1 Conclusion :

1.1 Future work :

There are many directions to be studied to improve the SCHMTs. For sake of time and personal interests we limit ourself for the remaining time of this PhD program to four main sub-projects at either short or long term.

1.1.1 Inference :

Even though theoretically SCHMTs are much more versatile, this document only reports classification task. We would like to test the other inference options offered by our model.

An interesting direction would be to use sensitivity analysis (see Sub-section **??**) to prune the tree and remove the least informative nodes and branches of the scattering convolutional network. Reducing the complexity of the tree would have several advantages. First a smaller tree would reduce the computational time required to perform inference. Second a smaller tree is expected to perform better when used for discrimination as the pruning will transform the over-complete representation generated by the SCHMT into a simpler, less redundant representation.

Another interesting task would be generation. This means generating the most likely image associated to a given class. This is a two steps process. First one needs to infer the most likely scattering tree given a class. This can be done fairly easily by using the SCHMT model's learned parameters. One has expressed the most likely scattering convolutional network given this class. The second step is to invert the scattering transform to generate an image. Unfortunately, the invertibility of the scattering transform is still an open question [Cheng et al., 2015].

All the building blocs for implementing the sensitivity analysis are ready, we just need to define the cleverest way to implement it. Hence we estimate at about a few weeks the implementation time —Finish mid-month 16. Then testing should take at least a month as the experiments will be fairly long to run (training HMTs for all possible pruning) —Finish end of month 17. Good results in pruning would provide sufficient material for a conference article, especially since it can be combined with other improvements (see subsection 1.1.2). Regarding data generation we prefer waiting until Séphane Mallat has properly defined the invertibility of the scattering transform before taking any action on this.

1.1.2 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 provides 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 $\mathbf{X}$, the corresponding targets $\mathbf{Y}$ and the model's parameters Θ, the task at hand 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 [Attias, 2000] [Wainwright and Jordan, 2008], and even to Hidden Markov Trees [Dasgupta and Carin, 2006] [Olariu et al., 2009].

Recently Variational Bayes has been improved by leveraging methods borrowed to the optimisation community. Hoffman et al. [2013] have developed a scalable modification of VB using stochastic gradients for optimisation —Stochastic Variational Bayes (SVB). Even more recently, Foti et al. [2014] have developed a method to apply SVB 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 to apply it to scattering convolutional hidden Markov trees. This project can be broken down into two major steps :

— First we plan to implement a variational version of the EM algorithm for SCHMTs following [Olariu et al., 2009]. Here again the algorithm has been developed for binary regular trees, but can be adapted to the non-binary non-regular trees we are interested in. We thus forecast a month for the implementation and at least two months for testing and experiments. Hopefully the VB version of SCHMTs will provide good enough results for a conference paper —Implementation finished by the end of month 16, experiments by the end of month 18.

— Second we want 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 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 will provide enough material to lead to both a conference and a journal paper —Theory finished by end of month 18 and experiments by the end of month 20.

Variational Bayes provides a powerful framework to simplify and improve the quality of the parameters learned for our graphical model. However the quality of approximation made in the variational Bayes framework is highly correlated to the chosen variational distribution family. Recently Ranganath et al. [2015] have developed a framework where the variational approximation is augmented with a prior on its parameters. This method offers a mean to learn the variational distribution family. We would like to adapt this framework to hidden Markov models and more specifically to SVB-SCHMTs —Theory finished by end of month 21, coding by end of month 22, experiments by end of month 24.

1.1.3 Hierarchical graphical models :

One of the richness of graphical models is that they can be incorporated into a broader networks as a nodes of this structure. This idea leads to hierarchical graphical models [Fine et al., 1998]. Using such an architecture opens tremendous number of possibilities. In Section **??** we have mentioned that the UDRC dataset also provided meta-data among which are the

spatial coordinates of the sensor for each imagery ; a hierarchical graphical model would allow us to leverage those information by modelling the acquisition area as a graphical model whose node would be the outcome of the SCHMTs. Another possibility would be to work with an array of sensor providing information on the same target

To a certain extend this project can be run in parallel to the "variational Bayes" one. We aim at developing a variational framework for this type of model and a procedure for training the full hierarchical model at once —as opposed to sequential training of each "layer" of the hierarchy. The year to come will be used to develop the theory and then a long time will be dedicated to experiments and testing —Finish expected at end of month 28.

1.1.4 Bayesian neural network :

Finally we would like to consider other architectures for the graphical model and make it include the representation learning step. This would be possible using Bayesian neural networks and probabilistic back-propagation [Hernández-Lobato and Adams, 2015] or a variational approach [Blundell et al., 2015]. Those neural networks encode the full distribution over their weights —instead of the standard point wise estimate— and thus possess all the advantages of graphical models.

The idea is to try to adapt the methods developed for SCHMTs to those networks. To do so we plan on studying the theory of those networks over the next year and to adapt some of our improvement to them during the last year of the PhD program —Month 25 and onward.

Bibliography :

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

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

Attias, H. (2000). A variational bayesian framework for graphical models. *Advances in neural information processing systems*, 12(1-2) :209–215.

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

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

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

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

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, J. (2012). Operators commuting with diffeomorphisms.

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

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

Cheng, X., Chen, X., and Mallat, S. (2015). Deep haar scattering networks. *arXiv preprint arXiv :1509.09187*.

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

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.

Dasgupta, N. and Carin, L. (2006). Texture analysis with variational hidden markov trees. *IEEE transactions on signal processing*, 54(6) :2352–2356.

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

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, 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 datasets. `https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/326061/DefenceReporter_Winter2009.pdf`. Accessed : 04-11-2015.

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

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

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

Foti, N., Xu, J., Laird, D., and Fox, E. (2014). Stochastic variational inference for hidden markov models. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 27*, pages 3599–3607. Curran Associates, Inc.

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

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

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

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

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*.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1) :1303–1347.

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

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

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

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

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

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

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

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

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

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

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

Olariu, V., Coca, D., Billings, S. A., Tonge, P., Gokhale, P., Andrews, P. W., and Kadirkamanathan, V. (2009). Modified variational bayes em estimation of hidden markov tree model of cell lineages. *Bioinformatics*, 25(21) :2824–2830.

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

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

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.

Ranganath, R., Tran, D., and Blei, D. M. (2015). Hierarchical variational models. *arXiv preprint arXiv :1511.02386*.

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

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

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

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

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