Optimal recurrent neural network model in paraphrase detection*

A. N. Smerdov, O. Yu. Bakhteev, V. V. Strijov

Abstract: This paper addresses the problem of the optimal recurrent neural network selection. It asserts the neural network evidence lower bound as the optimal criteria for selection. It investigates variational inference methods to approximate the posterior distribution of the network parameters. As a particular case the normal distribution of the parameters with different types of the covariance matrix is investigated. The authors propose a method of pruning parameters with the highest probability density in zero to increase the model marginal likelihood. As an illustrative example, a computational experiment of multiclass classification on SemEval 2015 dataset have been carried out.

Keywords: deep learning; recurrent neural network; neural network pruning; variational approach.

References

[1] Sutskever, I., O. Vinyals, and Q. V. Le. 2014. Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*. 27:3104–3112. Available at: `https://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf` (accessed December 29, 2017).

[2] Bishop, C. M. Pattern recognition and machine learning. Springer. 2006. 758 p.

[3] Kuznetsov, M. P., A. A. Tokmakova, and V. V. Strijov. 2016. Analytic and stochastic methods of structure parameter estimation. *Informatica*. 27(3):607–624.

*This research was supported by RFBR, project 16-07-01160, and by Government of the Russian Federation, agreement 05.Y09.21.0018.

[4] Popova, M. S. and V. V. Strijov. 2015. Selection of optimal physical activity classification model using measurements of accelerometer. *Informatics and Applications*. 9(1):76–86.

[5] Sanborn, A. and J. Skryzalin. 2015. Deep learning for semantic similarity. *Deep Learning for Natural Language Processing — Stanford, CA, USA: Stanford University*. CS224d:1–7. Available at: `https://cs224d.stanford.edu/reports/SanbornAdrian.pdf` (accessed December 29, 2017).

[6] Pennington, J., R. Socher, and C. D. Manning. 2014. Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing*. 12:1532–1543. `https://nlp.stanford.edu/pubs/glove.pdf` (accessed December 29, 2017).

[7] Rong, X. 2014. Word2vec parameter learning explained. *Arxiv*. Available at: `https://arxiv.org/abs/1411.2738` (accessed December 29, 2017).

[8] Shi, T. and Z. Liu. 2014. Linking GloVe with word2vec. *Arxiv*. Available at: `http://arxiv.org/abs/1411.5595` (accessed December 29, 2017).

[9] Zolotov, V. and D. Kung. 2017. Analysis and optimization of fastText linear text classifier. *Arxiv*. Available at: `https://arxiv.org/ftp/arxiv/papers/1702/1702.05531.pdf` (accessed December 29, 2017).

[10] Graves, A. 2011. Practical variational inference for neural networks. *Advances in Neural Information Processing Systems*. 24:2348–2356. Available at: `http://papers.nips.cc/paper/4329-practical-variational-inference-for-neural-networks.pdf` (accessed December 29, 2017).

[11] LeCun, Y., J. S. Denker., and S. A. Solla. 1989. Optimal brain damage. *Proceedings of Neural Information Processing Systems*. 2:598–605. Available at: `https://papers.nips.cc/paper/250-optimal-brain-damage.pdf` (accessed December 29, 2017).

[12] Hassibi, B., D. G. Stork, and G. J. Wolff. 1992. Optimal brain surgeon and general network pruning. *Neural Computation*. 4:1–8.

[13] Dataset of sentences with different types of similarity. Available at: `http://alt.qcri.org/semeval2015/task2/index.php?id=data-and-tools` (accessed December 29, 2017).

[14] Smerdov, A. N. Computational experiment code. Available at: `https://sourceforge.net/p/mlalgorithms/code/HEAD/tree/Group474/Smerdov2017Paraphrase/code/` (accessed December 29, 2017).

[15] Glove python library. Available at: `https://github.com/stanfordnlp/GloVe` (accessed December 29, 2017).

Выбор оптимальной модели рекуррентной сети в задачах поиска парафраза*

А. Н. Смердов[1], О. Ю. Бахтеев[2], В. В. Стрижов[3]

Аннотация: В работе рассматривается задача выбора оптимальной рекуррентной нейронной сети. В качестве критерия оптимальности используется нижняя оценка правдоподобия модели. Исследование сконцентрировано на применении вариационного подхода к аппроксимации апостериорного распределения параметров модели. Частным случаем аппроксимации выступает нормальное распределение параметров с различными видами матрицы ковариаций. Для увеличения правдоподобия модели предлагается метод удаления параметров с наибольшей плотностью вероятности в нуле. В качестве иллюстративного примера рассматривается задача многоклассовой классификации на выборке пар схожих и несхожих предложений SemEval 2015.

Ключевые слова: глубокое обучение; выбор оптимальной модели; рекуррентная нейросеть; разреживание нейросети; вариационный вывод.

1 Введение

Целью работы является выбор оптимальной нейросетевой модели из класса рекуррентных нейронных сетей. Рекуррентной нейросетью называется нейросеть со связью между нейронами одного слоя. В качестве критерия оптимальности используется нижняя оценка правдоподобия модели.

Число параметров в моделях глубокого обучения может достигать миллионов [1]. Большое число параметров влечет сложность оптимизации параметров и переобучение моделей [2]. Предлагается уменьшить число параметров рекуррентной сети. Это

*Работа выполнена при финансовой поддержке РФФИ (проект 16-07-01160) и правительства Российской Федерации (соглашение № 05.Y09.21.0018).

[1]Московский физико-технический институт, anton.smerdov1@gmail.com

[2]Московский физико-технический институт, bakhteev@phystech.edu

[3]Вычислительный центр им. А. А. Дородницына, Федеральный исследовательский центр «Информатика и управление» Российской Академии Наук, strijov@ccas.ru

обеспечит бóльшую устойчивость модели и снизит время оптимизации ее параметров. Для решения поставленной задачи используются как байесовские методы [3], так и методы прореживания переусложненной нейросети, наращивания простой нейросети и их комбинации [4].

Для построения модели рекуррентной сети рассматривается модель из [5], решающая задачу определения сходства предложений. Модель принимает на вход векторизованные представления слов. Векторизация выполняется с помощью алгоритма GloVe, основанного на факторизации матрицы слов-контекстов и использовании весовой функции для уменьшения значимости редких слов [6]. Альтернативой этому алгоритму выступает линейная модель Word2vec, комбинирующая в себе Continuous Bag-of-Words, skip-gram, negative sampling [7]. Несмотря на разные подходы к проблеме, GloVe и Word2vec оптимизируют схожие функционалы [8]. Упрощенной линейной моделью Word2vec, предназначенной для классификации документов, является fastText — метод, работающий на символьных n-граммах [9].

В работе предлагается подход, основанный на получении вариационной нижней оценки правдоподобия модели. Подобная задача решалась в [10] аппроксимацией апостериорного распределения нормальным, получением аналитических формул для нижней границы правдоподобия модели и удалением параметров с наибольшей плотностью вероятности в нуле. Описанный ниже подход продолжает это исследование. Априорное и апостериорное распределение параметров аппроксимируются нормальным со скалярным, диагональным и блочным видами матрицы ковариаций. После оптимизации гиперпараметров выполняется прореживание сети.

Предлагаемый подход сравнивается с методом удаления параметров Optimal Brain Damage, базирующимся на анализе функции ошибки [11]. Его обобщенной версией выступает алгоритм Optimal Brain Surgeon [12], не предполагающий диагонального вида гессиана функции ошибки.

Вычислительный эксперимент проводится на выборке размеченных пар предложений SemEval 2015. Для каждой пары предложений из корпуса дана экспертная оценка их семантической близости. Требуется построить модель, оценивающую семантическую близость двух предложений. Проблема рассматривается как задача многоклассовой классификации аналогично [5]. Критерием качества служит F1-мера, учитывающая как полноту, так и точность предсказаний. В качестве базовой модели рассматривается пара соединенных рекуррентных сетей с общим вектором параметров и softmax-классификатором на выходе.

2 Задача выбора оптимальной нейросетевой модели

Для построения выборки используем набор пар предложений SemEval 2015 [13]. Каждому слову сопоставим вектор размерности n. Обозначим через l число слов в самом длинном предложении. Предложения длины, меньшей l, дополним нулевыми векторами. Построим выборку

$$\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}, i = 1, \ldots, N,$$

где $\mathbf{x}_i = [\mathbf{x}_i^1, \mathbf{x}_i^2]$ — пары последовательностей векторов слов, соответствующих i-й паре предложений, $\mathbf{x}_i^1, \mathbf{x}_i^2 \in \mathbb{R}^{n \times l}$; $y_i \in Y = \{0, \ldots, 5\}$ — экспертная оценка семантической близости.

Требуется построить модель $f(\mathbf{w}) : \mathbb{R}^{n \times l} \times \mathbb{R}^{n \times l} \to Y$, сопоставляющую паре предложений $\mathbf{x}_i^1$ и $\mathbf{x}_i^2$ класс семантической близости, где $\mathbf{w} \in \mathbb{W} \subseteq \mathbb{R}^s$ — пространство параметров модели. Искомая модель выбирается из множества $\mathfrak{F}$ рекуррентных нейронных сетей с функцией активации tanh. Модель

$$f(\mathbf{w}) : \mathbb{R}^{n \times l} \times \mathbb{R}^{n \times l} \to Y$$

принадлежит искомому классу моделей $\mathfrak{F}$, если существуют такие матрицы перехода $\mathbf{W} \in \mathbb{R}^{n \times m}, \mathbf{U} \in \mathbb{R}^{n \times n}, \mathbf{V} \in \mathbb{R}^{(|Y| \times 2n)}$ и вектор смещения $\mathbf{b} \in \mathbb{R}^n$, что для j-х элементов $\mathbf{x}_{ij}^1, \mathbf{x}_{ij}^2 \in \mathbb{R}^m$ последовательностей $\mathbf{x}_i^1$ и $\mathbf{x}_i^2$ определены векторы скрытого слоя $\mathbf{h}_{ij}^1, \mathbf{h}_{ij}^2 \in \mathbb{R}^n$:

$$\mathbf{h}_{ij}^1 = \tanh(\mathbf{W} \cdot \mathbf{x}_{ij}^1 + \mathbf{U} \cdot \mathbf{h}_{i,j-1}^1 + \mathbf{b}), \tag{1}$$

$$\mathbf{h}_{ij}^2 = \tanh(\mathbf{W} \cdot \mathbf{x}_{ij}^2 + \mathbf{U} \cdot \mathbf{h}_{i,j-1}^2 + \mathbf{b}). \tag{2}$$

Для определения класса семантической близости используются последние значения скрытого слоя $\mathbf{h}_{il}^1$ и $\mathbf{h}_{il}^2$, сконкатенированные в один вектор. После l-й итерации пару предложений будем относить к классу с наибольшим значением, полученным после l-й итерации, $j = 1, \ldots, l$:

$$y = \arg\max_{k \in Y} \left(\mathbf{V} \begin{bmatrix} \mathbf{h}_{il}^1 \\ \mathbf{h}_{il}^2 \end{bmatrix} \right)_k, \tag{3}$$

где $(\cdot)_k$ — k-я компонента вектора. Для каждой модели и соответствующего ей вектора параметров $\mathbf{w} \in \mathbb{W}$ определим логарифмическую функцию правдоподобия выборки $L_{\mathfrak{D}}(\mathfrak{D}, \mathbf{f}, \mathbf{w})$:

$$L_{\mathfrak{D}}(\mathfrak{D}, \mathbf{f}, \mathbf{w}) = \log p(\mathbf{y}|\mathbf{x}, \mathbf{f}, \mathbf{w}) = \log p(\mathfrak{D}|\mathbf{f}, \mathbf{w}) = \sum_{(\mathbf{x}_i, y_i) \in \mathfrak{D}} \log p(y_i|\mathbf{x}_i, \mathbf{f}, \mathbf{w}), \tag{4}$$

где $p(\mathbf{y}|\mathbf{x}, \mathbf{f}, \mathbf{w})$ — апостериорная вероятность вектора $\mathbf{y}$ при заданных $\mathbf{x}, \mathbf{f}, \mathbf{w}$. Здесь и далее используется обозначение $p(\mathbf{x}|\mathbf{y}) = p(\mathfrak{D})$.

Оптимальная модель $\mathbf{f}$ находится максимизацией логарифма ее правдоподобия:

$$L_{\mathbf{f}}(\mathfrak{D}, \mathbf{f}) = L_{\mathbf{f}}(\mathfrak{D}|\mathbf{f}) = \log p(\mathbf{y}|\mathbf{x}, \mathbf{f}) = \log p(\mathfrak{D}|\mathbf{f}) = \log \int_{\mathbf{w} \in \mathbb{W}} p(\mathfrak{D}|\mathbf{w}, \mathbf{f}) p(\mathbf{w}|\mathbf{f}) d\mathbf{w}. \tag{5}$$

Апостериорное распределение параметров модели находится из уравнения:

$$p(\mathbf{w}|\mathfrak{D}, \mathbf{f}) = \frac{p(\mathfrak{D}|\mathbf{w}, \mathbf{f}) p(\mathbf{w}|\mathbf{f})}{p(\mathfrak{D}|\mathbf{f})}. \tag{6}$$

Приблизим интеграл (5) вариационной нижней оценкой. Воспользуемся оценкой [2] (разделы 10.2-10.4), полученной из неравенства Йенсена:

$$L_{\mathbf{f}}(\mathfrak{D},\mathbf{f}) = \log \int\limits_{\mathbf{w}\in\mathbb{W}} p(\mathfrak{D}|\mathbf{w})p(\mathbf{w}|\mathbf{f})d\mathbf{w} =$$
$$= \int\limits_{\mathbf{w}\in\mathbb{W}} p(\mathbf{w}|\mathfrak{D},\mathbf{f}) \log \frac{p(\mathbf{w}|\mathfrak{D},\mathbf{f})}{p(\mathfrak{D}|\mathbf{f},\mathbf{w})} d\mathbf{w} + D_{\mathrm{KL}}\big(p(\mathfrak{D}|\mathbf{w})||p(\mathfrak{D}|\mathbf{f})\big),$$

где $D_{\mathrm{KL}}\big(q(\mathbf{w})||p(\mathbf{w})\big)$ — расстояние Кульбака–Лейблера между $q(\mathbf{w})$ и $p(\mathbf{w})$,

$$D_{\mathrm{KL}}\big(q(\mathbf{w})||p(\mathbf{w})\big) = -\int_{\mathbf{w}} q(\mathbf{w}) \log \frac{q(\mathbf{w})}{p(\mathbf{w})} d\mathbf{w}.$$

Учитывая неотрицательность расстояния Кульбака–Лейблера, получаем

$$L_{\mathbf{f}}(\mathfrak{D},\mathbf{f}) \geq \int_{\mathbf{w}\in\mathbb{W}} p(\mathbf{w}|\mathfrak{D},\mathbf{f}) \log \frac{p(\mathbf{w}|\mathfrak{D},\mathbf{f})}{p(\mathfrak{D}|\mathbf{f},\mathbf{w})} d\mathbf{w}. \tag{7}$$

Упростим интеграл в левой части (7):

$$\int\limits_{\mathbf{w}\in\mathbb{W}} p(\mathbf{w}|\mathfrak{D},\mathbf{f}) \log \frac{p(\mathbf{w}|\mathfrak{D},\mathbf{f})}{p(\mathfrak{D}|\mathbf{f},\mathbf{w})} d\mathbf{w} =$$
$$= -D_{\mathrm{KL}}\big(p(\mathbf{w}|\mathfrak{D},\mathbf{f})||p(\mathbf{w}|\mathbf{f})\big) + \int\limits_{\mathbf{w}\in\mathbb{W}} p(\mathbf{w}|\mathfrak{D},\mathbf{f}) \log p(\mathbf{w}|\mathfrak{D},\mathbf{f}) d\mathbf{w}. \tag{8}$$

Обозначим сумму в левой части (8) через $-L(\mathfrak{D},\mathbf{f},\mathbf{w})$:

$$L(\mathfrak{D},\mathbf{f},\mathbf{w}) = \underbrace{D_{\mathrm{KL}}\big(p(\mathbf{w}|\mathfrak{D},\mathbf{f})||p(\mathbf{w}|\mathbf{f})\big)}_{L_{\mathbf{w}}(\mathfrak{D},\mathbf{f},\mathbf{w})} - \underbrace{\int_{\mathbf{w}\in\mathbb{W}} p(\mathbf{w}|\mathfrak{D},\mathbf{f}) \log p(\mathfrak{D}|\mathbf{f},\mathbf{w}) d\mathbf{w}}_{L_E(\mathfrak{D},\mathbf{f})}. \tag{9}$$

Первое слагаемое формулы (9) интерпретируется как минимальная длина описания распределения $p(\mathbf{w}|\mathfrak{D},\mathbf{f})$ с помощью $p(\mathbf{w}|\mathbf{f})$. Эту величину назовем сложностью модели $L_{\mathbf{w}}(\mathfrak{D},\mathbf{f},\mathbf{w})$:

$$L_{\mathbf{w}}(\mathfrak{D},\mathbf{f},\mathbf{w}) = D_{\mathrm{KL}}\big(p(\mathbf{w}|\mathfrak{D},\mathbf{f})||p(\mathbf{w}|\mathbf{f})\big). \tag{10}$$

Второе слагаемое формулы (9) есть минус матожидание правдоподобия выборки $L_{\mathfrak{D}}$ (4), и оно тем меньше, чем выше правдоподобие выборки, поэтому интерпретируется как функционал ошибки $L_E(\mathfrak{D},\mathbf{f})$ в ходе вычислительного эксперимента:

$$L_E(\mathfrak{D},\mathbf{f}) = \mathsf{E}_{\mathbf{w}\sim p(\mathbf{w}|\mathfrak{D},\mathbf{f})} L_{\mathfrak{D}}(\mathbf{y},\mathfrak{D},\mathbf{f},\mathbf{w}). \tag{11}$$

Запишем суммарную функцию потерь $L(\mathfrak{D}, \mathbf{f}, \mathbf{w})$ как сумму функционала сложности модели $L_{\mathbf{w}}(\mathfrak{D}, \mathbf{f}, \mathbf{w})$ и функционала ошибки $L_E(\mathfrak{D}, \mathbf{f})$:

$$L(\mathfrak{D}, \mathbf{f}, \mathbf{w}) = L_{\mathbf{w}}(\mathfrak{D}, \mathbf{f}, \mathbf{w}) + L_E(\mathfrak{D}, \mathbf{f}). \tag{12}$$

Искомая модель минимизирует суммарный функционал потерь

$$\mathbf{f} = \arg\min_{\mathbf{f}\in\mathfrak{F}} L(\mathfrak{D}, \mathbf{f}, \mathbf{w}). \tag{13}$$

3 Предлагаемое решение оптимизационной задачи

Так как апостериорное распределение $p(\mathbf{w}|\mathfrak{D}, \mathbf{f})$ (6) невозможно получить аналитически, минимизация функционала потерь $L(\mathfrak{D}, \mathbf{f}, \mathbf{w})$ (12) затруднена. Для решения этой проблемы применим вариационный подход. Он заключается в аппроксимации неизвестного распределения распределением из известного класса. В качестве приближения $p(\mathbf{w}|\mathfrak{D}, \mathbf{f})$ выберем нормальное распределение:

$$p(\mathbf{w}|\mathfrak{D}, \mathbf{f}) \sim \mathcal{N}(\mathbf{m}, \mathbf{A}_{\text{ps}}^{-1}),$$

где $\mathbf{m}, \mathbf{A}_{\text{ps}}^{-1}$ — вектор средних и матрица ковариации этого распределения. Априорное распределение $p(\mathbf{w}|\mathbf{f})$ вектора параметров $\mathbf{w}$ будем считать нормальным с параметрами $\boldsymbol{\mu}$ и $\mathbf{A}_{\text{pr}}^{-1}$:

$$p(\mathbf{w}|\mathbf{f}) \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{A}_{\text{pr}}^{-1}),$$

где $\boldsymbol{\mu}$ — вектор средних, $\mathbf{A}_{\text{pr}}^{-1}$ — матрица ковариаций. Расстояние Кульбака–Лейблера между нормальными распределениями $\mathcal{N}(\boldsymbol{\mu}, \mathbf{A}_{\text{pr}}^{-1})$ и $\mathcal{N}(\mathbf{m}, \mathbf{A}_{\text{ps}}^{-1})$ вычисляется по формуле

$$D_{\text{KL}}\big(\mathcal{N}(\boldsymbol{\mu}, \mathbf{A}_{\text{pr}}^{-1})||\mathcal{N}(\mathbf{m}, \mathbf{A}_{\text{ps}}^{-1})\big) = \frac{1}{2}\left(\log\frac{|\mathbf{A}_{\text{ps}}^{-1}|}{|\mathbf{A}_{\text{pr}}^{-1}|} - d + \operatorname{tr}(\mathbf{A}_2\mathbf{A}_{\text{pr}}^{-1}) + (\boldsymbol{\mu}-\mathbf{m})^{\mathsf{T}}\mathbf{A}_2(\boldsymbol{\mu}-\mathbf{m})\right). \tag{14}$$

Рассмотрим частные случаи вида матриц ковариаций $\mathbf{A}_{\text{pr}}^{-1}$ и $\mathbf{A}_{\text{ps}}^{-1}$. Так как априори нет предпочтений при выборе параметров, то априорное распределение для всех параметров считаем одинаковым, т. е. вектор средних $\boldsymbol{\mu} = \mu\mathbf{1}$, матрица ковариаций скалярна: $\mathbf{A}_{\text{pr}}^{-1} = \sigma\mathbf{I}$. После получения информации о выборке получаем апостериорный вектор средних $\mathbf{m}$.

Алгоритм решения оптимизационной задачи заключается в выполнении градиентного шага при заданном априорном распределении, вычислении апостериорного распределения и аппроксимации нового априорного распределения полученным апостериорным. Рассмотрим различные виды апостериорной матрицы ковариаций $\mathbf{A}_{\text{ps}}^{-1}$.

1. Матрица ковариаций скалярна: $\mathbf{A}_{\text{ps}}^{-1} = \alpha\mathbf{I}$. В этом случае

$$D_{\text{KL}}\big(\mathcal{N}(\boldsymbol{\mu}, \mathbf{A}_{\text{pr}}^{-1})||\mathcal{N}(\mathbf{m}, \mathbf{A}_{\text{ps}}^{-1})\big) = \sum_{i=1}^{W}\big(\log\frac{\sigma}{\alpha} + \frac{(\mu-m_i)^2+\alpha^2+\sigma^2}{2\sigma^2}\big).$$

По значениям параметров α и $\mathbf{m}$ апостериорного распределения вычислим параметры априорного. Число элементов вектора $\mathbf{m}$ обозначим W. Из условия $\frac{\partial}{\partial\mu}D_{\mathrm{KL}} = \sum_{i=1}^{W}\frac{\mu-m_i}{\sigma^2} = 0$ получаем выражения для μ на следующей итерации $\hat{\mu} = \frac{1}{W}\sum_{i=1}^{W} m_i$. Аналогично $\frac{\partial}{\partial\sigma^2}D_{\mathrm{KL}} = \sum_{i=1}^{W}\frac{1}{2\sigma^2} - \frac{(\mu-m_i)^2+\alpha^2}{2\sigma^4} = 0 \;\Rightarrow\; \hat{\sigma}^2 = \frac{1}{W}\sum_{i=1}^{W}(\mu-m_i)^2 + \alpha^2$.

2. Матрица ковариаций диагональна: $\mathbf{A}_{\mathrm{ps}}^{-1} = \mathrm{diag}(\boldsymbol{\sigma}^2)$. В этом случае

$$D_{\mathrm{KL}}\big(\mathcal{N}(\boldsymbol{\mu}, \mathbf{A}_{\mathrm{pr}}^{-1})||\mathcal{N}(\mathbf{m}, \mathbf{A}_{\mathrm{ps}}^{-1})\big) = \sum_{i=1}^{W}(\log\frac{\sigma}{\sigma_i} + \frac{(\mu-m_i)^2+\sigma_i^2+\sigma^2}{2\sigma^2}).$$

Значения параметров априорного распределения для следующей итерации вычисляются следующим образом:

$$\text{из } \frac{\partial}{\partial\mu}D_{\mathrm{KL}} = \sum_{i=1}^{W}\frac{\mu-m_i}{\sigma^2} = 0 \text{ получаем } \hat{\mu} = \frac{1}{W}\sum_{i=1}^{W} m_i,$$

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

Оптимизация параметров сводится к следующему алгоритму:

Инициализировать $\boldsymbol{\sigma} = \mathbf{1},\ \mathbf{m} = \mathbf{0},\ \mu = 0,\ \sigma^2 = 1$.

Повторять:

Сделать градиентный шаг $\boldsymbol{\sigma} := \boldsymbol{\sigma} - \eta\nabla\boldsymbol{\sigma},\ \mathbf{m} := \mathbf{m} - \eta\nabla\mathbf{m},\ \mathbf{w} := \mathbf{w} - \eta\nabla\mathbf{w}$.

Обновить параметры априорного распределения $\mu := \hat{\mu},\ \sigma^2 := \hat{\sigma}^2$.

Пока значение L не стабилизируется.

Таблица 1: Результаты вычислительного эксперимента

Classificator	F1-measure, валидация	F1-measure, test
Logistic Regression	0,286	0,286
SVC	0,290	0,290
DecisionTreeClassifier	0,316	0,316
KNeighborsClassifier	0,322	0,322
RNN	0,393	0,362
RNN+variational, I, I	—	0,311
RNN+variational, D, I	—	0,330

4 Удаление параметров из сети

Введем множество индексов активных параметров модели $\mathcal{A} = \{i|w_i \neq 0\}$. Для увеличения правдоподобия модели предлагается уменьшить ее сложность, т. е. уменьшить число параметров $|\mathcal{A}|$. Для удаления выберем параметры, имеющие наибольшую

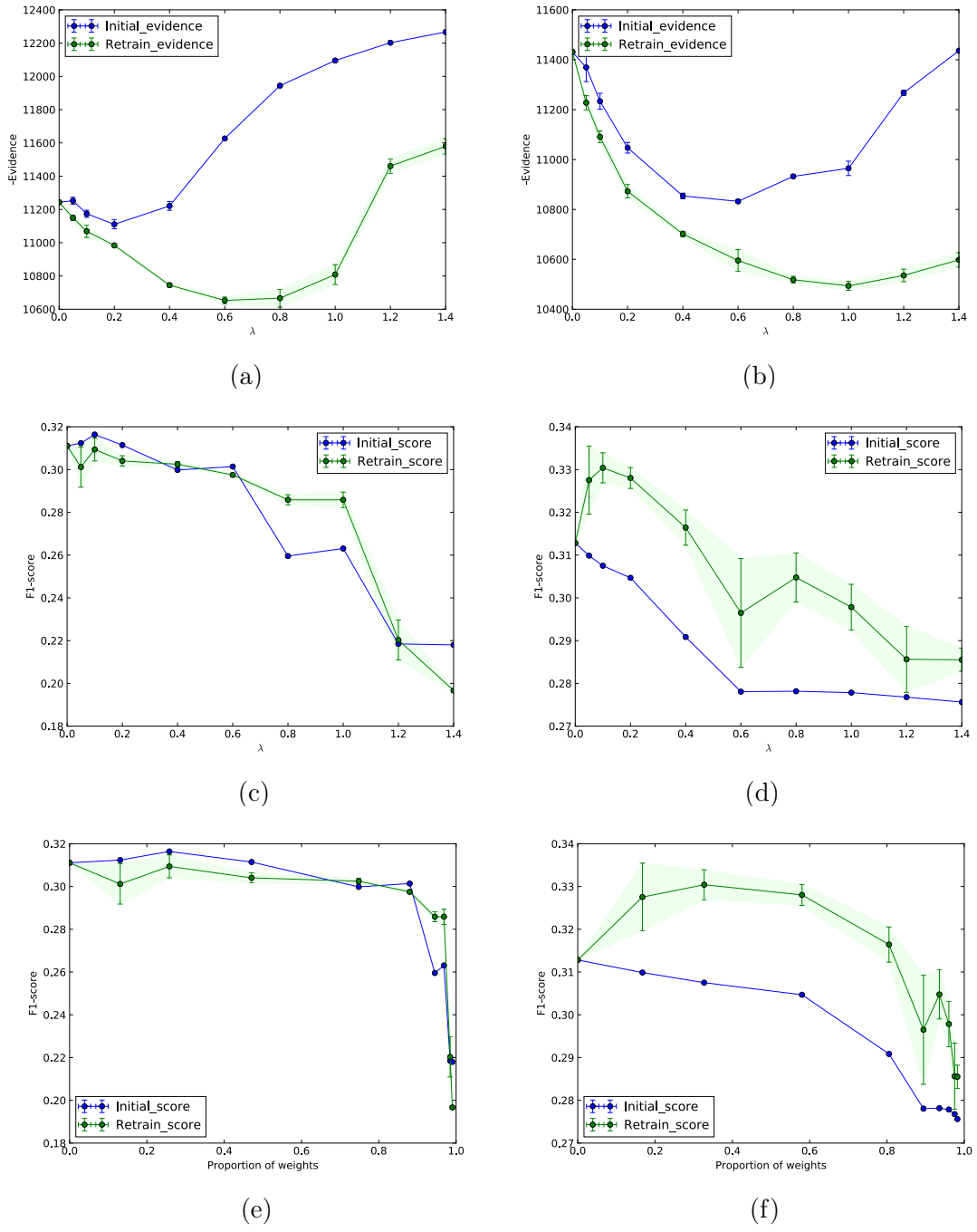

Рис. 1: Зависимость нижней оценки правдоподобия модели и F1-меры от λ для скалярной(a, c, e) и диагональной(b, d, f) матриц

плотность апостериорной вероятности ρ в нуле. Если апостериорная матрица ковариаций скалярна, то

$$\rho_i = \exp\left(-\frac{\mu_i^2}{2\sigma^2}\right). \tag{15}$$

Чем больше ρ, тем меньше $|\frac{\mu_i}{\sigma}|$, поэтому удаляются параметры со значением $|\frac{\mu_i}{\sigma}| < \lambda$, где λ — пороговое значение. Варьируя пороговое значение λ, выбираем оптимальное число неудаленных параметров. Для диагонального вида матрицы ковариаций критерий удаления параметров записывается как $|\frac{\mu_i}{\sigma_i}| < \lambda$.

5 Вычислительный эксперимент

Цель эксперимента — проверка работоспособности предложенного алгоритма и сравнение результатов с ранее полученными. В качестве данных использовалась выборка SemEval 2015, состоящая из 8331 пары схожих и несхожих предложений. Слова преобразовывались в векторы размерности 50 при помощи алгоритма GloVe [15]. Для базовых алгоритмов тренировочная, валидационная и тестовая выборки составили 70%, 15% и 15% соответственно. Для рекуррентной нейронной сети, полученной вариационным методом, валидационная выборка отсутствовала, а тренировочная и тестовая выборки составили 85% и 15% соответственно. Критерием качества была выбрана F1-мера. В качестве базовых алгоритмов использовались линейная регрессия, метод ближайших соседей, решающее дерево и модификация метода опорных векторов SVC. Базовые алгоритмы взяты из библиотеки sklearn. Дополнительно были построены рекуррентная нейросеть с одним скрытым слоем [5] и нейросеть с одним скрытым слоем и вариационной оптимизацией параметров [10, 14].

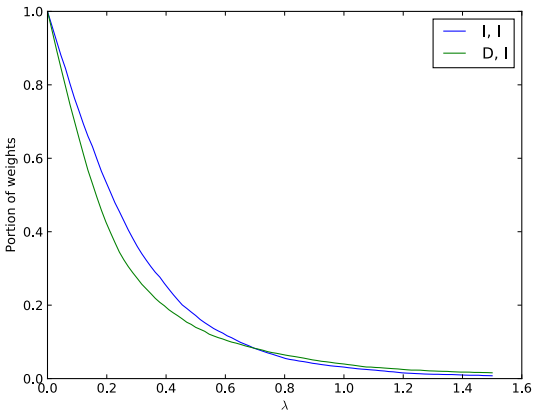

Рис. 2: Доля неудаленных параметров сети в зависимости от порогового значения λ для скалярного (I) и диагонального (D) вида апостериорной матрицы ковариаций

На рис. 1a и 1b представлена зависимость оценки правдоподобия L (12) от параметра λ. Для обоих случаев существует оптимальное значение λ, минимизирующее L; модели с таким параметром будут оптимальными. На рис. 1c, 1d, 1e и 1f отображены зависимости качества модели от λ и доли выброшенных параметров. Видно, что даже при удалении большинства параметров из сети качество предсказаний меняется несущественно, что говорит о слишком большом числе параметров исходной модели.

Из рис. 2 видно, что при малых λ из сети с диагональной апостериорной матрицей ковариаций удаляется больше весов, а при больших λ — меньше, что говорит о лучшем отборе параметров такой моделью.

6 Заключение

С помощью вариационного байесовского подхода был построен набор моделей глубинного обучения с оптимальной нижней оценкой правдоподобия, отличающихся различными предположениями о виде априорного и апостериорного распределения параметров. Из случайности распределения параметров был получен критерий их удаления, что позволило увеличить нижнюю оценку правдоподобия моделей. Результаты полученных нейросетей в вычислительном эксперименте оказались близки к результатам других алгоритмов.

Список литературы

[1] *Sutskever I., Vinyals O, Le Q. V.* Sequence to sequence learning with neural networks // Advances in Neural Information Processing Systems, 2014. P. 3104–3112. `https://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf`

[2] *Bishop C. M.* Pattern Recognition and Machine Learning. — Springer, 2006.

[3] *Kuznetsov M. P., Tokmakova A. A., Strijov V. V.* Analytic and stochastic methods of structure parameter estimation // Informatica, 2016, Vol. 27.3, P. 607–624.

[4] *Попова М. С., Стрижов В. В.* Выбор оптимальной модели классификации физической активности по измерениям акселерометра // Информатика и ее применения, 2015. Т. 9. Вып. 1. С. 76–86.

[5] *Sanborn A., Skryzalin J.* Deep Learning for Semantic Similarity // CS224d: Deep Learning for Natural Language Processing — Stanford, CA, USA: Stanford University, 2015. `https://cs224d.stanford.edu/reports/SanbornAdrian.pdf`.

[6] *Pennington J., Socher R., Manning C. D.* Glove: Global vectors for word representation // Proceedings of the Empiricial Methods in Natural Language Processing, 2014. Vol. 12. `https://nlp.stanford.edu/pubs/glove.pdf`.

[7] *Rong X.* Word2vec parameter learning explained // Arxiv, 2014. https://arxiv.org/abs/1411.2738.

[8] *Shi T., Liu Z.* Linking GloVe with word2vec // Arxiv, 2014. http://arxiv.org/abs/1411.5595.

[9] *Zolotov V., Kung D.* Analysis and optimization of fastText linear text classifier // Arxiv, 2017. https://arxiv.org/ftp/arxiv/papers/1702/1702.05531.pdf.

[10] *Graves A.* Practical variational inference for neural networks // Advances in Neural Information Processing Systems, 2011. P. 2348–2356. http://papers.nips.cc/paper/4329-practical-variational-inference-for-neural-networks.pdf.

[11] *Le Cun Y., Denker J. S., Solla S. A.* Optimal Brain Damage // Proceedings of NIPS-89, 1989. Vol. 2. P. 598–605. https://papers.nips.cc/paper/250-optimal-brain-damage.pdf.

[12] *Hassibi B., Stork D. G., Wolff G. J.* Optimal brain surgeon and general network pruning // IEEE International Conference on Neural Networks. — IEEE, 1993. P. 293–299.

[13] Выборка пар предложений различной степени похожести. http://alt.qcri.org/semeval2015/task2/index.php?id=data-and-tools.

[14] *Смердов А. Н.* Код вычислительного эксперимента. https://sourceforge.net/p/mlalgorithms/code/HEAD/tree/Group474/Smerdov2017Paraphrase/code/.

[15] Библиотека Glove, python. https://github.com/stanfordnlp/GloVe.