1 Conclusion :

1.1 Future work :

Despite showing good performances when the learning phase has converged properly, SCHMT's performances are still undermined by a sometime poor learning quality. This observation is one of the main drivers for the upcoming work. The other main focus is to keep a well define probabilistic framework in order to be able to express the uncertainty of our model and of our predictions. Those concerns yields to considering two main lines of research for upcoming work both articulated around the variational Bayes approximation to inference.

1.1.1 Variational Bayes for hidden Markov trees :

The version of the EM algorithm considered in this report uses full Bayesian inference. Despite providing interesting results this methods is computationally expensive and only provide with a point-wise estimate of the model's parameters. A way around those issues is to use the Variational Bayes (VB) approximation for inference [Wainwright and Jordan, 2008].

The idea behind variational Bayes is to substitute the -most of the time- intractable posterior inference evaluation by an easier to solve optimization problem. This is done by approximating the target posterior by a variational distribution from a simpler parametric family. Meaning that, given a set of data $\mathbf{X}$ and the corresponding targets $\mathbf{Y}$ as well as the model's parameters Θ, the task at end is now the minimization of the Kullback-Leibler divergence between the true target posterior $P(\mathbf{Y}|\mathbf{X}, \Theta)$ and the variational distribution $q(\Theta, \mathbf{Y})$. This trick has been applied successfully to a variety of Bayesian models ***TODO-1***, and even to Hidden Markov Trees ***TODO-2***.

Recently Variational Bayes has been improved by leveraging methods borrowed to the optimisation community. ***TODO-3*** developed a scalable modification to VB using stochastic gradients for optimisation -Stochastic Variational Bayes (SVB). Even more recently, ***TODO-4*** developed a method to apply SVI to hidden Markov Models.

We propose over the next few months to develop a stochastic variational Bayes version of the EM algorithm for hidden Markov trees and apply it to scattering convolutional hidden Markov trees. This project can be breakdown into two major steps :

— First we first plan to implement a variational version of the EM algorithm for SCHMTs following [**?**]. Here again the algorithm has been developed for binary regular trees, but can simply be adapted to the non-binary non-regular trees we are interested in. We thus plan a month for the implementation and at least another month for testing and experiments. Hopefully the VB version of SCHMTs will provide good enough results for a conference paper.

- The logical next step is to develop a stochastic version of the variational Bayes expectation maximization algorithm for hidden Markov trees, leading to the development of SVB-SCHMTs. This step is expected to require more work than the previous one as the SVB framework as not yet been developed for HMTs. However we can follow the framework defined by JB. Durand [Durand et al., 2004] when adapting [Devijver, 1985] smoothed-EM for HMM to HMTs to inspire us. We plan to spend about a month on developing the algorithm -can be done in parallel of coding VB-SCHMTS, another month for coding it and then a couple months for experiments. Hopefully the SVB-SCHMTs would provide enought material to lead to both a conference and a journal paper.

Beside a potential improvement of the performance, variational inference would also provide a estimated distribution for the model's parameter, thus allowing us to have access to a measure of uncertainty for the learned model and thus discard models according to the uncertainty on their parameters.

Another interesting direction to follow for future works is to integrate the scattering convolutional hidden Markov tree into a hierarchical graphical models [Fine et al., 1998]. This framework would allow the use the SCHMT model as a node of a wider probabilistic graphical model. Using such an architecture yields to a tremendous number of possibilities. The performance in the segmentation task could be improved by adding a layer of graphical model encoding the spatial dependencies between the different labels in the scene. One could also use hierarchical models to describe a network of sensors each providing information on a targeted scene or integrate multiple source of information into the final prediction.

Finally another interesting lead would be to consider other architectures for the graphical model and even make it includes the representation learning step. This would be possible using Bayesian neural networks and probabilistic back-propagation [Hernández-Lobato and Adams, 2015] or a Bayesian flavour of back-propagation [Blundell et al., 2015].

Bibliography :

[Abdel-Hamid et al., 2012] Abdel-Hamid, O., Mohamed, A.-r., Jiang, H., and Penn, G. (2012). Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4277–4280. IEEE.

[Ackley et al., 1985] Ackley, D. H., Hinton, G. E., and Sejnowski, T. J. (1985). A learning algorithm for boltzmann machines*. *Cognitive science*, 9(1) :147–169.

[Andén and Mallat, 2011] Andén, J. and Mallat, S. (2011). Multiscale scattering for audio classification. In *ISMIR*, pages 657–662.

[Baker et al., 2014] Baker, M. J., Trevisan, J., Bassan, P., Bhargava, R., Butler, H. J., Dorling, K. M., Fielden, P. R., Fogarty, S. W., Fullwood, N. J., Heys, K. A., et al. (2014). Using fourier transform ir spectroscopy to analyze biological materials. *Nature protocols*, 9(8) :1771–1791.

[Baum et al., 1970] Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The annals of mathematical statistics*, pages 164–171.

[Beyer et al., 1999] Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U. (1999). When is "nearest neighbor" meaningful ? In *Database Theory—ICDT'99*, pages 217–235. Springer.

[Bishop, 2006] Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

[Blundell et al., 2015] Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015). Weight uncertainty in neural networks. *arXiv preprint arXiv :1505.05424*.

[Bracewell, 1965] Bracewell, R. (1965). The fourier transform and iis applications. *New York*.

[Brailean and Katsaggelos, 1996] Brailean, J. C. and Katsaggelos, A. K. (1996). Recursive map displacement field estimation and its applications. In *Image Processing, 1996. Proceedings., International Conference on*, volume 1, pages 917–920. IEEE.

[Bruna, 2012] Bruna, J. (2012). Operators commuting with diffeomorphisms.

[Bruna and Mallat, 2010] Bruna, J. and Mallat, S. (2010). Classification with scattering operators. *arXiv preprint arXiv :1011.3023*.

[Bruna and Mallat, 2013] Bruna, J. and Mallat, S. (2013). Invariant scattering convolution networks. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8) :1872–1886.

[Chudacek et al., 2014] Chudacek, V., Talmon, R., Anden, J., Mallat, S., Coifman, R., Abry, P., and Doret, M. (2014). Low dimensional manifold embedding for scattering coefficients of intrapartum fetale heart rate variability. In *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, pages 6373–6376. IEEE.

[Cover and Hart, 1967] Cover, T. M. and Hart, P. E. (1967). Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1) :21–27.

[Crouse et al., 1998] Crouse, M. S., Nowak, R. D., and Baraniuk, R. G. (1998). Wavelet-based statistical signal processing using hidden markov models. *Signal Processing, IEEE Transactions on*, 46(4) :886–902.

[De Chazal et al., 2000] De Chazal, P., Celler, B., and Reilly, R. (2000). Using wavelet coefficients for the classification of the electrocardiogram. In *Engineering in Medicine and Biology Society, 2000. Proceedings of the 22nd Annual International Conference of the IEEE*, volume 1, pages 64–67. IEEE.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

[Devijver, 1985] Devijver, P. A. (1985). Baum's forward-backward algorithm revisited. *Pattern Recognition Letters*, 3(6) :369–373.

[DeVore et al., 1992] DeVore, R. A., Jawerth, B., and Lucier, B. J. (1992). Image compression through wavelet transform coding. *Information Theory, IEEE Transactions on*, 38(2) :719–746.

[Donoho, 1993] Donoho, D. L. (1993). Unconditional bases are optimal bases for data compression and for statistical estimation. *Applied and computational harmonic analysis*, 1(1) :100–115.

[Dstl, 2009] Dstl (2009). Dstl datasets. `https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/326061/DefenceReporter_Winter2009.pdf`. Accessed : 04-11-2015.

[Duncan et al., 2000] Duncan, T. E., Hu, Y., and Pasik-Duncan, B. (2000). Stochastic calculus for fractional brownian motion i. theory. *SIAM Journal on Control and Optimization*, 38(2) :582–612.

[Durand et al., 2004] Durand, J.-B., Goncalves, P., and Guédon, Y. (2004). Computational methods for hidden markov tree models-an application to wavelet trees. *Signal Processing, IEEE Transactions on*, 52(9) :2551–2560.

[Ephraim and Merhav, 2002] Ephraim, Y. and Merhav, N. (2002). Hidden markov processes. *Information Theory, IEEE Transactions on*, 48(6) :1518–1569.

[Fine et al., 1998] Fine, S., Singer, Y., and Tishby, N. (1998). The hierarchical hidden markov model : Analysis and applications. *Machine learning*, 32(1) :41–62.

[Forney Jr, 1973] Forney Jr, G. D. (1973). The viterbi algorithm. *Proceedings of the IEEE*, 61(3) :268–278.

[Gilks, 2005] Gilks, W. R. (2005). *Markov chain monte carlo*. Wiley Online Library.

[Haveliwala and Kamvar, 2003] Haveliwala, T. and Kamvar, S. (2003). The second eigenvalue of the google matrix. *Stanford University Technical Report*.

[Heckerman, 1998] Heckerman, D. (1998). *A tutorial on learning with Bayesian networks*. Springer.

[Hernández-Lobato and Adams, 2015] Hernández-Lobato, J. M. and Adams, R. P. (2015). Probabilistic backpropagation for scalable learning of bayesian neural networks. *arXiv preprint arXiv :1502.05336*.

[Hinton et al., 2012] Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv :1207.0580*.

[Hörmander, 1971] Hörmander, L. (1971). Fourier integral operators. i. *Acta mathematica*, 127(1) :79–183.

[Jarrett et al., 2009] Jarrett, K., Kavukcuoglu, K., Ranzato, M., and LeCun, Y. (2009). What is the best multi-stage architecture for object recognition ? In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 2146–2153. IEEE.

[Jordan, 2002] Jordan, A. (2002). On discriminative vs. generative classifiers : A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14 :841.

[Kemeny and Snell, 1960] Kemeny, J. G. and Snell, J. L. (1960). *Finite markov chains*, volume 356. van Nostrand Princeton, NJ.

[Kingsbury, 2001] Kingsbury, N. (2001). Complex wavelets for shift invariant analysis and filtering of signals. *Applied and computational harmonic analysis*, 10(3) :234–253.

[Kreutz-Delgado et al., 2003] Kreutz-Delgado, K., Murray, J. F., Rao, B. D., Engan, K., Lee, T.-W., and Sejnowski, T. J. (2003). Dictionary learning algorithms for sparse representation. *Neural computation*, 15(2) :349–396.

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

[LeCun, 2015] LeCun, Y. (2015). Personal webpage : State of the art on mnist. `http://yann.lecun.com/exdb/mnist/`. Accessed : 04-11-2015.

[LeCun and Bengio, 1995] LeCun, Y. and Bengio, Y. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10).

[LeCun et al., 2010] LeCun, Y., Kavukcuoglu, K., and Farabet, C. (2010). Convolutional networks and applications in vision. In *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, pages 253–256. IEEE.

[Lee et al., 1996] Lee, N., Huynh, Q., and Schwartz, S. (1996). New method of linear time-frequency analysis for signal detection. In *Time-Frequency and Time-Scale Analysis, 1996., Proceedings of the IEEE-SP International Symposium on*, pages 13–16. IEEE.

[Lin and Zha, 2008] Lin, T. and Zha, H. (2008). Riemannian manifold learning. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(5) :796–809.

[Lohmiller and Slotine, 1998] Lohmiller, W. and Slotine, J.-J. E. (1998). On contraction analysis for non-linear systems. *Automatica*, 34(6) :683–696.

[Mallat, 1999] Mallat, S. (1999). *A wavelet tour of signal processing*. Academic press.

[Mallat, 2012] Mallat, S. (Oct. 2012). Group invariant scattering. *Communications in Pure and Applied Mathematics*, 65(10) :1331–1398.

[Margaritis, 2003] Margaritis, D. (2003). *Learning Bayesian network model structure from data*. PhD thesis, US Army.

[Nelson and Kingsbury, 2010] Nelson, J. and Kingsbury, N. (2010). Fractal dimension based sand ripple suppression for mine hunting with sidescan sonar. In *International conference on synthetic aperture sonar and synthetic aperture radar*.

[Nilsson, 1998] Nilsson, N. J. (1998). *Artificial intelligence : a new synthesis*. Morgan Kaufmann.

[Oyallon and Mallat, 2014] Oyallon, E. and Mallat, S. (2014). Deep roto-translation scattering for object classification. *arXiv preprint arXiv :1412.8659*.

[Oyallon et al., 2013] Oyallon, E., Mallat, S., and Sifre, L. (2013). Generic deep networks with wavelet scattering. *arXiv preprint arXiv :1312.5940*.

[Platt et al., 1999] Platt, J. et al. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3) :61–74.

[Rish et al., 2002] Rish, I., Brodie, M., and Ma, S. (2002). Efficient fault diagnosis using probing. In *AAAI Spring Symposium on Information Refinement and Revision for Decision Making*.

[Schölkopf et al., 1997] Schölkopf, B., Sung, K.-K., Burges, C. J., Girosi, F., Niyogi, P., Poggio, T., and Vapnik, V. (1997). Comparing support vector machines with gaussian kernels to radial basis function classifiers. *Signal Processing, IEEE Transactions on*, 45(11) :2758–2765.

[Sifre and Mallat, 2013] Sifre, L. and Mallat, S. (2013). Rotation, scaling and deformation invariant scattering for texture discrimination.

[Simard et al., 2003] Simard, P. Y., Steinkraus, D., and Platt, J. C. (2003). Best practices for convolutional neural networks applied to visual document analysis. In *null*, page 958. IEEE.

[Smolensky, 1986] Smolensky, P. (1986). Information processing in dynamical systems : Foundations of harmony theory.

[Szegedy et al., 2013] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv :1312.6199.*

[Ulusoy and Bishop, 2005] Ulusoy, I. and Bishop, C. M. (2005). Generative versus discriminative methods for object recognition. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 258–265. IEEE.

[Viterbi, 1967] Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE Transactions on*, 13(2) :260–269.

[Wainwright and Jordan, 2008] Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2) :1–305.

[Waldspurger et al., 2015] Waldspurger, I., d'Aspremont, A., and Mallat, S. (2015). Phase recovery, maxcut and complex semidefinite programming. *Mathematical Programming*, 149(1-2) :47–81.

[Zhang et al., 2012] Zhang, Z., Wang, J., and Zha, H. (2012). Adaptive manifold learning. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(2) :253–265.