CS 760 Fall 2017: Example Final Project Topics

Yingyu Liang
yliang@cs.wisc.edu
Computer Sciences Department, University of Wisconsin-Madison

This note provides some example research topics for the final projects in the course CS 760 Machine Learning. It should be emphasized that these examples are limited to the knowledge scope and interests of the author, and only reflect a small fraction of the ongoing research topics in the active field of machine learning. Also, these topics are discussed at a high level and should be regarded as general directions rather than concrete research problems.

1 Theory

These topics focus on building formal models and providing theoretical analysis and guarantees.

1.1 Topic: Generalization bounds for deep learning

Recent studies showed some surprising observations about the generalization of deep neural networks (Zhang et al., 2016): The same architecture and regularizations lead to small generalization gaps when trained on natural images but large generalization gap when trained on random noisy inputs. The experimental results suggest that for random inputs the deep neural networks are simply memorizing the input and thus have poor generalization, while for natural images things are drastically different. However, traditional generalization bounds are independent of data distributions and thus fail to clearly explain this phenomenon. How to bound the generalization gap in a data-dependent or even training algorithm dependent way?

Example work: (Mou et al., 2017; Kawaguchi et al., 2017)

1.2 Topic: Distributed learning

Due to the explosion of data and the increasing capacity of learning models, many large scale machine learning applications adopts distributed computing to speed up the training. Theoretical guarantees under fairly mild conditions are known, and practical distributed learning platforms are deployed. However, much remains unclear for distributed learning. In many scenarios (especially when the computation is distributed over many machines), the speedup is far away from the ideal. Two key obstacles are the communication cost and that some machines are lagging behind. How to model these in a clean way and design algorithms to address them?

Example work: (Balcan et al., 2014; Pan et al., 2016; Recht et al., 2011)

1.3 Topic: Privacy and Fairness in machine learning

Machine learning methods are usually designed with performance (accuracy, speed, robustness to noise, etc) in mind. However, when they are deployed as practical systems, they may face social norms such as privacy and fairness. How to formalize privacy and fairness in the machine learning setting? How can a machine learning method be successfully deployed in applications without leaking sensitive information in the training data? How can one build a machine learning system whose predictions will guarantee fairness for the people affected? These are active research directions in recent years.

Privacy has been well understood to a deep extent. Especially, the notion of differential privacy is very successful both theoretically and practically. See (Dwork and Roth, 2014; Dwork and Cynthia, 2006).

Fairness has been studied before but recently has attracted a lot more focus. There are still debates about the formal definition, and much needs to be done to design efficient algorithms to achieve fairness. See (Kusner et al., 2017; Hardt et al., 2016; Dwork et al., 2011).

2 Empirical Studies

These topics focus on the empirical properties and performance of machine learning methods, e.g., when and why they work and how to improve them.

2.1 Topic: Unsupervised learning of representations

In supervised deep learning, the hidden layers are regarded as new representations of the input, which can allow to learn better classifiers/predictors. By using sufficiently many labeled data, recent deep neural network models can often get representations that leads to state-of-the-art performance. Sometimes the representations are also semantically meaningful, e.g., over images, the lower level hidden layers can discover line patterns that are also important for human vision systems, and higher level hidden layers can discover more complicated patterns. Are there similar phenomena for other type of data? Are these representations a summaries of the input data according to some general and unified principles? Can one design unsupervised learning systems that do not require human knowledge but use the general principles to discover the same representations?

Example work: (Hinton and Salakhutdinov, 2006; Le, 2013; Salakhutdinov, 2015)

2.2 Topic: Adversarial examples and robustness

Recent studies show a surprising weakness of some deep learning models: adding very small adversarial noise to the input can lead to drastic change in the output of the model. This caught a lot of attention since it presents an obstacle for the deployment of the such learning systems in practical scenarios, e.g., such adversarial attacks to the perception module of a self driving car can cause the machine to make severe mistakes and lead to great threats to properties and humans. Can we improve the learning methods so that the resulting systems are robust to such attacks? Will such robustness come with sacrifice of the performance or improve the performance?

Example work: (Nguyen et al., 2014; Lu et al.; Mądry et al.)

2.3 Topic: Interpretability and model extraction

As the applications that use machine learning become more and more complicated, the machine learning models used necessarily become more and more sophisticated (e.g., to encode more complicated prior knowledge, to achieve better performance, etc.). However, this comes with a price: such machine learning are hard to interpret and have to be treated mostly as black boxes even for the system designer. Can we design principle methods to interpret the black boxes? Can this be done efficiently, exploiting existing understanding about the models?

Example work: (Wei Koh and Liang; Kim et al.; Ribeiro et al.)

2.4 Topic: GAN for discrete data

Generative Adversarial Network (GAN) is a recently proposed framework for training deep generative models, whose goal is to take as input random noise and output data with distribution close to that of real world data. This framework achieves promising results, especially for producing real looking images. However, applying it to discrete data (such as those in Natural Language Processing) is much more difficult, and one

key reason is the difficulty to get meaningful gradient updates on discrete data. Can one design a general principle, which can be used to upgrade the framework and achieve desired performance on discrete data?

Existing work: (Rajeswar et al., 2017; Zhang et al., 2017)

3 Applications

These topics focus on practical applications on real world data, e.g., what are some useful properties of the data from the application, how to design a new system or improve the performance of existing systems for the application.

3.1 Topic: Sentence/paragraph/document embeddings

Many recent advances in Natural Language Processing build on top of word embeddings, i.e., mapping words as vectors in some low dimensional space. These embeddings are built via unsupervised methods and are (almost) universal in the sense that they can be used in many different NLP tasks. A natural question is whether one can do embedding for larger piece of text like sentences, paragraphs, or even documents. Although many existing methods implicitly get embeddings for them, they are either supervised or for specific tasks. Unsupervised learning of a universal embedding scheme is a largely open question.

Example work: (Arora et al., 2017; Wieting and Gimpel, 2017; Gan et al.)

3.2 Topic: Sentiment analysis

Sentiment analysis aims to determine the sentiment in pieces of text. It is practically important in many scenarios such as product recommendation, and is also theoretically interesting since it connects computers, text, and the elusive emotions of humans. Recent popular approaches of using word embeddings have known drawbacks due to that many word embedding approaches lead to similar embeddings for antonyms. And much of the effort is devoted to analyzing simple sentiment types, such as negative and positive, but little has been achieved for fine grain sentiment types. Can one develop word embeddings more suitable for sentiment analysis? How to analyze fine grain sentiment types?

Example work: (Socher et al.; Mozetic et al., 2016)

3.3 Topic: Image stylization

An interesting application of deep neural networks is to transform a given image to one in a certain artistic style, like that of Vincent Willem van Gogh. Besides a demonstration of the power of deep networks and a piece of amusing anecdote, it also gives potential hints to how the networks decompose the styles and other content of the images, and even raises the far reaching question about whether computers can do creative tasks instead of just intelligent tasks. Are such stylization simply by imposing learned stylized patterns onto the original images, or involving something deeper? Besides styles, can one extract and transfer even more abstract contents like sentiment types out from images?

Example work: (Luan et al., 2017; Gatys et al., 2015)

References

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A Simple but Tough-to-Beat Baseline for Sentence Embeddings. In *ICLR*, 2017.

Maria-Florina Balcan, Vandana Kanchanapally, Yingyu Liang, and David Woodruff. Improved Distributed Principal Component Analysis. aug 2014. URL http://arxiv.org/abs/1408.5823.

Cynthia Dwork and Cynthia. Differential Privacy. In *Proceedings of the 33rd international conference on Automata, Languages and Programming - Volume Part II*, pages 1–12. Springer-Verlag, 2006. ISBN 3-540-35907-9, 978-3-540-35907-4. doi: 10.1007/11787006_1. URL http://link.springer.com/10.1007/11787006_1.

Cynthia Dwork and Aaron Roth. The Algorithmic Foundations of Differential Privacy. *Theoretical Computer Science*, 9:3–4, 2014. doi: 10.1561/0400000042. URL http://www.cis.upenn.edu/ aaroth/Papers/privacybook.pdf.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. Fairness Through Awareness. apr 2011. URL http://arxiv.org/abs/1104.3913.

Zhe Gan, Yunchen Pu, Ricardo Henao, Chunyuan Li, Xiaodong He, and Lawrence Carin. Learning Generic Sentence Representations Using Convolutional Neural Networks. URL http://people.ee.duke.edu/ lcarin/sent2vec_emnlp2017.pdf.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A Neural Algorithm of Artistic Style. aug 2015. URL http://arxiv.org/abs/1508.06576.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of Opportunity in Supervised Learning. oct 2016. URL http://arxiv.org/abs/1610.02413.

G. E. Hinton and R. R. Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(July):504–507, 2006. ISSN 0036-8075. doi: 10.1126/science.1127647.

Kenji Kawaguchi, Leslie Pack Kaelbling, and Yoshua Bengio. Generalization in Deep Learning. 2017. URL https://arxiv.org/pdf/1710.05468.pdf.

Been Kim, Rajiv Khanna, and Oluwasanmi Koyejo. Examples are not Enough, Learn to Criticize! Criticism for Interpretability. URL http://people.csail.mit.edu/beenkim/papers/KIM2016NIPS_MMD.pdf.

Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual Fairness. mar 2017. URL http://arxiv.org/abs/1703.06856.

Quoc V. Le. Building high-level features using large scale unsupervised learning. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8595–8598. IEEE, may 2013. ISBN 978-1-4799-0356-6. doi: 10.1109/ICASSP.2013.6639343. URL http://ieeexplore.ieee.org/document/6639343/.

Jiajun Lu, Theerasit Issaranon, and David Forsyth. SafetyNet: Detecting and Rejecting Adversarial Examples Robustly. URL https://arxiv.org/pdf/1704.00103.pdf.

Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep Photo Style Transfer. mar 2017. URL http://arxiv.org/abs/1703.07511.

Aleksander Mądry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. URL https://arxiv.org/pdf/1706.06083.pdf.

Wenlong Mou, Liwei Wang, Xiyu Zhai, and Kai Zheng. Generalization Bounds of SGLD for Non-convex Learning: Two Theoretical Viewpoints. 2017. URL https://arxiv.org/pdf/1707.05947.pdf.

Igor Mozetic, Miha Grcar, and Jasmina Smailovic. Multilingual Twitter Sentiment Classification: The Role of Human Annotators. feb 2016. doi: 10.1371/journal.pone.0155036. URL http://arxiv.org/abs/1602.07563 http://dx.doi.org/10.1371/journal.pone.0155036.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. dec 2014. URL http://arxiv.org/abs/1412.1897.

Xinghao Pan, Maximilian Lam, Stephen Tu, Dimitris Papailiopoulos, Ce I Zhang Michael Jordan, Kannan Ramchandran, Chris Re, and Benjamin Recht. CYCLADES: Conflict-free Asynchronous Machine Learning. 2016.

Sai Rajeswar, Sandeep Subramanian, Francis Dutil, Christopher Pal, and Aaron Courville. Adversarial Generation of Natural Language. may 2017. URL http://arxiv.org/abs/1705.10929.

B Recht, C Re, S J Wright, and F Niu. Hogwild: {A} Lock-Free Approach to Parallelizing Stochastic Gradient Descent. In Peter Bartlett, Fernando Pereira, Richard Zemel, John Shawe-Taylor, and Kilian Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 693–701, 2011. URL http://books.nips.cc/nips24.html.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why Should I Trust You? Explaining the Predictions of Any Classifier. doi: 10.1145/2939672.2939778. URL http://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf.

Ruslan Salakhutdinov. Learning Deep Generative Models. *Annual Review of Statistics and Its Application*, 2:361–385, 2015. doi: 10.1146/annurev-statistics-010814-020120. URL https://www.cs.cmu.edu/ rsalakhu/papers/annrev.pdf.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. URL https://nlp.stanford.edu/ socherr/EMNLP2013_RNTN.pdf.

Pang Wei Koh and Percy Liang. Understanding Black-box Predictions via Influence Functions. URL https://arxiv.org/pdf/1703.04730.pdf.

John Wieting and Kevin Gimpel. Revisiting Recurrent Networks for Paraphrastic Sentence Embeddings. apr 2017. URL http://arxiv.org/abs/1705.00364.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. nov 2016. URL http://arxiv.org/abs/1611.03530.

Yizhe Zhang, Zhe Gan, Kai Fan, Zhi Chen, Ricardo Henao, Dinghan Shen, and Lawrence Carin. Adversarial Feature Matching for Text Generation. jun 2017. URL http://arxiv.org/abs/1706.03850.