Learning from Large, Complex Data with Applications in Natural Language Processing

Machine learning techniques have been widely applied in many domains. For example, they have been shown to be effective in resolving ambiguity and modeling linguistic structure in human languages, leading to substantial advancements in information extraction, question answering, and machine translation. However, in many cases, high accuracy requires training on large amounts of data, exploiting various sources of learning signals and/or exploring complex input and output interactions. *My research goal is to design practical algorithms to efficiently learn expressive models from large-scale data and massive quantities of knowledge.* My study benefits both the theory and the practice of Machine Learning (ML) and helps application domains such as Natural Language Processing (NLP), where researchers often encounter large, complex data.

My studies address large, complex data from three main angles: algorithm design, modeling, and applications. With respect to the algorithm design, I have contributed to the development of fast, theoretically justified learning algorithms for various binary classification [4, 15, 17, 16, 22, 13, 21, 6] and structured prediction (joint prediction) [8, 10, 5, 3, 2] models. Some of these algorithms are implemented in a linear classification package, LIBLINEAR [14],[1] which *has been downloaded more than 100,000 times* since 2007. Others of these algorithms are implemented in an industrial scale machine learning library, Vowpal Wabbit.[2] These software packages are commonly used by machine learning researchers and practitioners. From the modeling perspective, I proposed approximate and latent structured models for problems with complex output structure [9, 20]. This led to an efficient, principled, and linguistically motivated approach to an important NLP task, coreference resolution [7]. This was the best performing corefernece system at the time of its publication; the approach has since become mainstream and inspired follow-up work. I also design algorithms for learning from implicit supervision [2], structured bandit feedbacks [5], and human interactions [1]. These algorithms leverage different weak learning signals while training a model to overcome the difficulty of obtaining high-quality full annotations of structured labels. In the area of applications, I have worked on several key challenges in NLP, including coreference resolution [7], grammar correction [19, 18], vector space models [11], word embeddings [2], phrase suggestions [1], dependency parsing [3], semantic parsing [2], and relation extraction [12]. I also participated in prestigious shared task competitions in DM and NLP (e.g., KDDCUP 09 and CoNLL Shared Task 11, 12, 13, 14) and developed systems that ranked among the top systems in all cases. Studying applications has equipped me with a better understanding of the inadequacies of current machine learning approaches, and inspired me to design practical inference and learning methods at scale. I outline below key directions I have addressed in my research so far.

Learning with Large-Scale Data. In the past few years, the size of data has grown drastically and many companies and researchers report the need to learn on extremely large data sets. My earlier work [4, 15, 22] shows that when data can be fully loaded into memory, training a linear

[1] I implemented the main framework including the default solver and four other solvers based on my papers [4, 15, 17, 22].

[2] I contributed to the joint prediction system based on my works [5, 3].

model on sparse data with millions of samples and features takes only a few minutes. However, when the data is too large, training a model is difficult. My key contribution to date in this area is in proposing a selective block minimization framework [21, 6] for data that cannot fit in memory. The solver loads and trains on a block of samples at a time, and caches informative samples in memory during the training process. By placing more emphasis on learning from the data that is already in memory, the method enjoys provably fast global convergence and reduces the training time from half a day to half an hour. This line of research was awarded the *KDD Best Paper Award* in 2010 and won the *Yahoo! Key Scientific Challenges Award* in 2011.

Efficient Learning and Inference for Joint Prediction Models. Many machine learning applications (e.g., language parsing and information extraction) involve making joint predictions over a set, or a sequence, of mutually dependent output variables. Such problems are often referred to as structured prediction problems or joint prediction problems. When solving these problems, it is important to make consistent decisions that take the interdependencies among output variables into account. There are two popular families of approaches for modeling such problems, (1) graphical models and (2) learning to search. Over the past few years, I have contributed to both types of approaches. Graphical models, including Structured SVM, Structured Perceptron, and Conditional Random Fields, use a graph structure to characterize the correlations between variables. I have accelerated learning algorithms for graphical models from two orthogonal directions – asynchronous parallelism [8] and amortization [10]. I also studied the problem of structured prediction under test-time budget constraints [2] and proposed an approach based on selectively acquiring computationally costly features during test-time in order to reduce the computational cost of prediction with minimal performance degradation. My study shows that the policy for selecting features can be reduced to a series of structured learning problems, resulting in efficient training using existing algorithms. In the other modeling approach, the learning to search method casts the complex decision problem into a sequence of sub-decisions via a search space and learns a model to incrementally construct the complex output. I designed a novel learning algorithm to train a model when supervision signals are not perfect [5]. In addition, I proposed to cast learning to search into a credit assignment compiler [3] with a new programming abstraction for representing a search space. Together with several algorithmic improvements, this radically reduces both the complexity of programming and the running time. Based on these research results, I taught tutorials at NAACL 15 and AAAI 16. I also organized a workshop on *Structured Prediction for NLP* at EMNLP 16, a top-tier NLP conference. The same workshop has been accepted for next year's conference (EMNLP 17) due to its success.

Learning and Inference with Highly Complex Structure. Many applications involve highly complex structure. In such situations, learning and inference are intractable. My work suggests general and specific solutions to problems of this type. In [9], I proposed an approximate semi-supervised learning framework that uses piecewise training for estimating the model weights and a dual decomposition approach for solving the inference problem. The approach is general, and can incorporate domain-specific constraints. In [20], I designed an auxiliary structure to simplify a complex structured prediction model for supervised clustering problems. The presented Latent Left-Linking model (L3M) is a probabilistic model, where each item can link to a previous item with a certain probability according to a feature-based item similarity function. L3M learns the item similarity metric jointly with clustering, and yields accurate predictions. I have applied L3M to the coreference resolution task and achieved state-of-the art performance [7].

Research Agenda: Directions for Future Work

My research goal is to design efficient learning and inference algorithms for big and complex data. This would lead to a unified learning framework to facilitate a range of applications in NLP and other fields. I have already made significant strides toward this goal and intend to focus on further improving and expanding the techniques described above over the next few years. In particular, I am excited to work on the following directions in the near future.

Learning Joint Prediction Models with Auxiliary Supervision. Despite the success of joint prediction models, training them requires an extensive collection of training data. However, annotating structured outputs is difficult and often requires expertise and domain knowledge, making it costly to obtain high-quality annotations, which limits the performance of the model. On the other hand, an enormous amount of textual and user-generated content has accumulated due to the growth of the Internet and the digitization of library archives. Although these data do not directly annotate the desired output structures, they provide related information that can be formed into auxiliary learning signals. However, the huge volume and heterogeneous and noisy nature of the unstructured data present challenges to the existing learning algorithms. Based on my pilot study on learning an algebra word problem solver from different types of learning signals [2], I am interested in addressing the technical challenges of leveraging auxiliary supervision and designing learning algorithms that take heterogeneous learning signals as input. The proposed algorithms will be evaluated on a broad range of NLP applications.

Proactive Knowledge-based Completion. I am interested in developing robust learning systems to handle large web data. In the past decade, enormous quantities of textual and structured data have accumulated due to the growth of the Internet. This provides great opportunities and challenges to the machine learning research community. On one hand, the information embedded in texts, knowledge bases, and encyclopedia sites provides tremendous help in developing accurate learning systems. On the other hand, the huge, heterogeneous, and noisy nature of web data challenges the existing approaches. I would like to tackle this challenge and build an efficient system to automatically extract and refine knowledge from web data. This world knowledge is potentially useful for NLP applications such as coreference resolution, information extraction, and question answering.

Joint Prediction Model with Representation Learning. Structured data such as natural language data is inherently relational; there is a need to develop methods that learn from relational data to support better abstractions. I am interested in studying deep learning and tensor approaches to these problems. In my previous work, I have developed efficient tensor-based methods for identifying multiple relations between words [11], and for extracting relation tuples based on a large knowledge base [12]. Building on these preliminary studies, I plan to design a general framework for exploring entity relations, with the goal of learning structured embeddings of knowledge bases and discovering new relations missing from the database. I also interested in exploring how to combine representation learning (including deep learning) and joint prediction approaches. These two families of approaches are complementary. Joint prediction models are strong in modeling explicit correlations between labels, while representation learning can capture implicit correlations and long-distance dependencies. I am excited about designing the next generation of learning algorithms that leverage the strengths of both approaches.

Detecting and Avoiding Discrimination in NLP systems. Fairness and avoidance of discrimination are critically important in decision-making. Various laws have been passed to prevent discrimination on the basis of attributes such as gender or race. However, modern ML and NLP systems are often not aware of such regulations when making decisions, and they run the risk of amplifying biases present in data and potentially harming minorities. In many cases, even if the demographic information (e.g., gender or race) is not directly provided, it can still be intuited from the data based on implicit associations. In my preliminary study [2] featured on *NPR* and in *MIT Technology Review*, I discovered that such a danger is posed by word embedding, which is a popular framework for representing text data as vectors. Even word embeddings focused on news articles exhibit female/male gender stereotypes to a disturbing extent. Similarly, a system that automatically generates captions for images produces “A woman is cooking” even though the image clearly presents a man cooking. An automated resume filtering system unconsciously selects candidates based on their gender and race due to implicit associations. In the future, I intent to study how we can identify and quantify the implicit associations embedded in NLP systems and design models that make fair decisions by removing inappropriate implicit associations while maintaining desired ones.

Collaborations with Other Research Fields. Given my previous success in developing open-source libraries, I am enthusiastic about developing machine learning libraries, NLP softwares, and system demos and use them to facilitate collaborations with people outside my area. I have several projects collaborated with faculty members in different areas, including computer systems, information retrieval, and computer vision. For example, in one project, I collaborated with a systems professor in designing statistical models for generating programming language. In general, I am excited about applying my expertise in ML and NLP to other areas to solve real-world learning problems.

References

[1] K. Arnold, K.-W. Chang, and A. Kalai. Learning to suggest phrases. In *AAAI workshop on Human-Aware Artificial Intelligence*, 2017.

[2] T. Bolukbasi, K.-W. Chang, J. Wang, and V. Saligrama. Structured prediction with test-time budget constraints. In *AAAI*, 2017.

[3] K.-W. Chang, H. He, H. D. III, J. Langford, and S. Ross. A credit assignment compiler for joint prediction. In *NIPS*, 2016.

[4] K.-W. Chang, C.-J. Hsieh, and C.-J. Lin. Coordinate descent method for large-scale L2-loss linear SVM. *JMLR*, 9:1369–1398, 2008.

[5] K.-W. Chang, A. Krishnamurthy, A. Agarwal, H. Daume III, and J. Langford. Learning to search better than your teacher. In *ICML*, 2015.

[6] K.-W. Chang and D. Roth. Selective block minimization for faster convergence of limited memory large-scale linear models. In *KDD*, 2011.

[7] K.-W. Chang, R. Samdani, and D. Roth. A constrained latent variable model for coreference resolution. In *EMNLP*, 2013.

[8] K.-W. Chang, V. Srikumar, and D. Roth. Multi-core structural svm training. In *ECML*, 2013.

[9] K.-W. Chang, S. Sundararajan, and S. S. Keerthi. Tractable semi-supervised learning of complex structured prediction models. In *ECML*, pages 176–191, 2013.

[10] K.-W. Chang, S. Upadhyay, G. Kundu, and D. Roth. Structural learning with amortized inference. In *AAAI*, 2015.

[11] K.-W. Chang, W.-t. Yih, and C. Meek. Multi-relational latent semantic analysis. In *EMNLP*, 2013.

[12] K.-W. Chang, W.-t. Yih, B. Yang, and C. Meek. Typed tensor decomposition of knowledge bases for relation extraction. In *EMNLP*, 2014.

[13] Y.-W. Chang, C.-J. Hsieh, K.-W. Chang, M. Ringgaard, and C.-J. Lin. Training and testing low-degree polynomial data mappings via linear SVM. *JMLR*, 11:1471–1490, 2010.

[14] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *JMLR*, 9:1871–1874, 2008.

[15] C.-J. Hsieh, K.-W. Chang, C.-J. Lin, S. S. Keerthi, and S. Sundararajan. A dual coordinate descent method for large-scale linear SVM. In *ICML*, 2008.

[16] F.-L. Huang, C.-J. Hsieh, K.-W. Chang, and C.-J. Lin. Iterative scaling and coordinate descent methods for maximum entropy. *JMLR*, 11:815–848, 2010.

[17] S. S. Keerthi, S. Sundararajan, K.-W. Chang, C.-J. Hsieh, and C.-J. Lin. A sequential dual method for large scale multi-class linear SVMs. In *ACM KDD*, 2008.

[18] A. Rozovskaya, K.-W. Chang, M. Sammons, and D. Roth. The university of illinois system in the conll-2013 shared task. In *CoNLL Shared Task*, 2013.

[19] A. Rozovskaya, K.-W. Chang, M. Sammons, D. Roth, and N. Habash. The illinois-columbia system in the conll-2014 shared task. In *CoNLL Shared Task*, 2014.

[20] R. Samdani, K.-W. Chang, and D. Roth. A discriminative latent variable model for online clustering. In *ICML*, 2014.

[21] H.-F. Yu, C.-J. Hsieh, K.-W. Chang, and C.-J. Lin. Large linear classification when data cannot fit in memory. In *KDD*, 2010.

[22] G.-X. Yuan, K.-W. Chang, C.-J. Hsieh, and C.-J. Lin. A comparison of optimization methods and software for large-scale l1-regularized linear classification. *JMLR*, 11:3183–3234, 2010.