My research focuses on developing models for scientific machine learning with applications in computational fluid dynamics (e.g. turbulence modeling) and numerical analysis (e.g. efficient numerical solvers) via three principled approaches:

- Optimization (momentum-based neural networks, neural ordinary differential equations)
- Numerical analysis (fast multipole transformers, graph neural diffusion)
- Statistical modeling (deep generative models, mixture framework for transformers)

Below, I classify and summarize my works according to these approaches and the applications.

Approach 1 (Optimization): Momentum-based Deep Learning Model, Neural Ordinary Differential Equations

1. **Tan M. Nguyen**, Richard G. Baraniuk, Andrea Bertozzi, Stanley J. Osher, Bao Wang. **“MomentumRNN: Integrating Momentum into Recurrent Neural Networks”**. Conference on Neural Information Processing Systems (NeurIPS), 2020.

Summary: Designing deep neural networks is an art that often involves an expensive search over candidate architectures. To overcome this for recurrent neural nets (RNNs), we establish a connection between the hidden state dynamics in an RNN and gradient descent (GD). We then integrate momentum into this framework and propose a new family of RNNs, called MomentumRNNs. We theoretically prove and numerically demonstrate that MomentumRNNs alleviate the vanishing gradient issue in training RNNs. We study the momentum long-short term memory (MomentumLSTM) and verify its advantages in convergence speed and accuracy over its LSTM counterpart across a variety of benchmarks. We also demonstrate that MomentumRNN is applicable to many types of recurrent cells, including those in the state-of-the-art orthogonal RNNs. Finally, we show that other advanced momentum-based optimization methods, such as Adam and Nesterov accelerated gradients with a restart, can be easily incorporated into the MomentumRNN framework for designing new recurrent cells with even better performance.

2. Hedi Xia, Vai Suliafu, Hangjie Ji, **Tan M. Nguyen**, Andrea Bertozzi, Stanley J. Osher, Bao Wang. **“Heavy Ball Neural Ordinary Differential Equations”**. Conference on Neural Information Processing Systems (NeurIPS), 2021.

Summary: We propose heavy ball neural ordinary differential equations (HBNODEs), leveraging the continuous limit of the classical momentum accelerated gradient descent, to improve neural ODEs (NODEs) training and inference. HBNODEs have two properties that imply practical advantages over NODEs: (i) The adjoint state of an HBNODE also satisfies an HBNODE, accelerating both forward and backward ODE solvers, thus significantly reducing the number of function evaluations (NFEs) and improving the utility of the trained models. (ii) The spectrum of HBNODEs is well structured, enabling effective learning of long-term dependencies from complex sequential data. We verify the advantages of HBNODEs over NODEs on benchmark tasks, including image classification, learning complex dynamics, and sequential modeling. Our method requires remarkably fewer forward and backward NFEs, is more accurate, and learns long-term dependencies more effectively than the other ODE-based neural network models.

3. Nghia Nguyen*, **Tan M. Nguyen ***, Huyen Vo, Stanley J. Osher, Thieu Vo. **“Nesterov Neural Ordinary Differential Equations”**. Under review, Conference on Neural Information Processing Systems (NeurIPS), 2022.

Summary: We propose the Nesterov neural ordinary differential equations (NesterovNODEs) whose layers solve the second-order ordinary differential equations (ODEs) limit of Nesterov’s accelerated gradient (NAG) method. Taking the advantage of the convergence rate O(k^2) of the NAG scheme, NesterovNODEs speed up training and inference by reducing the number of function evaluations (NFEs) needed to solve the ODEs. We also prove that the adjoint state of a NesterovNODE also satisfies a NesterovNODE, thus accelerating both forward and backward ODE solvers and allowing the model to be scaled up for large-scale tasks. We empirically corroborate the advantage of NesterovNODEs on a wide range of practical applications including point cloud separation, image classification, and sequence modeling. Compared to NODEs, NesterovNODEs require a significantly smaller number of NFEs while achieving better accuracy across our experiments.

4. **Tan M. Nguyen**, Richard G. Baraniuk, Mike Kirby, Stanley J. Osher, Bao Wang. **"Momentum Transformer: Closing the Performance Gap Between Self-attention and Its Linearization"**. Under review, Mathematical and Scientific Machine Learning (MSML), 2022.

Summary: Transformers have achieved remarkable success in sequence modeling and beyond but suffer from quadratic computational and memory complexities with respect to the length of the input sequence. Leveraging techniques include sparse and linear attention and hashing tricks; efficient transformers have been proposed to reduce the quadratic complexity of transformers but significantly degrade the accuracy. In response, we first interpret the linear attention and residual connections in computing the attention map as gradient descent steps. We then introduce momentum into these components and propose the momentum transformer, which utilizes momentum to improve the accuracy of linear transformers while maintaining linear memory and computational complexities. Furthermore, we develop an adaptive strategy to compute the momentum value for our model based on the optimal momentum for quadratic optimization. This adaptive momentum eliminates the need to search for the optimal momentum value and further enhances the performance of the momentum transformer. A range of experiments on both autoregressive and non-autoregressive tasks, including image generation and machine translation, demonstrate that the momentum transformer outperforms popular linear transformers in training efficiency and accuracy.

Approach 2 (Numerical Analysis): Fast Multipole Transformers, Graph Neural Diffusion

1. **Tan M. Nguyen**, Vai Suliafu, Stanley J. Osher, Long Chen, Bao Wang. **"FMMformer: Efficient and Flexible Transformer via Decomposed Near-field and Far-field Attention"**. Conference on Neural Information Processing Systems (NeurIPS), 2021.

Summary: We propose FMMformers, a class of efficient and flexible transformers inspired by the celebrated fast multipole method (FMM) for accelerating interacting particle simulation. FMM decomposes particle-particle interaction into near-field and far-field components and then performs direct and coarse-grained computation, respectively. Similarly, FMMformers decompose the attention into near-field and far-field attention, modeling the near-field attention by a banded matrix and the far-field attention by a low-rank matrix. Computing the attention matrix for FMMformers requires linear complexity in computational time and memory footprint with respect to the sequence length. In contrast, standard transformers suffer from quadratic complexity. We analyze and validate the advantage of FMMformers over the standard transformer on the Long Range Arena and language modeling benchmarks. FMMformers can even outperform the standard transformer in terms of accuracy by a significant margin. For instance, FMMformers achieve an average classification accuracy of 60.74% over the five Long Range Arena tasks, which is significantly better than the standard transformer's average accuracy of 58.70%.

2. Matthew Thorpe*, **Tan M. Nguyen***, Hedi Xia*, Thomas Strohmer, Andrea Bertozzi, Stanley J. Osher, Bao Wang. **"GRAND++: Graph Neural Diffusion with a Source Term"**. International Conference on Learning Representations (ICLR), 2022.

Summary: We propose GRAph Neural Diffusion with a source term (GRAND++) for graph deep learning with a limited number of labeled nodes, i.e., low-labeling rate. GRAND++ is a class of continuous-depth graph deep learning architectures whose theoretical underpinning is the diffusion process on graphs with a source term. The source term guarantees two interesting theoretical properties of GRAND++: (i) the representation of graph nodes, under the dynamics of GRAND++, will not converge to a constant vector over all nodes even as the time goes to infinity, which mitigates the over-smoothing issue of graph neural networks and enables graph learning in very deep architectures. (ii) GRAND++ can provide accurate classification even when the model is trained with a very limited number of labeled training data. We experimentally verify the above two advantages on various graph deep learning benchmark tasks, showing a significant improvement over many existing graph neural networks.

Approach 3 (Statistical Modeling): Deep Generative Models, Mixture Framework for Transformers

1. Ankit B Patel, **Tan M. Nguyen**, Richard Baraniuk. **"A Probabilistic Framework for Deep Learning"**. Conference on Neural Information Processing Systems (NeurIPS), 2016.

Summary: We develop a probabilistic framework for deep learning based on the Deep Rendering Mixture Model (DRMM), a new generative probabilistic model that explicitly capture variations in data due to latent task nuisance variables. We demonstrate that max-sum inference in the DRMM yields an algorithm that exactly reproduces the operations in deep convolutional neural networks (DCNs), providing a first principles derivation. Our framework provides new insights into the successes and shortcomings of DCNs as well as a principled route to their improvement. DRMM training via the Expectation-Maximization (EM) algorithm is a powerful alternative to DCN back-propagation, and initial training results are promising. Classification based on the DRMM and other variants outperforms DCNs in supervised digit classification, training 2-3x faster while achieving similar accuracy. Moreover, the DRMM is applicable to semi-supervised and unsupervised learning tasks, achieving results that are state-of-the-art in several categories on the MNIST benchmark and comparable to state of the art on the CIFAR10 benchmark.

2. Yujia Huang, James Gornet, Sihui Dai, Zhiding Yu, **Tan M. Nguyen**, Doris Tsao, Anima Anandkumar. **"Neural Networks with Recurrent Generative Feedback"**. Conference on Neural Information Processing Systems (NeurIPS), 2020.

Summary: Neural networks are vulnerable to input perturbations such as additive noise and adversarial attacks. In contrast, human perception is much more robust to such perturbations. The Bayesian brain hypothesis states that human brains use an internal generative model to update the posterior beliefs of the sensory input. This mechanism can be interpreted as a form of self-consistency between the maximum a posteriori (MAP) estimation of an internal generative model and the external environment. Inspired by such hypothesis, we enforce self-consistency in neural networks by incorporating generative recurrent feedback. We instantiate this design on convolutional neural networks (CNNs). The proposed framework, termed Convolutional Neural Networks with Feedback (CNN-F), introduces a generative feedback with latent variables to existing CNN architectures, where consistent predictions are made through alternating MAP inference under a Bayesian framework. In the experiments, CNN-F shows considerably improved adversarial robustness over conventional feedforward CNNs on standard benchmarks.

3. **Tan M. Nguyen***, Nhat Ho*, Ankit B. Patel, Anima Anandkumar, Michael I. Jordan, Richard G. Baraniuk. **"A Bayesian Perspective of Convolutional Neural Networks through a Deconvolutional Generative Model"**. Under review, Journal of Machine Learning Research.

Summary: Inspired by the success of Convolutional Neural Networks (CNNs) for supervised prediction in images, we design the Deconvolutional Generative Model (DGM), a new probabilistic generative model whose inference calculations correspond to those in a given CNN architecture. The DGM uses a CNN to design the prior distribution in the probabilistic model. Furthermore, the DGM generates images from coarse to finer scales. It introduces a small set of latent variables at each scale, and enforces dependencies among all the latent variables via a conjugate prior distribution. This conjugate prior yields a new regularizer based on paths rendered in the generative model for training CNNs–the Rendering Path Normalization (RPN). We demonstrate that this regularizer improves generalization, both in theory and in practice. In addition, likelihood estimation in the DGM yields training losses for CNNs, and inspired by this, we design a new loss termed as the Max-Min cross entropy which outperforms the traditional cross-entropy loss for object classification. The Max-Min cross entropy suggests a new deep network architecture, namely the Max-Min network, which can learn from less labeled data while maintaining good prediction performance. Our experiments demonstrate that the DGM with the RPN and the Max-Min architecture exceeds or matches the-state-of-art on benchmarks including SVHN, CIFAR10, and CIFAR100 for semi-supervised and supervised learning tasks.

4. Tam Nguyen*, **Tan M. Nguyen***, Dung Le, Khuong Nguyen, Anh Tran, Richard G. Baraniuk, Nhat Ho*, Stanley J. Osher*. **"Improving Transformers with Probabilistic Attention Keys"**. International Conference on Machine Learning (ICML), 2022.

Summary: Multi-head attention is a driving force behind state-of-the-art transformers which achieve remarkable performance across a variety of natural language processing (NLP) and computer vision tasks. It has been observed that for many applications, those attention heads learn redundant embedding, and most of them can be removed without degrading the performance of the model. Inspired by this observation, we propose Transformer with a Mixture of Gaussian Keys (Transformer-MGK), a novel transformer architecture that replaces redundant heads in transformers with a mixture of keys at each head. These mixtures of keys follow a Gaussian mixture model and allow each attention head to focus on different parts of the input sequence efficiently. Compared to its conventional transformer counterpart, Transformer-MGK accelerates training and inference, has fewer parameters, and requires less FLOPs to compute while achieving comparable or better accuracy across tasks. Transformer-MGK can also be easily extended to use with linear attentions. We empirically demonstrate the advantage of Transformer-MGK in a range of practical applications including language modeling and tasks that involve very long sequences. On the Wikitext-103 and Long Range Arena benchmark, Transformer-MGKs with 4 heads attain comparable or better performance to the baseline transformers with 8 heads.

5. **Tan M. Nguyen***, Tam Nguyen*, Hai Do, Khai Nguyen, Vishwanath Saragadam, Minh Pham, Khuong Nguyen, Nhat Ho, Stanley J. Osher. **"FiSHFormer: Transformer with a Finite Admixture of Shared Heads"**. Under review, Conference on Neural Information Processing Systems (NeurIPS), 2022.

Summary: Transformers with multi-head self-attention have achieved remarkable success in sequence modeling and beyond. However, they suffer from high computational and memory complexities for computing the attention matrix at each head. Recently, it has been shown that those attention matrices lie on a low-dimensional manifold and, thus, are redundant. We propose the Transformer with a Finite Admixture of Shared Heads (FiSHformers), a novel class of efficient and flexible transformers that allow the sharing of attention matrices between attention heads. At the core of FiSHformer is a novel finite admixture model of shared heads (FiSH) that samples attention matrices from a set of global attention matrices. The number of global attention matrices is much smaller than the number of local attention matrices that they generate. FiSHformers directly learn these global attention matrices rather than the local ones as in other transformers, thus significantly improving the computational and memory efficiency of the model. We empirically verify the advantages of the FiSHformer over the baseline transformers on a wide range of practical applications including language modeling, machine translation, and image classification. On the WikiText-103, IWSLT'14 De-En and WMT'14 En-De, FiSHformers use much fewer floating-point operations per second (FLOPs), memory, and parameters compared to the baseline transformers.

6. **Tan M. Nguyen***, Tam Nguyen*, Long Bui, Hai Do, Dung Le, Hung Tran-The, Khuong Nguyen, Richard G. Baraniuk, Nhat Ho, Stanley J. Osher. **"A Probabilistic Framework for Pruning Transformers via a Finite Admixture of Keys"**. Under review, European Conference on Computer Vision (ECCV), 2022.

Summary: Pairwise dot product-based self-attention is key to the success of transformers which achieve state-of-the-art performance across a variety of applications in language and vision, but are costly to compute. However, it has been shown that most attention scores and keys in transformers are redundant and can be removed without loss of accuracy. In this paper, we develop a novel probabilistic framework for pruning attention scores and keys in transformers. We first formulate an admixture model of attention keys whose input data to be clustered are attention queries. We show that attention scores in self-attention correspond to the posterior distribution of this model when attention keys admit a uniform prior distribution. We then relax this uniform prior constraint and let the model learn these priors from data, resulting in a new Finite Admixture of Keys (FiAK). The learned priors in FiAK are used for pruning away redundant attention scores and keys in the baseline transformers, improving the diversity of attention patterns that the models capture. We corroborate the efficiency of transformers pruned with FiAK on practical tasks including ImageNet object classification, COCO object detection, and WikiText-103 language modeling. Our experiments demonstrate that transformers pruned with FiAK yield similar or even better accuracy than the baseline dense transformers while being much more efficient in terms of memory and computational cost.

7. **Tan M. Nguyen***, Minh Pham*, Tam Nguyen, Khai Nguyen, Stanley J. Osher, Nhat Ho. **"Transformer with Fourier Integral Attentions"**. Under review, Conference on Neural Information Processing Systems (NeurIPS), 2022.

Summary: Multi-head attention empowers the recent success of transformers, the state-of-the-art models that have achieved remarkable success in sequence modeling and beyond. These attention mechanisms compute the pairwise dot products between the queries and keys, which results from the use of unnormalized Gaussian kernels with the assumption that the queries follow a mixture of Gaussian distribution. There is no guarantee that this assumption is valid in practice. In response, we first interpret attention in transformers as a nonparametric kernel regression. We then propose the FourierFormer, a new class of transformers in which the dot-product kernels are replaced by the novel generalized Fourier integral kernels. Different from the dot-product kernels, where we need to choose a good covariance matrix to capture the dependency of the features of data, the generalized Fourier integral kernels can automatically capture such dependency and remove the need to tune the covariance matrix. We theoretically prove that our proposed Fourier integral kernels can efficiently approximate any key and query distributions. Compared to the conventional transformers with dot-product attention, FourierFormers attain better accuracy and reduce the redundancy between attention heads. We empirically corroborate the advantages of FourierFormers over the baseline transformers in a variety of practical applications including language modeling and image classification.

Applications: Turbulence Modeling, Efficient Numerical Solvers

1. Gavin D. Portwood, Peetak P. Mitra, Mateus Dias Ribeiro, **Tan M. Nguyen**, Balasubramanya T. Nadiga, Juan A. Saenz, Michael Chertkov, Animesh Garg, Anima Anandkumar, Andreas Dengel, Richard G. Baraniuk, David P. Schmidt. **"Turbulence Forecasting via Neural ODE"**. NeurIPS Workshop on Machine Learning and the Physical Sciences, 2019.

Summary: Fluid turbulence is characterized by strong coupling across a broad range of scales. Furthermore, besides the usual local cascades, such coupling may extend to interactions that are non-local in scale-space. As such the computational demands associated with explicitly resolving the full set of scales and their interactions, as in the Direct Numerical Simulation (DNS) of the Navier-Stokes equations, in most problems of practical interest are so high that reduced modeling of scales and interactions is required before further progress can be made. While popular reduced models are typically based on phenomenological modeling of relevant turbulent processes, recent advances in machine learning techniques have energized efforts to further improve the accuracy of such reduced models. In contrast to such efforts that seek to improve an existing turbulence model, we propose a machine learning (ML) methodology that captures, de novo, underlying turbulence phenomenology without a pre-specified model form. To illustrate the approach, we consider transient modeling of the dissipation of turbulent kinetic energy—a fundamental turbulent process that is central to a wide range of turbulence models—using a Neural ODE approach. After presenting details of the methodology, we show that this approach out-performs state-of-the-art approaches.

2. **Tan M. Nguyen**, Animesh Garg, Richard G Baraniuk, Anima Anandkumar. **"InfoCNF: An Efficient Conditional Continuous Normalizing Flow with Adaptive Solvers"**. ICML Workshop on Invertible Neural Nets and Normalizing Flows, 2019.

Summary: Continuous Normalizing Flows (CNFs) have emerged as promising deep generative models for a wide range of tasks thanks to their invertibility and exact likelihood estimation. However, conditioning CNFs on signals of interest for conditional image generation and downstream predictive tasks is inefficient due to the highdimensional latent code generated by the model, which needs to be of the same size as the input data. In this paper, we propose InfoCNF, an efficient conditional CNF that partitions the latent space into a class-specific supervised code and an unsupervised code that shared among all classes for efficient use of labeled information. Since the partitioning strategy (slightly) increases the number of function evaluations (NFEs), InfoCNF also employs gating networks to learn the error tolerances of its ordinary differential equation (ODE) solvers for better speed and performance. We show empirically that InfoCNF improves the test accuracy over the baseline while yielding comparable likelihood scores and reducing the NFEs on CIFAR10. Furthermore, applying the same partitioning strategy in InfoCNF on time-series data helps improve extrapolation performance.

Other Papers

1. Bao Wang*, **Tan M. Nguyen***, Andrea L. Bertozzi, Richard G. Baraniuk, Stanley J. Osher. **"Scheduled Restart Momentum for Accelerated Stochastic Gradient Descent"**. SIAM Journal on Imaging Sciences, 2022.

Summary: Stochastic gradient descent (SGD) with constant momentum and its variants such as Adam are the optimization algorithms of choice for training deep neural networks (DNNs). Since DNN training is incredibly computationally expensive, there is great interest in speeding up the convergence. Nesterov accelerated gradient (NAG) improves the convergence rate of gradient descent (GD) for convex optimization using a specially designed momentum; however, it accumulates error when an inexact gradient is used (such as in SGD), slowing convergence at best and diverging at worst. In this paper, we propose Scheduled Restart SGD (SRSGD), a new NAG-style scheme for training DNNs. SRSGD replaces the constant momentum in SGD by the increasing momentum in NAG but stabilizes the iterations by resetting the momentum to zero according to a schedule. Using a variety of models and benchmarks for image classification, we demonstrate that, in training DNNs, SRSGD significantly improves convergence and generalization; for instance in training ResNet200 for ImageNet classification, SRSGD achieves an error rate of 20.93% vs. the benchmark of 22.13%. These improvements become more significant as the network grows deeper. Furthermore, on both CIFAR and ImageNet, SRSGD reaches similar or even better error rates with significantly fewer training epochs compared to the SGD baseline.

2. Yue Wang, Jianghao Shen, Ting-Kuei Hu, Pengfei Xu, **Tan M. Nguyen**, Richard Baraniuk, Zhangyang Wang, Yingyan Lin. **"Dual Dynamic Inference: Enabling More Efficient, Adaptive, and Controllable Deep Inference"**. IEEE Journal of Selected Topics in Signal Processing, 2020.

Summary: State-of-the-art convolutional neural networks (CNNs) yield record-breaking predictive performance, yet at the cost of high-energy-consumption inference, that prohibits their widely deployments in resource-constrained Internet of Things (IoT) applications. We propose a dual dynamic inference (DDI) framework that highlights the following aspects: 1) we integrate both input-dependent and resource-dependent dynamic inference mechanisms under a unified framework in order to fit the varying IoT resource requirements in practice. DDI is able to both constantly suppress unnecessary costs for easy samples, and to halt inference for all samples to meet hard resource constraints enforced; 2) we propose a flexible multigrained learning to skip (MGL2S) approach for input-dependent inference which allows simultaneous layer-wise and channelwise skipping; 3) we extend DDI to complex CNN backbones such as DenseNet and show that DDI can be applied towards optimizing any specific resource goals including inference latency and energy cost. Extensive experiments demonstrate the superior inference accuracy-resource trade-off achieved by DDI, as well as the flexibility to control such a trade-off as compared to existing peer methods. Specifically, DDI can achieve up to 4 times computational savings with the same or even higher accuracy as compared to existing competitive baselines.