Research Statement of Tan Minh Nguyen

The growth in scope and complexity of modern datasets 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 in developing **machine learning models** from three principled approaches: 1) **optimization**, 2) **differential equation**, and 3) **statistical modeling**.

From an optimization viewpoint, my goal is to establish connections between deep learning architectures and optimization methods. These connections bring a rich mathematical background in optimization techniques such as robustness and convergence guarantees to the design of deep learning architectures. From a differential equation 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 utilize these principled models to develop large-scale natural language processing and computer vision systems and to solve challenging mathematical modeling problems.

I shall outline major focused areas of my research below. References are numbered as in my CV.

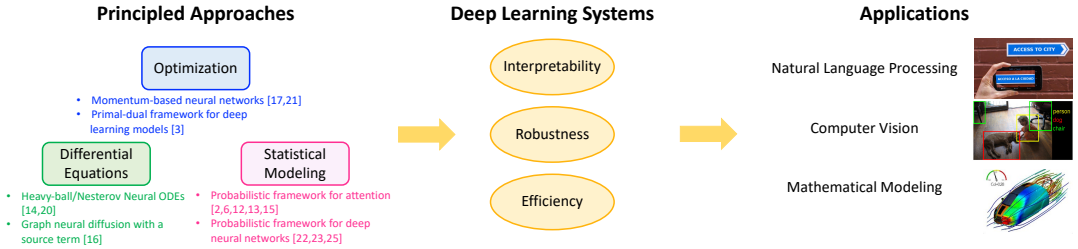

Figure 1: **My research focuses on the interplay of the interpretability, robustness, and efficiency of machine learning models from three principled approaches: 1) optimization, 2) differential equation, and 3) statistical modeling**

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 frameworks to improve them.

Momentum-based neural networks. Recurrent 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 backpropagation through time algorithm and thus may fail to learn long-term dependencies. In [1], I develop a gradient descent analogy of the recurrent cell. I then propose to integrate the momentum used for accelerating gradient dynamics into the recurrent cell, which results in the MomentumRNN. These new MomentumRNN models can alleviate the vanishing gradient problem and accelerate training. The design principle can also 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]. I also extend my momentum framework with an adaptive momentum to design linear attention in transformers [3].

A primal-dual framework for transformers and neural networks. Transformers are among the state-of-the-art models for sequential processing tasks. These models rely on the attention mechanism, particularly self-attention, as fundamental building blocks for their modeling to capture the contextual representation. Despite their remarkable success, a coherent principled framework for synthesizing attention layers has remained elusive. In [4], I derive self-attention as the support vector expansion of a given support vector regression 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 novel classes of attention. My previous work on using the fast multipole method in computing attention [5] can be combined with this primal-dual framework to enhance the model's efficiency.

2 Differential equation frameworks for designing deep learning models

I connect deep learning architectures to the continuous-time limits of accelerated gradient methods and the diffusion process. I then use methods in differential equation to improve these models.

Momentum-based neural ordinary differential equations. Neural ordinary differential equations (NeuralODEs) are gaining currency due to their ability to learn from irregularly-sampled sequential data and to model complex dynamical systems. Despite their advantages, 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 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. Compared to the NeuralODE, our proposed Nesterov NeuralODE has two advantages shared with our previous work in [7]. First, the adjoint equation used in training a Nesterov NeuralODE is a Nesterov NeuralODE, thus accelerating both forward and backward propagation. Second, the spectrum of the Nesterov NeuralODE is well-structured, enabling the model to capture long-term dependencies.

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 well-known problem of GNNs is their over-smoothing issue, i.e., increasing the depth of GNNs often results in a significant decrease in their performance. Moreover, the accuracy of GNNs drops severely when trained with a limited amount of labeled data. My collaborators and I develop new continuous-depth GNNs that overcome the over-smoothing issue and achieve better accuracy in low-labeling rate regimes [8]. We present a random walk interpretation of GNNs, revealing an inevitable over-smoothing phenomenon. Based on our random walk viewpoint, we propose the 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

In the final aspect of my research, I develop generative probabilistic models underlying self-attention in transformers and deep neural networks (DNNs).

Probabilistic frameworks for attention mechanism in transformers and deep neural networks. In [9], I develop a probabilistic framework underlying attention mechanism in transformers. 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. Given this framework, I propose to use a mixture of Gaussian keys (MGK) 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. In [10], I extend the MGK to a new finite admixture of keys (FiAK) for pruning redundant attention scores in transformers. In [11], I propose a new finite admixture of shared heads (FiSH) that generates 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 use the novel generalized Fourier integral kernels I develop to automatically capture the dependency of the data features. In [13], my collaborators and I propose a novel robust transformer framework based on robust attention arising from the robust kernel density estimators. In [14–16], I develop generative probabilistic models underlying DNNs.

4 Future directions

Optimization, differential equation, statistical modeling, and deep learning for scientific and engineering applications. My previous works provide interpretations of deep learning models from mathematical perspectives. I plan to pursue this research direction with a strong focus on developing *interpretable, robust, and efficient deep learning models for scientific and engineering applications*. In the latest work with my postdoc advisor and collaborators, I find an interesting connection between the proximal and Moreau envelope in proximal methods with a solution to the Burgers' Hamilton-Jacobi equation and the attention mechanism in transformers. Our result builds a bridge that connects essential research areas, ranging from optimization, control theory, and diffusion models to optimal transport, mixture models, and machine learning. This bridge enables the integration of deep learning models into (constrained) optimization and control equations governing scientific problems and engineering systems and sheds light on the design principles of deep learning architectures. A particular application that I am interested in is developing *a new class of efficient and robust physics-constrained deep learning models based on the neural differential-algebraic equations and constrained optimization for solving large-scale, high-dimensional, or underdetermined inverse problems*.

Convex methods for large-scale deep learning models. Large-scale diffusion models have recently 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. 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 gigantic size and trained on 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.

Adaptation of large-scale foundation models. The rise of foundation models, such as the diffusion models, empowers a recent paradigm shift in machine learning and artificial intelligence. These very large-scale models are trained on broad data and can be adapted to a variety of downstream tasks via transfer learning, alleviating the cost of designing and training the model from scratch for new applications. As the size of the pre-trained foundation models grows more substantially, fine-tuning all model parameters and deploying the fine-tuned model, for example, in embedded systems, becomes less feasible. To address these issues, I am interested in developing *principled and efficient adaptations of large-scale foundation models for downstream applications*. In particular, my multiresolution transformers [17] can be extended to provide a flexible framework for adaptation at different resolutions. Also, my optimization frameworks can be combined with meta-learning to enable efficient adaptation.

Reasoning models. 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 *from principled approaches*, as in my recent work [18], *for applications in science and engineering*.

* in references indicates equal contribution

References

[1] **Tan M Nguyen**, Richard Baraniuk, Andrea Bertozzi, Stanley Osher, and Bao Wang. Momentumrnn: Integrating momentum into recurrent neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

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

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

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

[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 *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[6] Nghia Nguyen*, **Tan M Nguyen***, Huyen Vo, Stanley Osher, and Thieu Vo. Improving neural ordinary differential equations with nesterovs accelerated gradient method. 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 *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 Baraniuk, Nhat Ho, and Stanley Osher. Improving transformers with probabilistic attention keys. In *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 Baraniuk, Nhat Ho, and Stanley 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 Osher. Improving transformer with an admixture of attention heads. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[12] **Tan M Nguyen***, Minh Pham*, Tam Nguyen, Khai Nguyen, Stanley Osher, and Nhat Ho. Fourierformer: Transformer meets generalized fourier integral theorem. 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. *Under review, International Conference on Learning Representations (ICLR)*, 2023.

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

[15] **Tan M Nguyen***, Nhat Ho*, Ankit Patel, Anima Anandkumar, Michael I Jordan, and Richard 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)*, 2020.

[17] **Tan M Nguyen***, Tho Tran*, Tam Nguyen, Minh Pham, Nhat Ho, and Stanley Osher. Transformers with multiresolution attention heads. *Under review, International Conference on Learning Representations*, 2023.

[18] Anh Do*, Duy Dinh*, **Tan M Nguyen***, Khuong Nguyen, Stanley Osher, and Nhat Ho. Improving generative flow networks with path regularization. *Under review, International Conference on Learning Representations (ICLR)*, 2023.