Poster No.1

Bayesian active learning of molecular properties with small initial data

Minyoung Ha

Materials Research Center,

Samsung Advanced Institute of Technology (SAIT)

Designing novel materials with limited data remains a critical challenge in materials science. Since large-scale data is often unavailable in early-stage research, materials discovery typically relies on small, proof-of-concept datasets and a trial-and-error process guided by intuition. To address this, we propose a two-stage Bayesian active learning workflow for molecular property prediction (MPP) and design of experiments (DOE), working with an initial dataset as small as 50 points.

Our approach utilizes a message-passing graph neural network (GNN) to map molecular structures into fixed-length graph embeddings, followed by a Gaussian process regressor (GPR) to predict molecular properties. Extensive virtual experiments were conducted for various tasks across multiple molecular domains to compare the MPP and DOE performances of graph embedding methods and GPR kernel choices. We demonstrate that system-agnostic embeddings such as random graph embeddings provide a solid baseline for a wide range of problems, while graph embeddings from a pre-trained GNN significantly enhance performance if a large database of associated molecular properties is available. Interestingly, although MPP and DOE performances are generally correlated, strong cross-validation performance in MPP does not always translate to effective DOE. These findings suggest new strategies for data-driven and data-efficient materials discovery, especially in the early, data-scarce stages of research, where traditional approaches are less effective.

Poster No.2

Interpreting Bayesian Methods Through the Geometry of Parameter Spaces

Cheongjae Jang

Hanyang University

In naive maximum a posteriori (MAP) estimation, the mode of the posterior distribution may shift depending on the choice of coordinates in the parameter space. This result is undesirable as it suggests that estimation outcomes can be influenced by reparametrization, which is irrelevant for inference. This issue can be addressed by appropriately considering the geometry of the parameter space. In this presentation, we will interpret various Bayesian methods through the lens of parameter space geometry and discuss prospective directions for future research.

Poster No.3

DiffAlign: Flexible Molecular Alignment Using Diffusion

Iljung Kim

Department of Computer Science, Hanyang University, Seoul, Republic of Korea

Flexible molecular alignment is fundamental in computational chemistry and structural biology, serving as cornerstone for understanding molecular interactions and function. Recent methods approach the flexible molecular alignment problem as a score optimization task, leaving room for improvements through artificial intelligence. To address this, we introduce DiffAlign, a diffusion model that frames flexible molecular alignment as a generative modeling problem. DiffAlign is trained using an SE(3) equivariant conditional diffusion approach. By training the model within a conditional diffusion framework, we can leverage classifier-free guidance, enabling DiffAlign to align one molecule to another with high precision. DiffAlign achieves a root-mean-square deviation (RMSD) of 0.33 for self-alignment, significantly improving upon the 0.59 RMSD obtained by traditional state-of-the-art methods for DUD-E INHA dataset. Additionally, DiffAlign demonstrates a major efficiency advantage, solving the alignment problem in just 2.6 seconds on average, compared to 41 seconds for previous state-of-the-art methods.

Poster No.4

Quantifying Parameter Importance in Neural Networks: Beyond Shapley Value approach

Do-Hoon Kim

Department of Computer Science
Hanyang University, Korea

Traditional interpretation and explanation methods of neural network (NN) models have predominantly focused on feature importance, addressing the question "Which features are most important to the model predictions?" However, our primary interest in model interpretation lies in '**parameter importance**', seeking to answer, "Which parameters are most crucial to the model itself?" Regardless of how important feature information is, it cannot influence model predictions without parameters to convey that information. Assigning importance values to parameters, while considering their inherent characteristics, is crucial for accurate model optimization and effective model compression and fundamental for developing reliable and efficient machine learning systems. The Shapley value, originating from cooperative game theory in economics, has been adopted for parameter importance value due to its solid foundation in fairly distributing contributions. However, we argue that the Shapley value, which averages parameter contributions across all possible combinations of parameters, where only a portion of the parameters is utilized in the model, overlooks critical interactions and additional information within these combinations. This limitation is particularly evident in scenarios where parameter interactions lead to overlapping information or enable feature information flow within specific parameter combinations. We introduce a novel perspective that emphasizes the roles of individual parameters within the context of all possible parameter combinations in NNs. Our main contribution is the development of criteria that enable informed decisions about retaining or removing specific parameters based on their roles in NNs. Additionally, we offer a more comprehensive understanding of parameter significance, extending beyond the traditional approaches provided by the Shapley value.

Poster No.5

A Unified Confidence Sequence for Generalized Linear Models, with Applications to Bandits

Junghyun Lee[1], Se-Young Yun[1], Kwang-Sung Jun[2]

[1]*Kim Jaechul Graduate School of AI, KAIST*

[2]*Department of Computer Science, University of Arizona*

We present a unified likelihood ratio-based confidence sequence (CS) for *any* (self-concordant) generalized linear models (GLMs) that is guaranteed to be convex and numerically tight. We show that this is on par or improves upon known CSs for various GLMs, including Gaussian, Bernoulli, and Poisson. In particular, for the first time, our CS for Bernoulli has a $\mathrm{poly}(S)$-free radius where S is the norm of the unknown parameter. Our first technical novelty is its derivation, which utilizes a time-uniform PAC-Bayesian bound with a uniform prior/posterior, despite the latter being a rather unpopular choice for deriving CSs. As a direct application of our new CS, we propose a simple and natural optimistic algorithm called **OFUGLB** applicable to *any* generalized linear bandits (**GLB**; Filippi et al., 2010). Our analysis shows that the celebrated optimistic approach simultaneously attains state-of-the-art regrets for various self-concordant (not necessarily bounded) **GLB**s, and even $\mathrm{poly}(S)$-free for bounded **GLB**s, including logistic bandits. The regret analysis, our second technical novelty, follows from combining our new CS with a new proof technique that completely avoids the previously widely used self-concordant control lemma (Lemma 9, Faury et al., 2020). We numerically verify that **OFUGLB** significantly outperforms, or is at least at par with, prior algorithms for logistic and Poisson bandits.

Poster No.6

Eigenvalue-Based Preprocessing for Tissue Extraction from Pathology Image Slides

DoHyun Lim

Department of Computer Science

Hanyang University

The objective of this study is to effectively extract patches containing only tissue from pathology image slides. Pathology image slides include not only tissue but also unwanted elements such as background and markings, necessitating precise preprocessing for accurate tissue extraction. To address this, we devised a preprocessing method utilizing the eigenvalues and eigenvectors derived from the covariance matrix calculated from the image's RGB values.

We hypothesized that the brightness difference between tissue and background induces the greatest variability. Consequently, the direction with the largest variance corresponds to the first principal eigenvector, and its magnitude corresponds to the first eigenvalue. Visualization of the RGB pixel data in a 3D scatter plot confirmed a trend where brightness increases in the direction of the first eigenvector.

Notably, tissue patches exhibited significantly larger first eigenvalues compared to background patches, demonstrating effective separation between tissue and background. Additionally, we observed clustering of regions representing key features within the tissue in the 3D scatter plot, indicating successful extraction of data that reflects important tissue characteristics. This preprocessing method enabled accurate extraction of tissue patches.

This study validates the efficacy of the eigenvalue-based method in the preprocessing stage of pathology image analysis and is expected to lay the foundation for the development of a Foundation Model optimized for pathology in the future.

Poster No.7

Provable Benefit of Cutout and CutMix for Feature Learning

Junsoo Oh

KAIST AI

Patch-level data augmentation techniques such as Cutout and CutMix have demonstrated significant efficacy in enhancing the performance of image-based tasks. However, a comprehensive theoretical understanding of these methods remains elusive. In this paper, we study two-layer neural networks trained using three distinct methods: vanilla training without augmentation, Cutout training, and CutMix training. Our analysis focuses on a feature-noise data model, which consists of several label-dependent features of varying rarity and label-independent noises of differing strengths. Our theorems demonstrate that Cutout training can learn features with low frequencies that vanilla training cannot, while CutMix training can even learn rarer features that Cutout cannot capture. From this, we establish that CutMix yields the highest test accuracy among the three. Our novel analysis reveals that CutMix training makes the network learn all features and noise vectors ``evenly'' regardless of the rarity and strength, which provides an interesting insight into understanding patch-level augmentation.

Poster No.8

Machine Learning Force Fields for Ionic Liquids

Anseong Park, Won Bo Lee*

School of Chemical and Biological Engineering, Seoul National University, Republic of Korea,

wblee@snu.ac.kr

The development of machine learning force fields (MLFFs) offers a promising alternative, aiming to reproduce potential energy surfaces (PES) based on DFT data. However, the quality of MLFFs largely depends on the expertise of researchers in preparing training datasets and tuning hyperparameters. Unlike traditional FFs, which are systematically built and transferable, MLFFs may face uncertainties in covering rare events during simulations, especially in systems with diverse atomic types. In this study, DeePMD was used to construct MLFFs, and the results were compared to classic and polarizable FF MD simulations. The findings reveal that incorporating non-equilibrated (nEQ) datasets enhances MLFF performance, yet discrepancies with polarizable FF results raise questions about the correctness and completeness of MLFF simulations in capturing complex behaviors.

Poster No.9

DASH: Warm-Starting Neural Network Training Without Loss of Plasticity Under Stationarity

Baekrok Shin

KAIST AI

Warm-starting neural networks by initializing them with previously learned weights is appealing, as practical neural networks are often deployed under a continuous influx of new data. However, this often leads to a phenomenon known as *loss of plasticity*, where the network loses its ability to learn new information and thereby shows worse generalization performance than those trained from scratch. While this issue has been actively studied in non-stationary data distributions (e.g., in reinforcement learning), it surprisingly occurs even when the data distribution remains stationary, and its underlying mechanism is poorly understood. To address this gap, we develop a learning framework that emulates real-world neural network training scenarios. Under this framework, we identify noise memorization as the primary cause of the loss of plasticity when warm-starting the neural networks in stationary data distributions. Motivated by this discovery, we propose an effective method called **Direction-Aware SHrinking (DASH)** to mitigate the loss of plasticity. DASH aims to selectively forget previously memorized noise while aiming to preserve learned features, based on the loss gradient computed from newly introduced data. We validate our approach in vision tasks using diverse datasets, models, and optimizers, demonstrating consistent improvements in test accuracy and training efficiency.

Poster No.10

Machine learning integrated Fourier transform for enhanced signal analysis

Jihye Kim[1], Woo Seok Lee[2], Jaehyun Jung[3] and Taegeun Song[1]

[1]*Department of Data Information and Physics, Kongju National University, Korea*

[2] *1st Bioterapuritics, Korea*

[1] *Department of Applied Mathatics, Kongju National University, Korea*

We propose a novel approach for performing Fourier transforms using a single hidden-layer perceptron with a sine activation function. The proposed approach demonstrates remarkable capabilities, including the ability to identify frequency components even with incomplete cycle data and an effectively infinite observation time akin to the delta peak. Due to its mathematical equivalence to the Fourier transform, the proposed method benefits from being effective even when overfitting occurs, by taking advantage of its inherent properties to improve results. As a practical application, we have applied this method to a quantum interferometer and suggest its potential use for missing data prediction and super-resolution tasks.

Poster No.11

Towards Calibrated Robust Fine-Tuning of Vision-Language Models

Changdae Oh, Hyesu Lim, Mijoo Kim, Dongyoon Han, Sangdoo Yun, Jaegul Choo, Alexander G Hauptmann, Zhi-Qi Cheng, **Kyungwoo Song**

University of Wisconsin–Madison, KAIST, Chung-Ang University, NAVER AI Lab, Carnegie Mellon University, ***Yonsei University***

Improving out-of-distribution (OOD) generalization through in-distribution (ID) adaptation is a primary goal of robust fine-tuning methods beyond the naive fine-tuning approach. However, despite decent OOD generalization performance from recent robust fine-tuning methods, OOD confidence calibration for reliable machine learning has not been fully addressed. This work proposes a robust fine-tuning method that improves both OOD accuracy and calibration error in Vision Language Models (VLMs). Firstly, we show that both types of errors have a shared upper bound consisting of two terms of ID data: 1) calibration error and 2) the smallest singular value of the input covariance matrix. Based on this insight, we design a novel framework that conducts fine-tuning with a constrained multimodal contrastive loss enforcing a larger smallest singular value, which is further aided by the self-distillation of a moving averaged model to achieve well-calibrated prediction. Starting from an empirical validation of our theoretical statements, we provide extensive experimental results on ImageNet distribution shift benchmarks that demonstrate the effectiveness of our method.