Research Statement

Hangfeng He

Moving Beyond Scale-Driven Learning

Machine learning, especially deep learning, has been recognized as a monumentally successful approach to many data-intensive applications across a broad range of domains. Despite the great success achieved, recent progress mainly relies on scaling up existing learning methods with regard to the size of models or training data, consuming enormous time and energy. Therefore, my research focuses on *moving beyond scale-driven learning* to avoid large-scale training data and overly complicated models.

My prior work has been driven by two problems: ***alleviating the supervision bottleneck*** and ***interpreting the behaviors of deep neural networks (DNNs)***. The former reduces the demand for task-specific data, and the latter helps to design simple and efficient models. I aim to identify applications that can be enabled or improved by new frameworks [3, 7, 8, 10, 11], provide theoretical understanding of deep learning phenomena [4, 5, 6], and design practical algorithms with theoretical guarantees [1, 2, 9]. My research profile spans the entire ***theory-to-application spectrum*** from theoretical advances to progress in real-world applications.

Incidental Supervision for Natural Language Understanding (NLU). The supervised learning paradigm, where direct supervision signals are available in high-quality and large amounts, has been struggling to fulfill needs in many AI applications, such as complex question answering (QA). Incidental supervision is thus proposed to alleviate the supervision bottleneck by exploiting weak signals that exist in the data and the environment, independently of the tasks at hand. Such signals could include unlabeled texts, partial labels, noisy labels, constraints, prior knowledge, cross-lingual signals, cross-domain or cross-task annotations, which are widely used in NLU and other domains. A specific example of named entity recognition is shown in Figure 1. However, much is unknown about the benefits of incidental signals, and we don't have a good protocol to use these signals efficiently. Therefore, my goal in this direction is to first provide a better understanding of incidental signals, and then design more efficient algorithms to *collect, select, and use* them.

Among various incidental signals, cross-task signals are widely used, but the corresponding benefits are not well explained, especially in deep learning. My recent work [1] introduces a notion of representation-based task distance to better understand the benefits of cross-task signals. This finer-grained notion of task distance allows one to reason in a theoretically principled way about several critical aspects of cross-task learning, such as the choice of the source data and the impact of fine-tuning. Based on minimizing the representation-based task distance between source and target tasks, a weighted training algorithm is further proposed to improve the sample efficiency of cross-task learning with limited target data. The algorithm has theoretical guarantees with a non-asymptotic generalization bound, and it improves the performance of BERT in both pre-training and joint training paradigms on several NLU tasks.

Annotating NLU tasks often requires significant linguistic expertise. My other work in this direction focuses on utilizing incidental signals to help NLU tasks. I proposed QuASE [7], a method to exploit information from cheap signals, i.e., crowdsourced QA pairs, and help tasks which

Figure 1: An example of named entity recognition with various incidental supervision signals: unlabeled signals, partial signals in structured outputs, noisy signals, cross-domain signals, cross-lingual signals, cross-task signals, constraints in structured learning, and prior knowledge.

typically require large datasets annotated by experts, such as semantic role labeling, named entity recognition, and textual entailment. This work provides supporting evidence to an important alternative to supervising NLU tasks: ***using natural language to annotate natural language***. I have also proposed a unified informativeness measure to facilitate selecting appropriate incidental signals for a given target task [9]. Our measure can quantify the benefits of various incidental signals and combinations of them. Additionally, I exploit other types of incidental signals to improve NLU, such as partial signals for structured learning [11], cross-modal signals for spatial relations [3], and QA signals for nominal semantic role labeling [10].

Interpretability of Deep Neural Networks. Despite the great success of DNNs in numerous machine learning tasks, there is still no comprehension of how they work. My goal is hence to demystify these surprisingly effective black-box models, visualize the essential elements, and enable the principled design of network architectures and training. In particular, I intend to first discover some important characteristics of DNNs, and then develop a mathematically tractable surrogate model while maintaining these characteristics to interpret the behaviors of DNNs.

In my recent work [6], I develop a simple and interpretable model termed the ***layer-peeled model*** to analyze well-trained DNNs. In particular, I isolate the topmost layer from the remainder of DNNs and impose certain constraints separately on the two parts of DNNs. The effectiveness of this model is evidenced by its ability to reproduce a known empirical pattern, i.e., neural collapse (a highly symmetric geometry of features and classifiers), in balanced training where each class is equal-sized. Moving to imbalanced datasets, layer-peeled model predicts a hitherto unknown phenomenon termed *minority collapse*. When minority collapse occurs, the classifiers corresponding to the minority classes would collapse to a single vector, which fundamentally limits the performance of deep learning models when the imbalance level is above a threshold.

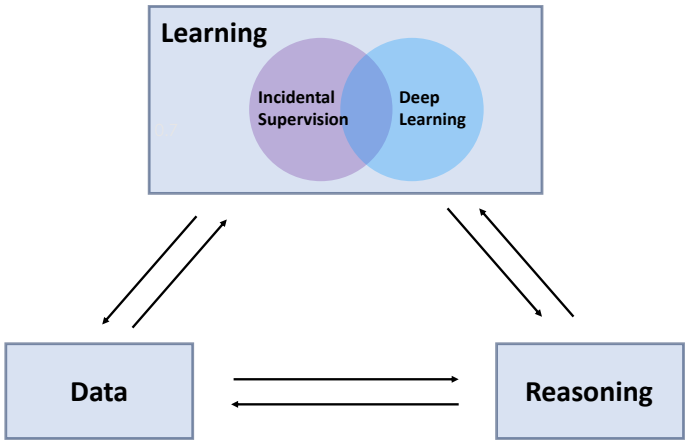

Figure 2: An illustration of the three areas of my research agenda: ***learning, reasoning, and data.***

Importantly, minority collapse is a rare example in the large volume of deep learning research, as it was first predicted by the layer-peeled model before its confirmation by considerable amounts of computational experiments.

Another line of my research in this direction centers around one important geometric property of DNNs – ***local elasticity*** [8]. Extensive experiments show that neural networks are locally elastic in the sense that, if the network weights are updated using an image by stochastic gradient descent, the prediction at another image is only significantly perturbed if the two images are semantically similar. This property has inspired several work on the generalization ability of DNNs, including better neural tangent kernels using label awareness [2], and better generalization bounds with locally elastic stability [5].

Research Agenda: Directions for Future Work

My previous work mainly lies in the learning realm. However, I aspire to expand my research goal to more areas, including comprehending the mechanism of reasoning and analyzing the structure of data. A general overview of my research agenda is depicted in Figure 2.

Reasoning in Natural Language. Among the 125 important questions collected by Science for its $125th$ anniversary, one of the questions is – "What are the limits of learning by machines?" Specifically, abstract reasoning is still beyond any machine, though computers have a wealth of information on the Web to draw on. Furthermore, reasoning in natural language tends to decompose a sophisticated task into simple components, which simplifies the model and requires less training data. Bearing in mind the potential significance of reasoning, I try to work on it in two general aspects.

One is to ***build efficient reasoning systems*** to solve tasks that require complex reasoning. For instance, given the following question,

Did Aristotle use a laptop?

we can see that this question requires *implicit* decomposition into reasoning steps, which is in contrast to the multi-step question that *explicitly* specifies the reasoning process, "Was Aristotle alive when the laptop was invented?". Moreover, to answer this question, we need to first decide the perspective that determines the answer: Was Aristotle rich enough to possess a laptop? Was Aristotle wise enough so that he was able to use a laptop? As long as there is one perspective where we find negation, the answer should be no. In order to decipher the exact reasoning processes for difficult questions like the above-mentioned one, I have a strong desire to work on tasks that require complex reasoning.

Existing reasoning formulations fall short in explaining or improving reasoning systems in NLU. It is still unclear how to ***theoretically formulate decomposition and composition***. For example, given a complicated problem, we need to first decompose the problem into components, and then combine these interrelated components based on their solutions. I would like to explore theoretical formulations for reasoning, while taking properties of natural language into more account. This initiative relies on collaborations with researchers in statistics and logic.

Structured Data Modeling. It is believed that building the theory of deep learning requires a comprehension of the intricate interplay between the model (e.g., DNNs), the algorithm, and the structure of the data. While there is plenty of work that analyzes models and algorithms, the structure of data is still far from being understood. I plan to work on structured data modeling and use the modeling to design more efficient algorithms for real-world applications.

I would like to place the emphasis on textual and visual data for structure modeling. I am concerned with textual data, because one needs to carefully handle the ***variability*** and ***ambiguity*** of natural language. Besides, natural language has a syntactic ***hierarchy***, from word to sentence to document, and also includes various ***perspectives***, such as topic, sentiment, and stance; while current representations only rely on modeling the co-occurrence of words, without fully utilizing the richness of language. To model the structure of visual data, its ***sparsity*** and ***invariance*** should be considered. The ***part-whole hierarchy*** of visual data makes the problem even more difficult. We lack desirable theoretical tools to describe image features, because each distinguishable component of an object can be used to represent the object and the dimension of features can be different. In a word, these difficulties in modeling the structure of textual or visual data indicate that copious open problems in this research remain to be explored. The problem of modeling the structure of data can be better addressed by collaborating with researchers of different backgrounds, such as linguistics, statistics, logic, and geometry.

References

[1] Shuxiao Chen, Koby Crammer, **Hangfeng He**, Dan Roth, and Weijie J Su **(alphabetical order)**. Weighted training for cross-task learning. *arXiv preprint arXiv:2105.14095*, 2021.

[2] Shuxiao Chen, **Hangfeng He**, and Weijie Su **(alphabetical order)**. Label-aware neural tangent kernel: Toward better generalization and local elasticity. In *Advances in Neural Information Processing Systems*, pages 15847–15858, 2020.

[3] Soham Dan, **Hangfeng He**, and Dan Roth. Understanding spatial relations through multiple modalities. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2368–2372, 2020.

[4] Zhun Deng, **Hangfeng He**, Jiaoyang Huang, and Weijie Su. Towards understanding the dynamics of the first-order adversaries. In *International Conference on Machine Learning*, pages 2484–2493, 2020.

[5] Zhun Deng, **Hangfeng He**, and Weijie Su. Toward better generalization bounds with locally elastic stability. In *International Conference on Machine Learning*, pages 2590–2600, 2021.

[6] Cong Fang, **Hangfeng He**, Qi Long, and Weijie J Su **(alphabetical order)**. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43), 2021.

[7] **Hangfeng He**, Qiang Ning, and Dan Roth. QuASE: Question-answer driven sentence encoding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8743–8758, 2020.

[8] **Hangfeng He** and Weijie Su. The local elasticity of neural networks. In *International Conference on Learning Representations*, 2020.

[9] **Hangfeng He**, Mingyuan Zhang, Qiang Ning, and Dan Roth. Foreseeing the benefits of incidental supervision. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1782–1800, 2021.

[10] Ayal Klein, Jonathan Mamou, Valentina Pyatkin, Daniela Stepanov, **Hangfeng He**, Dan Roth, Luke Zettlemoyer, and Ido Dagan. QANom: Question-answer driven SRL for nominalizations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3069–3083, 2020.

[11] Qiang Ning, **Hangfeng He**, Chuchu Fan, and Dan Roth. Partial or complete, that's the question. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2190–2200, 2019.