel \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{A}_1^{-1}))}_{L_w(\mathcal{D}, \mathbf{f})} + \underbrace{\int_{\mathbf{w}} \rho_2(\mathbf{w}) \log p(\mathbf{y}|\mathcal{D}, \mathbf{f}, \mathbf{w}) d\mathbf{w}}_{-L_E(\mathbf{y}, \mathcal{D}, \mathbf{f}, \mathbf{w})}. \end{aligned}$$

Второе слагаемое является матожиданием правдоподобия выборки $L_{\mathcal{D}}(\mathbf{y}, \mathcal{D}, \mathbf{f}, \mathbf{w}) = \sum_{(\mathbf{a}, \mathbf{b}, y) \in \mathcal{D}} \log p(y|\mathbf{a}, \mathbf{b}, \mathbf{w}, \mathbf{f})$:

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

Оптимизируемый функционал записывается в виде

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

оптимальная модель находится из выражения

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

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

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

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

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

Скалярная априорная дисперсия и вектор средних, апостериорная матрица ковариаций **диагональна**:

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

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

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

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

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:

- 1 Инициализация $\mathbf{m} = \mathbf{0}$, $\sigma = \mathbf{1}$, $\mu = 0$, $\sigma^2 = 1$
- 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

Проверка метода для задачи поиска парафраз

Цели вычислительного эксперимента

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

Данные

Вычислительный эксперимент проводился на выборке пар предложений разной степени схожести SemEval 2015. Тренировочная, валидационная и тестовая выборки составили 70%, 15% и 15% соответственно.

Векторизация слов для использования алгоритмами проводилась методом GloVe.

Функционалом качества была выбрана F1-мера:

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

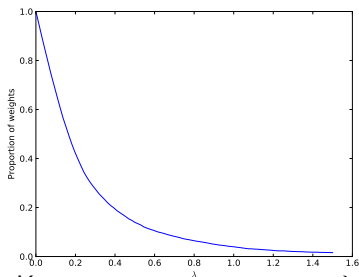
Результаты вычислительного эксперимента:

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

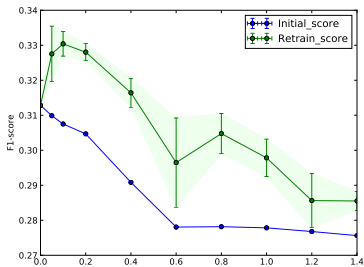
Результаты вычислительного эксперимента

Чем больше плотность вероятности в нуле $\rho(0) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp(-\frac{\mu_i^2}{2\sigma_i^2})$, тем меньше важность параметра.

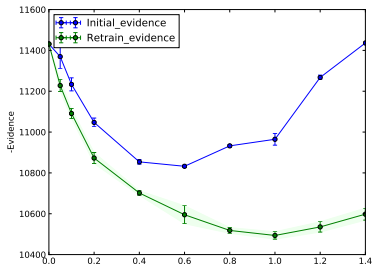
Обозначим отношение сигнал-шум за $\lambda = \left| \frac{\mu_i}{\sigma_i} \right|$, тогда $\rho(0) \sim \exp(-\frac{\lambda^2}{2})$.
Параметры с большим значением λ могут быть удалены.



Количество параметров от λ .



Зависимость F1-меры от λ .



Зависимость правдоподобия от λ .

- Предложена реализация байесовского вывода для задачи поиска парафраз.
- Минимизация правдоподобия модели не приводит к переобучению.
- Алгоритм удаления параметров позволяет упростить структуру модели без существенных потерь качества.

Публикации

А. Н. Смердов, О. Ю. Бахтеев, В. В. Стрижов. Выбор оптимальной модели рекуррентной сети в задачах поиска парафраз // Информатика и её применения, 2019. Том 13, выпуск 2.