Research Statement

Suraj Srinivas

Modern progress in machine learning is driven by large neural network architectures capable of handling vast amounts of data, resulting in remarkable capabilities across a diverse range of tasks. Yet, the immensity and complexity of these models present another challenge: understanding, iterating, and improving over them becomes difficult. Unlike mature engineering disciplines that advance through first-principles thinking, large-scale deep learning remains predominantly experimentally driven, guided by ad-hoc heuristics for critical aspects such as model architecture choice and the design of learning methods. As a result, well-performing deep learning systems are more akin to discoveries rather than deliberate designs, leading to an inadequate understanding of the fundamental principles underlying their remarkable generalization behavior.

Building a principled view of deep learning models is vital for driving future progress for several reasons. Firstly, this can make the process of model building more **computationally efficient** and streamlined, reducing the need for time-consuming and costly hyperparameter searches. Secondly, a principled understanding of model behavior is instrumental for enabling **trustworthy machine learning**, which can help us to answer questions about how models arrive at their outputs in a human-interpretable manner. Finally, a principled view of machine intelligence can reveal the principles behind natural intelligence, such as animal or human intelligence.

To this end, my research agenda is centered on improving our **mathematical and scientific understanding** of deep neural network models and the associated learning algorithms. I aim to make progress towards this high-level goal by **identifying concrete applications** that can benefit from an improved scientific understanding, such as computationally efficient and trustworthy machine learning, and **developing principled algorithms** for these use cases. A common theme in my research involves **identifying connections** between different areas of machine learning, such that advances made in one field can help progress in others. Through my research, I aspire to bridge the gap between the highly advanced engineering achievements in machine learning and the nascent state of its scientific understanding, to pave the way for more robust, efficient, and trustworthy machine learning systems.

Prior Work

In the section below, I discuss my previous work in the areas of computationally efficient deep learning and trustworthy machine learning, including interpretability and model robustness.

Computationally Efficient Deep Learning

Model Pruning with Limited Data Standard neural networks have highly redundant weights, making it possible to prune a large number of weight or neurons without sacrificing predictive accuracy. As a beginning masters student, I started gaining intuition for deep learning by visualizing neural activations in neural networks, when I observed many instances of nearly duplicate features. Inspired by this observation, we devised the first pruning method that removed duplicate neurons in standard pre-trained vision models (Srinivas and Babu, 2015). Critical to this approach was a 'surgery' step that updated model weights to account for the neuron removal, without using training data. More recently, we generalized this approach (Halabi et al., 2022) by casting pruning as submodular optimization, leading to more efficient pruning algorithms under limited data settings.

Incorporating Sparsity in Neural Networks When sufficient data is available for pruning, one can combine pruning with model training in several ways. One method we proposed is to relax the underlying discrete pruning optimization problem into a continuous one such that it is conductive to gradient-based learning, which we found to generalize to both neuron pruning (Srinivas and Babu, 2016b) and weight pruning (Srinivas et al., 2017) leading to sparse neural network models. Another method to achieve this is to alternate discrete pruning and model training, an algorithm called "iterative magnitude pruning". While this algorithm was initially proposed as a heuristic, we view pruning through the lens of compressed sensing and related it to projected gradient descent (PGD), which yields optimal sparse solutions in some simple settings. Inspired by this connection, we improved iterative magnitude pruning by proposing cyclical pruning (Srinivas et al., 2022a), an algorithm that better approximates the optimal PGD solution and obtains improved compression ratios as a result.

Trustworthy Machine Learning

Trustworthy machine learning is a rather broad area concerned with increasing model trust for end users and stakeholders. Within this, my work so far has been in the areas of post hoc interpretability and robust machine learning.

Understanding Feature Attribution Methods Feature attribution methods are common in post hoc interpretability, which aims to find important input features that drive model behavior for a given input instance. However, the plethora of methods that exist in literature are ad-hoc heuristics, making it difficult to identify unifying formal goals for this field. To this end, we proposed the local function approximation framework (Han et al., 2022) as a means of unification of feature methods. This enables us to view several feature attribution methods, including several ones proposed as heuristics, as instances of a single underlying framework. This also helped us identify a no-free lunch theorem showing that no feature attribution method can fully achieve the goal of local interpretability absent additional specifications.

Feature attribution methods are also often motivated via specific secondary properties. We showed (Srinivas and Fleuret, 2019) that two such properties – sensitivity and completeness – were impossible to satisfy simultaneously, highlighting a fundamental limitation of feature attribution methods. To alleviate this, we proposed the full-gradient representation which extracted feature importances at each layer of deep neural networks, and provably satisfied both the above properties.

Explaining Input-Gradient Attribution via Generative Modelling An important yet unexplored question in interpretability is the question of why standard models are interpretable wrt their feature attributions in the first place when such interpretability is formally independent of generalization (Srinivas and Fleuret, 2021). We explained this by invoking the latent energy-based generative properties $p(x \mid y)$ of softmax-based discriminate models $p(y \mid x)$ and showed that the training of this latent generative model specifically using score-matching (Hyvärinen, 2005), dictates gradient interpretability. Overall, this work shows that commonly used input-gradient methods do not capture information related to the discriminative model $p(y \mid x)$, but a latent generative model $p(x \mid y)$, and we must rethink its usage in practice.

Robustness by Regulating Local Geometry While standard robust models often sacrifice predictive accuracy, this is not the case for the task of distillation, where we showed that ideas from robustness can make distillation more sample efficient (Srinivas and Fleuret, 2018). Specifically we showed that Jacobian information at data points helps capture the local neighborhood around that data point, and thus provides more information to the student model to learn from the teacher model. We also showed applications in transfer learning using the same technique, which had similar benefits. Our work on explaining the structure of local geometry of deep models (Srinivas and Fleuret, 2021) also showed that robust models have similar generative modelling properties to model trained explicitly with generative regularizers. To explore this further, we investigated the inductive biases of robust models, and found that robust models tend often to have smaller curvature (in terms of Hessian norms) than non-robust ones (Srinivas et al., 2022b), and We used this observation to propose new ways to training robust models based explicitly penalizing an efficient upper bound on model curvature.

Future Work

In the future, I would like to focus my research toward novel model architectures that improve data and compute efficiency of models. In addition, large models require improved interpretability methods, as current paradiagms have been shown to be conceptually ineffective (Lipton, 2018) to address questions involving spurious correlations and precisely characterizing model behaviour. Toward this end, I propose three broad directions for future work: mechanistic interpretability and feature disentanglement, understanding memorization and data attribution in large models, and characterizing implicit biases of model architectures.

Mechanistic Interpretability and Feature Disentanglement

Understanding Memorization and Data Attribution in Generative Models

Computational Efficiency via Adaptive Computation Overall, these new lines of research can lead to a new generation of neural network models that are simultaneously more compute efficient and interpretable, enabling us to further scale large models while still being able to provide insight about their behaviour.

References

Bai, S., Kolter, J. Z., and Koltun, V. (2019). Deep equilibrium models. *Advances in Neural Information Processing Systems*, 32.

Bubeck, S. and Sellke, M. (2021). A universal law of robustness via isoperimetry. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*.

Chen, R. T., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. (2018). Neural ordinary differential equations. *Advances in neural information processing systems*, 31.

Grathwohl, W., Wang, K.-C., Jacobsen, J.-H., Duvenaud, D., Norouzi, M., and Swersky, K. (2020). Your classifier is secretly an energy based model and you should treat it like one. In *International Conference on Learning Representations*.

Graves, A. (2016). Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983*.

Halabi, M. E., Srinivas, S., and Lacoste-Julien, S. (2022). Data-efficient structured pruning via submodular optimization. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.

Han, T., Srinivas, S., and Lakkaraju, H. (2022). Which explanation should i choose? a function approximation perspective to characterizing post hoc explanations. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(Apr):695–709.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Mnih, V., Heess, N., Graves, A., et al. (2014). Recurrent models of visual attention. *Advances in neural information processing systems*, 27.

Poole, B., Lahiri, S., Raghu, M., Sohl-Dickstein, J., and Ganguli, S. (2016). Exponential expressivity in deep neural networks through transient chaos. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Srinivas, S. and Babu, R. V. (2015). Data-free parameter pruning for deep neural networks. In *Proceedings of the British Machine Vision Conference 2015 (BMVC 2015)*, pages 7–10.

Srinivas, S. and Babu, R. V. (2016a). Generalized dropout. *arXiv preprint arXiv:1611.06791*.

Srinivas, S. and Babu, R. V. (2016b). Learning the architecture of deep neural networks. In *Proceedings of the British Machine Vision Conference 2016 (BMVC 2016)*.

Srinivas, S. and Fleuret, F. (2018). Knowledge transfer with Jacobian matching. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4723–4731. PMLR.

Srinivas, S. and Fleuret, F. (2019). Full-gradient representation for neural network visualization. In *Advances in Neural Information Processing Systems*, pages 4126–4135.

Srinivas, S. and Fleuret, F. (2021). Rethinking the role of gradient-based attribution methods for model interpretability. In *International Conference on Learning Representations*.

Srinivas, S., Kuzmin, A., Nagel, M., van Baalen, M., Skliar, A., and Blankevoort, T. (2022a). Cyclical pruning for sparse neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2762–2771.

Srinivas, S., Matoba, K., Lakkaraju, H., and Fleuret, F. (2022b). Efficient training of low-curvature neural networks. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.

Srinivas, S., Subramanya, A., and Venkatesh Babu, R. (2017). Training sparse neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.