Deep Bayesian Learning and Understanding

Jen-Tzung Chien
Department of Electrical and Computer Engineering
National Chiao Tung University, Hsinchu, Taiwan
jtchien@nctu.edu.tw

1 Motivation

Given the current growth in research and related emerging technologies in machine learning and deep learning, it is timely to introduce this tutorial to a large number of researchers and practitioners who are attending COLING 2018 and working on statistical models, deep neural networks, sequential learning and natural language understanding. To the best of our knowledge, there is no similar tutorial presented in previous ACL/COLING/EMNLP/NAACL. This three-hour tutorial will concentrate on a wide range of theories and applications and systematically present the recent advances in deep Bayesian and sequential learning which are impacting the communities of computational linguistics, human language technology and machine learning for natural language processing.

2 Tutorial description

This tutorial introduces the advances in deep Bayesian learning with abundant applications for natural language understanding ranging from speech recognition (Saon and Chien, 2012; Chan et al., 2016) to document summarization (Chang and Chien, 2009), text classification (Blei et al., 2003; Zhang et al., 2015), text segmentation (Chien and Chueh, 2012), information extraction (Narasimhan et al., 2016), image caption generation (Vinyals et al., 2015; Xu et al., 2015), sentence generation (Li et al., 2016b), dialogue control (Zhao and Eskenazi, 2016; Li et al., 2016a), sentiment classification, recommendation system, question answering (Sukhbaatar et al., 2015) and machine translation (Bahdanau et al., 2014), to name a few. Traditionally, "deep learning" is taken to be a learning process where the inference or optimization is based on the real-valued deterministic model. The "semantic structure" in words, sentences, entities, actions and documents drawn from a large vocabulary may not be well expressed or correctly optimized in mathematical logic or computer programs. The "distribution function" in discrete or continuous latent variable model for natural language may not be properly decomposed or estimated in model inference. This tutorial addresses the fundamentals of statistical models and neural networks, and focus on a series of advanced Bayesian models and deep models including hierarchical Dirichlet process (Teh et al., 2006), Chinese restaurant process (Blei et al., 2010), hierarchical Pitman-Yor process (Teh, 2006), Indian buffet process (Ghahramani and Griffiths, 2005), recurrent neural network (Mikolov et al., 2010; Van Den Oord et al., 2016), long short-term memory (Hochreiter and Schmidhuber, 1997; Cho et al., 2014), sequence-to-sequence model (Sutskever et al., 2014), variational auto-encoder (Kingma and Welling, 2014), generative adversarial network (Goodfellow et al., 2014), attention mechanism (Chorowski et al., 2015; Seo et al., 2016), memory-augmented neural network (Graves et al., 2014; Graves et al., 2014), stochastic neural network (Bengio et al., 2014; Miao et al., 2016), predictive state neural network (Downey et al., 2017), policy gradient (Yu et al., 2017) and reinforcement learning (Mnih et al., 2015). We present how these models are connected and why they work for a variety of applications on symbolic and complex patterns in natural language. The variational inference and sampling method are formulated to tackle the optimization for complicated models (Rezende et al., 2014). The word and sentence embeddings, clustering and co-clustering are merged with linguistic and semantic constraints. A series of case studies are presented to tackle different issues in deep Bayesian learning and understanding. At last, we point out a number of directions and outlooks for future studies.

3 Tutorial outline

- Introduction
  - Motivation and background
  - Probabilistic models
  - Neural networks
  - Modern natural language models
- Bayesian Learning
  - Inference and optimization
  - Variational Bayesian (VB) inference
  - Monte Carlo Markov chain (MCMC) inference
  - Bayesian nonparametrics (BNP)
  - Hierarchical theme and topic model
  - Hierarchical Pitman-Yor-Dirichlet process
  - Nested Indian buffet process
- Deep Learning
  - Deep unfolded topic model
  - Gated recurrent neural network
  - Bayesian recurrent neural network (RNN)
    (Coffee Break)
  - Sequence-to-sequence learning
  - Convolutional neural network (CNN)
  - Dilated recurrent neural network
  - Generative adversarial network (GAN)
  - Variational auto-encoder (VAE)
- Advances in Deep Sequential Learning
  - Memory-augmented neural network
  - Neural variational text processing
  - Neural discrete representation learning
  - Recurrent ladder network
  - Stochastic recurrent network
  - Predictive-state recurrent neural network
  - Sequence generative adversarial network
  - Deep reinforcement learning & understanding
- Summarization and Future Trend

4 Description of tutorial content

The presentation of this tutorial is arranged into five parts. First of all, we share the current status of researches on natural language understanding, statistical modeling and deep neural network and explain the key issues in deep Bayesian learning for discrete-valued observation data and latent semantics. A new paradigm called the symbolic neural learning is introduced to extend how data analysis is performed from language processing to semantic learning and memory networking. Secondly, we address a number of Bayesian models ranging from latent variable model to VB inference (Chien and Chang, 2014; Chien and Chueh, 2011; Chien, 2015b), MCMC sampling (Watanabe and Chien, 2015) and BNP learning (Chien, 2016; Chien, 2015a; Chien, 2018) for hierarchical, thematic and sparse topics from natural language. In the third part, a series of deep models including deep unfolding (Chien and Lee, 2018), Bayesian RNN (Gal and Ghahramani, 2016; Chien and Ku, 2016), sequence-to-sequence learning (Graves et al., 2006; Gehring et al., 2017), CNN (Kalchbrenner et al., 2014; Xingjian et al., 2015; Dauphin et al., 2017), GAN (Tsai and Chien, 2017) and VAE are introduced. The coffee break is arranged within this part. Next, the fourth part focuses on a variety of advanced studies which illustrate how deep Bayesian learning is developed to infer the sophisticated recurrent models for natural language understanding. In particular, the memory network (Weston et al., 2015; Chien and Lin, 2018), neural variational learning (Serban et al., 2017; Chung et al., 2015), neural discrete representation (Jang et al., 2016; Maddison et al., 2016; van den Oord et al., 2017), recurrent ladder network (Rasmus et al., 2015; Prémont-Schwarz et al., 2017; Sønderby et al., 2016), stochastic neural network (Fraccaro et al., 2016; Goyal et al., 2017; Shabanian et al., 2017), Markov recurrent neural network (Venkatraman et al., 2017; Kuo and Chien, 2018), sequence GAN (Yu et al., 2017) and reinforcement learning (Tegho et al., 2017) are introduced in various deep models which open a window to more practical tasks, e.g. reading comprehension, sentence generation, dialogue system, question answering and machine translation. In the final part, we spotlight on some future directions for deep language understanding which can handle the challenges of big data, heterogeneous condition and dynamic system. In particular, deep learning, structural learning, temporal modeling, long history representation and stochastic learning are emphasized. Slides of this tutorial are available at http://chien.cm.nctu.edu.tw/home/coling/.

5 Instructor

Jen-Tzung Chien received his Ph.D. degree in electrical engineering from National Tsing Hua University, Hsinchu, Taiwan, in 1997. He is now with the Department of Electrical and Computer Engineering and the Department of Computer Science at the National Chiao Tung University, Hsinchu, where he is currently a Chair Professor. He was a visiting researcher with the IBM T. J. Watson Research Center, Yorktown Heights, NY, in 2010. His research interests include machine learning, deep learning, natural language processing and computer vision. Dr. Chien served as the associate editor of the IEEE Signal Processing Letters in 2008-2011, the guest editor of the IEEE Transactions on Audio, Speech and Language Processing in 2012, the organization committee member of ICASSP 2009, ISCSLP 2016, the area coordinator of Interspeech 2012, EUSIPCO 2017, 2018, the program chair of ISCSLP 2018, the general chair of MLSP 2017, and currently serves as an elected member of IEEE Machine Learning for Signal Processing Technical Committee. He received the Best Paper Award of IEEE Automatic Speech Recognition and Understanding Workshop in 2011 and the AAPM Farrington Daniels Paper Award in 2018. He has published extensively including the book "Bayesian Speech and Language Processing", Cambridge University Press, 2015. He was the tutorial speaker for APSIPA 2013, ISCSLP 2014, Interspeech 2013, 2016 and ICASSP 2012, 2015 and 2017. (http://chien.cm.nctu.edu.tw/)

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Yoshua Bengio, Eric Thibodeau-Laufer, Guillaume Alain, and Jason Yosinski. 2014. Deep generative stochastic networks trainable by backprop. In *Proc. of International Conference on Machine Learning*.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(3):993–1022, Jan.

David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. 2010. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57(2). Article 7.

W. Chan, N. Jaitly, Q. Le, and O. Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4960–4964.

Ying-Lang Chang and Jen-Tzung Chien. 2009. Latent Dirichlet learning for document summarization. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1689–1692.

Jen-Tzung Chien and Ying-Lan Chang. 2014. Bayesian sparse topic model. *Journal of Signal Processing Systems*, 74(3):375–389.

Jen-Tzung Chien and Chuang-Hua Chueh. 2011. Dirichlet class language models for speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(3):482–495.

Jen-Tzung Chien and Chuang-Hua Chueh. 2012. Topic-based hierarchical segmentation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):55–66.

Jen-Tzung Chien and Yuan-Chu Ku. 2016. Bayesian recurrent neural network for language modeling. *IEEE Transactions on Neural Networks and Learning Systems*, 27(2):361–374.

Jen-Tzung Chien and Chao-Hsi Lee. 2018. Deep unfolding for topic models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(2):318–331.

Jen-Tzung Chien and Ting-An Lin. 2018. Supportive attention in end-to-end memory networks. In *Proc. of IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6.

Jen-Tzung Chien. 2015a. Hierarchical Pitman-Yor-Dirichlet language model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(8):1259–1272.

Jen-Tzung Chien. 2015b. Laplace group sensing for acoustic models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(5):909–922.

Jen-Tzung Chien. 2016. Hierarchical theme and topic modeling. *IEEE Transactions on Neural Networks and Learning Systems*, 27(3):565–578.

Jen-Tzung Chien. 2018. Bayesian nonparametric learning for hierarchical and sparse topics. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(2):422–435.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proc. of Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734.

Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. In *Advances in Neural Information Processing Systems*, pages 577–585.

Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. 2015. A recurrent latent variable model for sequential data. In *Advances in Neural Information Processing Systems*, pages 2980–2988.

Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. Language modeling with gated convolutional networks. In *Proc. of International Conference on Machine Learning*, pages 933–941.

Carlton Downey, Ahmed Hefny, Byron Boots, Geoffrey J Gordon, and Boyue Li. 2017. Predictive state recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 6055–6066.

Marco Fraccaro, Søren Kaae Sø nderby, Ulrich Paquet, and Ole Winther. 2016. Sequential neural models with stochastic layers. In *Advances in Neural Information Processing Systems*, pages 2199–2207.

Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 1019–1027.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *Proc. of International Conference on Machine Learning*, pages 1243–1252.

Zoubin Ghahramani and Thomas L. Griffiths. 2005. Infinite latent feature models and the Indian buffet process. In *Advances in Neural Information Processing Systems*, volume 18, pages 475–482.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680.

Anirudh Goyal, Alessandro Sordoni, Marc-Alexandre Côté, Nan Ke, and Yoshua Bengio. 2017. Z-forcing: Training stochastic recurrent networks. In *Advances in Neural Information Processing Systems 30*, pages 6713–6723.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proc. of International Conference on Machine Learning*, pages 369–376.

Alex Graves, Greg Wayne, and Ivo Danihelka. 2014. Neural Turing machines. *arXiv preprint arXiv:1410.5401*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with Gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. In *Proc. of Annual Meeting of the Association for Computational Linguistics*, pages 655–665.

Diederik P Kingma and Max Welling. 2014. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*.

Che-Yu Kuo and Jen-Tzung Chien. 2018. Markov recurrent neural networks. In *Proc. of IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6.

Jiwei Li, Alexander H Miller, Sumit Chopra, Marc'Aurelio Ranzato, and Jason Weston. 2016a. Learning through dialogue interactions by asking questions. *arXiv preprint arXiv:1612.04936*.

Jiwei Li, Will Monroe, Alan Ritter, and Dan Jurafsky. 2016b. Deep reinforcement learning for dialogue generation. In *Proc. of Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. 2016. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*.

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *Proc. of International Conference on Machine Learning*, pages 1727–1736.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Proc. of Annual Conference of International Speech Communication Association*, pages 1045–1048, Sep.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

Karthik Narasimhan, Adam Yala, and Regina Barzilay. 2016. Improving information extraction by acquiring external evidence with reinforcement learning. In *Proc. of Conference on Empirical Methods in Natural Language Processing*, pages 2355–2365.

Isabeau Prémont-Schwarz, Alexander Ilin, Tele Hao, Antti Rasmus, Rinu Boney, and Harri Valpola. 2017. Recurrent ladder networks. In *Advances in Neural Information Processing Systems*, pages 6011–6021.

Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. 2015. Semi-supervised learning with ladder networks. In *Advances in Neural Information Processing Systems*, pages 3546–3554.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *Proc. of International Conference on Machine Learning*, pages 1278–1286.

George Saon and Jen-Tzung Chien. 2012. Large-vocabulary continuous speech recognition systems: A look at some recent advances. *IEEE Signal Processing Magazine*, 29(6):18–33, Nov.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Iulian V. Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proc. of AAAI Conference on Artificial Intelligence*, pages 3295–3301.

Samira Shabanian, Devansh Arpit, Adam Trischler, and Yoshua Bengio. 2017. Variational Bi-LSTMs. *arXiv preprint arXiv:1711.05717*.

Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. 2016. Ladder variational autoencoders. In *Advances in Neural Information Processing Systems*, pages 3738–3746.

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. 2015. End-to-end memory networks. In *Advances in Neural Information Processing Systems*, pages 2440–2448.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112. .

Christopher Tegho, Paweł Budzianowski, and Milica Gašić. 2017. Uncertainty estimates for efficient neural network-based dialogue policy optimisation. *arXiv preprint arXiv:1711.11486*.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006. Hierarchical Dirichlet process. *Journal of the American Statistical Association*, 101(476):1566–1581, Dec.

Yee Whye Teh. 2006. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proc. of International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistics*, pages 985–992.

Jen-Chieh Tsai and Jen-Tzung Chien. 2017. Adversarial domain separation and adaptation. In *Proc. of IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6.

Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, pages 6309–6318.

Arun Venkatraman, Nicholas Rhinehart, Wen Sun, Lerrel Pinto, Martial Hebert, Byron Boots, Kris Kitani, and J Bagnell. 2017. Predictive-state decoders: Encoding the future into recurrent networks. In *Advances in Neural Information Processing Systems*, pages 1172–1183.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164.

Shinji Watanabe and Jen-Tzung Chien. 2015. *Bayesian Speech and Language Processing*. Cambridge University Press.

Jason Weston, Sumit Chopra, and Antoine Bordes. 2015. Memory networks. In *Proc. of International Conference on Learning Representation*.

Shi Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems*, pages 802–810.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proc. of International Conference on Machine Learning*, pages 2048–2057.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. SeqGAN: sequence generative adversarial nets with policy gradient. In *Prof. of AAAI Conference on Artificial Intelligence*, volume 31, pages 2852–2858.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, pages 649–657.

Tiancheng Zhao and Maxine Eskenazi. 2016. Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning. *arXiv preprint arXiv:1606.02560*.