Speakers

1 Jeffrey Adie

2 Francis Bach

3 Prasanna Balaprakash

4 Mikhail Belkin

5 Andrea Bertolini

6 Steven Brunton

7 Erik Cambria

8 Li Cheng

9 Michael Choi

10 Anca Dragan

11 Jesper Dramsch

12 Weinan E

13 Jianqing Fan

14 Yang Feng

15 Cornelia Fermüller

16 Sebastian Goldt

17 Boris Hanin

18 Hamed Hassani

19 Hui Ji

20 Nikita Kazeev

21 Mohammad Emtiyaz Khan

22 Pang Wei Koh

23 Gitta Kutyniok

24 Qianxiao Li

25 Min Lin

26 Lydia Liu

27 Seth Neel

28 Juan-Pablo Ortega

29 Guillaume Sartoretti

30 Ohad Shamir

31 Harold Soh

32 Jascha Sohl-Dickstein

33 Linda Tan

34 Vincent Tan

35 Yan Shuo Tan

36 Andrey Ustyuzhanin

37 Rene Vidal

38 Yu-Ping Wang

39 Greg Yang

40 Haizhao Yang

41 Angela Yao

42 Joey Zhou Tianyi

43 Pan Zhou

44 Enrique Zuazua

Abstracts

Machine Learning and Its Applications

(17–28 Oct 2022)

1 Jeffrey Adie

NVIDIA, Singapore
Applied machine learning for climate and weather modelling.

Abstract

Climate change is the defining problem of our time and an existential threat to humanity. A key goal of the climate and weather community is to provide the best possible predictions to policy makers but long-term prediction is inherently challenging to get right and uncertainties can limit action to fight climate change. Current estimates are that the required traditional computational capabilities will not be feasible before 2060 at the earliest. Artificial Intelligence however gives us a new and powerful method to substantially accelerate time to solution with recent advances in machine learning. In this talk, I will discuss various applications for ML in Climate and weather modelling and present a new disruptive breakthrough from NVIDIA in data-driven forecasting.

2 Francis Bach

INRIA/ENS, France
Information theory through kernel methods

Abstract

Estimating and computing entropies of probability distributions are key computational tasks throughout data science. In many situations, the underlying

distributions are only known through the expectation of some feature vectors, which has led to a series of works within kernel methods. In this talk, I will explore the particular situation where the feature vector is a rank-one positive definite matrix, and show how the associated expectations (a covariance matrix) can be used with information divergences from quantum information theory to draw direct links with the classical notions of Shannon entropies.

3 Prasanna Balaprakash

Argonne National Laboratory, USA
Scalable automated machine learning with DeepHyper

Abstract

In recent years, deep neural networks (DNNs) have achieved considerable success in learning complex nonlinear relationships between features and targets from large datasets. Nevertheless, designing high-performing DNN architecture for a given data set is an expert-driven, time-consuming, trial-and-error manual task. A major bottleneck in the construction of DNNs is the vast search space of architectures that need to be explored in the face of new data sets. Moreover, DNNs typically require user-specified values for hyperparameters, which strongly influence performance factors such as training time and prediction accuracy. In this talk, we will introduce DeepHyper [1], a scalable automated machine learning package for developing a diverse set of deep neural network ensembles and leveraging them for improved prediction and uncertainty quantification in scientific machine learning applications. DeepHyper provides an infrastructure that targets experimental research in neural architecture search (NAS) and hyperparameter search (HPS) methods, scalability, and portability across different U.S. Department of Energy supercomputers.

References

[1] https://deephyper.readthedocs.io/en/latest/

4 Mikhail Belkin

University of California, San Diego, USA
Neural networks, wide and deep, singular kernels and Bayes optimality

Abstract

Wide and deep neural networks are used in many important practical setting. In this talk I will discuss some aspects of width and depth related to optimization and generalization. I will first discuss what happens when neural networks become infinitely wide, giving a general result for the transition to linearity (i.e., showing that neural networks become linear functions of parameters) for a broad class of wide neural networks corresponding to directed graphs.

I will then proceed to the question of depth, showing equivalence between infinitely wide and deep fully connected networks trained with gradient descent and Nadaraya-Watson predictors based on certain singular kernels. Using this connection we show that for certain activation functions these wide and deep networks are (asymptotically) optimal for classification but, interestingly, never for regression.

Based on joint work with Chaoyue Liu, Adit Radhakrishnan, Caroline Uhler and Libin Zhu.

5 Andrea Bertolini

Sant'Anna School of Advanced Studies – Pisa, Italy
Bringing innovation to the market.
Regulatory approaches to AI and the case of Europe

Abstract

The lecture will present a brief account of the evolution of the regulatory attempts in the field of advanced technologies, in particular AI. To this end, it will focus on some grounding concepts such as (i) what is regulation and (ii) what is its purpose; (iii) a bottom-up versus a top-down approach, and (iv) (the viability and limits of) technology neutral approaches; (v) the role of standardization; (vi) the Brussels-effect.

Based on such an account, it will then discuss the most relevant attempt to-date: the AI Act, presented by the European Commission in April 2021,

commenting on its structures and founding principles, describing its potential weaknesses and strengths. A similar analysis will then briefly discuss civil liability and the obligations to repay damages.

6 Steven Brunton

University of Washington, USA
Machine learning for scientific discovery, with examples in fluid mechanics

Abstract

This work describes how machine learning may be used to develop accurate and efficient nonlinear dynamical systems models for complex natural and engineered systems. We explore the sparse identification of nonlinear dynamics (SINDy) algorithm, which identifies a minimal dynamical system model that balances model complexity with accuracy, avoiding overfitting. This approach tends to promote models that are interpretable and generalizable, capturing the essential "physics" of the system. We also discuss the importance of learning effective coordinate systems in which the dynamics may be expected to be sparse. This sparse modeling approach will be demonstrated on a range of challenging modeling problems in fluid dynamics, and we will discuss how to incorporate these models into existing model-based control efforts. Because fluid dynamics is central to transportation, health, and defense systems, we will emphasize the importance of machine learning solutions that are interpretable, explainable, generalizable, and that respect known physics.

7 Erik Cambria

Nanyang Technological University, Singapore
Neurosymbolic AI for Social Media Applications

Abstract

With the recent developments of deep learning, AI research has gained new vigor and prominence. However, machine learning still faces three big challenges: (1) it requires a lot of training data and is domain-dependent; (2)

different types of training or parameter tweaking leads to inconsistent results; (3) the use of black-box algorithms makes the reasoning process uninterpretable. At SenticNet, we address such issues in the context of NLP via sentic computing, a neurosymbolic approach that aims to bridge the gap between statistical NLP and the many other disciplines necessary for understanding human language such as linguistics, commonsense reasoning, and affective computing. Sentic computing is both top-down and bottom-up: top-down because it leverages symbolic models such as semantic networks and conceptual dependency representations to encode meaning; bottom-up because it uses subsymbolic methods such as deep neural networks and multiple kernel learning to infer syntactic patterns from data.

8 Li Cheng

National University of Singapore, Singapore
Bayesian Fixed-domain Asymptotics for Covariance Parameters in Gaussian Process Regression

Abstract

Gaussian process regression models are widely used in machine learning, computer experiments, and Bayesian inference for spatially referenced data. We discuss some recent advances in the theory of Bayesian Gaussian process regression models without and with the nugget effect from measurement error. We mainly study the Bayesian estimation of parameters in the Gaussian process covariance function under the fixed-domain asymptotics regime, as well as its implications on the Bayesian posterior prediction. For the model without nugget, also known as universal kriging, we derive the limiting joint posterior distribution of the microergodic parameter and the range parameter. We further show that the Bayesian kriging predictor satisfies the posterior asymptotic efficiency in linear prediction. For the more challenging model with a nugget, we propose a new framework to derive the Bayesian posterior contraction rates for the microergodic parameter and the nugget. We illustrate these theoretical results with numerical experiments and real data analysis.

9 Michael Choi

National University of Singapore, Singapore
Landscape modification meets spin systems: from torpid to rapid mixing and tunneling in the low-temperature regime

Abstract

This talk centers around a technique that we call landscape modification. The core idea is that the Hamiltonian function is suitably modified in a way for rapid mixing while maintaining proximity with the original target distribution. We first present model-independent results that give rapid mixing and tunneling in the low-temperature regime. Building upon these results, we investigate the effect of landscape modification on some prototypical statistical physics models including the Ising model on various deterministic and random graphs as well as Derrida's random energy model. This talk highlights a novel use of the geometry and structure of the landscape, in particular the global minimum value or its estimate, to the design of accelerated samplers or optimizers.

10 Anca Dragan

University of California, Berkeley, USA
Challenges in learning for and from interaction with people

Abstract

AI agents that interact with people encounter fascinating challenges from a learning perspective. They have heightened sample complexity and robustness bars. We do not have great human simulators so they have to learn in real interaction, and they have to do so safely. They need to optimize for what people want, even though people are themselves suboptimal in ways that we do not understand, let alone formally model. Their actions influence human behavior, beliefs, and even preferences. In this talk, we will touch upon some of these challenges and on ways to make progress in human-AI interaction.

11 Jesper Dramsch

ECMWF, Germany
Integrating machine learning into operational weather forecasts at the ECMWF

Abstract

The European Centre for Medium-Range Weather Forecasts is in the process of delivering a 10-year strategy for machine learning in the numerical weather and climate prediction. This strategy involves specific objectives regarding the prediction quality and performance of operational forecasts, as both accuracy and the timely delivery of weather forecasts are central to the work at ECMWF. Weather forecasting provides a unique opportunity to combine knowledge from multiple domains including geosciences, simulations, and statistics with data-driven machine learning models. This talk will present a general overview of different areas where machine learning can improve these objectives and then go into specific applications and their challenges. Specifically, I will discuss different post-processing techniques of forecasts that provide unique challenges regarding the amount of data, mathematical foundations and domain knowledge to approach this topic, which provides insights into bridging the gap between classically trained machine learning experts and decades of knowledge and conventions from subject matter experts.

12 Weinan E

Peking University, China
Towards a mathematical theory of machine learning

Abstract

Given a machine learning model, what are the class of functions that can be approximated by this particular model efficiently, in the sense that the convergence rate for the approximation, estimation and optimization errors does not deteriorate as dimensionality goes up? We address this question for three classes of machine learning models: The random feature model, two-layer neural networks and the residual neural network model. We will also discuss, in the over-parametrized regime, how different optimization algorithms pick different global minimum. During the process, we will also

summarize the current status of the theoretical foundation of deep learning, and discuss some of the key open questions.

13 Jianqing Fan

Princeton University, USA
The Efficacy of Pessimism in Asynchronous Q-Learning

Abstract

Motivated by the recent advances in offline reinforcemen learning, we develop an algorithmic framework that incorporates the principle of pessimism into asynchronous Q-learning, which penalizes infrequently-visited state-action pairs based on suitable lower confidence bounds (LCBs). This framework leads to, among other things, improved sample efficiency and enhanced adaptivity in the presence of near-expert data. Our approach permits the observed data in some important scenarios to cover only partial state-action space, which is in stark contrast to prior theory that requires uniform coverage of all state-action pairs. When coupled with the idea of variance reduction, asynchronous Q-learning with LCB penalization achieves near-optimal sample complexity, provided that the target accuracy level is small enough. In comparison, prior works were suboptimal in terms of the dependency on the effective horizon even when i.i.d. sampling is permitted. Our results deliver the first theoretical support for the use of pessimism principle in the presence of Markovian non-i.i.d. data. (Joint work with Yuling Yan, Gen Li, and Yuxin Chen)

14 Yang Feng

New York University, USA
Transfer learning under high-dimensional generalized linear models

Abstract

In this work, we study the transfer learning problem under high-dimensional generalized linear models (GLMs), which aim to improve the fit on *target* data by borrowing information from useful *source* data. Given which sources to transfer, we propose a transfer learning algorithm on GLM, and derive

its ℓ_1/ℓ_2-estimation error bounds as well as a bound for a prediction error measure. The theoretical analysis shows that when the target and source are sufficiently close to each other, these bounds could be improved over those of the classical penalized estimator using only target data under mild conditions. When we don't know which sources to transfer, an *algorithm-free* transferable source detection approach is introduced to detect informative sources. The detection consistency is proved under the high-dimensional GLM transfer learning setting. We also propose an algorithm to construct confidence intervals of each coefficient component, and the corresponding theories are provided. Extensive simulations and a real-data experiment verify the effectiveness of our algorithms. We implement the proposed GLM transfer learning algorithms in a new R package `glmtrans`, which is available on CRAN.

15 Cornelia Fermüller

University of Maryland, USA
Bio-inspired visual motion analysis

Abstract

Visual motion interpretation is the core of many real-world AI applications, including self-driving cars, robotics, augmented reality, and human motion analysis. Classical computational approaches are based on computing correspondence, i.e., matching image points, in consecutive image frames which are then used to reconstruct scene models. However, in biology, we find systems with low computational power that do not compute correspondence but are very efficient in using visual motion. Their principles have not been translated to our computational approaches yet. In this talk I will describe work along three directions on using visual motion signals for geometric motion tasks and ask where machine learning should be used vs classic optimization in such tasks. First, neuromorphic event-based sensors which do not record image frames but temporal information about scene changes provide us with data in the form of point clouds in space time that approximate continuous motion. Exploiting the advantages of this data, we developed scene segmentation algorithms that function in the most challenging scenarios. Second, by changing the sequence of computations, and estimating 3D motion from robust filter output, we have developed optimization and machine learning

algorithms that are robust and generalize better to new scenarios. Third, experiments on visual illusions will be shown that give an indication of the motion computations in the early visual processes in nature and point to directions for improving current motion computations.

16 Sebastian Goldt

SISSA, Italy
What do neural networks learn? On the interplay between data structure and representation learning

Abstract

Neural networks are powerful feature extractors - but which features do they extract from their data? And how does the structure in the data shape the representations they learn? We investigate these questions by introducing several synthetic data models, each of which accounts for a salient feature of modern data sets: low intrinsic dimension of images [1],symmetries and non-Gaussian statistics [2], and finally sequence memory [3]. Using tools from statistics and statistical physics, we will show how the learning dynamics and the representations are shaped by the statistical properties of the training data.

References

[1] Goldt, Mézard, Krzakala, Zdeborová (2020) Physical Review X 10 (4), 041044 [arXiv:1909.11500]

[2] Ingrosso & Goldt (2022) [arXiv:2202.00565]

[3] Seif, Loos, Tucci, Roldán, Goldt [arXiv:2205.14683]

17 Boris Hanin

Princeton University, USA
Exactly Solving Bayesian Interpolation with Deep Linear Networks

Abstract

This talk concerns Bayesian interpolation with an overparameterized depth L linear network of input dimension N_0 and width N equipped with a Gaussian prior on its weights and a MSE log-likelihood on P training datapoints. I will present a recent set results, joint with Alexander Zlokapa (MIT Physics), in which we obtain an exact representation for the posterior distribution of the predictor, which holds for any choices of N_0, N, L, P, in terms of a class of special functions known as Meijer G-functions. I will describe how these expressions give new and surprisingly rich insights into the interplay of depth, width and number of datapoints in the triple scaling limit in which N_0, P, N, L all tend to infinity as fixed limiting ratios P / N_0, L / N, and P / N.

18 Hamed Hassani

University of Pennsylvania, USA
The Curse of Overparametrization in (adversarial) Robustness

Abstract

Successful deep learning models often involve training neural network architectures that contain more parameters than the number of training samples. Such overparametrized models have been extensively studied in recent years, and the virtues of overparametrization have been established from both the statistical perspective, via the double-descent phenomenon, and the computational perspective via the structural properties of the optimization landscape.

Despite the remarkable success of deep learning architectures in the overparametrized regime, it is also well known that these models are highly vulnerable to small adversarial perturbations in their inputs. Even when adversarially trained, their performance on perturbed inputs (robust generalization) is considerably worse than their best attainable performance on benign inputs (standard generalization). It is thus imperative to understand how overparametrization fundamentally affects robustness.

In this talk, I will provide a precise characterization of the role of overparametrization on robustness by focusing on simple linear regression models as well as random features regression models (two-layer neural networks with random first layer weights). I consider a regime where the sample size, the input dimension and the number of parameters grow in proportion to each other, and derive an asymptotically exact formula for the robust general-

ization error when the model is adversarially trained. Our developed theory reveals fundamental tradeoffs between robustness and accuracy as well as the non-trivial effect of overparametrization on robustness. As one of the main messages, I show that for adversarially trained models, high overparametrization can hurt robust generalization.

19 Hui Ji

National University of Singapore, Singapore
Self-supervised Deep Learning for Inverse Problems in Imaging

Abstract

In recent years, deep learning has emerged as a highly successful tool in many domains, including inverse imaging problems. Most existing successful deep learning methods are based on supervised learning, which requires many ground-truth images for training a deep neural network (DNN). Such a prerequisite on training datasets limits their applicability in data-limited domains, e.g., medicine and science. This talk will introduce a series of works on self-supervised learning for solving inverse imaging problems, which teaches a DNN to predict images from their noisy and partial measurements without seeing any related image. The main ingredient in these works is the neutralization of Bayesian inference with DNN-based over-parametrization of images. Coming as a surprise to most, even without given any training data, the proposed self-supervised method can compete well against supervised learning methods in many real-world imaging tasks.

20 Nikita Kazeev

National University of Singapore, Singapore
Fast and accurate 2D material property learning by graph neural networks using sparse defect representation

Abstract

Two-dimensional materials represent great potential and challenge at the same time. Such materials can exhibit various desired quantum emission, electronic, and optoelectronic properties to be used in solid-state devices.

Our particular interest is the possibility of obtaining desirable properties via controlled defect introduction [1]. The search space for such crystals is enormous, and proper ab-initio computations become prohibitively expensive. It makes essential to have a way to predict the properties of a crystal with a specific defect configuration. We propose a machine learning approach for rapidly estimating 2D material properties given lattice structure and defect configuration. The method suggests a sparse representation of a 2D material configuration that allows for a neural network to train quickly and accurately. We compare it with the state-of-the-art approaches [2, 3] and demonstrate a decrease in mean absolute error up to 8.5 times and a robustness increase. Also, our approach is an order of magnitude more resource-efficient than its contenders both for the training and inference part. The proposed representation can be easily expanded to 3D cases.

References

[1] I. Aharonovich, D. Englund, and M. Toth. "Solid-state single-photon emitters". In: Nature Photonics 10.10 (2016), pp.631–641.

[2] Kristof T Schütt et al. "Schnet–a deep learning architecture for molecules and materials". In: The Journal of Chemical Physics 148.24 (2018), p. 241722.

[3] Johannes Klicpera, Florian Becker, and Stephan Günnemann. GemNet: Universal Directional Graph Neural Networks for Molecules. 2021. doi: 10.48550/ARXIV.2106.08903. url: https://arxiv.org/abs/2106.08903.

21 Mohammad Emtiyaz Khan

RIKEN, Japan
The Bayesian learning rule for adaptive AI

Abstract

Humans and animals have a natural ability to autonomously learn and quickly adapt to their surroundings. How can we design AI systems that do the same? In this talk, I will present Bayesian principles to bridge such gaps between humans and AI. I will show that a wide-variety of machine-learning

algorithms are instances of a single learning-rule called the Bayesian learning rule. The rule unravels a dual perspective yielding new adaptive mechanisms for machine-learning based AI systems. My hope is to convince the audience that Bayesian principles are indispensable for an AI that learns as efficiently as we do.

22 Pang Wei Koh

Stanford University, USA
Benchmarks and methods for real-world distribution shifts

Abstract

Distribution shifts – where the training distribution differs from the test distribution – can substantially degrade the accuracy of machine learning (ML) systems deployed in the wild. In this talk, I will discuss our recent work on building benchmarks and methods for mitigating these distribution shifts. First, I will describe WILDS – a benchmark of in-the-wild distribution shifts spanning applications such as pathology, conservation, remote sensing, and drug discovery – and show how existing methods that are state-of-the-art on synthetic distribution shifts can still fail to be robust on these real-world shifts. I will then discuss and analyze simple data augmentation approaches that leverage prior knowledge to achieve state-of-the-art performance on two of the WILDS datasets.

23 Gitta Kutyniok

Ludwig-Maximilians-Universität München, Germany
Reliable AI: Vision or Illusion?

Abstract

Artificial intelligence is currently leading to one breakthrough after the other, both in public life with, for instance, autonomous driving and speech recognition, and in the sciences in areas such as medical diagnostics or molecular dynamics. However, one current major drawback is the lack of reliability of such methodologies.

In this lecture we will first provide an introduction into this vibrant research area, focussing specifically on deep neural networks. We will then

present some recent advances, in particular, concerning explainability. Finally, we will discuss fundamental limitations of deep neural networks and related approaches in terms of computability, which seriously affects their reliability.

24 Qianxiao Li

National University of Singapore, Singapore
Approximation Theory of Deep Learning from the Dynamical Systems Viewpoint

Abstract

In this talk, we present some recent results on the approximation theory of deep learning from the dynamical systems viewpoint. This viewpoint highlights a key new aspect of modern deep learning, namely the presence of compositional/dynamical structures. We first discuss mathematical frameworks to study the capacity of deep feed-forward architectures for function approximation. Next, we discuss approximation theories of modern architectures for sequence modelling, including recurrent neural networks, dilated convolutional networks (WaveNet), and encoder-decoder structures. These analyses reveal some interesting connections between approximation, dynamics, memory, sparsity and low rank phenomena that may guide the practical selection and design of these network architectures.

25 Min Lin

SEA, Singapore
A Deep Learning Approach to Kohn-Sham Density Functional Theory

Abstract

Kohn-Sham Density Functional Theory (KS-DFT) has been traditionally solved by the Self-Consistent Field (SCF) method. Behind the SCF loop is the physics intuition of solving a system of independent single electron wave functions under an effective potential. In this work, we demonstrate that direct energy minimization can be a competitive alternative with techniques borrowed from deep learning. With the wave function orthogonality constraint reparameterized as a feed-forward computation, we can rely on the

automatic differentiation engine to compute the gradient of the total energy w.r.t the wave-function parameters. Reparameterized constraint enables us to convert the integration which is usually based on full batch into minibatch SGD. We demonstrate that amortizing the integration over the optimization steps makes the computation more scalable. We further demonstrate that with the above changes, new families of basis functions can be enabled and are worth exploring as the next step.

26 Lydia Liu

University of California, Berkeley, USA
Lost in translation: reimagining the machine learning life cycle in education

Abstract

Machine learning (ML) techniques are prevalent in education, predicting student dropout (Tamhane et al, 2014) to admissions (Waters and Miikkulainen, 2013). Concurrently, a number of algorithmic solutions in education have led to negative and disparate outcomes. Such unintended consequences demonstrate a continued gap between intent and impact in ML for education domain applications ("ML4Ed"). To address the divide between intent and impact in existing ML research, we present qualitative insights from interviews with education domain experts, grounded in ML for education (ML4Ed) papers published in preeminent applied ML conferences over the past decade. Our central research goal is to critically examine whether the stated or implied societal objectives of these papers are aligned with the ML problem formulation, objectives, and interpretation of results. Our work joins a growing number of meta-analytical studies as well as critical analyses of the societal impact of ML. Specifically, by engaging education researchers in discussions of current machine learning research, we used a cross-disciplinary lens to identify fairness blind spots in machine learning research applied to education contexts. Our two main findings were the misalignment of machine learning experts and education researchers with respect to the problem formulation and the limits of prediction tasks. Based on our findings, we propose an extended ML life cycle that highlights two translational challenges in ML4Ed: the translation of education goals via ML problem formulation and the translation of predictions to interventions.

27 Seth Neel

Harvard Business School, USA
Adaptive Machine Unlearning

Abstract

Data deletion algorithms aim to remove the influence of deleted data points from trained models at a cheaper computational cost than fully retraining those models. First in the convex setting, by leveraging techniques from convex optimization and reservoir sampling, we give the first data deletion algorithms that are able to handle an arbitrarily long sequence of adversarial updates while promising both per-deletion run-time and steady-state error that do not grow with the length of the update sequence. For sequences of deletions, most prior work gives valid guarantees only for sequences that are chosen independently of the models that are published. If people choose to delete their data as a function of the published models (because they don't like what the models reveal about them, for example), then the update sequence is adaptive. This becomes a particularly meaningful distinction in the non-convex setting. We give a general reduction from deletion guarantees against adaptive sequences to deletion guarantees against non-adaptive sequences, using differential privacy and its connection to max information. Combined with ideas from prior work which give guarantees for non-adaptive deletion sequences, this leads to extremely flexible algorithms able to handle arbitrary model classes and training methodologies, giving strong provable deletion guarantees for adaptive deletion sequences. We show in theory how prior work for non-convex models fails against adaptive deletion sequences, and use this intuition to design a practical attack against the SISA algorithm of Bourtoule et al. [2021] on CIFAR-10, MNIST, Fashion-MNIST.

28 Juan-Pablo Ortega

Nanyang Technological University, Singapore
Transport in reservoir computing

Abstract

Reservoir computing systems are essentially dynamical systems influenced by an exogenous input. Such systems are extensively used in biologically

inspired information processing, and are the state-of-the art techniques for several machine learning tasks. If the statistics of the response or output of the system depends discontinuously on the distribution of the inputs, a fundamental challenge arises in applications where inherent changes in the input stochastic source or noise is expected. This problem can be experimentally demonstrated by showing that altering input statistics can drastically affect the statistics of the response. In this talk we explain how we solve this instability problem by providing sufficient conditions under which both the marginals and the joint distributions of the response depend continuously on that of the input. To prove our results, we establish the existence of an invariant measure and show that its dependence on the input process is continuous when the processes are endowed with the Wasserstein distance. The main tool in these developments is the characterization of those invariant measures as fixed points of naturally defined Foias operators that appear in this context and which are examined extensively in the paper. These fixed points are obtained by imposing a newly introduced stochastic state contractivity on the driven system that is readily verifiable in examples. Stochastic state contractivity can be satisfied by systems that are not state-contractive, which is a need typically evoked to guarantee the echo state property in reservoir computing. As a result, it may actually be satisfied even if the echo state property is missing. This is a joint work with G Manjunath and the preprint is posted in the ArXiv.

29 Guillaume Sartoretti

National University of Singapore, Singapore

Distributed Learning Based Scalable Collaboration in Robotic Multi-Agent Systems

Abstract

Preliminary abstract: My work has dealt with the curse of dimensionality in high degree-of-freedom (DOF) robot systems: as the number of agents (robots or DOFs) in the system grows, so does the combinatorial complexity of coordinating them. There are many solutions to managing this complexity growth, and my work has favored distributed, and more recently decentralized approaches, whether it be for a team of mobile robots or a single articulated robot. Specifically, I have embraced advances in distributed

reinforcement learning (dRL) to let multiple agents learn a common decentralized policy in a time-efficient manner. This has produced collaborative policies that naturally scale to an arbitrary numbers of agents, while remaining near-optimal. In this talk, I will present dRL based approaches to the problem of 1) one-shot and lifelong multi-agent path finding (e.g., for warehouse automation), 2) collective robotic construction of 3D structures, and 3) controlling the posture of an articulated robots during locomotion over steep or unstructured terrains. I will present experiments on autonomous ground vehicles and on a hexapod robot that help validate the learned policies, and finally briefly go over some of my lab's ongoing projects.

30 Ohad Shamir

Weizmann Institute of Science, Israel
Implicit bias in machine learning

Abstract

Most practical algorithms for supervised machine learning boil down to optimizing the average performance over a training dataset. However, it is increasingly recognized that although the optimization objective is the same, the manner in which it is optimized plays a decisive role in the properties of the resulting predictor. For example, when training large neural networks, there are generally many weight combinations that will perfectly fit the training data. However, gradient-based training methods somehow tend to reach those which, for example, do not overfit; are brittle to adversarially crafted examples; or have other unusual properties. In this talk, I'll describe several recent theoretical and empirical results related to this question.

31 Harold Soh

National University of Singapore, Singapore
Machine Learning for Human-Robot Interaction

Abstract

My group — the Collaborative, Learning, and Adaptive Robots (CLeAR) lab — seeks to improve people's lives through intelligent robotics. In the

past 5 years, our central focus has been on developing physical and social skills for robots. This talk will give an overview of our work starting with our contributions towards giving robots a physical skill: the sense of touch. We'll detail our work on event-driven touch sensing and perception. Next, we'll delve into our work on giving robots a social skill: a sense of trust; we'll cover machine learning models that capture how humans trust robots across multiple tasks and also models for task-oriented multi-modal communication.

32 Jascha Sohl-Dickstein

Google, USA
Understanding infinite width neural networks

Abstract

As neural networks become wider their accuracy improves, and their behavior becomes easier to analyze theoretically. I will give an introduction to a rapidly growing body of work which examines the learning dynamics and distribution over functions induced by infinitely wide, randomly initialized, neural networks. Core results that I will discuss include: that the distribution over functions computed by a wide neural network often corresponds to a Gaussian process with a particular compositional kernel, both before and after training; that the predictions of wide neural networks are linear in their parameters throughout training; that the posterior distribution over parameters also takes on a simple form in wide Bayesian networks. These results provide for surprising capabilities – for instance, the evaluation of test set predictions which would come from an infinitely wide trained neural network without ever instantiating a neural network, or the rapid training of 10,000+ layer convolutional networks. I will argue that this growing understanding of neural networks in the limit of infinite width is foundational for future theoretical and practical understanding of deep learning.

Neural Tangents: `https://github.com/google/neural-tangents`

33 Linda Tan

National University of Singapore, Singapore
Analytic natural gradient updates for Cholesky factor in Gaussian variational approximation

Abstract

Stochastic gradient methods have enabled variational inference for high-dimensional models and large datasets. However, the steepest ascent direction in the parameter space of a statistical model is actually given by the natural gradient which premultiplies the widely used Euclidean gradient by the inverse of the Fisher information matrix. Use of natural gradients can improve convergence, but inverting the Fisher information matrix is daunting in high-dimensions. In Gaussian variational approximation, natural gradient updates of the mean and precision matrix of the Gaussian distribution can be derived analytically, but do not ensure the precision matrix remains positive definite. To tackle this issue, we consider Cholesky decomposition of the covariance or precision matrix, and derive analytic natural gradient updates of the Cholesky factor, which depend only on the first derivative of the log posterior density. Efficient natural gradient updates of the Cholesky factor are also derived under sparsity constraints representing different posterior correlation structures. As Adam's adaptive learning rate does not seem to pair well with natural gradients, we propose using stochastic normalized natural gradient ascent with momentum. The efficiency of proposed methods are demonstrated using generalized linear mixed models.

34 Vincent Tan

National University of Singapore, Singapore
Minimax Optimal Fixed-Budget Best Arm Identification in Linear Bandits

Abstract

We study the problem of best arm identification in linear bandits in the fixed-budget setting. By leveraging properties of the G-optimal design and incorporating it into the arm allocation rule, we design a parameter-free algorithm, Optimal Design-based Linear Best Arm Identification (OD-LinBAI). We provide a theoretical analysis of the failure probability of OD-LinBAI. Instead of all the optimality gaps, the performance of OD-LinBAI depends only on the gaps of the top d arms, where d is the effective dimension of the linear bandit instance. Complementarily, we present a minimax lower bound for this problem. The upper and lower bounds show that OD-LinBAI is minimax optimal up to constant multiplicative factors in the exponent, which is

a significant theoretical improvement over existing methods (e.g., BayesGap, Peace, LinearExploration and GSE), and settles the question of ascertaining the difficulty of learning the best arm in the fixed-budget setting. Finally, numerical experiments demonstrate considerable empirical improvements over existing algorithms on a variety of real and synthetic datasets.

This is joint work with Junwen Yang (IORA, NUS)

35 Yan Shuo Tan

National University of Singapore, Singapore
Understanding and overcoming the statistical limitations of decision trees

Abstract

Decision trees are important both as interpretable models, amenable to high-stakes decision-making, and as building blocks of ensemble methods such as random forests and gradient boosting. Their statistical properties, however, are not yet well understood. In particular, it is unclear why there is a prediction performance gap between them and powerful but uninterpretable machine learning methods. In this talk, we discuss how to bridge this gap partially via Hierarchical Shrinkage (HS), a post-hoc algorithm which regularizes the tree not by altering its structure, but by shrinking the prediction over each leaf toward the sample means over each of its ancestors. Furthermore, we discuss generalization lower bounds that reveal some of the inductive biases of tree-based methods, and how HS helps to overcome some of it.

36 Andrey Ustyuzhanin

National University of Singapore, Singapore
Black-Box Optimization with Local Generative Surrogates

Abstract

In this talk I'll describe a novel method for gradient-based optimization of black-box simulators using differentiable local surrogate models. In fields such as physics and engineering, many processes are modeled with non-differentiable simulators with intractable likelihoods. Optimization of these

forward models is particularly challenging, especially when the simulator is stochastic. To address such cases, we introduce the use of deep generative models to iteratively approximate the simulator in local neighborhoods of the parameter space. We demonstrate that these local surrogates can be used to approximate the gradient of the simulator, and thus enable gradient-based optimization of simulator parameters. In cases where the dependence of the simulator on the parameter space is constrained to a low dimensional submanifold, we observe that our method attains minima faster than baseline methods, including Bayesian optimization, numerical optimization, and approaches using score function gradient estimators.

37 Rene Vidal

Johns Hopkins University, USA
Explainable AI via Semantic Information Pursuit

Abstract

There is a significant interest in developing ML algorithms whose final predictions can be explained in domain-specific terms that are understandable to a human. Providing such an "explanation" can be crucial for the adoption of ML algorithms in risk-sensitive domains such as healthcare. This has motivated a number of approaches that seek to provide explanations for existing ML algorithms in a post-hoc manner. However, many of these approaches have been widely criticized for a variety of reasons and no clear methodology exists for developing ML algorithms whose predictions are readily understandable by humans. To address this challenge, we develop a method for constructing high performance ML algorithms that are "explainable by design". Namely, our method makes its prediction by asking a sequence of domain- and task-specific yes/no queries about the data (akin to the game "20 questions"), each having a clear interpretation to the end-user. We then minimize the expected number of queries needed for accurate prediction on any given input. This allows for human interpretable understanding of the prediction process by construction, as the questions which form the basis for the prediction are specified by the user as interpretable concepts about the data. Experiments on vision and NLP tasks demonstrate the efficacy of our approach and its superiority over post-hoc explanations. Joint work

with Aditya Chattopadhyay, Stewart Slocum, Benjamin Haeffele and Donald Geman.

38 Yu-Ping Wang

Tulane University, USA
Interpretable multimodal deep learning with application to biomedical data fusion

Abstract

Deep network-based data fusion models have been developed to integrate complementary information from multi-modal datasets while capture their complex relationships. This is particularly useful in biomedical domain, where multi-modal data such as imaging and multi-omics are ubiquitous and the integration of these heterogenous data can lead to novel biological findings. However, deep learning models are often difficult to interpret, bringing about challenges for uncovering biological mechanisms using these models. In this work, we develop an interpretable multimodal deep learning-based fusion model to perform automated disease diagnosis and result interpretation simultaneously. We name it Grad-CAM guided convolutional collaborative learning (gCAM-CCL), which is achieved by combining intermediate feature maps with gradient-based weights in a multi-modal convolution network. The gCAM-CCL model can generate interpretable activation maps to quantify pixel-level contributions of the input features. Moreover, the estimated activation maps are class-specific, which can therefore facilitate the identification of biomarkers underlying different groups. Finally, we apply and validate the gCAM-CCL model in a study of brain development with integrative analysis of brain imaging and genomics data. We demonstrate its successful application to both the classification of cognitive function group and the discovery of underlying biological mechanisms.

39 Greg Yang

Microsoft Research, USA
Feature Learning Infinite-Width Neural Networks Outperform Finite Ones

Abstract

Two popular memes (in the ML community at large) exist about infinite-width neural networks (NN) of general architecture: 1) they are kernel machines, and 2) they underperform finite NNs. Meme 1) is contradicted by the existence of the feature learning limit, or μ-limit, of wide NNs. (Yang and Hu, 2020) has dispelled meme 2) on 1-hidden-layer linear NNs by showing the opposite for their μ-limit. However, further evidence in the more relevant deep nonlinear setting has not been forthcoming because of -limit's computational difficulty. Here, we show that the μ-limit of a form of projected gradient descent is efficiently computable. On CIFAR10 and Omniglot and for deep relu MLP, this limit outperforms finite NNs (trained normally without projection), thereby finally dispelling meme (2) for deep nonlinear models.

40 Haizhao Yang

University of Maryland, USA
Finite expression method for solving high-dimensional PDEs

Abstract

Designing efficient and accurate numerical solvers for high-dimensional partial differential equations (PDE) remains a challenging and important topic in computational science and engineering, mainly due to the "curse of dimensionality" in designing numerical schemes that scales in dimension. This talk introduces a new methodology that seeks an approximate PDE solution in the space of functions with finitely many analytic expressions and, hence, this methodology is named as the finite expression method (FEX). It is proved in approximation theory that FEX can avoid the curse of dimensionality. As a proof of concept, a deep reinforcement learning method is proposed to implement FEX for various high-dimensional PDEs in different dimensions, achieving high and even machine accuracy with a memory complexity polynomial in dimension and an amenable time complexity. An approximate solution with finite analytic expressions also provides interpretible insights of the ground truth PDE solution, which can further help to advance the understanding of physical systems and design postprocessing techniques for a refined solution.

41 Angela Yao

National University of Singapore, Singapore
Is Classification All You Need for Computer Vision?

Abstract

Classification and regression are two fundamental tasks of machine learning. The choice between the two usually depends on the categorical or continuous nature of the target output. Curiously, in computer vision, specifically with deep learning, regression-type problems such as depth estimation, age estimation, crowd-counting and pose estimation, often yield better performance when formulated as a classification task.

The phenomenon of classification outperforming regression on inherently continuous estimation tasks naturally begs the question – why? In this talk, I will highlight some possible causes based on some task-specific investigations for pose estimation and crowd-counting related to label accuracy and strength of supervision. I will then introduce a more general comparison between classification and regression from a learning point of view. Our findings suggest that the key difference lies in the learned feature spaces from the different losses used in classification versus regression.

42 Joey Zhou Tianyi

*A*STAR, Singapore*
Trusted Multi-view Classification

Abstract

Existing multi-view classification algorithms focus on promoting accuracy by exploiting different views, typically integrating them into common representations for follow-up tasks. Although effective, it is also crucial to ensure the reliability of both the multi-view integration and the final decision, especially for noisy, corrupted and out-of-distribution data. Dynamically assessing the trustworthiness of each view for different samples could provide reliable integration. This can be achieved through uncertainty estimation. With this in mind, we propose a novel multi-view classification algorithm, termed trusted multi-view classification (TMC), providing a new paradigm for multi-view learning by dynamically integrating different views at an evidence level. The

proposed TMC can promote classification reliability by considering evidence from each view. Specifically, we introduce the variational Dirichlet to characterize the distribution of the class probabilities, parameterized with evidence from different views and integrated with the Dempster-Shafer theory. The unified learning framework induces accurate uncertainty and accordingly endows the model with both reliability and robustness against possible noise or corruption. Both theoretical and experimental results validate the effectiveness of the proposed model in accuracy, robustness and trustworthiness.

43 Pan Zhou

SEA, Singapore

Adan: Adaptive Nesterov Momentum Algorithm for Faster Optimizing Deep Models

Abstract

Adaptive gradient algorithms borrow the moving average idea of heavy ball acceleration to estimate accurate first- and second-order moments of the gradient for accelerating convergence. However, Nesterov acceleration which converges faster than heavy ball acceleration in theory and in many empirical cases is much less investigated under the adaptive gradient setting. In this work, we propose the ADAptive Nesterov momentum algorithm, Adan for short, to speed up the training of deep neural networks effectively. Adan first reformulates the vanilla Nesterov acceleration to develop a new Nesterov momentum estimation (NME) method, which avoids the extra computation and memory overhead of computing gradient at the extrapolation point. Then Adan adopts NME to estimate the first- and second-order moments of the gradient in adaptive gradient algorithms for convergence acceleration. Besides, we prove that on the nonconvex stochastic problems (e.g. deep learning problems), for Adan, its stochastic gradient complexity to find an approximate first-order stationary point can match the best-known lower bound. Extensive experimental results show that Adan surpasses the corresponding SoTA optimizers on both CNNs and transformers, and sets new SoTAs for many popular networks and frameworks, e.g. ResNet, ConvNext, ViT, Swin, MAE, LSTM, TransformerXL and BERT. More surprisingly, Adan can use half of the training cost (epochs) of SoTA optimizers to achieve higher or comparable performance on ViT and ResNet, etc, and also shows great tol-

erance to a large range of minibatch size, e.g. from 1k to 32k. We hope Adan can contribute to the development of deep learning by reducing training cost and relieving engineering burden of trying different optimizers on various architectures.

44 Enrique Zuazua

Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
Control and machine learning

Abstract

In this lecture we shall present some recent results on the interplay between control and Machine Learning, and more precisely, Supervised Learning and Universal Approximation.

We adopt the perspective of the simultaneous or ensemble control of systems of Residual Neural Networks (ResNets). Roughly, each item to be classified corresponds to a different initial datum for the Cauchy problem of the ResNets, leading to an ensemble of solutions to be driven to the corresponding targets, associated to the labels, by means of the same control.

We present a genuinely nonlinear and constructive method, allowing to show that such an ambitious goal can be achieved, estimating the complexity of the control strategies.

This property is rarely fulfilled by the classical dynamical systems in Mechanics and the very nonlinear nature of the activation function governing the ResNet dynamics plays a determinant role. It allows deforming half of the phase space while the other half remains invariant, a property that classical models in mechanics do not fulfill.

The turnpike property is also analyzed in this context, showing that a suitable choice of the cost functional used to train the ResNet leads to more stable and robust dynamics.

This lecture is inspired in joint work, among others, with Borjan Geshkovski (MIT), Carlos Esteve (Cambridge), Domènec Ruiz-Balet (IC, London) and Dario Pighin (Sherpa.ai).