Федеральное государственное автономное образовательное учреждение
высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»
Физтех-школа Прикладной Математики и Информатики
Кафедра интеллектуальных систем

Направление подготовки / специальность: 03.04.01 Прикладные математика и физика
Направленность (профиль) подготовки: Математическая физика, компьютерные технологии и математическое моделирование в экономике

ОБУЧЕНИЕ С ЭКСПЕРТОМ ДЛЯ ПОСТРОЕНИЯ ИНТЕРПРЕТИРУЕМЫХ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ

(магистерская диссертация)

Студент:
Грабовой Андрей Валериевич

_____
(подпись студента)

Научный руководитель:
Стрижов Вадим Викторович,
д-р физ.-мат. наук

_____
(подпись научного руководителя)

Консультант *(при наличии)***:**
_____
_____
(подпись консультанта)

Москва 2021

Содержание

Аннотация

Исследуется проблема понижения сложности аппроксимирующих моделей с целью повышения их интерпретируемости. В рамках данного исследования рассматриваются методы, которые используют экспертную информацию о данных с целью получения простых, более интерпретируемых моделей. Предлагаются методы, основанные на дистилляции моделей глубокого обучения, где модель учителя рассматривается в качестве эксперта. Также предлагаются методы основанные на экспертном описании задачи, что позволяет строить специальные признаковые описания объектов. Теоретические результаты анализируются в вычислительном эксперименте на синтетических выборках и реальных данных. В качестве реальных данных рассматривается популярные выборки, такие как MNIST, FashionMNIST и Twitter Sentiment Analysis.

Ключевые слова: выбор модели; байесовский вывод; дистилляция модели; локальные преобразования; преобразования вероятностных пространств; релевантноть параметров.

Список литературы

[1] *Грабовой А. В., Бахтеев О. Ю., Стрижов В. В.* Определение релевантности параметров нейросети // Информатика и ее применения, 2019. Т. 13. Вып. 2. С. 62–70.

[2] *Грабовой А. В., Бахтеев О. Ю., Стрижов В. В.* Введение отношения порядка на множестве параметров аппроксимирующих моделей // Информатика и ее применения, 2020. Т. 14. Вып. 2. С. 58–65.

[3] *A V Grabovoy, V V Strijov* Quasi-Periodic Time Series Clustering for Human Activity Recognition // Lobachevskii Journal of Mathematics, 2020 Pp. 333–339

[4] *Грабовой А. В., Стрижов В. В.* Анализ выбора априорного распределения для смеси экспертов // Журнал Вычислительной Математики и Математической Физики, 2021, С. 1149–1161.

[5] *Huang, Zehao and Wang, Naiyan* Like What You Like: Knowledge Distill via Neuron Selectivity Transfer // arXiv e-prints, 2017.

[6] *S. Aeberhard* Wine Data Set, 1991.

[7] *Linting Xue and Noah Constant and Adam Roberts and Mihir Kale and Rami Al-Rfou and Aditya Siddhant and Aditya Barua and Colin Raffel.* mT5: A massively multilingual pre-trained text-to-text transformer // arXiv preprinted, 2021.

[8] *He K., Zhang X., Ren S., Sun J.* Deep Residual Learning for Image Recognition // Proc. of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016. P. 770–778.

[9] *Graves A.* Practical Variational Inference for Neural Networks // Advances in Neural Information Processing Systems, 2011. Vol. 24. P. 2348–2356.

[10] *Vapnik V., Izmailov R.* Learning Using Privileged Information: Similarity Control and Knowledge Transfer // Journal of Machine Learning Research. 2015. No 16. P. 2023–2049.

[11] *Sutskever I., Vinyals O., Le Q.* Sequence to Sequence Learning with Neural Networks // Advances in Neural Information Processing Systems, 2014. Vol. 2. P. 3104–3112.

[12] *Li C., Chen C., Carlson D., Carin L.* Preconditioned Stochastic Gradient Langevin Dynamics for Deep Neural Networks // Thirtieth AAAI Conference on Artificial Intelligence. — Phoenix, USA, 2016. P. 1788–1794.

[13] *Tibshirani R.* Regression shrinkage and selection via the Lasso // Journal of the Royal Statistical Society, 1996. Vol. 58. P. 267–288.

[14] *Zou H., Hastie T.* Regularization and variable selection via the Elastic Net // Journal of the Royal Statistical Society, 2005. Vol. 67. P. 301–320.

[15] *Srivastava N., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov R.* Dropout: A Simple Way to Prevent Neural Networks from Overfitting // Journal of Machine Learning Research, 2014. Vol. 15. P. 1929–1958.

[16] *Molchanov D., Ashukha A., Vetrov D.* Variational Dropout Sparsifies Deep Neural Networks // 34th International Conference on Machine Learning. — Sydney, Australia, 2017. Vol. 70. P. 2498–2507.

[17] *LeCun Y., Denker J., Solla S.* Optimal Brain Damage // Advances in Neural Information Processing Systems, 1989. Vol. 2. P. 598–605.

[18] *Mandt S., Hoffman M., Blei D.* Stochastic Gradient Descent as Approximate Bayesian Inference // Journal Of Machine Learning Research, 2017. Vol. 18. P. 1–35.

[19] *Harrison D., Rubinfeld D.* Hedonic prices and the demand for clean air // Journal of Environmental Economics and Management, 1991. Vol. 5. P. 81–102.

[20] *Maclaurin D., Duvenaud D., Adams R.* Gradient-based Hyperparameter Optimization Through Reversible Learning // Proceedings of the 32th International Conference on Machine Learning, 2015. Vol. 37. P. 2113–2122.

[21] *Luketina J., Berglund M., Raiko T., Greff K.* Scalable Gradient-based Tuning of Continuous Regularization Hyperparameters // Proceedings of the 33th International Conference on Machine Learning, 2016. Vol. 48. P. 2952–2960.

[22] *Bishop C.* Pattern Recognition and Machine Learning, 2006. Pp. 396.

[23] *Neychev R., Katrutsa A., Strijov V.* Robust selection of multicollinear features in forecasting // Factory Laboratory, 2016. Vol. 82. P. 68–74.

[24] *Neal A., Radford M.* Bayesian Learning for Neural Networks, 1995.

[25] *Louizos C., Ullrich K., Welling M.* Bayesian Compression for Deep Learning, 2017. P. 3288–3298.

[26] *J. R. Kwapisz, G. M. Weiss, S. A. Moore* Activity Recognition using Cell Phone Accelerometers // Proceedings of the Fourth International Workshop on Knowledge Discovery from Sensor Data, 2010. Vol. 12. P. 74–82.

[27] *W. Wang, H. Liu, L. Yu, F. Sun* Activity Recognition using Cell Phone Accelerometers // Joint Conference on Neural Networks, 2014. P. 1185–1190.

[28] *A. D. Ignatov, V. V. Strijov* Human activity recognition using quasiperiodic time series collected from a single tri-axial accelerometer. // Multimedial Tools and Applications, 2015.

[29] *A. Olivares, J. Ramirez, J. M. Gorris, G. Olivares, M. Damas* Detection of (in)activity periods in human body motion using inertial sensors: A comparative study. // Sensors, 12(5):5791–5814, 2012.

[30] *Y. G. Cinar and H. Mirisaee* Period-aware content attention RNNs for time series forecasting with missing values // Neurocomputing, 2018. Vol. 312. P. 177–186.

[31] *A. P. Motrenko, V. V. Strijov* Extracting fundamental periods to segment biomedical signals // Journal of Biomedical and Health Informatics, 2015, 20(6). P. 1466 - 1476.

[32] *Y. P. Lukashin* Adaptive methods for short-term forecasting // Finansy and Statistik, 2003.

[33] *И. П. Ивкин, М. П. Кузнецов* Алгоритм классификации временных рядов акселерометра по комбинированному признаковому описанию. // Машинное обучение и анализ данных, 2015.

[34] *V. V. Strijov, A. M. Katrutsa* Stresstes procedures for features selection algorithms. // Schemometrics and Intelligent Laboratory System, 2015.

[35] *I. Borg, P. J. F. Groenen* Modern Multidimensional Scaling. — New York: Springer, 2005. 540 p.

[36] *Д. Л. Данилова, А. А. Жигловский* Главные компоненты временных рядов: метод "Гусеница". — Санкт-Петербурскиий университет, 1997.

[37] *Tianqi C., Carlos G.* XGBoost: A Scalable Tree Boosting System // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016.

[38] *Xi C., Hemant I.* Random Forests for Genomic Data Analysis // Genomics. 2012. Issues. 99. No. 6. P. 323–329.

[39] *Esen Y. S., Wilson J., Gader P. D.* Twenty Years of Mixture of Experts // IEEE Transactions on Neural Networks and Learning Systems. 2012. Issues. 23. No. 8. P. 1177–1193.

[40] *Rasmussen C. E., Ghahramani Z.* Infinite Mixtures of Gaussian Process Experts // Advances in Neural Information Processing Systems 14. 2002. P. 881–888.

[41] *Shazeer N., Mirhoseini A., Maziarz K.* Outrageously large neural networks: the sparsely-gated mixture-of-experts layer // International Conference on Learning Representations. 2017.

[42] *Jordan M. I.* Hierarchical mixtures of experts and the EM algorithm // Neural Comput. 1994. Vol. 6, No. 2. P. 181–214.

[43] *Jordan M. I., Jacobs R. A.* Hierarchies of adaptive experts // In Advances in Neural Information Processing Systems. 1991. P. 985–992.

[44] *Lima C., Coelho A., Zuben F. J.* Hybridizing mixtures of experts with support vector machines: Investigation into nonlinear dynamic systems identification // Inf. Sci. 2007. Vol. 177. No. 10. P. 2049–2074.

[45] *Cao L.* Support vector machines experts for time series forecasting // Neurocomputing. 2003. Vol. 51. P. 321–339.

[46] *Yumlu M. S., Gurgen F. S., Okay N.* Financial time series prediction using mixture of experts // In Proc. 18th Int. Symp. Comput. Inf. Sci. 2003. P. 553–560.

[47] *Cheung Y. M., Leung W. M., Xu L.* Application of mixture of experts model to financial time series forecasting // On Proc. Int. Conf. Neural Netw. Signal Process. 1995. P. 1–4.

[48] *Weigend A. S., Shi S.* Predicting daily probability distributions of S&P500 returns // J. Forecast. 2000. Vol. 19. No. 4. P. 375–392.

[49] *Ebrahimpour R., Moradian M. R., Esmkhani A., Jafarlou F. M.* Recognition of Persian handwritten digits using characterization loci and mixture of experts // J. Digital Content Technol. Appl. 2009. Vol. 3. No. 3. P. 42–46.

[50] *Estabrooks A., Japkowicz N.* A mixture-of-experts framework for text classification //In Proc. Workshop Comput. Natural Lang. Learn., Assoc. Comput. Linguist. 2001. P. 1–8.

[51] *Mossavat S., Amft O., Petkov Vries B., Kleijn W.* A Bayesian hierarchical mixture of experts approach to estimate speech quality // In Proc. 2nd Int. Workshop Qual. Multimedia Exper. 2010. P. 200–205.

[52] *Peng F., Jacobs R. A., Tanner M. A.* Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition // J. Amer. Stat. Assoc. 1996. Vol. 91. No. 435. P. 953–960.

[53] *Tuerk A.* The state based mixture of experts HMM with applications to the recognition of spontaneous speech. Ph.D. thesis. Cambridge: Univ. Cambridge, 2001.

[54] *Sminchisescu C., Kanaujia A., Metaxas D.* Discriminative density propagation for visual tracking // IEEE Trans. Pattern Anal. Mach. Intell. 2007. Vol. 29. No. 11. P. 2030–2044.

[55] *Bowyer K., Hollingsworth K., Flynn P.* A Survey of Iris Biometrics Research: 2008–2010.

[56] *Matveev I.* Detection of iris in image by interrelated maxima of brightness gradient projections // Appl.Comput. Math. 2010. Vol. 9. No. 2. P. 252–257.

[57] *Matveev I., Simonenko I.*. Detecting precise iris boundaries by circular shortest path method // Pattern Recognition and Image Analysis. 2014. Vol. 24. P. 304–309.

[58] *Dempster A. P., Laird N. M., Rubin D. B.* Maximum Likelihood from Incomplete Data via the EM Algorithm // Journal of the Royal Statistical Society. Series B (Methodological). 1977. Vol. 39. No. 1 P. 1–38.

[59] Akhtar N, Mian A (2018) Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey. IEEE Access 6:14410–14430

[60] Han X, Yao M, Debayan D, Hui L, Ji-Liang T, Anil J (2020) Adversarial Attacks and Defenses in Images, Graphs and Text: A Review. International Journal of Automation and Computing 17:151–178

[61] Ribeiro M, Singh S, Guestrin C (2016) Why Should I Trust You?": Explaining the Predictions of Any Classifier. KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 1135–1144

[62] Salamani D, Gadatsch S, Golling T, Stewart G, Ghosh A, Rousseau D, Hasib A, Schaarschmidt J (2018) Deep Generative Models for Fast Shower Simulation in ATLAS. 2018 IEEE 14th International Conference on e-Science (e-Science) https://doi.org/10.1109/eScience.2018.00091

[63] Demidenko E (2007) Sample size determination for logistic regression revisited. *Statist. Med.* 26:3385–3397.

[64] Joseph L, Berger R, Be'lisle P (1995) Bayesian and mixed bayesian likelihood criteria for sample size determination. *Statistician* 16:769–781.

[65] Joseph L, Wolfson D, Berger R (1997) Sample size calculations for binomial proportions via highest posterior density intervals. *Statistical Medicine* 44:143–154.

[66] Kloek T (1975). Note on a large-sample result in specification analysis. *Econometrica* 43:933–936.

[67] Lindley D (1997) The choice of sample size. *The Statistician* 46:129–138.

[68] Motrenko A, Strijov V, Weber G (2014) Sample size determination for logistic regression. *Journal of Computational and Applied Mathematics* 255:743–752.

[69] Quinlan J (1992) Learning with continuous classes. *Proc. 5th Australian Joint Conference on AI* 343–348.

[70] Qumsiyeh M (2013) Using the bootstrap for estimation the sample size in statistical experiments. *Journal of modern applied statistical methods* 8:305–321.

[71] Rubin D, Stern H (1998) Sample size determination using posterior predictive distributions. *Sankhya: The Indian Journal of Statistics Special Issue on Bayesian Analysis* 60:161–175.

[72] Self S, Mauritsen R (1988) Power/sample size calculations for generalized linear models. *Biometrics* 44:79–86.

[73] Self S, Mauritsen R, Ohara J (1992) Power calculations for likelihood ratio tests in generalized linear models. *Biometrics* 48:31–39.

[74] Shieh G (2000) On power and sample size calculations for likelihood ratio tests in generalized linear models. *Biometrics* 56:1192–1196.

[75] Shieh G (2005) On power and sample size calculations for Wald tests in generalized linear models. *Journal of Statistical Planning and Inference* 128:43–59.

[76] Wang F, Gelfand A (2002) A Simulation-based Approach to Bayesian Sample Size Determination for Performance under a Given Model and for Separating Models. *Statistical Science* 17:193–208.

[77] *Devlin J., Chang M., Lee K., Toutanova K.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // arXiv preprint arXiv:1810.04805. 2018.

[78] *Бахтеев О. Ю., Стрижов В. В.* Выбор моделей глубокого обучения субоптимальной сложности // АиТ. 2018. № 8. С. 129–147.

[79] *Madala H., Ivakhnenko A.* Inductive Learning Algorithms for Complex Systems Modeling. Boca Raton: CRC Press Inc., 1994.

[80] *Hochreiter S., Schmidhuber J.* Long short-term memory // Neural Computation. 1997. V. 9. No 8. P. 1735–1780.

[81] *Alex Krizhevsky and Vinod Nair and Geoffrey Hinton* CIFAR-10 (Canadian Institute for Advanced Research) // `http://www.cs.toronto.edu/~kriz/cifar.html`

[82] *Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.* Imagenet: A large-scale hierarchical image database // IEEE conference on computer vision and pattern recognition, 2009. P. 248–255.

[83] *Kui Ren and Tianhang Zheng and Zhan Qin and Xue Liu* Adversarial Attacks and Defenses in Deep Learning // Engineering, 2020. P. 346–360.

[84] *Alex Krizhevsky, Ilya Sutskever, Geoffrey Hinton* ImageNet Classification with Depp Convolutional Neural Networks // NIPS, 2012.

[85] *Karen Simonyan and Andrew Zisserman* Very Deep Convolutional Networks for Large-Scale Image Recognition // NIPS, 2014.

[86] *Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A., Kaiser L., Polosukhin I.* Attention Is All You Need // In Advances in Neural Information Processing Systems. 2017. V. 5. P. 6000–6010.

[87] *Tom B. Brown et al* GPT3: Language Models are Few-Shot Learners // arXiv preprinted, 2020.

[88] *Yang, Ziqing and Cui, Yiming and Chen, Zhipeng and Che, Wanxiang and Liu, Ting and Wang, Shijin and Hu, Guoping* TextBrewer: An Open-Source Knowledge Distillation Toolkit for Natural Language Processing // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2020. P. 9–16.

[89] *Hinton G., Vinyals O., Dean J.* Distilling the Knowledge in a Neural Network // NIPS Deep Learning and Representation Learning Workshop. 2015.

[90] *LeCun Y., Cortes C., Burges C.* The MNIST dataset of handwritten digits, 1998. http://yann.lecun.com/exdb/mnist/index.html.

[91] *Lopez-Paz D., Bottou L., Scholkopf B., Vapnik V.* Unifying Distillation and Privileged Information // In International Conference on Learning Representations. Puerto Rico, 2016.

[92] *Xiao H., Rasul K., Vollgraf R.* Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms // arXiv preprint arXiv:1708.07747. 2017.

[93] *Wilson T., Kozareva Z., Nakov P., Rosenthal S., Stoyanov V., Ritter A.* SemEval-2013 Task 2: Sentiment Analysis in Twitter // Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). Atlanta, 2013. P. 312–320.

[94] *LeCun Y., Boser B., Denker J., Henderson D., Howard R., Hubbard W., Jackel L.* Backpropagation Applied to Handwritten Zip Code Recognition // Neural Computation. 1989. V. 1. No 4. P. 541–551.

[95] *Kingma D, Ba J.* Adam: A Method for Stochastic Optimization // arXiv preprint arXiv:1412.6980. 2014.