RESEARCH STATEMENT

Xuezhe Ma (Max)
Carnegie Mellon University

Vision: Recent decades have witnessed a phenomenal level of successes in machine learning (ML) algorithms and their applications in many domains such as natural language processing (NLP) and computer vision (CV). To make task-oriented predictions, the development of these ML systems heavily rely on extracting abstract informative features from real-world data that are represented in raw digital formats. For example, *syntactic*, *semantic* and *contextual* information are essentially important for a wide range of NLP tasks such as information extraction (IE). Text documents, however, are stored by individual tokens of words or characters, upon which these abstract information is hard to represent. Similar scenarios happen in image data — digital images are made up of pixels, from which it is difficult to extract abstract features such as *edge* or *shape* that are crucial for tasks, for instance, image classification. Classical approaches to extract instructive features mainly rely on task-specific expertise and heuristically designed hand-crafted features, in an iterative feature-selection process. There are two problems with that brute-force methodology: 1) the combinatorial nature of empirical feature selection process makes it expensive to handcraft features; 2) The development of these features is commonly task-, domain-, or even language-specific, preventing it from adapting to new tasks or domains.

I believe that representation learning techniques based on deep learning methods can fundamentally transform the conventional feature designing paradigm. Representation learning can, in principle, automatically learn representations that are mathematically and computationally convenient to process. Furthermore, beyond learning representations for specific tasks, representation learning allows us to identify and disentangle the underlying causal factors, to tease apart the underlying dependencies of the data, so that it becomes easier to understand, to classify, or to perform other tasks such as, even, controllable and interpretable data generation or manipulation. **My research focuses on fulfilling this transformation to enhance the *effectiveness*, *efficiency*, *controllability* and *interpretablility* of representation learning, by developing and analyzing deep learning techniques**. The **key contributions** of my research are as follows:

- *Supervised feature learning* to get rid of feature engineering in ML tasks. We advanced the state-of-the-art on linguistic structured predictions and cross-lingual transfer learning by *proposing a general deep neural architecture, BLSTM-CNNs, for learning representations of text*. BLSTM-CNNs provide sentence representations which are applicable across different structured prediction tasks, while eliminating the hand-crafted feature engineering by end-to-end learning (§1.1).
- *Representation learning* via deep generative models. We developed deep generative models to improve both data density estimation and latent representation learning for text and image data. The key idea is to learn the intrinsic structure and valuable information of data via modeling the data generation process (§1.2).
- *Controllable and interpretable representation learning*. We theoretically and empirically analyzed the variants of neural architectures and their impacts on the internal representation, and investigated the information represented in the internal layers of deep neural models to help understand how deep neural networks memorize and process information (§1.3).

1 Ph.D. Research on Representation Learning

1.1 Learning Representations for Supervised Machine Learning Tasks

Supervised representation learning (a.k.a. feature learning) aims to automatically learn representations that are mathematically and computationally convenient for machine learning algorithms on specific tasks, to replace hand-crafted feature engineering. Feature learning is challenging because it requires methods not only to support end-to-end learning of features directly from task-oriented raw (labeled) data, but also be applicable to a wide range of tasks.

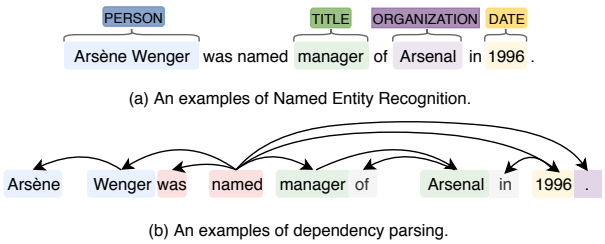

Figure 1: An example sentence of named entity recognition and dependency parsing.

My work on supervised representation learning is in the area of deep neural models for *linguistic structured predictions* (see Fig 1 for examples). Most of these work have been based on **a consistent deep neural architecture named bi-directional LSTM stacked**

with Convolutional neural networks (BLSTM-CNNs) that I introduced in [6] for encoding input sentences. BLSTM-CNNs is capable of capturing both character-level and word-level contextual information. Moreover, BLSTM-CNNs provide **sentence representations that are applicable across different structured prediction tasks, while supporting the kind of end-to-end learning, saving us from hand-crafted feature engineering.** By stacking different structured decoding layers on top of BLSTM-CNNs, we proposed deep neural models for linguistic structured prediction tasks including sequence labeling (BLSTM-CNNs-CRF) [6], graph-based dependency parsing (NeuroMST Parser) [7] and transition-based dependency parsing (Stack-Pointer Parser) [8]. We were excited to see these BLSTM-CNNs based models have been used as standard benchmark approaches for different linguistic structured prediction tasks across a wide variety of languages, and still remain at or near the state of the art. We also found that BLSTM-CNNs improve cross-lingual transferability for dependency parsing under zero-shot setting [1].

1.2 Learning Representations from/for Data Generation

> What I cannot create, I do not understand.
>
> *Richard P. Feynman*

Unlike supervised feature learning that learns representations for specific ML tasks, general-purpose representation learning aims to learn representations that help understand the intrinsic structures and valuable information of data that benefit various tasks and objectives. I believe that learning informative representations from data goes hand-in-hand with learning to generate the data themselves — the kinds of inductive bias imposed by a demand for generating data are often precisely those that encourage effective/informative representation learning. Generative models hold the promise to provide AI systems with a framework for all the many different intuitive concepts they need to understand, giving them the ability to reason about these concepts in the face of uncertainty. My research on this direction has focused on pushing the frontier of both generative models and representation learning — *learning representations from and for data generation*.

Enhancing capability and efficiency of generative flows for data generation: developing expressive flow architectures to improve the performance of density estimation on complex data and to accelerate generation process. *Generative flows* typically warp a simple distribution into a complex one by mapping points from the simple distribution to the complex data distribution through a chain of invertible transformations (see Fig. 2). **I proposed MACOW [9], a novel architecture of invertible transformations which leverages masked convolutional neural networks, for image generation**. MACOW enjoys the merits of stable training, efficient sampling and state-of-the-art performance of density estimation on multiple benchmarks of image generation. **The resulting model reduces the time complexity of image generation significantly from quadratic ($(h \times w)$) to linear ($O(h)$ or $O(w)$) with h and w being the height and wide of the input image.**

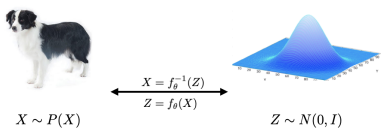

Figure 2: Diagram of generative flow.

Learning latent representation via variational auto-encoding: designing regularization methods for variational auto-encoders (VAEs) to leverage inductive bias in learned latent spaces. My work aimed to make the latent space in VAEs contain meaningful and desired information. **I proposed *mutual posterior-divergence regularization* [10], a novel regularization that is able to control the geometry of the latent space to accomplish meaningful representation learning.** This regularization technique turns out to be useful to tackle the *posterior collapse* issue in VAEs, and capable of learning meaningful latent representations. Experimentally, we showed that the learned latent representations can be directly applied to subsequent tasks such as image reconstruction, supervised image classification and unsupervised image clustering.

Learning complex latent representation using generative flows: utilizing generative flows to learn expressive latent representations. Previous work focused on learning compressed (low-dimensional) latent representations from data. **My recent work introduced FlowSeq, a model that incorporates generative flows in variational auto-encoding, to learn complex (high-dimensional) expressive latent representations for non-autoregressive sequence generation [11].** Empowered by the capability of generative flows in modeling complex distributions, we were excited to see FlowSeq can learn expressive latent representations which is capable of decoupling the dependencies between tokens in the target sequence, yielding effective and efficient non-autoregressive sequence generation. More importantly, the intrinsically modularized nature of FlowSeq opens up a whole new bag of possibilities of learning multilingual representations shared across different languages.

1.3 Learning Interpretable and Controllable Representations

The end-to-end training paradigm in deep neural models simplifies the feature engineering process while giving the model flexibility to optimize for the desired task. This, however, often comes at the expense of model interpretability, making it difficult to understand the role of its different components. Such deep neural models are sometimes perceived as "black-box", hindering research efforts and limiting their utility to society.

My work on interpretability has focused on three directions — (i) understanding how deep neural networks (DNNs) memorize and process information by investigating the internal representations learned by DNNs; ii) Theoretically and empirically analyzing the variants of neural architectures and their impacts on the internal representations; (iii) Learning controllable and effective representations by developing better neural architectures that optimizes the model interpretability.

Analyzing the impact of Dropout theoretically and empirically: My first work in this area theoretically investigated the impact of Dropout, a commonly used method to reduce over-fitting, on the training and inference phases of deep neural networks. **We formally studied the inference gap of dropout, and introduced the notion of (approximate) expectation-linearity to measure and characterize this gap [5]**. In particular, the proposed measure of the inference gap can be used to regularize the standard dropout training objective, consistently leading to improved performance on multiple benchmark datasets.

Investigating the internal representations of neural parsers: One part of my Ph.D dissertation empirically explored **the linguistic information represented in the internal representations in different neural dependency parsing models** [4], using the method of utilizing supervised learning tasks to probe the internal representations in end-to-end models [2]. **I proposed three groups of experiments to investigate the information flow in deep neural models, i.e. how the various types of information are processed and propagated across different layers.**

2 Future Directions

Two broad themes run through the work I have outlined above: **representation learning as a class of techniques for machine learning**, and **representation learning as an explanatory device for understanding and controlling learned models**. I believe that the learning techniques of controllability and interpretability are closely related to and mutually enhance the methodologies that encourage effective and robust generalization. Moving forward, I will continue working along the previously mentioned themes and also branch out to explore problems related to controllable and interpretable representation learning and its applications.

Advancing supervised representation learning for core NLP tasks: Supervised representation learning techniques have led to a number of impressive empirical successes on ML, NLP and CV. Recent works that exploit self-supervised pre-training models, followed by fine-tuning on specific task annotations, kept pushing the state-of-the-art performance on a wide range of NLP and CV tasks. Of course, there are still plenty of unsolved problems involving supervised representation learning! I'm especially interested in continuing to explore the supervised representation learning, probably equipped with pre-training models, on core NLP tasks involving linguistic structured inference procedures such as syntactic and semantic parsing, co-reference resolution and language generation. In particular, **I want to study how to capture linguistically interpretable representations and how to use them to enhance core NLP tasks, by combining representation learning technologies such as self-supervised pre-training and computational linguistic theories in linguistic structure learning and inference procedures.**

Learning Mathematically Interpretable Interlingual Representations: Previous work on encoding texts from different languages into shared interlingual representations is commonly achieved by parameter sharing and lexical overlap [3]. However, the learned interlingual representation is not entirely language-independent and can only capture shallow semantic information. In addition, these interlingual representations are highly unexplainable, and the success of applying them to NLP tasks often relies on the heuristic empirical results. I believe one promising direction in the next 5 to 10 years is to learn interpretable interlingual representations that capture globally language-independent semantic meaning of texts. In my future research, **I want to explore the possibilities of utilizing the intrinsically modularized nature and shared common prior space to learn universal interlingual representations.** Furthermore, **I hope to exploit the interlingual representation as an intermediate to enhance a broad range of computational approaches to multilingual NLP, in particular for resource-limited languages.** In addition, since the prior space in FlowSeq [11] is a well-defined mathematical space, **we attempt**

to investigate the linguistic structures and properties by mapping the corresponding sentences in different languages into the mathematical space.

Establishing theoretical framework towards controllable representation learning: Broadly, learning interpretable and controllable representation assists in identifying and decoupling underlying causal factors of data, making it feasible for controllable and interpretable data generation or manipulation. I think one of the key things we have lost in the era of deep learning for representation learning is **a well-established framework or theory to formally link various neural architectures with the learned representations.** Our analysis of representations learned from different neural architectures and objectives, such as supervised structured prediction and unsupervised generation, sheds preliminary empirical light on this. But there is a huge amount of work needed to establish a theoretical basis for a better analysis and understanding of representation learning. More generally, I suspect that the following two decades will be defined as much by what representation can accomplish for learning as by what learning can accomplish for representation.

References

[1] W. Ahmad, Z. Zhang, X. Ma, E. Hovy, K.-W. Chang, and N. Peng. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. In *Proceedings of NAACL-2019*, pages 2440–2452, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[2] Y. Belinkov. *On internal language representations in deep learning: An analysis of machine translation and speech recognition*. PhD thesis, Massachusetts Institute of Technology, 2018.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[4] X. Ma. *Neural Networks for Linguistic Structured Prediction and Its Interpretability*. PhD thesis, PhD Thesis, Carnegie Mellon University, 2020.

[5] X. Ma, Y. Gao, Z. Hu, Y. Yu, Y. Deng, and E. Hovy. Dropout with expectation-linear regularization. In *Proceedings of the 5th International Conference on Learning Representations (ICLR-2017)*, Toulon, France, April 2017.

[6] X. Ma and E. Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of ACL-2016*, pages 1064–1074, Berlin, Germany, August 2016. Association for Computational Linguistics.

[7] X. Ma and E. Hovy. Neural probabilistic model for non-projective mst parsing. In *Proceedings of IJCNLP-2017*, pages 59–69, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing.

[8] X. Ma, Z. Hu, J. Liu, N. Peng, G. Neubig, and E. Hovy. Stack-pointer networks for dependency parsing. In *Proceedings of ACL-2018*, pages 1403–1414. Association for Computational Linguistics, 2018.

[9] X. Ma, X. Kong, S. Zhang, and E. Hovy. Macow: Masked convolutional generative flow. In *Advances in Neural Information Processing Systems 33*. Curran Associates, Inc., 2019.

[10] X. Ma, C. Zhou, and E. Hovy. Mae: Mutual posterior-divergence regularization for variational autoencoders. In *Proceedings of the 7th International Conference on Learning Representations (ICLR-2019)*, New Orleans, Louisiana, USA, May 2019.

[11] X. Ma, C. Zhou, X. Li, G. Neubig, and E. Hovy. Flowseq: Non-autoregressive conditional sequence generation with generative flow. In *Proceedings of EMNLP-2019*, Hong Kong, November 2019.