PhD position in Statistics & Data Science
Deep Mixtures-of-Experts for Unsupervised Feature Learning

Subject:

Unsupervised learning of feature hierarchies is becoming of broader interest in statistical machine learning and artificial intelligence, with the increasing prevalence of deep learning (Lecun et al., 2015). Deep learning of feature hierarchies, see for example (Hinton and Salakhutdinov, 2006; Hinton et al., 2006; Bengio and LeCun, 2007; Ranzato et al., 2008; Bengio, 2009), produce high level representations by considering a hierarchical architecture of stacked neural network modules, learned in an unsupervised way. This gives deep learning architectures the ability to capture different levels of representations in the data, going from low-level features to high-levels abstractions; the learned representations allow therefore to obtain very accurate results in prediction, even when using standard linear classifiers including logistic regression of linear Support Vector Machines. Deep neural networks have been developed as biologically plausible "models" for approximating humans in different tasks including object recognition and hence have biological foundations to explain how and (somewhat less) why do they work. From a mathematical and a statistical point of view, there is an increasing interest to explain the ability of such paradigm in representations and to understand and explain why do deep learning algorithms work, see for example Mallat (2016); Patel et al. (2016).

The power of such models in representations with achieved biological and statistical foundations motivates this PhD research program, which aims to investigate deep learning within a statistical Mixtures-of-Experts modeling. Mixtures-of-Experts (MoE) models, introduced by Jacobs et al. (1991), are successful neural-network architectures for modeling heterogeneous data in many statistics and machine learning problems including regression, clustering and classification. MoE are based on the divide and conquer learning principle and, from a statistical modeling point of view, are defined as fully conditional mixture models (McLachlan and Peel., 2000) where both the mixing proportions, i.e, the gating network, and the components densities, i.e, the experts network, are functions of the inputs, chosen in such a way to sufficiently represent the data. We can distinguish MoE for time series data (Chamroukhi et al., 2009) and functional data (Chamroukhi et al., 2010), MoE which are robust to atypical data (Chamroukhi, 2016a; Nguyen and McLachlan, 2016) or MoE for data with possibly skewed and non-normal distributions (Chamroukhi, 2017, 2016b; Samé et al., 2011), sparse MoE for high-dimensional data (Chamroukhi and Huynh, 2018). This makes MoE more flexible and efficient for representations than standard unconditional mixture distributions, while having a neural-network interpretation. A general review of the MoE models and their applications can be found in Nguyen and Chamroukhi (2018); Yuksel et al. (2012).

MoE have been widely studied in the statistics and machine learning literature for clustering, regression and classification of heterogeneous data. There is a very recent emphasis on incorporating these models into the construction of deep networks for feature learning (Eigen et al., 2014), in different research area of Artificial Intelligence, including, *i)* language modeling (Shazeer et al., 2017) for machine translation, *ii)* computer vision (Zhao et al., 2017) for large-scale visual recognition, and *iii)*speech (enhancement) (Chazan et al., 2017). There is also an increasing focus on the scalability of such models and on their effective use in "real" large-scale scenarios. The objective of this thesis is to contribute to this timely subject by proposing novel deep MoE models for unsupervised learning of feature hierarchies with desirable statistical properties, and the expected results of this thesis are the following:

- propose statistical deep MoE models with desirable statistical properties and study the approximation capabilities of the deep architecture;
- propose learning algorithms for representation and inference algorithms for prediction, that are effective at large-scale and study their statistical and numerical properties within a parallel high-performance computing framework;
- Study the model within three main application domains: unsupervised time series and functional data representation (Chamroukhi and Nguyen, 2018), large-scale unsupervised bio-acoustic signal representation (Bartcus et al., 2015), and the unsupervised representation of large-scale health care data (e.g fMRI imges).

Additional information:
PhD Director: Pr. Faïcel Chamroukhi: http://math.unicaen.fr/~chamroukhi/
Laboratory: The Lab of Mathematics Nicolas Oresme LMNO - UMR CNRS 6139, Caen
University: University of Caen-Normandy, Caen, France.
Required profile: Successful candidates should have a master degree in Mathematics with a major in Statistics, Machine learning, or statistical signal processing, or a closely related area. They should have strong skills in statistical learning and in programming with Matlab and/or R and/or Python. The PhD thesis as well all the research reports, including publications, will be written in English. So strong skills in English writing/speaking are needed. Expected skills include unsupervised learning, deep learning and distributed large-scale algorithms computing.
Starting date: September 2018 (for 36 months)
Application deadline: may 10, 2018
Salary: 1758 € gross per month (Granted by the French Ministry of Scientific Research). We also offer to the PhD student the possibility to teach courses in statistics/data analysis (up to 64 hours a year) for undergraduate students at the department of Mathematics & CS.
How to apply: Please send your application file (CV+transcripts of the last three academic years) in *A SINGLE .pdf FILE* to chamroukhi@unicaen.fr or fill out this form.
Please note that excellent international applications are very welcome! For this PhD research project, there is no any need to speak/understand French. For candidates who wish to learn French, free French courses are offered by the university/the doctoral school to foreign students.

References

Bartcus, M., Chamroukhi, F., and Glotin, H. (2015). Hierarchical dirichlet process hidden markov model for unsupervised bioacoustic analysis. In *2015 International Joint Conference on Neural Networks, IJCNN 2015, Killarney, Ireland, July 12-17, 2015*, pages 1–7.

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127. Also published as a book. Now Publishers, 2009.

Bengio, Y. and LeCun, Y. (2007). Scaling learning algorithms towards AI. In Bottou, L., Chapelle, O., DeCoste, D., and Weston, J., editors, *Large Scale Kernel Machines*. MIT Press.

Chamroukhi, F. (2016a). Robust mixture of experts modeling using the *t* distribution. *Neural Networks*, 79:20–36.

Chamroukhi, F. (2016b). Skew-normal mixture of experts. In *Neural Networks (IJCNN), 2016 International Joint Conference on Neuron Networks*, pages 3000–3007. IEEE.

Chamroukhi, F. (2017). Skew *t* mixture of experts. *Neurocomputing - Elsevier*, 266:390–408.

Chamroukhi, F. and Huynh, B. (2018). Regularised maximum-likelihood inference of mixture-of-experts models for high-dimensional data. In *The International Joint Conference on Neural Networks (IJCNN)*, page to appear, Rio, Brazil.

Chamroukhi, F. and Nguyen, H. D. (2018). Model-based clustering and classification of functional data. *arXiv:1803.00276v2*.

Chamroukhi, F., Samé, A., Govaert, G., and Aknin, P. (2009). Time series modeling by a regression approach based on a latent process. *Neural Networks*, 22(5-6):593–602.

Chamroukhi, F., Samé, A., Govaert, G., and Aknin, P. (2010). A hidden process regression model for functional data description. application to curve discrimination. *Neurocomputing*, 73(7-9):1210–1221.

Chazan, S. E., Goldberger, J., and Gannot, S. (2017). Speech enhancement using a deep mixture of experts. *arXiv:1703.09302*.

Eigen, D., Ranzato, M., and Sutskeve, I. (2014). Learning factored representations in a deep mixture of experts. *arXiv:1312.4314v3*.

Hinton, G. and Salakhutdinov, R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507.

Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87.

Lecun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Mallat, S. (2016). Understanding deep convolutional networks. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 374 2065:20150203.

McLachlan, G. J. and Peel., D. (2000). *Finite mixture models*. New York: Wiley.

Nguyen, H. D. and Chamroukhi, F. (2018). Practical and theoretical aspects of mixture-of-experts modeling: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, pages e1246–n/a.

Nguyen, H. D. and McLachlan, G. J. (2016). Laplace mixture of linear experts. *Computational Statistics & Data Analysis*, 93:177–191.

Patel, A. B., Nguyen, M. T., and Baraniuk, R. (2016). A probabilistic framework for deep learning. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 2558–2566. Curran Associates, Inc.

Ranzato, M. A., Boureau, Y., and Cun, Y. L. (2008). Sparse feature learning for deep belief networks. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems (NIPS) 20*, pages 1185–1192.

Samé, A., Chamroukhi, F., Govaert, G., and Aknin, P. (2011). Model-based clustering and segmentation of time series with changes in regime. *Advances in Data Analysis and Classification*, 5:301–321.

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., and Dean, J. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *ICLR 2017*. arXiv preprint arXiv:1701.06538.

Yuksel, S. E., W., J. N., and Gader, P. D. (2012). Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems*, 23(8):1177–1193.

Zhao, T., Jun Yu, Z. K., Zhang, W., and Fan, J. (2017). Deep mixture of diverse experts for large-scale visual recognition. *arXiv:1706.07901*.