RESEARCH STATEMENT

Sheng Liu (shengliu@nyu.edu)

Like what Alan Turing envisioned in the 1950s, computers today use machine learning to "simulate" the child's mind, a blank slate upon which knowledge and representations are gradually imprinted. However, machines today do not learn like a child. They learn from a massive amount of clean and curated data while a child learns from data in the real world, which are noisy and uncurated. My research aims to **build machine learning algorithms that work with real-world imperfect datasets and apply such algorithms to problems in healthcare**. Within this theme, I have focused on the following questions: When do failures happen in deploying machine learning models in the real world? Can we make learning algorithms robust to these failure modes? and importantly, in specialized fields, how to bring domain experts on board? To answer these questions,

- I found that when machines are supervised with noisy signals, they often fail and *memorize* the wrong signals, hurting generalization performance [1, 2, 3];
- But before *memorization*, I observed that correct information can still be inferred during *early-learning*.
- Based on these observations, on the machine learning side, I proposed robust algorithms that exploit the *early-learning*; I also incorporate these algorithms with domain knowledge to address real world problems [2, 5], especially problems in healthcare [7, 8].

Below, I describe my research experiences in detail and present my future research plans.

Early-Learning

Recently, over-parameterized deep networks, with increasingly more network parameters than training samples, have dominated the performances of modern machine learning. Nonetheless, the success of such models critically depends on the availability of clean training data; when the training data are noisy, over-parameterized networks tend to overfit and not generalize. To address this, we recognized that neural networks are able to infer useful information, even from the noisy examples during the early stage of training, we termed this as *early-learning* [1]. My research analyzed the learning dynamics in different problems:

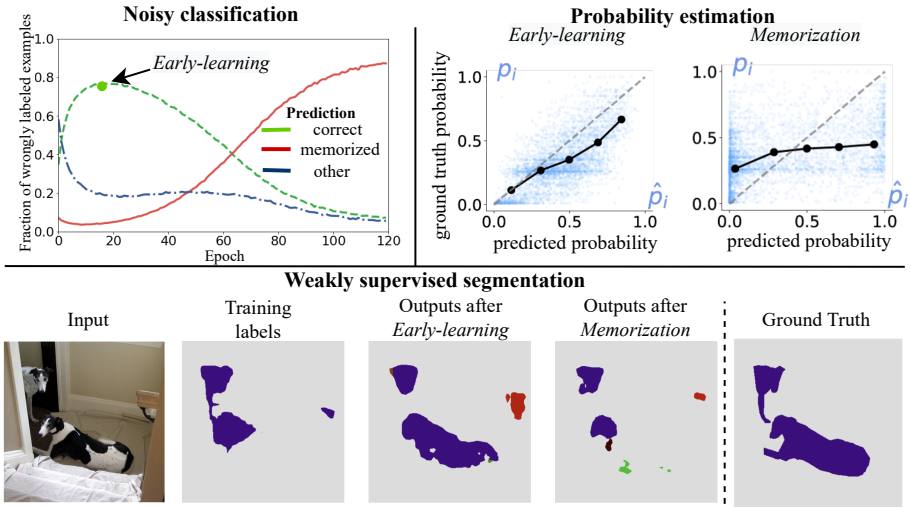

Figure 1: *Early-learning* occurs in various machine learning problems. Neural networks can infer clean signals even when they are supervised by noisy signals during the *Early-learning* phase: for noisy classification, wrong labels can be predicted correctly [1]; for probability estimation, probabilities can be estimated from 0-1 outcomes [5]; for segmentation, annotation errors can be corrected by the segmentation model's outputs [2].

Early-Learning in noisy classification Due to the cost or difficulty of manual labeling and the demands of domain expertise, real-world datasets are often with lower-quality annotations. Such annotations inevitably contain numerous mistakes. When trained on noisy labels, we observed that **deep neural networks correctly predict the true class even trained with false labels after the *early-learning* phase**, before eventually overfitting the false labels. In [1], we established that ***early-learning* and overfitting are fundamental phenomena in high dimensions**, proving that they occur even for simple linear models. Motivated by these findings, we developed a new technique for noisy classification tasks, which steers the model toward inferred information after the *early-learning* phase. Such a technique was validated on many datasets to show robustness to high levels of synthetic and real-world noises.

Previously, *early-learning* is modeled for the whole dataset, In [3], we further modeled it for each sample when label noises are sample dependent. Such sample dependent noises are observed in Alzheimer's detection where the cognitive labels are derived based on different cognitive examinations that patients have taken, the labeling criteria thus depend on each patient. Therefore, we proposed to **model the *early-learning* for**

each sample and learn to separate noise from the true signals [3]. The method was inspired by the fact that label noise is sparse and incoherent with the network learned during *early-learning*. We theoretically justified our approach for exactly separating sparse corruption from the data when *early-learning* occurs. We further observed that *early-learning* is particularly evident for models trained with self-supervised learning methods, we thus propose to pretrain with self-supervised learning for noisy classification [4].

Early-learning in weakly supervised segmentation Weakly supervised semantic segmentation aims to perform segmentation based on weak supervision signals, such as image-level labels. Current state-of-the-art methods for this problem use a classification model to generate noisy pixel-level annotations, which are then used to train a segmentation model. In [2], **we studied the learning dynamics of deep segmentation networks trained on these generated noisy annotations**. We observed that different from noisy classification problems, *early-learning* phase does not occur simultaneously for all categories, therefore useful information need to be obtained at different stages across categories.

Early-learning in probability estimation Reliable probability estimation is important for problems with inherent (aleatoric) uncertainty. Models of such problems are trained on observed 0-1 outcomes with cross-entropy loss, because the ground-truth probabilities are typically unknown. When training neural networks on these 0-1 labels, we observed that they eventually overfit to the outcomes completely, with estimated probabilities collapse to 0 or 1. While before overfitting, in [5], **we observed and proved that calibrated probability is attained during the *early-learning* phase** and thus designed an algorithm to preserve the attained probability while maintaining the discrimination power of the model.

Memorization

The second direction of my research has been understanding the overfitting of overparametrized neural networks. Overfitting to features that can not generalize is often termed as *memorization*. My research has focused on examining the *memorization* effect in supervised learning. There are many losses that have been proposed to prevent *memorization*. For example, focal loss (FL) [13] and label smoothing (LS) [14] are designed with this aim. However, in [12], we showed through global solution and landscape analyses that **a broad family of loss functions**, including cross-entropy loss, mean square error loss, LS, and FL, **all produce equivalent features on training data**. *memorization* can occur regardless of loss functions.

To prevent *memorization*, we instead proposed to regularize the model's parameters. In [6], we demonstrated that promoting the orthogonality to the model's parameters helps prevent *memorization*. Therefore, **we proposed a normalization layer to encourage orthogonality of the convolutional layers**. The proposed normalization layer was shown to efficiently improve generalization and robustness against noise corruption, adversarial attacks, and data scarcity.

Robust Machine learning for healthcare

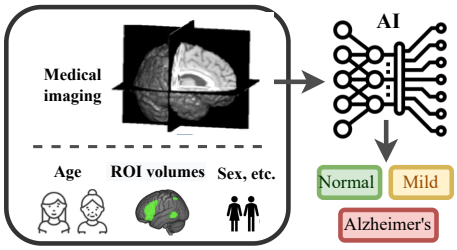

Figure 2: Overview of the AI framework for dementia's automatic diagnosis [7, 8].

I also proposed to contribute to the scientific discovery community with methods developed from the previous sections. I made efforts to understand Alzheimer's disease from the MRI data, showing the potential of machine learning in scientific discovery once the methods are robust for real-world data. As the most common dementia disease, Alzheimer's disease (AD) is the sixth leading cause of death in the U.S.; AD-related brain degeneration begins years before the clinical onset of symptoms, suggesting that early detection of AD might be possible from standard structural brain imaging scans. Unfortunately, both clinical and research-grade detection rates remain low.

In [7, 9], we focused on **learning to differentiate between cognitively normal aging, mild cognitive impairment, and AD, using structural brain MRI (T1-weighted scans)**. We proposed a 3D convolutional neural network architecture that achieves state-of-the-art performance for this task. Besides, we recognized that the diagnosis labels are extremely noisy; the dataset is also small (with only around 3000 scans). We performed analyses on how these issues result in failures of better generalization. We also argued that even though the dataset comes with several issues, useful and related information for the disease is still learned. A wide range of brain regions captured by the model had been reported to be associated with AD.

In order to **utilize prior domain knowledge**, we have further built a model based on the volumes and thickness of previously reported brain regions that are known to be implicated in disease progression [8]. We also engineered the data to extract the volume and thickness of brain regions from the scans, and **combined these human-crafted volumes with MRI scans to train an ensemble model**. This multimodal model achieves the best accuracy, robustness, and with better forecasting capacity.

Future research agenda

My past research has outlined the importance of making AI algorithms robust to real-world datasets. My long-term goal is to further explore the specific challenges arising in the realm of applying machine learning for scientific recovery in the real-world. With this philosophy in mind, I identify the following three research topics I am thrilled to pursue next. I am also interested in building AI frameworks for medical research that is robust to uncurated medical data of all kinds of modalities. With this philosophy in mind, I identify the following three research topics that I am thrilled to pursue next.

Data science meets machine learning. Thanks to the increasingly bigger size of data, machine learning models continuously improved in recent decades; their performances are pushed toward the limits with many human-curated datasets. Despite the success, challenges also come when data are big. Processing them could be time-consuming and expensive. This challenge is significant if machine learning is best trained on human-curated consistent data. The well-known benchmark dataset ImageNet [15] took more than two and a half years to complete. Is there a way to shorten this timeline? Can we design better data science tools to accelerate processing data for machine learning? In the future, I intend to work on improving data science for machine learning – automatic data engineering, domain-driven data augmentation, valuation, etc.

Self-supervised learning in-the-wild. Recent advances, such as contrastive self-supervised learning combined with strong data augmentation, present a promising avenue to learn transferable visual representations. However, many self-supervised learning methods are trained on ImageNet that are not completely "self-supervised". The training set of ImageNet, on which the representations are learned, is heavily curated and requires extensive human effort to create. Imagenet contains many categories and each one contains roughly the same number of images. Images collected in the wild, on the other hand, are often long and heavy-tailed with much more diverse content. Therefore this distribution shift and other issues may prohibit direct applications of these models in the wild. I propose to fill this gap, by studying the learning dynamic of self-supervised learning on datasets collected in the wild.

Multimodal learning for scientific research. Currently, I am working on validating the clinical value of the published machine learning algorithm for the early detection of Alzheimer's disease, closely collaborating with neuro-radiologists from NYU Langone hospital and NYU Alzheimer's Disease Research Center. I would like to continue to develop machine learning algorithms for scientific research with various modalities of data, such as different medical imaging types, clinical notes, genetic and clinical test data, etc. In the long term, I propose utilizing large-scale multi-modality datasets from larger populations to facilitate AI for scientific discovery.

References

[1] **Liu, S.**, Niles-Weed, J., Razavian, N., Fernandez-Granda, C. (2020). "Early-learning regularization prevents memorization of noisy labels". *Advances in neural information processing systems (NeurIPS)*, 33, 20331-20342.

[2] **Liu, S.**, Liu, K., Zhu, W., Shen, Y., Fernandez-Granda, C. (2022). "Adaptive early-learning correction for segmentation from noisy annotations". In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2606-2616).

[3] **Liu, S.**, Zhu, Z., Qu, Q., You, C. (2022). "Robust Training under Label Noise by Over-parameterization". In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, in *Proceedings of Machine Learning Research* 162:14153-14172.

[4] **Liu, S.**, Yi, L., She, Q., McLeod, A. I., Wang, B. (2022). On Learning Contrastive Representations for Learning with Noisy Labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 16682-16691).

[5] **Liu, S.**, Kaku, A., Zhu, W., Leibovich, M., Mohan, S., Yu, B., Zanna, L., Razavian, N. and Fernandez-Granda, C (2022). Deep probability estimation. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, in *Proceedings of Machine Learning Research* 162:13746-13781

[6] **Liu, S.**, Li, X., Zhai, Y., You, C., Zhu, Z., Fernandez-Granda, C., Qu, Q. (2021). Convolutional normalization: Improving deep convolutional network robustness and training. *Advances in Neural Information Processing Systems (NeurIPS)*, 35, 28919-28928.

[7] **Liu, S.**, Yadav, C., Fernandez-Granda, C., Razavian, N. (2020, April). On the design of convolutional neural networks for automatic detection of Alzheimer's disease. In *Machine Learning for Health Workshop at NeurIPS* (pp. 184-201). PMLR.

[8] **Liu, S.**, Masurkar, A.V., Rusinek, H., Chen, J., Zhang, B., Zhu, W., Fernandez-Granda, C., Razavian, N. (2022). Generalizable Deep Learning Model for Early Alzheimer's Disease Detection from Structural MRIs. Nature Scientific reports, 12, 17106

[9] **Liu, S.**, Masurkar, A.V., Rusinek, H., Chen, J., Zhang, B., Zhu, W., Fernandez-Granda, C., Razavian, N. (2022). Development of a Deep Learning Model for Early Alzheime's Disease Detection from Structural MRIs and External Validation on an Independent Cohort. preprint.

[10] **Liu, S.**, Zhang, X.,Sekhar N., Wu Y., Singhal P., Fernandez-Granda C. (2023). Avoiding spurious correlations via logit correction. *International Conference on Learning Representations (ICLR)*.

[11] Bernstein B., **Liu, S.**, Papadaniil C., Fernandez-Granda C. Sparse recovery beyond compressed sensing: Separable nonlinear inverse problems. *IEEE transactions on information theory*66 (9), 5904-5926.

[12] Zhou, J., You, C., Li, X., Liu, K., **Liu, S.**, Qu, Q., Zhu, Z. (2022). Are All Losses Created Equal: A Neural Collapse Perspective.*Advances in Neural Information Processing Systems (NeurIPS)*.

[13] Lin, T. Y., Goyal, P., Girshick, R., He, K., Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980-2988).

[14] Müller, R., Kornblith, S., Hinton, G. E. (2019). When does label smoothing help?. *Advances in neural information processing systems*, 32.

[15] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248-255). Ieee.