RESEARCH STATEMENT

TAN MINH NGUYEN

The growth in scope and complexity of modern data sets and models presents the field of machine learning with numerous inferential and computational challenges, among them how to deal with various forms of interpretability, robustness, and efficiency. An overarching theme of my research focuses on the interplay of these issues from three principled approaches: 1) optimization, 2) differential equations, and 3) statistical modeling.

From an optimization viewpoint, my goal is to establish connections between deep learning architectures (e.g., recurrent neural networks (RNNs), convolutional neural networks (CNNs), transformers, and deep implicit models) and optimization methods (e.g., gradient descent variants, primal-dual algorithms, and proximal techniques). The connections bring a rich mathematical background in the optimization techniques such as robustness and convergence guarantees to the designing of deep learning architectures. From a differential equations approach, I am exploring the differential equations that govern the dynamics of deep learning models, as well as the numerical methods to solve these equations, to improve the efficiency and accuracy of the models. Using statistical modeling as a tool, I develop new generative models that shed light on the state-of-the-art transformers and various deep neural network architectures, suggesting new directions to improve these models in terms of robustness and efficiency. On the application side, I am interested in using my principled models for developing large-scale natural language processing and computer vision systems and for solving challenging problems in mathematical modeling.

In the following, I shall outline major focused areas of my research to design interpretable, robust, and efficient deep learning models using the three aforementioned principled approaches.

1 Optimization frameworks for designing deep learning models

I connect deep learning models to the gradient-based and primal-dual methods to shed light on the underlying mechanisms of those models and provide principled framework to improve them.

Momentum-based neural networks. RNNs are a popular class of neural networks that capture the dynamics of sequences via cycles in the network of nodes. RNNs tend to suffer from exploding or vanishing gradients during training by the error backpropagation through time algorithm and thus may fail to learn long-term dependencies. In [1], I develop a gradient descent (GD) analogy of the recurrent cell. I then propose to integrate the momentum that is used for accelerating gradient dynamics into the recurrent cell, which results in the MomentumRNN. By choosing the appropriate momentum coefficients, MomentumRNN can alleviate the vanishing gradient problem and accelerate training. My momentum framework for designing RNNs is principled with theoretical guarantees provided by the momentum-accelerated dynamical system for optimization and sampling. The design principle can be applied to many existing RNNs and generalized to other advanced momentum-based optimization methods, including Adam and Nesterov accelerated gradients with a restart [2]. In this direction, I also develop an adaptive strategy to compute the momentum coefficients based on the optimal momentum for quadratic optimization and extend my momentum framework to design linear attention in transformers [3].

A primal-dual framework for transformers and neural network. Like RNNs, transformers are among the state-of-the-art models for sequential processing tasks in natural language processing, computer vision and other important applications. These models rely on the attention mechanism and particularly self-attention that captures the contextual representation as fundamental building blocks for their modeling. Despite their remarkable success, a coherent principled framework for synthesizing attention layers has remained elusive. In [4], I derive the self-attention as the support vector expansion of a given support vector regression (SVR) problem. The primal representation of the regression function

has the form of a neural network layer. Thus, I establish a primal-dual connection between an attention layer in transformers and a neural network layer in deep neural networks. My framework suggests a principled approach to developing an attention mechanism: Starting from a neural network layer and a support vector regression problem, I derive the dual as a support vector expansion to attain the corresponding attention layer. I then employ this principled approach to invent two novel classes of attention: the Batch Normalized Attention (Attention-BN) and the Attention with Scaled Heads (Attention-SH). In the Attention-BN, the similarity between a query and a key is adjusted by the similarity between that key and all the keys in the attention. In the Attention-SH, each attention head approximates the output of the attention at a different resolution. My previous work on using the fast multipole method in computing attention [5] can be combined with the Attention-BN/SH to enhance the model efficiency.

2 Differential equations frameworks for designing deep learning models

Connecting deep architectures to the continuous-time limit of accelerated gradient methods and the diffusion process, I use methods in differential equations to improve these models.

Momentum-based neural ordinary differential equations. Continuous models like the the NeuralODEs are gaining currency due to their ability to learn from irregularly-sampled sequential data and to model complex dynamical systems. Despite their advantages and popularity, the drawback of NeuralODEs is also prominent. NeuralODEs require a very high number of steps to solve the ODEs in both training and inference. Another issue is that NeuralODEs often fail to effectively learn long-term dependencies in sequential data. Motivated by the quadratic convergence rate $\mathcal{O}(1/k^2)$ of the Nesterov's accelerated gradient (NAG) method, in [6], my collaborators and I leverage the continuous limit of the NAG scheme and propose the Nesterov NeuralODE to improve the efficiency of the NeuralODE training and inference. At the core of the Nesterov NeuralODE is replacing the first-order ODE in the NeuralODE with a Nesterov ODE, i.e., a second-order ODE with a time-dependent damping term. To improve the computational efficiency of the model, we convert this second-order ODE into an equivalent system of first-order differential-algebraic equations that are solved in both forward and backward propagations of the NesterovNODE. Our proposed Nesterov NeuralODE has two theoretical properties that imply practical advantages over the NeuralODE. First, the adjoint equation used for training a Nesterov NeuralODE is also a Nesterov NeuralODE, thus accelerating both forward and backward propagation. Second, the spectrum of the Nesterov NeuralODE is well-structured, alleviating the vanishing gradient issue in back-propagation and enabling the model to effectively learn long-term dependencies from sequential data. In this direction, in [7], my colleagues and I also propose the Heavy Ball NeuralODEs whose layers solve the second-order ordinary differential equations (ODEs) limit of the classical heavy ball momentum method to speed up the NeuralODE.

Overcoming the over-smoothing issue in graph neural networks using diffusion process with a source term. Graph neural networks (GNNs) are the backbone for deep learning on graphs, a class of deep learning models that directly operate on graph structures. A well-known problem of GNNs is that increasing the depth of GNNs often results in a significant drop in performance on various graph learning tasks. This performance degradation has been widely interpreted as the over-smoothing issue of GNNs. Moreover, the accuracy of existing GNNs drops severely when they are trained with a limited amount of labeled data. Working with my collaborators, I develop new continuous-depth GNNs that overcome the over-smoothing issue and achieve better accuracy in low-labeling rate regimes [8]. We first present a random walk interpretation of GNNs, revealing a potentially inevitable over-smoothing phenomenon. Based on our random walk viewpoint of GNNs, we propose graph neural diffusion with a source term (GRAND++) that corrects the bias arising from the diffusion process underlying GNNs. GRAND++ theoretically guarantees that: (i) under GRAND++ dynamics, the graph node features do not converge to a constant vector over all nodes even as the time goes to infinity, and (ii) GRAND++ can provide accurate prediction even when it is trained with the limited number of labeled nodes.

3 Statistical frameworks for designing deep learning models

The final aspect of my research is to employ statistical modeling tools to interpret and design deep learning models. In particular, I develop generative models underlying the self-attention mechanism in transformers and deep neural networks (DNNs).

A mixture model framework for attention mechanism in transformers. In [9], I develop a probabilistic framework underlying attention mechanism in transformers and propose a new transformer with a mixture of Gaussian keys (Transformer-MGK), that replaces redundant heads in transformers with a mixture of keys at each head. In particular, I derive a new Gaussian mixture model (GMM) for attention queries. Each Gaussian distribution in this mixture has an attention key as its mean. The posterior distributions of attention keys given attention queries in this mixture model correspond to the attention scores in self-attention, which capture the similarity between queries and keys. As a result, the GMM that I propose provides a principled probabilistic framework to study self-attention in transformers. Given this framework, I discover that a Gaussian distribution centered around each key has a limited capacity to capture the distribution of attention queries since this distribution can be asymmetric, skewed, or even multimodal. Therefore, I further propose to use a mixture of Gaussian keys to increase the representation power of the model so that attention keys can explain the queries better, resulting in the Transformer-MGK with more diverse heads. I then derive the hard E-step inference, soft E-step inference, and the learning algorithm for Transformer-MGK. Along this direction, in [10], I extend the mixture of keys in Transformer-MGK to a new finite admixture of keys (FiAK) for pruning redundant attention scores in transformers. In another work [11], from the observation that attention matrices, which are matrices of attention scores, in transformers lie on a low-dimensional manifold, I propose a new finite admixture of shared heads (FiSH) that generates many local attention matrices from a small set of global attention matrices, thus reducing both computational and memory costs.

In [12], I extend my mixture model framework for self-attention to a nonparametric kernel regression model and propose the FourierFormer, a new class of transformers that the novel generalized Fourier integral kernels I develop to automatically capture the dependency of the data features and remove the need to tune the covariance matrix. Along this nonparametric kernel regression direction, in the joint work with my collaborators [13], I propose a novel robust transformer framework based on robust attention arising from the robust kernel density estimators associated with the robust kernel regression problem. In addition to transformers, in [14–16], my collaborators and I develop a new generative probabilistic model that explicitly captures variation due to latent variables and provides insights into both the successes and shortcomings of DNNs, as well as a principled route to their design.

4 Future research plans

Recently, very large-scale diffusion models have shown remarkable performance in image and video generation tasks. The backbone of these models is transformers, but they are trained on a colossal amount of multimodal data collected from various sources. This sheer number of data from different modalities and the gigantic model size make optimizing the model exceedingly challenging. The success of the diffusion models is rooted in the convex loss designed to train them. For future work and looking forward to a high-impact research direction, I will focus on convex methods for designing and training diffusion models and future deep learning architectures of more substantial size that are trained on more and more heterogeneous data. Due to the increasing complexity of the models and the training data, convexity is key to designing and optimizing the next generation of deep learning models.

The principled models that I have described above can be classified into the class of system 1, which are models that do pattern recognition. Taking a step toward artificial general intelligence–the state at which intelligent agents can understand or learn intellectual tasks–I will incorporate reasoning ability into my models. These models are classified into the class of system 2. I am interested in developing these system 2 models from principled approaches as in my recent work [17].

References

[1] **Tan M Nguyen**, Richard Baraniuk, Andrea Bertozzi, Stanley Osher, and Bao Wang. Momentumrnn: Integrating momentum into recurrent neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1924–1936. Curran Associates, Inc., 2020.

[2] Bao Wang*, **Tan M Nguyen***, Andrea L Bertozzi, Richard G Baraniuk, and Stanley J Osher. Scheduled restart momentum for accelerated stochastic gradient descent. *SIAM Journal on Imaging Sciences*, 2022.

[3] **Tan M Nguyen**, Richard G Baraniuk, Mike Kirby, Stanley J Osher, and Bao Wang. Momentum transformer: Closing the performance gap between self-attention and its linearization. In *Mathematical and Scientific Machine Learning (MSML), 2022*, 2022.

[4] **Tan M Nguyen**, Tam Nguyen, Nhat Ho, Andrea Bertozzi, Richard G Baraniuk, and Stanley J Osher. A primal-dual framework for transformers and neural networks. In *Submitted to The Eleventh International Conference on Learning Representations (ICLR)*, 2023. under review.

[5] **Tan M Nguyen**, Vai Suliafu, Stanley Osher, Long Chen, and Bao Wang. Fmmformer: Efficient and flexible transformer via decomposed near-field and far-field attention. In M. Ranzato, A. Beygelzimer, K. Nguyen, P. S. Liang, J. W. Vaughan, and Y. Dauphin, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 29449–29463. Curran Associates, Inc., 2021.

[6] Nghia Nguyen*, **Tan M Nguyen***, Huyen Vo, Stanley J Osher, and Thieu Vo. Nesterov neural differential equations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[7] Hedi Xia, Vai Suliafu, Hangjie Ji, **Tan M Nguyen**, Andrea Bertozzi, Stanley Osher, and Bao Wang. Heavy ball neural ordinary differential equations. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[8] Matthew Thorpe*, **Tan Minh Nguyen***, Hedi Xia*, Thomas Strohmer, Andrea Bertozzi, Stanley Osher, and Bao Wang. GRAND++: Graph neural diffusion with a source term. In *International Conference on Learning Representations (ICLR)*, 2022.

[9] Tam Nguyen*, **Tan M Nguyen***, Dung Le, Khuong Nguyen, Anh Tran, Richard G Baraniuk, Nhat Ho, and Stanley J Osher. Improving transformers with probabilistic attention keys. *International Conference on Machine Learning (ICML)*, 2022.

[10] **Tan M Nguyen***, Tam Nguyen*, Long Bui*, Hai Do, Dung Le, Hung Tran-The, Khuong Nguyen, Richard G Baraniuk, Nhat Ho, and Stanley J Osher. A probabilistic framework for pruning transformers via a finite admixture of keys. *Under review, Transactions on Machine Learning Research (TMLR)*, 2022.

[11] **Tan M Nguyen***, Tam Nguyen*, Hai Do, Khai Nguyen, Vishwanath Saragadam, Minh Pham, Khuong Nguyen, Nhat Ho, and Stanley J Osher. Fishformer: Transformer with a finite admixture of shared heads, 2022.

[12] **Tan M Nguyen***, Minh Pham*, Tam Nguyen, Khai Nguyen, Stanley J Osher, and Nhat Ho. Transformer with fourier integral attentions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[13] Xing Han*, Tongzheng Ren*, **Tan M Nguyen***, Khai Nguyen, Joydeep Ghosh, and Nhat Ho. Robustify transformers with robust kernel density estimation. In *Submitted to The Eleventh International Conference on Learning Representations (ICLR)*, 2023. under review.

[14] Ankit B Patel, **Tan M Nguyen**, and Richard Baraniuk. A probabilistic framework for deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016.

[15] **Tan M Nguyen***, Nhat Ho*, Ankit Patel, Anima Anandkumar, Michael I Jordan, and Richard G Baraniuk. A bayesian perspective of convolutional neural networks through a deconvolutional generative model. *Under review, Journal of Machine Learning Research (JMLR)*, 2022.

[16] Yujia Huang, James Gornet, Sihui Dai, Zhiding Yu, **Tan M Nguyen**, Doris Tsao, and Anima Anandkumar. Neural networks with recurrent generative feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 535–545, 2020.

[17] Anh Do*, Duy Dinh*, **Tan M Nguyen***, Khuong Nguyen, Stanley Osher, and Nhat Ho. Improving generative flow networks with path regularization. In *Submitted to The Eleventh International Conference on Learning Representations (ICLR)*, 2023. under review.