801. Prevalence of Negative Transfer in Continual Reinforcement Learning: Analyses and a Simple Baseline

链接：https://iclr.cc/virtual/2025/poster/30073 abstract：

802. SPDIM: Source-Free Unsupervised Conditional and Label Shift Adaptation in EEG

链接：https://iclr.cc/virtual/2025/poster/30495 abstract：

803. Maximizing the Potential of Synthetic Data: Insights from Random Matrix Theory

链接：https://iclr.cc/virtual/2025/poster/30188 abstract： Synthetic data has gained attention for training large language models, but poor-quality data can harm performance (see, e.g., Shumailov et al. (2023); Seddik et al. (2024)). A potential solution is data pruning, which retains only high-quality data based on a score function (human or machine feedback). Previous work Feng et al. (2024) analyzed models trained on synthetic data as sample size increases. We extend this by using random matrix theory to derive the performance of a binary classifier trained on a mix of real and pruned synthetic data in a high dimensional setting. Our findings identify conditions where synthetic data could improve performance, focusing on the quality of the generative model and verification strategy. We also show a smooth phase transition in synthetic label noise, contrasting with prior sharp behavior in infinite sample limits. Experiments with toy models and large language models validate our theoretical results.

804. Clique Number Estimation via Differentiable Functions of Adjacency Matrix Permutations

链接：https://iclr.cc/virtual/2025/poster/30469 abstract： Estimating the clique number in a graph is central to various applications, e.g., community detection, graph retrieval, etc. Existing estimators often rely on non-differentiable combinatorial components. Here, we propose a full differentiable estimator for clique number estimation, which can be trained from distant supervision of clique numbers, rather than demonstrating actual cliques.Our key insight is a formulation of the maximum clique problem (MCP) as a maximization of the size of fully dense square submatrix, within a suitably row-column-permuted adjacency matrix.We design a differentiable mechanism to search for permutations that lead to the discovery of such dense blocks.However, the optimal permutation is not unique, which leads to the learning of spurious permutations. To tackle this problem, we view the MCP problem as a sequence of subgraph matching tasks, each detecting progressively larger cliques in a nested manner. This allows effective navigation through suitable node permutations.These steps result in MxNet, an end-to-end differentiable model, which learns to predict clique number without explicit clique demonstrations, with the added benefit of interpretability. Experiments on eight datasets show the superior accuracy of our approach.

805. UniDetox: Universal Detoxification of Large Language Models via Dataset Distillation

链接：https://iclr.cc/virtual/2025/poster/28932 abstract： We present UniDetox, a universally applicable method designed to mitigate toxicity across various large language models (LLMs).Previous detoxification methods are typically model-specific, addressing only individual models or model families, and require careful hyperparameter tuning due to the trade-off between detoxification efficacy and language modeling performance. In contrast, UniDetox provides a detoxification technique that can be universally applied to a wide range of LLMs without the need for separate model-specific tuning. Specifically, we propose a novel and efficient dataset distillation technique for detoxification using contrastive decoding. This approach distills detoxifying representations in the form of synthetic text data, enabling universal detoxification of any LLM through fine-tuning with the distilled text. Our experiments demonstrate that the detoxifying text distilled from GPT-2 can effectively detoxify larger models, including OPT, Falcon, and LLaMA-2. Furthermore, UniDetox eliminates the need for separate hyperparameter tuning for each model, as a single hyperparameter configuration can be seamlessly applied across different models. Additionally, analysis of the detoxifying text reveals a reduction in politically biased content, providing insights into the attributes necessary for effective detoxification of LLMs.

806. Charting the Design Space of Neural Graph Representations for Subgraph Matching

链接：https://iclr.cc/virtual/2025/poster/30929 abstract： Subgraph matching is vital in knowledge graph (KG) question answering, molecule design, scene graph, code and circuit search, etc.Neural methods have shown promising results for subgraph matching.Our study of recent systems suggests refactoring them into a unified design space for graph matching networks.Existing methods occupy only a few isolated patches in this space, which remains largely uncharted.We undertake the first comprehensive exploration of this space, featuring such axes as attention-based vs. soft permutation-based interaction

between query and corpus graphs, aligning nodes vs. edges, and the form of the final scoring network that integrates neural representations of the graphs.Our extensive experiments reveal that judicious and hitherto-unexplored combinations of choices in this space lead to large performance benefits.Beyond better performance, our study uncovers valuable insights and establishes general design principles for neural graph representation and interaction, which may be of wider interest.

807. SyllableLM: Learning Coarse Semantic Units for Speech Language Models

链接：https://iclr.cc/virtual/2025/poster/28995 abstract： Language models require tokenized inputs. However, tokenization strategies for continuous data like audio and vision are often based on simple heuristics such as fixed sized convolutions or discrete clustering, which do not necessarily align with the semantic structure of the data. For speech in particular, the high resolution of waveforms (16,000 samples/second or more) presents a significant challenge as speech-based language models have had to use several times more tokens per word than text-based language models. In this work, we introduce a controllable self-supervised technique to merge speech representations into coarser syllable-like units while still preserving semantic information. We do this by 1) extracting noisy boundaries through analyzing correlations in pretrained encoder losses and 2) iteratively improving model representations with a novel distillation technique. Our method produces controllable-rate semantic units at as low as 5Hz and 60bps and achieves SotA in syllabic segmentation and clustering. Using these coarse tokens, we successfully train SyllableLM, a Speech Language Model (SpeechLM) that matches or outperforms current SotA SpeechLMs on a range of spoken language modeling tasks. SyllableLM also achieves significant improvements in efficiency with a 30x reduction in training compute and a 4x wall-clock inference speedup. Our code and checkpoints are available at https://www.github.com/alanbaade/SyllableLM

808. Isometric Regularization for Manifolds of Functional Data

链接：https://iclr.cc/virtual/2025/poster/27800 abstract： While conventional data are represented as discrete vectors, Implicit Neural Representations (INRs) utilize neural networks to represent data points as continuous functions. By incorporating a shared network that maps latent vectors to individual functions, one can model the distribution of functional data, which has proven effective in many applications, such as learning 3D shapes, surface reflectance, and operators.However, the infinite-dimensional nature of these representations makes them prone to overfitting, necessitating sufficient regularization. Naïve regularization methods -- those commonly used with discrete vector representations -- may enforce smoothness to increase robustness but result in a loss of data fidelity due to improper handling of function coordinates. To overcome these challenges, we start by interpreting the mapping from latent variables to INRs as a parametrization of a Riemannian manifold. We then recognize that preserving geometric quantities -- such as distances and angles -- between the latent space and the data manifold is crucial. As a result, we obtain a manifold with minimal intrinsic curvature, leading to robust representations while maintaining high-quality data fitting. Our experiments on various data modalities demonstrate that our method effectively discovers a well-structured latent space, leading to robust data representations even for challenging datasets, such as those that are small or noisy.

809. A Deep Generative Learning Approach for Two-stage Adaptive Robust Optimization

链接：https://iclr.cc/virtual/2025/poster/30520 abstract： Two-stage adaptive robust optimization (ARO) is a powerful approach for planning under uncertainty, balancing first-stage decisions with recourse decisions made after uncertainty is realized. To account for uncertainty, modelers typically define a simple uncertainty set over which potential outcomes are considered. However, classical methods for defining these sets unintentionally capture a wide range of unrealistic outcomes, resulting in overly-conservative and costly planning in anticipation of unlikely contingencies. In this work, we introduce AGRO, a solution algorithm that performs adversarial generation for two-stage adaptive robust optimization using a variational autoencoder. AGRO generates high-dimensional contingencies that are simultaneously adversarial and realistic, improving the robustness of first-stage decisions at a lower planning cost than standard methods. To ensure generated contingencies lie in high-density regions of the uncertainty distribution, AGRO defines a tight uncertainty set as the image of "latent" uncertainty sets under the VAE decoding transformation. Projected gradient ascent is then used to maximize recourse costs over the latent uncertainty sets by leveraging differentiable optimization methods. We demonstrate the cost-efficiency of AGRO by applying it to both a synthetic production-distribution problem and a real-world power system expansion setting. We show that AGRO outperforms the standard column-and-constraint algorithm by up to 1.8% in production-distribution planning and up to 8% in power system expansion.

810. Budgeted Online Continual Learning by Adaptive Layer Freezing and Frequency-based Sampling

链接：https://iclr.cc/virtual/2025/poster/28988 abstract： The majority of online continual learning (CL) advocates single-epoch training and imposes restrictions on the size of replay memory. However, single-epoch training would incur a different amount of computations per CL algorithm, and the additional storage cost to store logit or model in addition to replay memory is largely ignored in calculating the storage budget. Arguing different computational and storage budgets hinder fair comparison among CL algorithms in practice, we propose to use floating point operations (FLOPs) and total memory size in Byte as a metric for

computational and memory budgets, respectively, to compare and develop CL algorithms in the same ‘total resource budget.’ To improve a CL method in a limited total budget, we propose adaptive layer freezing that does not update the layers for less informative batches to reduce computational costs with a negligible loss of accuracy. In addition, we propose a memory retrieval method that allows the model to learn the same amount of knowledge as using random retrieval in fewer iterations. Empirical validations on the CIFAR-10/100, CLEAR-10/100, and ImageNet-1K datasets demonstrate that the proposed approach outperforms the state-of-the-art methods within the same total budget. Furthermore, we validate its effectiveness in the Multi-modal Concept incremental Learning setup with multimodal large language models, such as LLaVA-1.5-7B. Code is available at https://github.com/snumprlab/budgeted-cl.

811. Decomposition Polyhedra of Piecewise Linear Functions

链接：https://iclr.cc/virtual/2025/poster/27910 abstract： In this paper we contribute to the frequently studied question of how to decompose a continuous piecewise linear (CPWL) function into a difference of two convex CPWL functions. Every CPWL function has infinitely many such decompositions, but for applications in optimization and neural network theory, it is crucial to find decompositions with as few linear pieces as possible. This is a highly challenging problem, as we further demonstrate by disproving a recently proposed approach by Tran and Wang [Minimal representations of tropical rational functions. Algebraic Statistics, 15(1):27–59, 2024]. To make the problem more tractable, we propose to fix an underlying polyhedral complex determining the possible locus of nonlinearity. Under this assumption, we prove that the set of decompositions forms a polyhedron that arises as intersection of two translated cones. We prove that irreducible decompositions correspond to the bounded faces of this polyhedron and minimal solutions must be vertices. We then identify cases with a unique minimal decomposition, and illustrate how our insights have consequences in the theory of submodular functions. Finally, we improve upon previous constructions of neural networks for a given convex CPWL function and apply our framework to obtain results in the nonconvex case.

812. REGENT: A Retrieval-Augmented Generalist Agent That Can Act In-Context in New Environments

链接：https://iclr.cc/virtual/2025/poster/29847 abstract： Building generalist agents that can rapidly adapt to new environments is a key challenge for deploying AI in the digital and real worlds. Is scaling current agent architectures the most effective way to build generalist agents? We propose a novel approach to pre-train relatively small policies on relatively small datasets and adapt them to unseen environments via in-context learning, without any finetuning. Our key idea is that retrieval offers a powerful bias for fast adaptation. Indeed, we demonstrate that even a simple retrieval-based 1-nearest neighbor agent offers a surprisingly strong baseline for today's state-of-the-art generalist agents. From this starting point, we construct a semi-parametric agent, REGENT, that trains a transformer-based policy on sequences of queries and retrieved neighbors. REGENT can generalize to unseen robotics and game-playing environments via retrieval augmentation and in-context learning, achieving this with up to 3x fewer parameters and up to an order-of-magnitude fewer pre-training datapoints, significantly outperforming today's state-of-the-art generalist agents.

813. Understanding Virtual Nodes: Oversquashing and Node Heterogeneity

链接：https://iclr.cc/virtual/2025/poster/29859 abstract：

814. Language Guided Skill Discovery

链接：https://iclr.cc/virtual/2025/poster/28721 abstract：

815. Privacy-Aware Lifelong Learning

链接：https://iclr.cc/virtual/2025/poster/29446 abstract： Lifelong learning algorithms enable models to incrementally acquire new knowledge without forgetting previously learned information. Contrarily, the field of machine unlearning focuses on explicitly forgetting certain previous knowledge from pretrained models when requested, in order to comply with data privacy regulations on the right-to-be-forgotten. Enabling efficient lifelong learning with the capability to selectively unlearn sensitive information from models presents a critical and largely unaddressed challenge with contradicting objectives. We address this problem from the perspective of simultaneously preventing catastrophic forgetting and allowing forward knowledge transfer during task-incremental learning, while ensuring exact task unlearning and minimizing memory requirements, based on a single neural network model to be adapted. Our proposed solution, privacy-aware lifelong learning (PALL), involves optimization of task-specific sparse subnetworks with parameter sharing within a single architecture. We additionally utilize an episodic memory rehearsal mechanism to facilitate exact unlearning without performance degradations. We empirically demonstrate the scalability of PALL across various architectures in image classification, and provide a state-of-the-art solution that uniquely integrates lifelong learning and privacy-aware unlearning mechanisms for responsible AI applications.

816. Model-based Offline Reinforcement Learning with Lower Expectile Q-Learning

链接：https://iclr.cc/virtual/2025/poster/29841 abstract： Model-based offline reinforcement learning (RL) is a compelling approach that addresses the challenge of learning from limited, static data by generating imaginary trajectories using learned models. However, these approaches often struggle with inaccurate value estimation from model rollouts. In this paper, we introduce a novel model-based offline RL method, Lower Expectile Q-learning (LEQ), which provides a low-bias model-based value estimation via lower expectile regression of λ-returns. Our empirical results show that LEQ significantly outperforms previous model-based offline RL methods on long-horizon tasks, such as the D4RL AntMaze tasks, matching or surpassing the performance of model-free approaches and sequence modeling approaches. Furthermore, LEQ matches the performance of state-of-the-art model-based and model-free methods in dense-reward environments across both state-based tasks (NeoRL and D4RL) and pixel-based tasks (V-D4RL), showing that LEQ works robustly across diverse domains. Our ablation studies demonstrate that lower expectile regression, λ-returns, and critic training on offline data are all crucial for LEQ.

817. Approximation algorithms for combinatorial optimization with predictions

链接：https://iclr.cc/virtual/2025/poster/30643 abstract： We initiate a systematic study of utilizing predictions to improve over approximation guarantees of classic algorithms, without increasing the running time. We propose a generic method for a wide class of optimization problems that ask to select a feasible subset of input items of minimal (or maximal) total weight. This gives simple (near-)linear-time algorithms for, e.g., Vertex Cover, Steiner Tree, Minimum Weight Perfect Matching, Knapsack, and Maximum Clique. Our algorithms produce an optimal solution when provided with perfect predictions and their approximation ratio smoothly degrades with increasing prediction error. With small enough prediction error we achieve approximation guarantees that are beyond the reach without predictions in given time bounds, as exemplified by the NP-hardness and APX-hardness of many of the above problems. Although we show our approach to be optimal for this class of problems as a whole, there is a potential for exploiting specific structural properties of individual problems to obtain improved bounds; we demonstrate this on the Steiner Tree problem. We conclude with an empirical evaluation of our approach.

818. Transformers Learn Low Sensitivity Functions: Investigations and Implications

链接：https://iclr.cc/virtual/2025/poster/30997 abstract： Transformers achieve state-of-the-art accuracy and robustness across many tasks, but an understanding of their inductive biases and how those biases differ from other neural network architectures remains elusive. In this work, we identify the sensitivity of the model to token-wise random perturbations in the input as a unified metric which explains the inductive bias of transformers across different data modalities and distinguishes them from other architectures. We show that transformers have lower sensitivity than MLPs, CNNs, ConvMixers and LSTMs, across both vision and language tasks. We also show that this low-sensitivity bias has important implications: i) lower sensitivity correlates with improved robustness; it can also be used as an efficient intervention to further improve the robustness of transformers; ii) it corresponds to flatter minima in the loss landscape; and iii) it can serve as a progress measure for grokking. We support these findings with theoretical results showing (weak) spectral bias of transformers in the NTK regime, and improved robustness due to the lower sensitivity.

819. Visually Consistent Hierarchical Image Classification

链接：https://iclr.cc/virtual/2025/poster/30832 abstract： Hierarchical classification predicts labels across multiple levels of a taxonomy, e.g., from coarse-level \textit{Bird} to mid-level \textit{Hummingbird} to fine-level \textit{Green hermit}, allowing flexible recognition under varying visual conditions. It is commonly framed as multiple single-level tasks, but each level may rely on different visual cues. Distinguishing \textit{Bird} from \textit{Plant} relies on {\it global features} like {\it feathers} or {\it leaves}, while separating \textit{Anna's hummingbird} from \textit{Green hermit} requires {\it local details} such as {\it head coloration}. Prior methods improve accuracy using external semantic supervision, but such statistical learning criteria fail to ensure consistent visual grounding at test time, resulting in incorrect hierarchical classification. We propose, for the first time, to enforce \textit{internal visual consistency} by aligning fine-to-coarse predictions through intra-image segmentation. Our method outperforms zero-shot CLIP and state-of-the-art baselines on hierarchical classification benchmarks, achieving both higher accuracy and more consistent predictions. It also improves internal image segmentation without requiring pixel-level annotations.

820. TSC-Net: Prediction of Pedestrian Trajectories by Trajectory-Scene-Cell Classification

链接：https://iclr.cc/virtual/2025/poster/29293 abstract： To predict future trajectories of pedestrians, scene is as important as the history trajectory since i) scene reflects the position of possible goals of the pedestrian ii) trajectories are affected by the semantic information of the scene. It requires the model to capture scene information and learn the relation between scenes and trajectories. However, existing methods either apply Convolutional Neural Networks (CNNs) to summarize the scene to a feature vector, which raises the feature misalignment issue, or convert trajectory to heatmaps to align with the scene map, which ignores the interactions among different pedestrians. In this work, we introduce the trajectory-scene-cell feature to represent both trajectories and scenes in one feature space. By decoupling the trajectory in temporal domain and the scene in spatial domain, trajectory feature and scene feature are re-organized in different types of cell feature, which well aligns trajectory and scene, and

allows the framework to model both human-human and human-scene interactions. Moreover, the Trajectory-Scene-Cell Network (TSC-Net) with new trajectory prediction manner is proposed, where both goal and intermediate positions of the trajectory are predict by cell classification and offset regression. Comparative experiments show that TSC-Net achieves the SOTA performance on several datasets with most of the metrics. Especially for the goal estimation, TSC-Net is demonstrated better on predicting goals for trajectories with irregular speed.

821. SIM: Surface-based fMRI Analysis for Inter-Subject Multimodal Decoding from Movie-Watching Experiments

链接：https://iclr.cc/virtual/2025/poster/29830 abstract： Current AI frameworks for brain decoding and encoding, typically train and test models within the same datasets. This limits their utility for cognitive training (neurofeedback) for which it would be useful to pool experiences across individuals to better simulate stimuli not sampled during training. A key obstacle to model generalisation is the degree of variability of inter-subject cortical organisation, which makes it difficult to align or compare cortical signals across participants. In this paper we address this through use of surface vision transformers, which build a generalisable model of cortical functional dynamics, through encoding the topography of cortical networks and their interactions as a moving image across a surface. This is then combined with tri-modal self-supervised contrastive (CLIP) alignment of audio, video, and fMRI modalities to enable the retrieval of visual and auditory stimuli from patterns of cortical activity (and vice-versa). We validate our approach on 7T task-fMRI data from 174 healthy participants engaged in the movie-watching experiment from the Human Connectome Project (HCP). Results show that it is possible to detect which movie clips an individual is watching purely from their brain activity, even for individuals and movies not seen during training. Further analysis of attention maps reveals that our model captures individual patterns of brain activity that reflect semantic and visual systems. This opens the door to future personalised simulations of brain function. Code \& pre-trained models will be made available at https://github.com/metrics-lab/sim.

822. Balanced Ranking with Relative Centrality: A multi-core periphery perspective

链接：https://iclr.cc/virtual/2025/poster/31168 abstract： Ranking of vertices in a graph for different objectives is one of the most fundamental tasks in computer science. It is known that traditional ranking algorithms can generate unbalanced ranking when the graph has underlying communities, resulting in loss of information, polarised opinions, and reduced diversity (Celis, Straszak \& Vishnoi [ICALP 2018]).In this paper, we focus on *unsupervised ranking* on graphs and observe that popular centrality measure based ranking algorithms such as PageRank may often generate unbalanced ranking here as well. We address this issue by coining a new approach, which we term *relative centrality*. Our approach is based on an iterative graph-dependent local normalization of the centrality score, which promotes balancedness while maintaining the validity of the ranking.We further quantify reasons behind this unbalancedness of centrality measures on a novel structure that we propose is called multi-core-periphery with communities (MCPC). We also provide theoretical and extensive simulation support for our approach towards resolving the unbalancedness in MCPC.Finally, we consider graph embeddings of 11 single-cell datasets. We observe that top-ranked as per existing centrality measures are better separable into the ground truth communities. However, due to the unbalanced ranking, the top nodes often do not contain points from some communities. Here, our relative-centrality-based approach generates a ranking that provides a similar improvement in clusterability while providing significantly higher balancedness.

823. Fréchet Wavelet Distance: A Domain-Agnostic Metric for Image Generation

链接：https://iclr.cc/virtual/2025/poster/29680 abstract： Modern metrics for generative learning like Fréchet Inception Distance (FID) and DINOv2-Fréchet Distance (FD-DINOv2) demonstrate impressive performance. However, they suffer from various shortcomings, like a bias towards specific generators and datasets. To address this problem, we propose the Fréchet Wavelet Distance (FWD) as a domain-agnostic metric based on the Wavelet Packet Transform ($\mathcal{W}_p$). FWD provides a sight across a broad spectrum of frequencies in images with a high resolution, preserving both spatial and textural aspects. Specifically, we use $\mathcal{W}_p$ to project generated and real images to the packet coefficient space. We then compute the Fréchet distance with the resultant coefficients to evaluate the quality of a generator. This metric is general-purpose and dataset-domain agnostic, as it does not rely on any pre-trained network, while being more interpretable due to its ability to compute Fréchet distance per packet, enhancing transparency. We conclude with an extensive evaluation of a wide variety of generators across various datasets that the proposed FWD can generalize and improve robustness to domain shifts and various corruptions compared to other metrics.

824. Test-time Adaptation for Image Compression with Distribution Regularization

链接：https://iclr.cc/virtual/2025/poster/29083 abstract： Current test- or compression-time adaptation image compression (TTA-IC) approaches, which leverage both latent and decoder refinements as a two-step adaptation scheme, have potentially enhanced the rate-distortion (R-D) performance of learned image compression models on cross-domain compression tasks, \textit{e.g.,} from natural to screen content images. However, compared with the emergence of various decoder refinement

variants, the latent refinement, as an inseparable ingredient, is barely tailored to cross-domain scenarios. To this end, we are interested in developing an advanced latent refinement method by extending the effective hybrid latent refinement (HLR) method, which is designed for \textit{in-domain} inference improvement but shows noticeable degradation of the rate cost in \textit{cross-domain} tasks. Specifically, we first provide theoretical analyses, in a cue of marginalization approximation from in- to cross-domain scenarios, to uncover that the vanilla HLR suffers from an underlying mismatch between refined Gaussian conditional and hyperprior distributions, leading to deteriorated joint probability approximation of marginal distribution with increased rate consumption. To remedy this issue, we introduce a simple Bayesian approximation-endowed \textit{distribution regularization} to encourage learning a better joint probability approximation in a plug-and-play manner. Extensive experiments on six in- and cross-domain datasets demonstrate that our proposed method not only improves the R-D performance compared with other latent refinement counterparts, but also can be flexibly integrated into existing TTA-IC methods with incremental benefits.

825. Learning Mask Invariant Mutual Information for Masked Image Modeling

链接：https://iclr.cc/virtual/2025/poster/29858 abstract：

826. Last Iterate Convergence of Incremental Methods as a Model of Forgetting

链接：https://iclr.cc/virtual/2025/poster/28465 abstract：

827. Parameter Expanded Stochastic Gradient Markov Chain Monte Carlo

链接：https://iclr.cc/virtual/2025/poster/28902 abstract： Bayesian Neural Networks (BNNs) provide a promising framework for modeling predictive uncertainty and enhancing out-of-distribution robustness (OOD) by estimating the posterior distribution of network parameters. Stochastic Gradient Markov Chain Monte Carlo (SGMCMC) is one of the most powerful methods for scalable posterior sampling in BNNs, achieving efficiency by combining stochastic gradient descent with second-order Langevin dynamics. However, SGMCMC often suffers from limited sample diversity in practice, which affects uncertainty estimation and model performance. We propose a simple yet effective approach to enhance sample diversity in SGMCMC without the need for tempering or running multiple chains. Our approach reparameterizes the neural network by decomposing each of its weight matrices into a product of matrices, resulting in a sampling trajectory that better explores the target parameter space. This approach produces a more diverse set of samples, allowing faster mixing within the same computational budget. Notably, our sampler achieves these improvements without increasing the inference cost compared to the standard SGMCMC. Extensive experiments on image classification tasks, including OOD robustness, diversity, loss surface analyses, and a comparative study with Hamiltonian Monte Carlo, demonstrate the superiority of the proposed approach.

828. The Utility and Complexity of In- and Out-of-Distribution Machine Unlearning

链接：https://iclr.cc/virtual/2025/poster/30223 abstract： Machine unlearning, the process of selectively removing data from trained models, is increasingly crucial for addressing privacy concerns and knowledge gaps post-deployment. Despite this importance, existing approaches are often heuristic and lack formal guarantees. In this paper, we analyze the fundamental utility, time, and space complexity trade-offs of approximate unlearning, providing rigorous certification analogous to differential privacy. For in-distribution forget data—data similar to the retain set—we show that a surprisingly simple and general procedure, empirical risk minimization with output perturbation, achieves tight unlearning-utility-complexity trade-offs, addressing a previous theoretical gap on the separation from unlearning ``for free" via differential privacy, which inherently facilitates the removal of such data. However, such techniques fail with out-of-distribution forget data—data significantly different from the retain set—where unlearning time complexity can exceed that of retraining, even for a single sample. To address this, we propose a new robust and noisy gradient descent variant that provably amortizes unlearning time complexity without compromising utility.

829. Looking Backward: Retrospective Backward Synthesis for Goal-Conditioned GFlowNets

链接：https://iclr.cc/virtual/2025/poster/28879 abstract： Generative Flow Networks (GFlowNets), a new family of probabilistic samplers, have demonstrated remarkable capabilities to generate diverse sets of high-reward candidates, in contrast to standard return maximization approaches (e.g., reinforcement learning) which often converge to a single optimal solution. Recent works have focused on developing goal-conditioned GFlowNets, which aim to train a single GFlowNet capable of achieving different outcomes as the task specifies. However, training such models is challenging due to extremely sparse rewards, particularly in high-dimensional problems. Moreover, previous methods suffer from the limited coverage of explored trajectories during training, which presents more pronounced challenges when only offline data is available. In this work, we propose a novel method called \textbf{R}etrospective \textbf{B}ackward \textbf{S}ynthesis (\textbf{RBS}) to address these critical problems. Specifically, RBS synthesizes new backward trajectories in goal-conditioned GFlowNets to enrich training trajectories with enhanced quality and diversity, thereby introducing copious learnable signals for effectively tackling the sparse reward problem. Extensive empirical results show that our method improves sample efficiency by a large margin and outperforms strong baselines on various standard evaluation benchmarks. Our codes are available at \url{https://github.com/tinnerhrhe/Goal-

Conditioned-GFN}.

830. Adaptive Gradient Clipping for Robust Federated Learning

链接：https://iclr.cc/virtual/2025/poster/31272 abstract： Robust federated learning aims to maintain reliable performance despite the presence of adversarial or misbehaving workers. While state-of-the-art (SOTA) robust distributed gradient descent (Robust-DGD) methods were proven theoretically optimal, their empirical success has often relied on pre-aggregation gradient clipping.However, existing static clipping strategies yield inconsistent results: enhancing robustness against some attacks while being ineffective or even detrimental against others.To address this limitation, we propose a principled adaptive clipping strategy, Adaptive Robust Clipping (ARC), which dynamically adjusts clipping thresholds based on the input gradients. We prove that ARC not only preserves the theoretical robustness guarantees of SOTA Robust-DGD methods but also provably improves asymptotic convergence when the model is well-initialized. Extensive experiments on benchmark image classification tasks confirm these theoretical insights, demonstrating that ARC significantly enhances robustness, particularly in highly heterogeneous and adversarial settings.

831. Neuroplastic Expansion in Deep Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/31169 abstract： The loss of plasticity in learning agents, analogous to the solidification of neural pathways in biological brains, significantly impedes learning and adaptation in reinforcement learning due to its non-stationary nature. To address this fundamental challenge, we propose a novel approach, Neuroplastic Expansion (NE), inspired by cortical expansion in cognitive science. NE maintains learnability and adaptability throughout the entire training process by dynamically growing the network from a smaller initial size to its full dimension. Our method is designed with three key components: (1) elastic neuron generation based on potential gradients, (2) dormant neuron pruning to optimize network expressivity, and (3) neuron consolidation via experience review to strike a balance in the plasticity-stability dilemma. Extensive experiments demonstrate that NE effectively mitigates plasticity loss and outperforms state-of-the-art methods across various tasks in MuJoCo and DeepMind Control Suite environments. NE enables more adaptive learning in complex, dynamic environments, which represents a crucial step towards transitioning deep reinforcement learning from static, one-time training paradigms to more flexible, continually adapting models.

832. Learning to Plan Before Answering: Self-Teaching LLMs to Learn Abstract Plans for Problem Solving

链接：https://iclr.cc/virtual/2025/poster/30031 abstract： In the field of large language model (LLM) post-training, the effectiveness of utilizing synthetic data generated by the LLM itself has been well-presented. However, a key question remains unaddressed: what essential information should such self-generated data encapsulate? Existing approaches only produce step-by-step problem solutions, and fail to capture the abstract meta-knowledge necessary for generalization across similar problems. Drawing insights from cognitive science, where humans employ high-level abstraction to simplify complex problems before delving into specifics, we introduce a novel self-training algorithm: LEarning to Plan before Answering (LEPA). LEPA trains the LLM to formulate anticipatory plans, which serve as abstract meta-knowledge for problem-solving, before engaging with the intricacies of problems. This approach not only outlines the solution generation path but also shields the LLM from the distraction of irrelevant details. During data generation, LEPA first crafts an anticipatory plan based on the problem, and then generates a solution that aligns with both the plan and the problem. LEPA refines the plan through self-reflection, aiming to acquire plans that are instrumental in yielding correct solutions. During model optimization, the LLM is trained to predict both the refined plans and the corresponding solutions. By efficiently extracting and utilizing the anticipatory plans, LEPA demonstrates remarkable superiority over conventional algorithms on various challenging natural language reasoning benchmarks.

833. Improved Regret Bounds for Linear Adversarial MDPs via Linear Optimization

链接：https://iclr.cc/virtual/2025/poster/31509 abstract： Learning Markov decision processes (MDP) in an adversarial environment has been a challenging problem. The problem becomes even more challenging with function approximation since the underlying structure of the loss function and transition kernel are especially hard to estimate in a varying environment. In fact, the state-of-the-art results for linear adversarial MDP achieve a regret of $\tilde{\mathcal{O}}({K^{6/7}})$ (K denotes the number of episodes), which admits a large room for improvement. In this paper, we propose a novel explore-exploit algorithm framework and investigate the problem with a new view, which reduces linear MDP into linear optimization by subtly setting the feature maps of the bandit arms of linear optimization. This new technique, under an exploratory assumption, yields an improved bound of $\tilde{\mathcal{O}}({K^{4/5}})$ for linear adversarial MDP without access to a transition simulator. The new view could be of independent interest for solving other MDP problems that possess a linear structure.

834. Zero-shot Model-based Reinforcement Learning using Large Language Models

链接：https://iclr.cc/virtual/2025/poster/27971 abstract： The emerging zero-shot capabilities of Large Language Models (LLMs) have led to their applications in areas extending well beyond natural language processing tasks. In reinforcement

learning, while LLMs have been extensively used in text-based environments, their integration with continuous state spaces remains understudied. In this paper, we investigate how pre-trained LLMs can be leveraged to predict in context the dynamics of continuous Markov decision processes. We identify handling multivariate data and incorporating the control signal as key challenges that limit the potential of LLMs' deployment in this setup and propose Disentangled In-Context Learning (DICL) to address them.We present proof-of-concept applications in two reinforcement learning settings: model-based policy evaluation and data-augmented off-policy reinforcement learning, supported by theoretical analysis of the proposed methods.Our experiments further demonstrate that our approach produces well-calibrated uncertainty estimates. We release the code at https://github.com/abenechehab/dicl.

835. Sufficient Context: A New Lens on Retrieval Augmented Generation Systems

链接：https://iclr.cc/virtual/2025/poster/30092 abstract： Augmenting LLMs with context leads to improved performance across many applications. Despite much research on Retrieval Augmented Generation (RAG) systems, an open question is whether errors arise because LLMs fail to utilize the context from retrieval or the context itself is insufficient to answer the query. To shed light on this, we develop a new notion of sufficient context, along with a method to classify instances that have enough information to answer the query. We then use sufficient context to analyze several models and datasets. By stratifying errors based on context sufficiency, we find that larger models with higher baseline performance (Gemini 1.5 Pro, GPT 4o, Claude 3.5) excel at answering queries when the context is sufficient, but often output incorrect answers instead of abstaining when the context is not. On the other hand, smaller models with lower baseline performance (Llama 3.1, Mistral 3, Gemma 2) hallucinate or abstain often, even with sufficient context. We further categorize cases when the context is useful, and improves accuracy, even though it does not fully answer the query and the model errs without the context. Building on our findings, we explore ways to reduce hallucinations in RAG systems, including a new selective generation method that leverages sufficient context information for guided abstention. Our method improves the fraction of correct answers among times where the model responds by 2--10% for Gemini, GPT, and Gemma. Code for our selective generation method and the prompts used in our autorater analysis are available on our github.

836. A Large-Scale 3D Face Mesh Video Dataset via Neural Re-parameterized Optimization

链接：https://iclr.cc/virtual/2025/poster/31483 abstract： We propose NeuFace, a 3D face mesh pseudo annotation method on videos via neural re-parameterized optimization. Despite the huge progress in 3D face reconstruction methods, generating reliable 3D face labels for in-the-wild dynamic videos remains challenging. Using NeuFace optimization, we annotate the per-view/-frame accurate and consistent face meshes on large-scale face videos, called the NeuFace-dataset. We investigate how neural re-parameterization helps to reconstruct image-aligned facial details on 3D meshes via gradient analysis. By exploiting the naturalness and diversity of 3D faces in our dataset, we demonstrate the usefulness of our dataset for 3D face-related tasks: improving the reconstruction accuracy of an existing 3D face reconstruction model and learning 3D facial motion prior.

837. Semantic Aware Representation Learning for Lifelong Learning

链接：https://iclr.cc/virtual/2025/poster/29333 abstract：

838. MrT5: Dynamic Token Merging for Efficient Byte-level Language Models

链接：https://iclr.cc/virtual/2025/poster/29408 abstract： Models that rely on subword tokenization have significant drawbacks, such as sensitivity to character-level noise like spelling errors and inconsistent compression rates across different languages and scripts. While character- or byte-level models like ByT5 attempt to address these concerns, they have not gained widespread adoption—processing raw byte streams without tokenization results in significantly longer sequence lengths, making training and inference inefficient. This work introduces MrT5 (MergeT5), a more efficient variant of ByT5 that integrates a token deletion mechanism in its encoder to dynamically shorten the input sequence length. After processing through a fixed number of encoder layers, a learned delete gate determines which tokens are to be removed and which are to be retained for subsequent layers. MrT5 effectively "merges" critical information from deleted tokens into a more compact sequence, leveraging contextual information from the remaining tokens. In continued pre-training experiments, we find that MrT5 can achieve significant gains in inference runtime with minimal effect on performance, as measured by bits-per-byte. Additionally, with multilingual training, MrT5 adapts to the orthographic characteristics of each language, learning language-specific compression rates. Furthermore, MrT5 shows comparable accuracy to ByT5 on downstream evaluations such as XNLI, TyDi QA, and character-level tasks while reducing sequence lengths by up to 75%. Our approach presents a solution to the practical limitations of existing byte-level models.

839. Min-K%++: Improved Baseline for Pre-Training Data Detection from Large Language Models

链接：https://iclr.cc/virtual/2025/poster/29228 abstract： The problem of pre-training data detection for large language models (LLMs) has received growing attention due to its implications in critical issues like copyright violation and test data

contamination. Despite improved performance, existing methods (including the state-of-the-art, Min-K%) are mostly developed upon simple heuristics and lack solid, reasonable foundations. In this work, we propose a novel and theoretically motivated methodology for pre-training data detection, named Min-K%++. Specifically, we present a key insight that training samples tend to be local maxima of the modeled distribution along each input dimension through maximum likelihood training, which in turn allow us to insightfully translate the problem into identification of local maxima. Then, we design our method accordingly that works under the discrete distribution modeled by LLMs, whose core idea is to determine whether the input forms a mode or has relatively high probability under the conditional categorical distribution. Empirically, the proposed method achieves new SOTA performance across multiple settings (evaluated with 5 families of 10 models and 2 benchmarks). On the WikiMIA benchmark, Min-K%++ outperforms the runner-up by 6.2% to 10.5% in detection AUROC averaged over five models. On the more challenging MIMIR benchmark, it consistently improves upon reference-free methods while performing on par with reference-based method that requires an extra reference model.

840. An Effective Manifold-based Optimization Method for Distributionally Robust Classification

链接：https://iclr.cc/virtual/2025/poster/28383 abstract： How to promote the robustness of existing deep learning models is a challenging problem for many practical classification tasks. Recently, Distributionally Robust Optimization (DRO) methods have shown promising potential to tackle this problem. These methods aim to construct reliable models by minimizing the worst-case risk within a local region (called "uncertainty set") around the empirical data distribution. However, conventional DRO methods tend to be overly pessimistic, leading to certain discrepancy between the real data distribution and the uncertainty set, which can degrade the classification performance. To address this issue, we propose a manifold-based DRO method that takes the geometric structure of training data into account for constructing the uncertainty set. Specifically, our method employs a carefully designed "game" that integrates contrastive learning with Jacobian regularization to capture the manifold structure, enabling us to solve DRO problems constrained by the data manifold. By utilizing a novel idea for approximating geodesic distance on manifolds, we also provide the theoretical guarantees for its robustness. Moreover, our proposed method is easy to implement in practice. We conduct a set of experiments on several popular benchmark datasets, where the results demonstrate our advantages in terms of accuracy and robustness.

841. MambaExtend: A Training-Free Approach to Improve Long Context Extension of Mamba

链接：https://iclr.cc/virtual/2025/poster/29980 abstract： The inherent quadratic complexity of the attention mechanism in transformer models has driven the research community to explore alternative architectures with sub-quadratic complexity, such as state-space models. Mamba has established itself as a leading model within this emerging paradigm, achieving state-of-the-art results in various language modeling benchmarks. However, despite its impressive performance, Mamba's effectiveness is limited by its pre-training context length, resulting in a pronounced degradation when the model is tasked with handling longer contexts. Our investigation reveals that Mamba's inability to generalize effectively to long contexts is primarily due to the out-of-distribution (OOD) discretization steps. To address this critical limitation, we introduce ***MambaExtend***, a novel framework designed to significantly enhance the context extension capabilities of Mamba. Specifically, MambaExtend leverages a ***training-free*** approach to calibrate *only* the scaling factors of discretization modules for different layers. We demonstrate both gradient-based and gradient-free zeroth-order optimization to learn the optimal scaling factors for each Mamba layer, requiring orders of magnitude fewer updates as opposed to the parameter fine-tuning-based alternatives. Using this approach, we achieve a training-free context extension of up to 32x, expanding the context from 2k to 64k tokens with minimal increases in perplexity. In contrast to existing fine-tuning methods, MambaExtend selectively calibrates the scaling factors, requiring up to $\mathbf{5.42 * 10^6} \times$ fewer parameter updates and incurring up to $\mathbf{3.87} \times$ lower peak memory usage, while delivering comparable or superior long-context performance across multiple tasks. Codes and checkpoints are available here^1.

842. Detecting Backdoor Samples in Contrastive Language Image Pretraining

链接：https://iclr.cc/virtual/2025/poster/30032 abstract： Contrastive language-image pretraining (CLIP) has been found to be vulnerable to poisoning backdoor attacks where the adversary can achieve an almost perfect attack success rate on CLIP models by poisoning only 0.01\% of the training dataset. This raises security concerns on the current practice of pretraining large-scale models on unscrutinized web data using CLIP. In this work, we analyze the representations of backdoor-poisoned samples learned by CLIP models and find that they exhibit unique characteristics in their local subspace, i.e., their local neighborhoods are far more sparse than that of clean samples. Based on this finding, we conduct a systematic study on detecting CLIP backdoor attacks and show that these attacks can be easily and efficiently detected by traditional density ratio-based local outlier detectors, whereas existing backdoor sample detection methods fail. Our experiments also reveal that an unintentional backdoor already exists in the original CC3M dataset and has been trained into a popular open-source model released by OpenCLIP. Based on our detector, one can clean up a million-scale web dataset (e.g., CC3M) efficiently within 15 minutes using 4 Nvidia A100 GPUs.

843. Learning to Solve Differential Equation Constrained Optimization

Problems

链接：https://iclr.cc/virtual/2025/poster/29404 abstract： Differential equations (DE) constrained optimization plays a critical role in numerous scientific and engineering fields, including energy systems, aerospace engineering, ecology, and finance, where optimal configurations or control strategies must be determined for systems governed by ordinary or stochastic differential equations. Despite its significance, the computational challenges associated with these problems have limited their practical use. To address these limitations, this paper introduces a learning-based approach to DE-constrained optimization that combines techniques from proxy optimization \citep{kotary2021end} and neural differential equations \citep{chen2019neural}. The proposed approach uses a dual-network architecture, with one approximating the control strategies, focusing on steady-state constraints, and another solving the associated DEs. This combination enables the approximation of optimal strategies while accounting for dynamic constraints in near real-time.Experiments across problems in energy optimization and finance modeling show that this method provides full compliance with dynamic constraints and it produces results up to 25 times more precise than other methods which do not explicitly model the system's dynamic equations.

844. Mitigating Spurious Correlations in Zero-Shot Multimodal Models

链接：https://iclr.cc/virtual/2025/poster/29449 abstract： Multimodal models or Vision Language Models (VLMs) have reshaped the paradigm in machine learning, offering zero-shot capabilities that require no additional training when adapted to new classification tasks. However, despite their advancements, spurious correlations still exist in VLMs. Existing approaches to tackle this issue often require target label annotations, contradicting the principle of zero-shot classification, or they primarily focus on a single modality, risking misalignment between text and image modalities. Others rely on extensive domain knowledge or large language models (LLMs) to characterize spurious features, making the performance sensitive to the generated prompts and undermining zero-shot capability. In response, we propose a new solution that tackles spurious correlations in VLMs within the zero-shot setting. Our approach utilizes a translation operation that preserves the latent space distribution to address issues of spurious correlations. In particular, our method is grounded in and inspired by a theoretical analysis, which identifies that the optimal translation directions are along the spurious vector. As VLMs unify two modalities, we compute spurious vectors from the text prompts and guide the translation for image embeddings, aligning the requirements for the fusion of different modalities in VLMs. We conducted experiments on benchmark datasets, which have shown significant improvements in worst-group accuracy. Additionally, our visualizations of VLMs further demonstrate the effectiveness of this intervention.

845. Representational Similarity via Interpretable Visual Concepts

链接：https://iclr.cc/virtual/2025/poster/28681 abstract： How do two deep neural networks differ in how they arrive at a decision? Measuring the similarity of deep networks has been a long-standing open question. Most existing methods provide a single number to measure the similarity of two networks at a given layer, but give no insight into what makes them similar or dissimilar. We introduce an interpretable representational similarity method (RSVC) to compare two networks. We use RSVC to discover shared and unique visual concepts between two models. We show that some aspects of model differences can be attributed to unique concepts discovered by one model that are not well represented in the other. Finally, we conduct extensive evaluation across different vision model architectures and training protocols to demonstrate its effectiveness.

846. Separation Power of Equivariant Neural Networks

链接：https://iclr.cc/virtual/2025/poster/29661 abstract： The separation power of a machine learning model refers to its ability to distinguish between different inputs and is often used as a proxy for its expressivity. Indeed, knowing the separation power of a family of models is a necessary condition to obtain fine-grained universality results. In this paper, we analyze the separation power of equivariant neural networks, such as convolutional and permutation-invariant networks.We first present a complete characterization of inputs indistinguishable by models derived by a given architecture. From this results, we derive how separability is influenced by hyperparameters and architectural choices—such as activation functions, depth, hidden layer width, and representation types. Notably, all non-polynomial activations, including ReLU and sigmoid, are equivalent in expressivity and reach maximum separation power. Depth improves separation power up to a threshold, after which further increases have no effect. Adding invariant features to hidden representations does not impact separation power. Finally, block decomposition of hidden representations affects separability, with minimal components forming a hierarchy in separation power that provides a straightforward method for comparing the separation power of models.

847. Test-Time Adaptation for Combating Missing Modalities in Egocentric Videos

链接：https://iclr.cc/virtual/2025/poster/31208 abstract： Understanding videos that contain multiple modalities is crucial, especially in egocentric videos, where combining various sensory inputs significantly improves tasks like action recognition and moment localization. However, real-world applications often face challenges with incomplete modalities due to privacy concerns, efficiency needs, or hardware issues. Current methods, while effective, often necessitate retraining the model entirely to handle missing modalities, making them computationally intensive, particularly with large training datasets. In this study, we propose a novel approach to address this issue at test time without requiring retraining. We frame the problem as a test-time adaptation task, where the model adjusts to the available unlabeled data at test time. Our method, MiDl~(Mutual information with self-Distillation), encourages the model to be insensitive to the specific modality source present during testing by minimizing the

mutual information between the prediction and the available modality. Additionally, we incorporate self-distillation to maintain the model's original performance when both modalities are available. MiDl represents the first self-supervised, online solution for handling missing modalities exclusively at test time. Through experiments with various pretrained models and datasets, MiDl demonstrates substantial performance improvement without the need for retraining.

848. Bundle Neural Network for message diffusion on graphs

链接：https://iclr.cc/virtual/2025/poster/28114 abstract： The dominant paradigm for learning on graphs is message passing. Despite being a strong inductive bias, the local message passing mechanism faces challenges such as over-smoothing, over-squashing, and limited expressivity. To address these issues, we introduce Bundle Neural Networks (BuNNs), a novel graph neural network architecture that operates via message diffusion on flat vector bundles — geometrically inspired structures that assign to each node a vector space and an orthogonal map. A BuNN layer evolves node features through a diffusion-type partial differential equation, where its discrete form acts as a special case of the recently introduced Sheaf Neural Network (SNN), effectively alleviating over-smoothing. The continuous nature of message diffusion enables BuNNs to operate at larger scales, reducing over-squashing. We establish the universality of BuNNs in approximating feature transformations on infinite families of graphs with injective positional encodings, marking the first positive expressivity result of its kind. We support our claims with formal analysis and synthetic experiments. Empirically, BuNNs perform strongly on heterophilic and long-range tasks, which demonstrates their robustness on a diverse range of challenging real-world tasks.

849. Neural Spacetimes for DAG Representation Learning

链接：https://iclr.cc/virtual/2025/poster/28100 abstract： We propose a class of trainable deep learning-based geometries called Neural SpaceTimes (NSTs), which can universally represent nodes in weighted Directed Acyclic Graphs (DAGs) as events in a spacetime manifold. While most works in the literature focus on undirected graph representation learning or causality embedding separately, our differentiable geometry can encode both graph edge weights in its spatial dimensions and causality in the form of edge directionality in its temporal dimensions. We use a product manifold that combines a quasi-metric (for space) and a partial order (for time). NSTs are implemented as three neural networks trained in an end-to-end manner: an embedding network, which learns to optimize the location of nodes as events in the spacetime manifold, and two other networks that optimize the space and time geometries in parallel, which we call a neural (quasi-)metric and a neural partial order, respectively. The latter two networks leverage recent ideas at the intersection of fractal geometry and deep learning to shape the geometry of the representation space in a data-driven fashion, unlike other works in the literature that use fixed spacetime manifolds such as Minkowski space or De Sitter space to embed DAGs. Our main theoretical guarantee is a universal embedding theorem, showing that any k-point DAG can be embedded into an NST with $1+\mathcal{O}(\log(k))$ distortion while exactly preserving its causal structure. The total number of parameters defining the NST is sub-cubic in k and linear in the width of the DAG. If the DAG has a planar Hasse diagram, this is improved to $\mathcal{O}(\log(k) + 2)$ spatial and 2 temporal dimensions. We validate our framework computationally with synthetic weighted DAGs and real-world network embeddings; in both cases, the NSTs achieve lower embedding distortions than their counterparts using fixed spacetime geometries.

850. Training-Free Message Passing for Learning on Hypergraphs

链接：https://iclr.cc/virtual/2025/poster/31030 abstract： Hypergraphs are crucial for modelling higher-order interactions in real-world data. Hypergraph neural networks (HNNs) effectively utilise these structures by message passing to generate informative node features for various downstream tasks like node classification. However, the message passing module in existing HNNs typically requires a computationally intensive training process, which limits their practical use. To tackle this challenge, we propose an alternative approach by decoupling the usage of hypergraph structural information from the model learning stage. This leads to a novel training-free message passing module, named TF-MP-Module, which can be precomputed in the data preprocessing stage, thereby reducing the computational burden. We refer to the hypergraph neural network equipped with our TF-MP-Module as TF-HNN. We theoretically support the efficiency and effectiveness of TF-HNN by showing that: 1) It is more training-efficient compared to existing HNNs; 2) It utilises as much information as existing HNNs for node feature generation; and 3) It is robust against the oversmoothing issue while using long-range interactions. Experiments based on seven real-world hypergraph benchmarks in node classification and hyperlink prediction show that, compared to state-of-the-art HNNs, TF-HNN exhibits both competitive performance and superior training efficiency. Specifically, on the large-scale benchmark, Trivago, TF-HNN outperforms the node classification accuracy of the best baseline by 10% with just 1% of the training time of that baseline.

851. The Value of Sensory Information to a Robot

链接：https://iclr.cc/virtual/2025/poster/28678 abstract： A decision-making agent, such as a robot, must observe and react to any new task-relevant information that becomes available from its environment. We seek to study a fundamental scientific question: what value does sensory information hold to an agent at various moments in time during the execution of a task? Towards this, we empirically study agents of varying architectures, generated with varying policy synthesis approaches (imitation, RL, model-based control), on diverse robotics tasks. For each robotic agent, we characterize its regret in terms of performance degradation when state observations are withheld from it at various task states for varying lengths of time. We find that sensory information is surprisingly rarely task-critical in many commonly studied task setups. Task characteristics such as stochastic dynamics largely dictate the value of sensory information for a well-trained robot; policy architectures such as planning vs. reactive control generate more nuanced second-order effects. Further, sensing efficiency is curiously correlated with task

proficiency: in particular, fully trained high-performing agents are more robust to sensor loss than novice agents early in their training. Overall, our findings characterize the tradeoffs between sensory information and task performance in practical sequential decision making tasks, and pave the way towards the design of more resource-efficient decision-making agents.

852. Manifolds, Random Matrices and Spectral Gaps: The geometric phases of generative diffusion

链接：https://iclr.cc/virtual/2025/poster/30036 abstract： In this paper, we investigate the latent geometry of generative diffusion models under the manifold hypothesis. For this purpose, we analyze the spectrum of eigenvalues (and singular values) of the Jacobian of the score function, whose discontinuities (gaps) reveal the presence and dimensionality of distinct sub-manifolds. Using a statistical physics approach, we derive the spectral distributions and formulas for the spectral gaps under several distributional assumptions, and we compare these theoretical predictions with the spectra estimated from trained networks. Our analysis reveals the existence of three distinct qualitative phases during the generative process: a trivial phase; a manifold coverage phase where the diffusion process fits the distribution internal to the manifold; a consolidation phase where the score becomes orthogonal to the manifold and all particles are projected on the support of the data. This `division of labor' between different timescales provides an elegant explanation of why generative diffusion models are not affected by the manifold overfitting phenomenon that plagues likelihood-based models, since the internal distribution and the manifold geometry are produced at different time points during generation.

853. QPM: Discrete Optimization for Globally Interpretable Image Classification

链接：https://iclr.cc/virtual/2025/poster/30260 abstract： Understanding the classifications of deep neural networks, e.g. used in safety-critical situations, is becoming increasingly important. While recent models can locally explain a single decision, to provide a faithful global explanation about an accurate model’s general behavior is a more challenging open task. Towards that goal, we introduce the Quadratic Programming Enhanced Model (QPM), which learns globally interpretable class representations. QPM represents every class with a binary assignment of very few, typically 5, features, that are also assigned to other classes, ensuring easily comparable contrastive class representations. This compact binary assignment is found using discrete optimization based on predefined similarity measures and interpretability constraints. The resulting optimal assignment is used to fine-tune the diverse features, so that each of them becomes the shared general concept between the assigned classes. Extensive evaluations show that QPM delivers unprecedented global interpretability across small and large-scale datasets while setting the state of the art for the accuracy of interpretable models.

854. SegLLM: Multi-round Reasoning Segmentation with Large Language Models

链接：https://iclr.cc/virtual/2025/poster/29732 abstract： We present SegLLM, a novel multi-round interactive reasoning segmentation model that enhances LLM-based segmentation by exploiting conversational memory of both visual and textual outputs. By leveraging a mask-aware multimodal LLM, SegLLM re-integrates previous segmentation results into its input stream, enabling it to reason about complex user intentions and segment objects in relation to previously identified entities, including positional, interactional, and hierarchical relationships, across multiple interactions. This capability allows SegLLM to respond to visual and text queries in a chat-like manner. Evaluated on the newly curated MRSeg benchmark, SegLLM outperforms existing methods in multi- round interactive reasoning segmentation by over 20%. Additionally, we observed that training on multi-round reasoning segmentation data enhances performance on standard single-round referring segmentation and localization tasks, resulting in a 5.5% increase in cIoU for referring expression segmentation and a 4.5% improvement in Acc@0.5 for referring expression localization.

855. BrainOOD: Out-of-distribution Generalizable Brain Network Analysis

链接：https://iclr.cc/virtual/2025/poster/31047 abstract： In neuroscience, identifying distinct patterns linked to neurological disorders, such as Alzheimer's and Autism, is critical for early diagnosis and effective intervention. Graph Neural Networks (GNNs) have shown promising in analyzing brain networks, but there are two major challenges in using GNNs: (1) distribution shifts in multi-site brain network data, leading to poor Out-of-Distribution (OOD) generalization, and (2) limited interpretability in identifying key brain regions critical to neurological disorders. Existing graph OOD methods, while effective in other domains, struggle with the unique characteristics of brain networks. To bridge these gaps, we introduce BrainOOD, a novel framework tailored for brain networks that enhances GNNs' OOD generalization and interpretability. BrainOOD framework consists of a feature selector and a structure extractor, which incorporates various auxiliary losses including an improved Graph Information Bottleneck (GIB) objective to recover causal subgraphs. By aligning structure selection across brain networks and filtering noisy features, BrainOOD offers reliable interpretations of critical brain regions. Our approach outperforms 16 existing methods and improves generalization to OOD subjects by up to 8.5%. Case studies highlight the scientific validity of the patterns extracted, which aligns with the findings in known neuroscience literature. We also propose the first OOD brain network benchmark, which provides a foundation for future research in this field. Our code is available at https://github.com/AngusMonroe/BrainOOD.

856. DebGCD: Debiased Learning with Distribution Guidance for Generalized

Category Discovery

链接：https://iclr.cc/virtual/2025/poster/30717 abstract： In this paper, we tackle the problem of Generalized Category Discovery (GCD). Given a dataset containing both labelled and unlabelled images, the objective is to categorize all images in the unlabelled subset, irrespective of whether they are from known or unknown classes. In GCD, an inherent label bias exists between known and unknown classes due to the lack of ground-truth labels for the latter. State-of-the-art methods in GCD leverage parametric classifiers trained through self-distillation with soft labels, leaving the bias issue unattended. Besides, they treat all unlabelled samples uniformly, neglecting variations in certainty levels and resulting in suboptimal learning. Moreover, the explicit identification of semantic distribution shifts between known and unknown classes, a vital aspect for effective GCD, has been neglected. To address these challenges, we introduce DebGCD, a Debiased learning with distribution guidance framework for GCD. Initially, DebGCD co-trains an auxiliary debiased classifier in the same feature space as the GCD classifier, progressively enhancing the GCD features. Moreover, we introduce a semantic distribution detector in a separate feature space to implicitly boost the learning efficacy of GCD. Additionally, we employ a curriculum learning strategy based on semantic distribution certainty to steer the debiased learning at an optimized pace. Thorough evaluations on GCD benchmarks demonstrate the consistent state-of-the-art performance of our framework, highlighting its superiority. Project page: https://visual-ai.github.io/debgcd/

857. Cached Multi-Lora Composition for Multi-Concept Image Generation

链接：https://iclr.cc/virtual/2025/poster/30998 abstract：

858. Robust Feature Learning for Multi-Index Models in High Dimensions

链接：https://iclr.cc/virtual/2025/poster/29175 abstract： Recently, there have been numerous studies on feature learning with neural networks, specifically on learning single- and multi-index models where the target is a function of a low-dimensional projection of the input. Prior works have shown that in high dimensions, the majority of the compute and data resources are spent on recovering the low-dimensional projection; once this subspace is recovered, the remainder of the target can be learned independently of the ambient dimension. However, implications of feature learning in adversarial settings remain unexplored. In this work, we take the first steps towards understanding adversarially robust feature learning with neural networks. Specifically, we prove that the hidden directions of a multi-index model offer a Bayes optimal low-dimensional projection for robustness against ℓ_2-bounded adversarial perturbations under the squared loss, assuming that the multi-index coordinates are statistically independent from the rest of the coordinates. Therefore, robust learning can be achieved by first performing standard feature learning, then robustly tuning a linear readout layer on top of the standard representations. In particular, we show that adversarially robust learning is just as easy as standard learning. Specifically, the additional number of samples needed to robustly learn multi-index models when compared to standard learning, does not depend on dimensionality.

859. Improving Graph Neural Networks by Learning Continuous Edge Directions

链接：https://iclr.cc/virtual/2025/poster/28711 abstract： Graph Neural Networks (GNNs) traditionally employ a message-passing mechanism that resembles diffusion over undirected graphs, which often leads to homogenization of node features and reduced discriminative power in tasks such as node classification. Our key insight for addressing this limitation is to assign fuzzy edge directions---that can vary continuously from node i pointing to node j to vice versa---to the edges of a graph so that features can preferentially flow in one direction between nodes to enable long-range information transmission across the graph. We also introduce a novel complex-valued Laplacian for directed graphs with fuzzy edges where the real and imaginary parts represent information flow in opposite directions. Using this Laplacian, we propose a general framework, called Continuous Edge Direction (CoED) GNN, for learning on graphs with fuzzy edges and prove its expressivity limits using a generalization of the Weisfeiler-Leman (WL) graph isomorphism test for directed graphs with fuzzy edges. Our architecture aggregates neighbor features scaled by the learned edge directions and processes the aggregated messages from in-neighbors and out-neighbors separately alongside the self-features of the nodes. Since continuous edge directions are differentiable, they can be learned jointly with the GNN weights via gradient-based optimization. CoED GNN is particularly well-suited for graph ensemble data where the graph structure remains fixed but multiple realizations of node features are available, such as in gene regulatory networks, web connectivity graphs, and power grids. We demonstrate through extensive experiments on both synthetic and real graph ensemble datasets that learning continuous edge directions significantly improves performance both for undirected and directed graphs compared with existing methods.

860. Edge Prompt Tuning for Graph Neural Networks

链接：https://iclr.cc/virtual/2025/poster/30726 abstract： Pre-training powerful Graph Neural Networks (GNNs) with unlabeled graph data in a self-supervised manner has emerged as a prominent technique in recent years. However, inevitable objective gaps often exist between pre-training and downstream tasks. To bridge this gap, graph prompt tuning techniques design and learn graph prompts by manipulating input graphs or reframing downstream tasks as pre-training tasks without fine-tuning the pre-trained GNN models. While recent graph prompt tuning methods have proven effective in adapting pre-trained GNN models for downstream tasks, they overlook the crucial role of edges in graph prompt design, which can significantly affect the quality of graph representations for downstream tasks.In this study, we propose EdgePrompt, a simple yet effective graph prompt tuning

method from the perspective of edges. Unlike previous studies that design prompt vectors on node features, EdgePrompt manipulates input graphs by learning additional prompt vectors for edges and incorporates the edge prompts through message passing in the pre-trained GNN models to better embed graph structural information for downstream tasks. Our method is compatible with prevalent GNN architectures pre-trained under various pre-training strategies and is universal for different downstream tasks.We provide comprehensive theoretical analyses of our method regarding its capability of handling node classification and graph classification as downstream tasks.Extensive experiments on ten graph datasets under four pre-training strategies demonstrate the superiority of our proposed method against six baselines. Our code is available at https://github.com/xbfu/EdgePrompt.

861. T2V-Turbo-v2: Enhancing Video Model Post-Training through Data, Reward, and Conditional Guidance Design

链接：https://iclr.cc/virtual/2025/poster/30564 abstract： In this paper, we focus on enhancing a diffusion-based text-to-video (T2V) model during the post-training phase by distilling a highly capable consistency model from a pretrained T2V model. Our proposed method, T2V-Turbo-v2, introduces a significant advancement by integrating various supervision signals, including high-quality training data, reward model feedback, and conditional guidance, into the consistency distillation process. Through comprehensive ablation studies, we highlight the crucial importance of tailoring datasets to specific learning objectives and the effectiveness of learning from diverse reward models for enhancing both the visual quality and text-video alignment. Additionally, we highlight the vast design space of conditional guidance strategies, which centers on designing an effective energy function to augment the teacher ODE solver. We demonstrate the potential of this approach by extracting motion guidance from the training datasets and incorporating it into the ODE solver, showcasing its effectiveness in improving the motion quality of the generated videos with the improved motion-related metrics from VBench and T2V-CompBench. Empirically, our T2V-Turbo-v2 establishes a new state-of-the-art result on VBench, with a Total score of 85.13, surpassing proprietary systems such as Gen-3 and Kling.

862. Rethinking Spiking Neural Networks from an Ensemble Learning Perspective

链接：https://iclr.cc/virtual/2025/poster/29191 abstract： Spiking neural networks (SNNs) exhibit superior energy efficiency but suffer from limited performance. In this paper, we consider SNNs as ensembles of temporal subnetworks that share architectures and weights, and highlight a crucial issue that affects their performance: excessive differences in initial states (neuronal membrane potentials) across timesteps lead to unstable subnetwork outputs, resulting in degraded performance. To mitigate this, we promote the consistency of the initial membrane potential distribution and output through membrane potential smoothing and temporally adjacent subnetwork guidance, respectively, to improve overall stability and performance. Moreover, membrane potential smoothing facilitates forward propagation of information and backward propagation of gradients, mitigating the notorious temporal gradient vanishing problem. Our method requires only minimal modification of the spiking neurons without adapting the network structure, making our method generalizable and showing consistent performance gains in 1D speech, 2D object, and 3D point cloud recognition tasks. In particular, on the challenging CIFAR10-DVS dataset, we achieved 83.20\% accuracy with only four timesteps. This provides valuable insights into unleashing the potential of SNNs.

863. Regret Bounds for Episodic Risk-Sensitive Linear Quadratic Regulator

链接：https://iclr.cc/virtual/2025/poster/29427 abstract： Risk-sensitive linear quadratic regulator is one of the most fundamental problems in risk-sensitive optimal control. In this paper, we study online adaptive control of risk-sensitive linear quadratic regulator in the finite horizon episodic setting. We propose a simple least-squares greedy algorithm and show that it achieves $\widetilde{\mathcal{O}}(\log N)$ regret under a specific identifiability assumption, where N is the total number of episodes. If the identifiability assumption is not satisfied, we propose incorporating exploration noise into the least-squares-based algorithm, resulting in an algorithm with $\widetilde{\mathcal{O}}(\sqrt{N})$ regret. To our best knowledge, this is the first set of regret bounds for episodic risk-sensitive linear quadratic regulator. Our proof relies on perturbation analysis of less-standard Riccati equations for risk-sensitive linear quadratic control, and a delicate analysis of the loss in the risk-sensitive performance criterion due to applying the suboptimal controller in the online learning process.

864. RefactorBench: Evaluating Stateful Reasoning in Language Agents Through Code

链接：https://iclr.cc/virtual/2025/poster/29864 abstract： Recent advances in language model (LM) agents and function calling have enabled autonomous, feedback-driven systems to solve problems across various digital domains. To better understand the unique limitations of LM agents, we introduce RefactorBench, a benchmark consisting of 100 large handcrafted multi-file refactoring tasks in popular open-source repositories. Solving tasks within RefactorBench requires thorough exploration of dependencies across multiple files and strong adherence to relevant instructions. Every task is defined by 3 natural language instructions of varying specificity and is mutually exclusive, allowing for the creation of longer combined tasks on the same repository. Baselines on RefactorBench reveal that current LM agents struggle with simple compositional tasks, solving only 22\% of tasks with base instructions, in contrast to a human developer with short time constraints solving 87\%. Through trajectory analysis, we identify various unique failure modes of LM agents, and further explore the failure mode of tracking past actions. By adapting a baseline agent to condition on representations of state, we achieve a 43.9\% improvement in solving

RefactorBench tasks. We further extend our state-aware approach to encompass entire digital environments and outline potential directions for future research. RefactorBench aims to support the study of LM agents by providing a set of real-world, multi-hop tasks within the realm of code.

865. MIA-Bench: Towards Better Instruction Following Evaluation of Multimodal LLMs

链接：https://iclr.cc/virtual/2025/poster/30836 abstract： Effective evaluation of Multimodal Large Language Models (MLLMs) is essential for understanding their capabilities and limitations. In this paper, we introduce MIA-Bench, a benchmark designed to assess MLLMs' ability to strictly adhere to complex instructions. Our benchmark comprises a diverse set of 400 image-prompt pairs, each crafted to challenge the models' compliance with layered instructions in generating accurate and contextually appropriate responses. Evaluation results from a wide array of state-of-the-art MLLMs reveal significant variations in performance, highlighting areas for improvement in instruction fidelity. Additionally, we create extra training data and explore supervised fine-tuning and direct preference optimization to enhance the models' ability to strictly follow instructions without compromising performance on other tasks. We hope this benchmark not only serves as a tool for measuring MLLM adherence to instructions, but also guides future developments in MLLM training methods.

866. metabench - A Sparse Benchmark of Reasoning and Knowledge in Large Language Models

链接：https://iclr.cc/virtual/2025/poster/31009 abstract： Large Language Models (LLMs) vary in their abilities on a range of tasks. Initiatives such as the Open LLM Leaderboard aim to quantify these differences with several large benchmarks (sets of test items to which an LLM can respond either correctly or incorrectly). However, high correlations within and between benchmark scores suggest that (1) there exists a small set of common underlying abilities that these benchmarks measure, and (2) items tap into redundant information and the benchmarks may thus be considerably compressed. We use data from n > 5000 LLMs to identify the most informative items of six benchmarks, ARC, GSM8K, HellaSwag, MMLU, TruthfulQA and WinoGrande (with d = 28,632 items in total). From them we distill a sparse benchmark, metabench, that has less than 3% of the original size of all six benchmarks combined. This new sparse benchmark goes beyond point scores by yielding estimators of the underlying benchmark-specific abilities. We show that these estimators (1) can be used to reconstruct each original individual benchmark score with, on average, 1.24% root mean square error (RMSE), (2) reconstruct the original total score with 0.58% RMSE, and (3) have a single underlying common factor whose Spearman correlation with the total score is r = 0.94.

867. Second-Order Fine-Tuning without Pain for LLMs: A Hessian Informed Zeroth-Order Optimizer

链接：https://iclr.cc/virtual/2025/poster/29118 abstract： Fine-tuning large language models (LLMs) is necessary for specific downstream tasks, but classic first-order optimizer entails prohibitive GPU memory because of the back propagation. Recent works such as MeZO have turned to zeroth-order optimizers for fine-tuning, which reduce substantial memory by using two forward passes. However, heterogeneous curvatures across different parameter dimensions in LLMs often cause model convergence instability or even failure. In this work, we propose HiZOO, a diagonal Hessian informed Zeroth-Order Optimizer , which is the first work to leverage the diagonal Hessian to enhance ZOO for fine-tuning LLMs. We provide theoretical proof for HiZOO and visualize the optimization trajectories on test functions to illustrate how it improves convergence in handling heterogeneous curvatures. Extensive experiments on various models (RoBERTa, OPT, Phi-2 and LLama3, with 350M$\sim$66B parameters) indicate that HiZOO significantly reduces training steps and enhances model accuracy, while keeping the memory advantage of ZOO. For example, on SST2 task HiZOO achieves $8\times$ speedup and better accuracy over MeZO across different models. We also propose HiZOO-L, which reduces the Hessian memory cost to 10\% of the MeZO, while maintaining almost same performance. Compared with ZO-Adam, HiZOO-L achieves a 4.3\% improvement, just using 50\% of the GPU memory. Code is available at https://anonymous.4open.science/r/HiZOO-27F8.

868. Bootstrapped Model Predictive Control

链接：https://iclr.cc/virtual/2025/poster/28718 abstract： Model Predictive Control (MPC) has been demonstrated to be effective in continuous control tasks. When a world model and a value function are available, planning a sequence of actions ahead of time leads to a better policy. Existing methods typically obtain the value function and the corresponding policy in a model-free manner. However, we find that such an approach struggles with complex tasks, resulting in poor policy learning and inaccurate value estimation. To address this problem, we leverage the strengths of MPC itself. In this work, we introduce Bootstrapped Model Predictive Control (BMPC), a novel algorithm that performs policy learning in a bootstrapped manner. BMPC learns a network policy by imitating an MPC expert, and in turn, uses this policy to guide the MPC process. Combined with model-based TD-learning, our policy learning yields better value estimation and further boosts the efficiency of MPC. We also introduce a lazy reanalyze mechanism, which enables computationally efficient imitation learning. Our method achieves superior performance over prior works on diverse continuous control tasks. In particular, on challenging high-dimensional locomotion tasks, BMPC significantly improves data efficiency while also enhancing asymptotic performance and training stability, with comparable training time and smaller network sizes. Code is available at https://github.com/wertyuilife2/bmpc.

869. Understanding Model Calibration - A gentle introduction and visual exploration of calibration and the expected calibration error (ECE)

链接：https://iclr.cc/virtual/2025/poster/31360 abstract：

870. MotionAura: Generating High-Quality and Motion Consistent Videos using Discrete Diffusion

链接：https://iclr.cc/virtual/2025/poster/29105 abstract： The spatio-temporal complexity of video data presents significant challenges in tasks such as compression, generation, and inpainting. We present four key contributions to address the challenges of spatiotemporal video processing. First, we introduce the 3D Mobile Inverted Vector-Quantization Variational Autoencoder (3D-MBQ-VAE), which combines Variational Autoencoders (VAEs) with masked modeling to enhance spatiotemporal video compression. The model achieves superior temporal consistency and state-of-the-art (SOTA) reconstruction quality by employing a novel training strategy with full frame masking. Second, we present MotionAura, a text-to-video generation framework that utilizes vector-quantized diffusion models to discretize the latent space and capture complex motion dynamics, producing temporally coherent videos aligned with text prompts. Third, we propose a spectral transformer-based denoising network that processes video data in the frequency domain using the Fourier Transform. This method effectively captures global context and long-range dependencies for high-quality video generation and denoising. Lastly, we introduce a downstream task of Sketch Guided Video Inpainting. This task leverages Low-Rank Adaptation (LoRA) for parameter-efficient fine-tuning. Our models achieve SOTA performance on a range of benchmarks. Our work offers robust frameworks for spatiotemporal modeling and user-driven video content manipulation.

871. Learning vector fields of differential equations on manifolds with geometrically constrained operator-valued kernels

链接：https://iclr.cc/virtual/2025/poster/29790 abstract： We address the problem of learning ordinary differential equations (ODEs) on manifolds. Existing machine learning methods, particularly those using neural networks, often struggle with high computational demands. To overcome this issue, we introduce a geometrically constrained operator-valued kernel that allows us to represent vector fields on tangent bundles of smooth manifolds. The construction of the kernel imposes the geometric constraints that are estimated from the data and ensures the computational feasibility for learning high dimensional systems of ODEs. Once the vector fields are estimated, e.g., by the kernel ridge regression, we need an ODE solver that guarantees the solution to stay on (or close to) the manifold. To overcome this issue, we propose a geometry-preserving ODE solver that approximates the exponential maps corresponding to the ODE solutions. We deduce a theoretical error bound for the proposed solver that guarantees the approximate solutions to lie on the manifold in the limit of large data. We verify the effectiveness of the proposed approach on high-dimensional dynamical systems, including the cavity flow problem, the beating and travelling waves in Kuramoto-Sivashinsky equations, and the reaction-diffusion dynamics.

872. Multi-Modal and Multi-Attribute Generation of Single Cells with CFGen

链接：https://iclr.cc/virtual/2025/poster/31082 abstract： Generative modeling of single-cell RNA-seq data is crucial for tasks like trajectory inference, batch effect removal, and simulation of realistic cellular data. However, recent deep generative models simulating synthetic single cells from noise operate on pre-processed continuous gene expression approximations, overlooking the discrete nature of single-cell data, which limits their effectiveness and hinders the incorporation of robust noise models. Additionally, aspects like controllable multi-modal and multi-label generation of cellular data remain underexplored. This work introduces CellFlow for Generation (CFGen), a flow-based conditional generative model that preserves the inherent discreteness of single-cell data. CFGen generates whole-genome multi-modal single-cell data reliably, improving the recovery of crucial biological data characteristics while tackling relevant generative tasks such as rare cell type augmentation and batch correction. We also introduce a novel framework for compositional data generation using Flow Matching. By showcasing CFGen on a diverse set of biological datasets and settings, we provide evidence of its value to the fields of computational biology and deep generative models.

873. Graph Neural Preconditioners for Iterative Solutions of Sparse Linear Systems

链接：https://iclr.cc/virtual/2025/poster/29521 abstract： Preconditioning is at the heart of iterative solutions of large, sparse linear systems of equations in scientific disciplines. Several algebraic approaches, which access no information beyond the matrix itself, are widely studied and used, but ill-conditioned matrices remain very challenging. We take a machine learning approach and propose using graph neural networks as a general-purpose preconditioner. They show attractive performance for many problems and can be used when the mainstream preconditioners perform poorly. Empirical evaluation on over 800 matrices suggests that the construction time of these graph neural preconditioners (GNPs) is more predictable and can be much shorter than that of other widely used ones, such as ILU and AMG, while the execution time is faster than using a Krylov method as the preconditioner, such as in inner-outer GMRES. GNPs have a strong potential for solving large-scale, challenging algebraic problems arising from not only partial differential equations, but also economics, statistics, graph, and optimization, to

name a few.

874. Steering LLMs' Behavior with Concept Activation Vectors

链接：https://iclr.cc/virtual/2025/poster/37630 abstract： Concept activation vectors have been shown to take effects in safety concepts, efficiently and effectively guiding a considerable number of open-source large language models (LLMs) to respond positively to malicious instructions. In this blog, we aim to explore the capability boundaries of concept activation vectors in guiding various behaviors of LLMs through more extensive experiments. Our experiments demonstrate that this reasoning technique can low-costly transfer text styles and improve performance on specific tasks such as code generation.

875. Distribution-Free Data Uncertainty for Neural Network Regression

链接：https://iclr.cc/virtual/2025/poster/28309 abstract： Quantifying uncertainty is an essential part of predictive modeling, especially in the context of high-stakes decision-making. While classification output includes data uncertainty by design in the form of class probabilities, the regression task generally aims only to predict the expected value of the target variable. Probabilistic extensions often assume parametric distributions around the expected value, optimizing the likelihood over the resulting explicit densities. However, using parametric distributions can limit practical applicability, making it difficult for models to capture skewed, multi-modal, or otherwise complex distributions. In this paper, we propose optimizing a novel nondeterministic neural network regression architecture for loss functions derived from a sample-based approximation of the continuous ranked probability score (CRPS), enabling a truly distribution-free approach by learning to sample from the target's aleatoric distribution, rather than predicting explicit densities. Our approach allows the model to learn well-calibrated, arbitrary uni- and multivariate output distributions. We evaluate the method on a variety of synthetic and real-world tasks, including uni- and multivariate problems, function inverse approximation, and standard regression uncertainty benchmarks. Finally, we make all experiment code publicly available.

876. Revolutionizing EMCCD Denoising through a Novel Physics-Based Learning Framework for Noise Modeling

链接：https://iclr.cc/virtual/2025/poster/27890 abstract： Electron-multiplying charge-coupled device (EMCCD) has been instrumental in sensitive observations under low-light situations including astronomy, material science, and biology. Despite its ingenious designs to enhance target signals overcoming read-out circuit noises, produced images are not completely noise free, which could still cast a cloud on desired experiment outcomes, especially in fluorescence microscopy.Existing studies on EMCCD's noise model have been focusing on statistical characteristics in theory, yet unable to incorporate latest advancements in the field of computational photography, where physics-based noise models are utilized to guide deep learning processes, creating adaptive denoising algorithms for ordinary image sensors.Still, those models are not directly applicable to EMCCD.In this paper, we intend to pioneer EMCCD denoising by introducing a systematic study on physics-based noise model calibration procedures for an EMCCD camera, accurately estimating statistical features of observable noise components in experiments, which are then utilized to generate substantial amount of authentic training samples for one of the most recent neural networks.A first real-world test image dataset for EMCCD is captured, containing both images of ordinary daily scenes and those of microscopic contents.Benchmarking upon the testset and authentic microscopic images, we demonstrate distinct advantages of our model against previous methods for EMCCD and physics-based noise modeling, forging a promising new path for EMCCD denoising.

877. Temporal Heterogeneous Graph Generation with Privacy, Utility, and Efficiency

链接：https://iclr.cc/virtual/2025/poster/28030 abstract： Nowadays, temporal heterogeneous graphs attract much research and industrial attention for building the next-generation Relational Deep Learning models and applications, due to their informative structures and features. While providing timely and precise services like personalized recommendations and question answering, this rich information also introduces extra exposure risk for each node in the graph. The distinctive local topology, the abundant heterogeneous features, and the time dimension of the graph data are more prone to expose sensitive information and narrow down the scope of victim candidates, which calls for well-defined protection techniques on graphs. To this end, we propose a Temporal Heterogeneous Graph Generator balancing Privacy, Utility, and Efficiency, named THePUff. More specifically, we first propose a differential privacy algorithm to perturb the input temporal heterogeneous graph for protecting privacy, and then utilize both the perturbed graph and the original one in a generative adversarial setting for THePUff to learn and generate privacy-guaranteed and utility-preserved graph data in an efficient manner. We further propose 6 new metrics in the temporal setting to measure heterogeneous graph utility and privacy. Finally, based on temporal heterogeneous graph datasets with up to 1 million nodes and 20 million edges, the experiments show that THePUff generates utilizable temporal heterogeneous graphs with privacy protected, compared with state-of-the-art baselines.

878. Memory Efficient Transformer Adapter for Dense Predictions

链接：https://iclr.cc/virtual/2025/poster/27922 abstract： While current Vision Transformer (ViT) adapter methods have shown promising accuracy, their inference speed is implicitly hindered by inefficient memory access operations, e.g., standard

normalization and frequent reshaping. In this work, we propose META, a simple and fast ViT adapter that can improve the model's memory efficiency and decrease memory time consumption by reducing the inefficient memory access operations. Our method features a memory-efficient adapter block that enables the common sharing of layer normalization between the self-attention and feed-forward network layers, thereby reducing the model's reliance on normalization operations. Within the proposed block, the cross-shaped self-attention is employed to reduce the model's frequent reshaping operations. Moreover, we augment the adapter block with a lightweight convolutional branch that can enhance local inductive biases, particularly beneficial for the dense prediction tasks, e.g., object detection, instance segmentation, and semantic segmentation. The adapter block is finally formulated in a cascaded manner to compute diverse head features, thereby enriching the variety of feature representations. Empirically, extensive evaluations on multiple representative datasets validate that META substantially enhances the predicted quality, while achieving a new state-of-the-art accuracy-efficiency trade-off. Theoretically, we demonstrate that META exhibits superior generalization capability and stronger adaptability.

879. Learning Robust Representations with Long-Term Information for Generalization in Visual Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/29765 abstract： Generalization in visual reinforcement learning (VRL) aims to learn agents that can adapt to test environments with unseen visual distractions. Despite advances in robust representations learning, many methods do not take into account the essential downstream task of sequential decision-making. This leads to representations that lack critical long-term information, impairing decision-making abilities in test environments. To tackle this problem, we propose a novel robust action-value representation learning (ROUSER) under the information bottleneck (IB) framework. ROUSER learns robust representations to capture long-term information from the decision-making objective (i.e., action values). Specifically, ROUSER uses IB to encode robust representations by maximizing their mutual information with action values for long-term information, while minimizing mutual information with state-action pairs to discard irrelevant features. As action values are unknown, ROUSER proposes to decompose robust representations of state-action pairs into one-step rewards and robust representations of subsequent pairs. Thus, it can use known rewards to compute the loss for robust representation learning. Moreover, we show that ROUSER accurately estimates action values using learned robust representations, making it applicable to various VRL algorithms. Experiments demonstrate that ROUSER outperforms several state-of-the-art methods in eleven out of twelve tasks, across both unseen background and color distractions.

880. MOCA: Self-supervised Representation Learning by Predicting Masked Online Codebook Assignments

链接：https://iclr.cc/virtual/2025/poster/31503 abstract：

881. Unlocking the Potential of Model Calibration in Federated Learning

链接：https://iclr.cc/virtual/2025/poster/29797 abstract：

882. Emerging Safety Attack and Defense in Federated Instruction Tuning of Large Language Models

链接：https://iclr.cc/virtual/2025/poster/28122 abstract：

883. GraphEval: A Lightweight Graph-Based LLM Framework for Idea Evaluation

链接：https://iclr.cc/virtual/2025/poster/30947 abstract：

884. DiffPuter: Empowering Diffusion Models for Missing Data Imputation

链接：https://iclr.cc/virtual/2025/poster/31061 abstract： Generative models play an important role in missing data imputation in that they aim to learn the joint distribution of full data. However, applying advanced deep generative models (such as Diffusion models) to missing data imputation is challenging due to 1) the inherent incompleteness of the training data and 2) the difficulty in performing conditional inference from unconditional generative models. To deal with these challenges, this paper introduces DiffPuter, a tailored diffusion model combined with the Expectation-Maximization (EM) algorithm for missing data imputation. DiffPuter iteratively trains a diffusion model to learn the joint distribution of missing and observed data and performs an accurate conditional sampling to update the missing values using a tailored reversed sampling strategy. Our theoretical analysis shows that DiffPuter's training step corresponds to the maximum likelihood estimation of data density (M-step), and its sampling step represents the Expected A Posteriori estimation of missing values (E-step). Extensive experiments across ten diverse datasets and comparisons with 17 different imputation methods demonstrate DiffPuter's superior performance. Notably, DiffPuter achieves an average improvement of 8.10\% in MAE and 5.64\% in RMSE compared to the most competitive existing method.

885. TabDiff: a Mixed-type Diffusion Model for Tabular Data Generation

链接：https://iclr.cc/virtual/2025/poster/28087 abstract： Synthesizing high-quality tabular data is an important topic in many data science tasks, ranging from dataset augmentation to privacy protection. However, developing expressive generative models for tabular data is challenging due to its inherent heterogeneous data types, complex inter-correlations, and intricate column-wise distributions. In this paper, we introduce TabDiff, a joint diffusion framework that models all mixed-type distributions of tabular data in one model. Our key innovation is the development of a joint continuous-time diffusion process for numerical and categorical data, where we propose feature-wise learnable diffusion processes to counter the high disparity of different feature distributions. TabDiff is parameterized by a transformer handling different input types, and the entire framework can be efficiently optimized in an end-to-end fashion. We further introduce a mixed-type stochastic sampler to automatically correct the accumulated decoding error during sampling, and propose classifier-free guidance for conditional missing column value imputation. Comprehensive experiments on seven datasets demonstrate that TabDiff achieves superior average performance over existing competitive baselines across all eight metrics, with up to 22.5% improvement over the state-of-the-art model on pair-wise column correlation estimations. Code is available at https://github.com/MinkaiXu/TabDiff.

886. Towards Faster Decentralized Stochastic Optimization with Communication Compression

链接：https://iclr.cc/virtual/2025/poster/30517 abstract： Communication efficiency has garnered significant attention as it is considered the main bottleneck for large-scale decentralized Machine Learning applications in distributed and federated settings. In this regime, clients are restricted to transmitting small amounts of compressed information to their neighbors over a communication graph. Numerous endeavors have been made to address this challenging problem by developing algorithms with compressed communication for decentralized non-convex optimization problems. Despite considerable efforts, current theoretical understandings of the problem are still very limited, and existing algorithms all suffer from various limitations. In particular, these algorithms typically rely on strong, and often infeasible assumptions such as bounded data heterogeneity or require large batch access while failing to achieve linear speedup with the number of clients. In this paper, we introduce MoTEF, a novel approach that integrates communication compression with $\textbf{Mo}$mentum $\textbf{T}$racking and $\textbf{E}$rror $\textbf{F}$eedback. MoTEF is the first algorithm to achieve an asymptotic rate matching that of distributed SGD under arbitrary data heterogeneity, hence resolving a long-standing theoretical obstacle in decentralized optimization with compressed communication. We provide numerical experiments to validate our theoretical findings and confirm the practical superiority of MoTEF.

887. MetaMetrics: Calibrating Metrics for Generation Tasks Using Human Preferences

链接：https://iclr.cc/virtual/2025/poster/28099 abstract： Understanding the quality of a performance evaluation metric is crucial for ensuring that model outputs align with human preferences. However, it remains unclear how well each metric captures the diverse aspects of these preferences, as metrics often excel in one particular area but not across all dimensions. To address this, it is essential to systematically calibrate metrics to specific aspects of human preference, catering to the unique characteristics of each aspect. We introduce MetaMetrics, a calibrated meta-metric designed to evaluate generation tasks across different modalities in a supervised manner. MetaMetrics optimizes the combination of existing metrics to enhance their alignment with human preferences. Our metric demonstrates flexibility and effectiveness in both language and vision downstream tasks, showing significant benefits across various multilingual and multi-domain scenarios. MetaMetrics aligns closely with human preferences and is highly extendable and easily integrable into any application. This makes MetaMetrics a powerful tool for improving the evaluation of generation tasks, ensuring that metrics are more representative of human judgment across diverse contexts.

888. Holistically Evaluating the Environmental Impact of Creating Language Models

链接：https://iclr.cc/virtual/2025/poster/31271 abstract： As the performance of artificial intelligence systems has dramatically increased, so too has the environmental impact of creating these systems. While many model developers release estimates of the power consumption and carbon emissions from the final training runs for their latest models, there is comparatively little transparency into the impact of model development, hardware manufacturing, and total water usage throughout. In this work, we estimate the real-world environmental impact of developing a series of language models, ranging from 20 million to 13 billion active parameters, trained on up to 5.6 trillion tokens each. When accounting for hardware manufacturing, model development, and our final training runs, we find that our series of models released 493 metric tons of carbon emissions, equivalent to powering about 98 homes in the United States for one year, and consumed 2.769 million liters of water, equivalent to about 24.5 years of water usage by a person in the United States, even though our data center is extremely water-efficient. We measure and report the environmental impact of our model development; to the best of our knowledge we are the first to do so for LLMs, and we find that model development, the impact of which is generally not disclosed by most model developers, amounted to ~50% of that of training. By looking at detailed time series data for power consumption, we also find that power usage throughout training is not consistent, fluctuating between ~15% and ~85% of our hardware's maximum power draw, with negative implications for grid-scale planning as demand continues to grow. We close with a discussion on the continued difficulty of

estimating the environmental impact of AI systems, and key takeaways for model developers and the public at large.

889. Generating Likely Counterfactuals Using Sum-Product Networks

链接：https://iclr.cc/virtual/2025/poster/28206 abstract： The need to explain decisions made by AI systems is driven by both recent regulation and user demand. The decisions are often explainable only post hoc. In counterfactual explanations, one may ask what constitutes the best counterfactual explanation. Clearly, multiple criteria must be taken into account, although "distance from the sample" is a key criterion. Recent methods that consider the plausibility of a counterfactual seem to sacrifice this original objective. Here, we present a system that provides high-likelihood explanations that are, at the same time, close and sparse. We show that the search for the most likely explanations satisfying many common desiderata for counterfactual explanations can be modeled using Mixed-Integer Optimization (MIO). We use a Sum-Product Network (SPN) to estimate the likelihood of a counterfactual. To achieve that, we propose an MIO formulation of an SPN, which can be of independent interest. The source code with examples is available at https://github.com/Epanemu/LiCE.

890. A Sanity Check for AI-generated Image Detection

链接：https://iclr.cc/virtual/2025/poster/29835 abstract： With the rapid development of generative models, discerning AI-generated content has evoked increasing attention from both industry and academia. In this paper, we conduct a sanity check on whether the task of AI-generated image detection has been solved. To start with, we present Chameleon dataset, consisting of AI-generated images that are genuinely challenging for human perception. To quantify the generalization of existing methods, we evaluate 9 off-the-shelf AI-generated image detectors on Chameleon dataset. Upon analysis, almost all models misclassify AI-generated images as real ones. Later, we propose AIDE AI-generated Image DEtector with Hybrid Features, which leverages multiple experts to simultaneously extract visual artifacts and noise patterns. Specifically, to capture the high-level semantics, we utilize CLIP to compute the visual embedding. This effectively enables the model to discern AI-generated images based on semantics and contextual information. Secondly, we select the highest and lowest frequency patches in the image, and compute the low-level patchwise features, aiming to detect AI-generated images by low-level artifacts, for example, noise patterns, anti-aliasing effects. While evaluating on existing benchmarks, for example, AIGCDetectBenchmark and GenImage, AIDE achieves +3.5% and +4.6% improvements to state-of-the-art methods, and on our proposed challenging Chameleon benchmarks, it also achieves promising results, despite the problem of detecting AI-generated images remains far from being solved.

891. Conformalized Survival Analysis for General Right-Censored Data

链接：https://iclr.cc/virtual/2025/poster/30112 abstract： We develop a framework to quantify predictive uncertainty in survival analysis, providing a reliable lower predictive bound (LPB) for the true, unknown patient survival time. Recently, conformal prediction has been used to construct such valid LPBs for type-I right-censored data, with the guarantee that the bound holds with high probability. Crucially, under the type-I setting, the censoring time is observed for all data points. As such, informative LPBs can be constructed by framing the calibration as an estimation task with covariate shift, relying on the conditionally independent censoring assumption. This paper expands the conformal toolbox for survival analysis, with the goal of handling the ubiquitous general right-censored setting, in which either the censoring or survival time is observed, but not both. The key challenge here is that the calibration cannot be directly formulated as a covariate shift problem anymore. Yet, we show how to construct LPBs with distribution-free finite-sample guarantees, under the same assumptions as conformal approaches for type-I censored data. Experiments demonstrate the informativeness and validity of our methods in simulated settings and showcase their practical utility using several real-world datasets.

892. Sensitivity-Constrained Fourier Neural Operators for Forward and Inverse Problems in Parametric Differential Equations

链接：https://iclr.cc/virtual/2025/poster/30461 abstract： Parametric differential equations of the form $\frac{\partial u}{\partial t} = f(u, x, t, p)$ are fundamental in science and engineering. While deep learning frameworks like the Fourier Neural Operator (FNO) efficiently approximate differential equation solutions, they struggle with inverse problems, sensitivity calculations $\frac{\partial u}{\partial p}$, and concept drift. We address these challenges by introducing a novel sensitivity loss regularizer, demonstrated through Sensitivity-Constrained Fourier Neural Operators (SC-FNO). Our approach maintains high accuracy for solution paths and outperforms both standard FNO and FNO with Physics-Informed Neural Network regularization. SC-FNO exhibits superior performance in parameter inversion tasks, accommodates more complex parameter spaces (tested with up to 82 parameters), reduces training data requirements, and decreases training time while maintaining accuracy. These improvements apply across various differential equations and neural operators, enhancing their reliability without significant computational overhead (30%–130% extra training time per epoch). Models and selected experiment code are available at: https://github.com/AMBehroozi/SC_Neural_Operators.

893. Do I Know This Entity? Knowledge Awareness and Hallucinations in Language Models

链接：https://iclr.cc/virtual/2025/poster/29377 abstract： Hallucinations in large language models are a widespread problem, yet the mechanisms behind whether models will hallucinate are poorly understood, limiting our ability to solve this problem. Using

sparse autoencoders as an interpretability tool, we discover that a key part of these mechanisms is entity recognition, where the model detects if an entity is one it can recall facts about. Sparse autoencoders uncover meaningful directions in the representation space, these detect whether the model recognizes an entity, e.g. detecting it doesn't know about an athlete or a movie. This shows that models can have self-knowledge: internal representations about their own capabilities. These directions are causally relevant: capable of steering the model to refuse to answer questions about known entities, or to hallucinate attributes of unknown entities when it would otherwise refuse. We demonstrate that despite the sparse autoencoders being trained on the base model, these directions have a causal effect on the chat model's refusal behavior, suggesting that chat finetuning has repurposed this existing mechanism. Furthermore, we provide an initial exploration into the mechanistic role of these directions in the model, finding that they disrupt the attention of downstream heads that typically move entity attributes to the final token.

894. See What You Are Told: Visual Attention Sink in Large Multimodal Models

链接：https://iclr.cc/virtual/2025/poster/30795 abstract： Large multimodal models (LMMs) "see" images by leveraging the attention mechanism between text and visual tokens in the transformer decoder. Ideally, these models should focus on key visual information relevant to the text token. However, recent findings indicate that LMMs have an extraordinary tendency to consistently allocate high attention weights to specific visual tokens, even when these tokens are irrelevant to the corresponding text. In this study, we investigate the property behind the appearance of these irrelevant visual tokens and examine their characteristics. Our findings show that this behavior arises due to the massive activation of certain hidden state dimensions, which resembles the attention sink found in language models. Hence, we refer to this phenomenon as the visual attention sink. In particular, our analysis reveals that removing the irrelevant visual sink tokens does not impact model performance, despite receiving high attention weights. Consequently, we recycle the attention to these tokens as surplus resources, redistributing the attention budget to enhance focus on the image. To achieve this, we introduce Visual Attention Redistribution (VAR), a method that redistributes attention in image-centric heads, which we identify as innately focusing on visual information. VAR can be seamlessly applied across different LMMs to improve performance on a wide range of tasks, including general vision-language tasks, visual hallucination tasks, and vision-centric tasks, all without the need for additional training, models, or inference steps. Experimental results demonstrate that VAR enables LMMs to process visual information more effectively by adjusting their internal attention mechanisms, offering a new direction to enhancing the multimodal capabilities of LMMs.

895. MeToken: Uniform Micro-environment Token Boosts Post-Translational Modification Prediction

链接：https://iclr.cc/virtual/2025/poster/28395 abstract： Post-translational modifications (PTMs) profoundly expand the complexity and functionality of the proteome, regulating protein attributes and interactions that are crucial for biological processes. Accurately predicting PTM sites and their specific types is therefore essential for elucidating protein function and understanding disease mechanisms. Existing computational approaches predominantly focus on protein sequences to predict PTM sites, driven by the recognition of sequence-dependent motifs. However, these approaches often overlook protein structural contexts. In this work, we first compile a large-scale sequence-structure PTM dataset, which serves as the foundation for fair comparison. We introduce the MeToken model, which tokenizes the micro-environment of each amino acid, integrating both sequence and structural information into unified discrete tokens. This model not only captures the typical sequence motifs associated with PTMs but also leverages the spatial arrangements dictated by protein tertiary structures, thus providing a holistic view of the factors influencing PTM sites. Designed to address the long-tail distribution of PTM types, MeToken employs uniform sub-codebooks that ensure even the rarest PTMs are adequately represented and distinguished. We validate the effectiveness and generalizability of MeToken across multiple datasets, demonstrating its superior performance in accurately identifying PTM types. The results underscore the importance of incorporating structural data and highlight MeToken's potential in facilitating accurate and comprehensive PTM predictions, which could significantly impact proteomics research.

896. The Foundations of Tokenization: Statistical and Computational Concerns

链接：https://iclr.cc/virtual/2025/poster/30592 abstract： Tokenization — the practice of converting strings of characters from an alphabet into sequences of tokens over a vocabulary — is a critical step in the NLP pipeline. The use of token representations is widely credited with increased model performance but is also the source of many undesirable behaviors, such as spurious ambiguity or inconsistency. Despite its recognized importance as a standard representation method in NLP, the theoretical underpinnings of tokenization are not yet fully understood. In particular, the impact of tokenization on language model estimation has been investigated primarily through empirical means. The present paper contributes to addressing this theoretical gap by proposing a unified formal framework for representing and analyzing tokenizer models. Based on the category of stochastic maps, this framework enables us to establish general conditions for a principled use of tokenizers and, most importantly, the necessary and sufficient conditions for a tokenizer model to preserve the consistency of statistical estimators. In addition, we discuss statistical and computational concerns crucial for designing and implementing tokenizer models, such as inconsistency, ambiguity, finiteness, and sequentiality. The framework and results advanced in this paper contribute to building robust theoretical foundations for representations in neural language modeling that can inform future theoretical and empirical research.

897. To Code or Not To Code? Exploring Impact of Code in Pre-training

链接：https://iclr.cc/virtual/2025/poster/27667 abstract： Including code in the pre-training data mixture, even for models not specifically designed for code, has become a common practice in LLMs pre-training. While there has been anecdotal consensus among practitioners that code data plays a vital role in general LLMs' performance, there is only limited work analyzing the precise impact of code on non-code tasks. In this work, we systematically investigate the impact of code data on general performance. We ask “what is the impact of code data used in pre-training on a large variety of downstream tasks beyond code generation”. We conduct extensive ablations and evaluate across a broad range of natural language reasoning tasks, world knowledge tasks, code benchmarks, and LLM-as-a-judge win-rates for models with sizes ranging from 470M to 2.8B parameters. Across settings, we find a consistent results that code is a critical building block for generalization far beyond coding tasks and improvements to code quality have an outsized impact across all tasks. In particular, compared to text-only pre-training, the addition of code results in up to relative increase of 8.2% in natural language (NL) reasoning, 4.2% in world knowledge, 6.6% improvement in generative win-rates, and a 12x boost in code performance respectively. Our work suggests investments in code quality and preserving code during pre-training have positive impacts.

898. Make Haste Slowly: A Theory of Emergent Structured Mixed Selectivity in Feature Learning ReLU Networks

链接：https://iclr.cc/virtual/2025/poster/31160 abstract： In spite of finite dimension ReLU neural networks being a consistent factor behind recent deep learning successes, a theory of feature learning in these models remains elusive. Currently, insightful theories still rely on assumptions including the linearity of the network computations, unstructured input data and architectural constraints such as infinite width or a single hidden layer. To begin to address this gap we establish an equivalence between ReLU networks and Gated Deep Linear Networks, and use their greater tractability to derive dynamics of learning. We then consider multiple variants of a core task reminiscent of multi-task learning or contextual control which requires both feature learning and nonlinearity. We make explicit that, for these tasks, the ReLU networks possess an inductive bias towards latent representations which are not strictly modular or disentangled but are still highly structured and reusable between contexts. This effect is amplified with the addition of more contexts and hidden layers. Thus, we take a step towards a theory of feature learning in finite ReLU networks and shed light on how structured mixed-selective latent representations can emerge due to a bias for node-reuse and learning speed.

899. MathGAP: Out-of-Distribution Evaluation on Problems with Arbitrarily Complex Proofs

链接：https://iclr.cc/virtual/2025/poster/30934 abstract： Large language models (LLMs) can solve arithmetic word problems with high accuracy, but little is known about how well they generalize to more complex problems. This is difficult to study, as (i) much of the available evaluation data has already been seen by the most capable models during training, and (ii) existing benchmarks do not capture how problem proofs may be arbitrarily complex in various ways. In this paper, we present a data-generation framework for evaluating LLMs on problems with arbitrarily complex arithmetic proofs, called MathGAP. MathGAP generates problem statements and chain-of-thought reasoning traces according to specifications about their arithmetic proof structure, enabling systematic studies on easy-to-hard generalization with respect to complexity of proof trees. Using MathGAP, we find that LLMs show a significant decrease in performance as proofs get deeper and wider. This effect is more pronounced in complex, nonlinear proof structures, which are challenging even for the most capable models. The models are also sensitive to simple changes in sentence ordering. However, they remain capable of solving some complex problems, suggesting that reasoning generalization is noisy.

900. POTEC: Off-Policy Contextual Bandits for Large Action Spaces via Policy Decomposition

链接：https://iclr.cc/virtual/2025/poster/29991 abstract： We study off-policy learning (OPL) of contextual bandit policies in large discrete action spaces where existing methods -- most of which rely crucially on reward-regression models or importance-weighted policy gradients -- fail due to excessive bias or variance. To overcome these issues in OPL, we propose a novel two-stage algorithm, called Policy Optimization via Two-Stage Policy Decomposition (POTEC). It leverages clustering in the action space and learns two different policies via policy- and regression-based approaches, respectively. In particular, we derive a novel low-variance gradient estimator that enables to learn a first-stage policy for cluster selection efficiently via a policy-based approach. To select a specific action within the cluster sampled by the first-stage policy, POTEC uses a second-stage policy derived from a regression-based approach within each cluster. We show that a local correctness condition, which only requires that the regression model preserves the relative expected reward differences of the actions within each cluster, ensures that our policy-gradient estimator is unbiased and the second-stage policy is optimal. We also show that POTEC provides a strict generalization of policy- and regression-based approaches and their associated assumptions. Comprehensive experiments demonstrate that POTEC provides substantial improvements in OPL effectiveness particularly in large and structured action spaces.

901. A General Framework for Off-Policy Learning with Partially-Observed

Reward

链接：https://iclr.cc/virtual/2025/poster/28461 abstract： Off-policy learning (OPL) in contextual bandits aims to learn a decision-making policy that maximizes the target rewards by using only historical interaction data collected under previously developed policies. Unfortunately, when rewards are only partially observed, the effectiveness of OPL degrades severely. Well-known examples of such partial rewards include explicit ratings in content recommendations, conversion signals on e-commerce platforms that are partial due to delay, and the issue of censoring in medical problems. One possible solution to deal with such partial rewards is to use secondary rewards, such as dwelling time, clicks, and medical indicators, which are more densely observed. However, relying solely on such secondary rewards can also lead to poor policy learning since they may not align with the target reward. Thus, this work studies a new and general problem of OPL where the goal is to learn a policy that maximizes the expected target reward by leveraging densely observed secondary rewards as supplemental data. We then propose a new method called Hybrid Policy Optimization for Partially-Observed Reward (HyPeR), which effectively uses the secondary rewards in addition to the partially observed target reward to achieve effective OPL despite the challenging scenario. We also discuss a case where we aim to optimize not only the expected target reward but also the expected secondary rewards to some extent; counter-intuitively, we will show that leveraging the two objectives is in fact advantageous also for the optimization of only the target reward. Along with statistical analysis of our proposed methods, empirical evaluations on both synthetic and real-world data show that HyPeR outperforms existing methods in various scenarios.

902. SeRA: Self-Reviewing and Alignment of LLMs using Implicit Reward Margins

链接：https://iclr.cc/virtual/2025/poster/27987 abstract： Direct alignment algorithms (DAAs), such as direct preference optimization (DPO), have become popular alternatives to Reinforcement Learning from Human Feedback (RLHF) due to their simplicity, efficiency, and stability. However, the preferences used by DAAs are usually collected before alignment training begins and remain unchanged (off-policy). This design leads to two problems where the policy model (1) picks up on spurious correlations in the dataset (as opposed to only learning alignment to human preferences), and (2) overfits to feedback on off-policy trajectories that have less likelihood of being generated by the updated policy model. To address these issues, we introduce Self-Reviewing and Alignment (SeRA), a cost-efficient and effective method that can be readily combined with existing DAAs. SeRA comprises of two components: (1) sample selection using implicit reward margin to alleviate over-optimization on such undesired features, and (2) preference bootstrapping using implicit rewards to augment preference data with updated policy models in a cost-efficient manner. Extensive experiments, including on instruction-following tasks, demonstrate the effectiveness and generality of SeRA in training LLMs with diverse offline preference datasets and and DAAs.

903. PolyNet: Learning Diverse Solution Strategies for Neural Combinatorial Optimization

链接：https://iclr.cc/virtual/2025/poster/29548 abstract： Reinforcement learning-based methods for constructing solutions to combinatorial optimization problems are rapidly approaching the performance of human-designed algorithms. To further narrow the gap, learning-based approaches must efficiently explore the solution space during the search process. Recent approaches artificially increase exploration by enforcing diverse solution generation through handcrafted rules, however, these rules can impair solution quality and are difficult to design for more complex problems. In this paper, we introduce PolyNet, an approach for improving exploration of the solution space by learning complementary solution strategies. In contrast to other works, PolyNet uses only a single-decoder and a training schema that does not enforce diverse solution generation through handcrafted rules. We evaluate PolyNet on four combinatorial optimization problems and observe that the implicit diversity mechanism allows PolyNet to find better solutions than approaches that explicitly enforce diverse solution generation.

904. Streamlining Prediction in Bayesian Deep Learning

链接：https://iclr.cc/virtual/2025/poster/28294 abstract： The rising interest in Bayesian deep learning (BDL) has led to a plethora of methods for estimating the posterior distribution. However, efficient computation of inferences, such as predictions, has been largely overlooked with Monte Carlo integration remaining the standard. In this work we examine streamlining prediction in BDL through a single forward pass without sampling. For this, we use local linearisation of activation functions and local Gaussian approximations at linear layers. Thus allowing us to analytically compute an approximation of the posterior predictive distribution. We showcase our approach for both MLP and transformers, such as ViT and GPT-2, and assess its performance on regression and classification tasks.Open-source library: https://github.com/AaltoML/SUQ.

905. Learning 3D Perception from Others' Predictions

链接：https://iclr.cc/virtual/2025/poster/29248 abstract： Accurate 3D object detection in real-world environments requires a huge amount of annotated data with high quality. Acquiring such data is tedious and expensive, and often needs repeated effort when a new sensor is adopted or when the detector is deployed in a new environment. We investigate a new scenario to construct 3D object detectors: learning from the predictions of a nearby unit that is equipped with an accurate detector. For example, when a self-driving car enters a new area, it may learn from other traffic participants whose detectors have been optimized for that area. This setting is label-efficient, sensor-agnostic, and communication-efficient: nearby units only need to

share the predictions with the ego agent (e.g., car). Naively using the received predictions as ground-truths to train the detector for the ego car, however, leads to inferior performance. We systematically study the problem and identify viewpoint mismatches and mislocalization (due to synchronization and GPS errors) as the main causes, which unavoidably result in false positives, false negatives, and inaccurate pseudo labels. We propose a distance-based curriculum, first learning from closer units with similar viewpoints and subsequently improving the quality of other units' predictions via self-training. We further demonstrate that an effective pseudo label refinement module can be trained with a handful of annotated data, largely reducing the data quantity necessary to train an object detector. We validate our approach on the recently released real-world collaborative driving dataset, using reference cars' predictions as pseudo labels for the ego car. Extensive experiments including several scenarios (e.g., different sensors, detectors, and domains) demonstrate the effectiveness of our approach toward label-efficient learning of 3D perception from other units' predictions.

906. Graph Sparsification via Mixture of Graphs

链接：https://iclr.cc/virtual/2025/poster/30843 abstract： Graph Neural Networks (GNNs) have demonstrated superior performance across various graph learning tasks but face significant computational challenges when applied to large-scale graphs. One effective approach to mitigate these challenges is graph sparsification, which involves removing non-essential edges to reduce computational overhead. However, previous graph sparsification methods often rely on a single global sparsity setting and uniform pruning criteria, failing to provide customized sparsification schemes for each node's complex local context.In this paper, we introduce Mixture-of-Graphs (MoG), leveraging the concept of Mixture-of-Experts (MoE), to dynamically select tailored pruning solutions for each node. Specifically, MoG incorporates multiple sparsifier experts, each characterized by unique sparsity levels and pruning criteria, and selects the appropriate experts for each node. Subsequently, MoG performs a mixture of the sparse graphs produced by different experts on the Grassmann manifold to derive an optimal sparse graph. One notable property of MoG is its entirely local nature, as it depends on the specific circumstances of each individual node. Extensive experiments on four large-scale OGB datasets and two superpixel datasets, equipped with five GNN backbones, demonstrate that MoG (I) identifies subgraphs at higher sparsity levels ($8.67\%\sim 50.85\%$), with performance equal to or better than the dense graph, (II) achieves $1.47-2.62\times$ speedup in GNN inference with negligible performance drop, and (III) boosts ``top-student'' GNN performance ($1.02\%\uparrow$ on RevGNN+\textsc{ogbn-proteins} and $1.74\%\uparrow$ on DeeperGCN+\textsc{ogbg-ppa}). The source code is available at \url{https://github.com/yanweiyue/MoG}.

907. Rethinking Graph Prompts: Unraveling the Power of Data Manipulation in Graph Neural Networks

链接：https://iclr.cc/virtual/2025/poster/31337 abstract： Graph Neural Networks (GNNs) have transformed graph learning but face challenges like distribution shifts, data anomalies, and adversarial vulnerabilities. Graph prompt emerges as a novel solution, enabling data transformation to align graph data with pre-trained models without altering model parameters. This paradigm addresses negative transfer, enhances adaptability, and bridges modality gaps. Unlike traditional fine-tuning, graph prompts rewrite graph structures and features through components like prompt tokens and insertion patterns, improving flexibility and efficiency. Applications in IoT, drug discovery, fraud detection, and personalized learning demonstrate their potential to dynamically adapt graph data. While promising, challenges such as optimal design, benchmarks, and gradient issues persist. Addressing these will unlock full potential of graph prompt to advance GNNs for complex real-world tasks.

908. Training Language Models to Self-Correct via Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/30501 abstract： Self-correction is a highly desirable capability of large language models (LLMs), yet it has consistently been found to be largely ineffective in modern LLMs. Current methods for training self-correction typically depend on either multiple models, a more advanced model, or additional forms of supervision. To address these shortcomings, we develop a multi-turn online reinforcement learning (RL) approach, SCoRe, that significantly improves an LLM's self-correction ability using entirely self-generated data. To build SCoRe, we first show that variants of supervised fine-tuning (SFT) on offline model-generated correction traces are insufficient for instilling self-correction behavior. In particular, we observe that training via SFT either suffers from a distribution mismatch between the training data and the model's own responses or implicitly prefers only a certain mode of correction behavior that is often not effective at test time. SCoRe addresses these challenges by training under the model's own distribution of self-generated correction traces and using appropriate regularization to steer the learning process into learning a self-correction strategy that is effective at test time as opposed to simply fitting high-reward responses for a given prompt. This regularization prescribes running a first phase of RL on a base model to generate a policy initialization that is less susceptible to collapse and then using a reward bonus to amplify self-correction during training. When applied to Gemini 1.0 Pro and 1.5 Flash models, we find that SCoRe achieves state-of-the-art self-correction performance, improving the base models' self-correction by 15.6% and 9.1% respectively on the MATH and HumanEval benchmarks.

909. Joint Fine-tuning and Conversion of Pretrained Speech and Language Models towards Linear Complexity

链接：https://iclr.cc/virtual/2025/poster/30727 abstract： Architectures such as Linformer and Mamba have recently emerged as competitive linear time replacements for transformers. However, corresponding large pretrained models are often unavailable, especially in non-text domains. To remedy this, we present a Cross-Architecture Layerwise Distillation (CALD)

approach that jointly converts a transformer model to a linear time substitute and fine-tunes it to a target task. We also compare several means to guide the fine-tuning to optimally retain the desired inference capability from the original model. The methods differ in their use of the target model and the trajectory of the parameters. In a series of empirical studies on language processing, language modeling, and speech processing, we show that CALD can effectively recover the result of the original model, and that the guiding strategy contributes to the result. Some reasons for the variation are suggested.

910. Generalization Guarantees for Representation Learning via Data-Dependent Gaussian Mixture Priors

链接：https://iclr.cc/virtual/2025/poster/28886 abstract： We establish in-expectation and tail bounds on the generalization error of representation learning type algorithms. The bounds are in terms of the relative entropy between the distribution of the representations extracted from the training and "test" datasets and a data-dependent symmetric prior, i.e., the Minimum Description Length (MDL) of the latent variables for the training and test datasets. Our bounds are shown to reflect the "structure" and "simplicity" of the encoder and significantly improve upon the few existing ones for the studied model. We then use our in-expectation bound to devise a suitable data-dependent regularizer; and we investigate thoroughly the important question of the selection of the prior. We propose a systematic approach to simultaneously learning a data-dependent Gaussian mixture prior and using it as a regularizer. Interestingly, we show that a weighted attention mechanism emerges naturally in this procedure. Our experiments show that our approach outperforms the now popular Variational Information Bottleneck (VIB) method as well as the recent Category-Dependent VIB (CDVIB).

911. How many samples are needed to train a deep neural network?

链接：https://iclr.cc/virtual/2025/poster/28262 abstract： Even though neural networks have become standard tools in many areas, many important statistical questions remain open. This paper studies the question of how much data are needed to train a ReLU feed-forward neural network. Our theoretical and empirical results suggest that the generalization error of ReLU feed-forward neural networks scales at the rate $1/\sqrt{n}$ in the sample size n-rather than the "parametric rate" $1/n$, which might be suggested by traditional statistical theories. Thus, broadly speaking, our results underpin the common belief that neural networks need "many" training samples. Along the way, we also establish new technical insights, such as the first lower bounds of the entropy of ReLU feed-forward networks.

912. Revisiting In-context Learning Inference Circuit in Large Language Models

链接：https://iclr.cc/virtual/2025/poster/27767 abstract： In-context Learning (ICL) is an emerging few-shot learning paradigm on Language Models (LMs) with inner mechanisms un-explored. There are already existing works describing the inner processing of ICL, while they struggle to capture all the inference phenomena in large language models. Therefore, this paper proposes a comprehensive circuit to model the inference dynamics and try to explain the observed phenomena of ICL. In detail, we divide ICL inference into 3 major operations: (1) Input Text Encode: LMs encode every input text (in the demonstrations and queries) into linear representation in the hidden states with sufficient information to solve ICL tasks. (2) Semantics Merge: LMs merge the encoded representations of demonstrations with their corresponding label tokens to produce joint representations of labels and demonstrations. (3) Feature Retrieval and Copy: LMs search the joint representations of demonstrations similar to the query representation on a task subspace, and copy the searched representations into the query. Then, language model heads capture these copied label representations to a certain extent and decode them into predicted labels. Through careful measurements, the proposed inference circuit successfully captures and unifies many fragmented phenomena observed during the ICL process, making it a comprehensive and practical explanation of the ICL inference process. Moreover, ablation analysis by disabling the proposed steps seriously damages the ICL performance, suggesting the proposed inference circuit is a dominating mechanism. Additionally, we confirm and list some bypass mechanisms that solve ICL tasks in parallel with the proposed circuit.

913. Neural Wave Equation for Irregularly Sampled Sequence Data

链接：https://iclr.cc/virtual/2025/poster/28570 abstract： Sequence labeling problems arise in several real-world applications such as healthcare and robotics. In many such applications, sequence data are irregularly sampled and are of varying complexities. Recently, efforts have been made to develop neural ODE-based architectures to model the evolution of hidden states continuously in time, to address irregularly sampled sequence data. However, they assume a fixed architectural depth and limit their flexibility to adapt to data sets with varying complexities. We propose the neural wave equation, a novel deep learning method inspired by the wave equation, to address this through continuous modeling of depth. Neural Wave Equation models the evolution of hidden states continuously across time as well as depth by using a non-homogeneous wave equation parameterized by a neural network. Through d'Alembert's analytical solution of the wave equation, we also show that the neural wave equation provides denser connections across the hidden states, allowing for better modeling capability. We conduct experiments on several sequence labeling problems involving irregularly sampled sequence data and demonstrate the superior performance of the proposed neural wave equation model.

914. Deep Learning Alternatives Of The Kolmogorov Superposition Theorem

链接：https://iclr.cc/virtual/2025/poster/29570 abstract： This paper explores alternative formulations of the Kolmogorov Superposition Theorem (KST) as a foundation for neural network design. The original KST formulation, while mathematically elegant, presents practical challenges due to its limited insight into the structure of inner and outer functions and the large number of unknown variables it introduces. Kolmogorov-Arnold Networks (KANs) leverage KST for function approximation, but they have faced scrutiny due to mixed results compared to traditional multilayer perceptrons (MLPs) and practical limitations imposed by the original KST formulation. To address these issues, we introduce ActNet, a scalable deep learning model that builds on the KST and overcomes some of the drawbacks of Kolmogorov's original formulation. We evaluate ActNet in the context of Physics-Informed Neural Networks (PINNs), a framework well-suited for leveraging KST's strengths in low-dimensional function approximation, particularly for simulating partial differential equations (PDEs). In this challenging setting, where models must learn latent functions without direct measurements, ActNet consistently outperforms KANs across multiple benchmarks and is competitive against the current best MLP-based approaches. These results present ActNet as a promising new direction for KST-based deep learning applications, particularly in scientific computing and PDE simulation tasks.

915. Understanding Long Videos with Multimodal Language Models

链接：https://iclr.cc/virtual/2025/poster/29788 abstract： Large Language Models (LLMs) have allowed recent LLM-based approaches to achieve excellent performance on long-video understanding benchmarks. We investigate how extensive world knowledge and strong reasoning skills of underlying LLMs influence this strong performance. Surprisingly, we discover that LLM-based approaches can yield surprisingly good accuracy on long-video tasks with limited video information, sometimes even with no video-specific information. Building on this, we explore injecting video-specific information into an LLM-based framework. We utilize off-the-shelf vision tools to extract three object-centric information modalities from videos, and then leverage natural language as a medium for fusing this information. Our resulting Multimodal Video Understanding (MVU) framework demonstrates state-of-the-art performance across multiple video understanding benchmarks. Strong performance also on robotics domain tasks establishes its strong generality. Code: github.com/kahnchana/mvu

916. Residual Stream Analysis with Multi-Layer SAEs

链接：https://iclr.cc/virtual/2025/poster/29317 abstract： Sparse autoencoders (SAEs) are a promising approach to interpreting the internal representations of transformer language models. However, SAEs are usually trained separately on each transformer layer, making it difficult to use them to study how information flows across layers. To solve this problem, we introduce the multi-layer SAE (MLSAE): a single SAE trained on the residual stream activation vectors from every transformer layer. Given that the residual stream is understood to preserve information across layers, we expected MLSAE latents to 'switch on' at a token position and remain active at later layers. Interestingly, we find that individual latents are often active at a single layer for a given token or prompt, but the layer at which an individual latent is active may differ for different tokens or prompts. We quantify these phenomena by defining a distribution over layers and considering its variance. We find that the variance of the distributions of latent activations over layers is about two orders of magnitude greater when aggregating over tokens compared with a single token. For larger underlying models, the degree to which latents are active at multiple layers increases, which is consistent with the fact that the residual stream activation vectors at adjacent layers become more similar. Finally, we relax the assumption that the residual stream basis is the same at every layer by applying pre-trained tuned-lens transformations, but our findings remain qualitatively similar. Our results represent a new approach to understanding how representations change as they flow through transformers. We release our code to train and analyze MLSAEs at https://github.com/tim-lawson/mlsae.

917. When Attention Sink Emerges in Language Models: An Empirical View

链接：https://iclr.cc/virtual/2025/poster/30847 abstract： Auto-regressive language Models (LMs) assign significant attention to the first token, even if it is not semantically important, which is known as attention sink. This phenomenon has been widely adopted in applications such as streaming/long context generation, KV cache optimization, inference acceleration, model quantization, and others. Despite its widespread use, a deep understanding of attention sink in LMs is still lacking. In this work, we first demonstrate that attention sinks exist universally in auto-regressive LMs with various inputs, even in small models. Furthermore, attention sink is observed to emerge during the LM pre-training, motivating us to investigate how optimization, data distribution, loss function, and model architecture in LM pre-training influence its emergence. We highlight that attention sink emerges after effective optimization on sufficient training data. The sink position is highly correlated with the loss function and data distribution. Most importantly, we find that attention sink acts more like key biases, storing extra attention scores, which could be non-informative and not contribute to the value computation. We also observe that this phenomenon (at least partially) stems from tokens' inner dependence on attention scores as a result of softmax normalization. After relaxing such dependence by replacing softmax attention with other attention operations, such as sigmoid attention without normalization, attention sinks do not emerge in LMs up to 1B parameters. The code is available at https://github.com/sail-sg/Attention-Sink.

918. Protecting against simultaneous data poisoning attacks

链接：https://iclr.cc/virtual/2025/poster/28204 abstract： Current backdoor defense methods are evaluated against a single attack at a time. This is unrealistic, as powerful machine learning systems are trained on large datasets scraped from the internet, which may be attacked multiple times by one or more attackers. We demonstrate that multiple backdoors can be simultaneously installed in a single model through parallel data poisoning attacks without substantially degrading clean accuracy. Furthermore, we show that existing backdoor defense methods do not effectively defend against multiple simultaneous attacks. Finally, we leverage insights into the nature of backdoor attacks to develop a new defense, BaDLoss (Backdoor Detection via

Loss Dynamics), that is effective in the multi-attack setting. With minimal clean accuracy degradation, BaDLoss attains an average attack success rate in the multi-attack setting of 7.98% in CIFAR-10, 10.29% in GTSRB, and 19.17% in Imagenette, compared to the average of other defenses at 63.44%, 74.83%, and 41.74% respectively. BaDLoss scales to ImageNet-1k, reducing the average attack success rate from 88.57% to 15.61%.

919. Rethinking Reward Modeling in Preference-based Large Language Model Alignment

链接：https://iclr.cc/virtual/2025/poster/28181 abstract： The Bradley-Terry (BT) model is a common and successful practice in reward modeling for Large Language Model (LLM) alignment. However, it remains unclear *why* this model --- originally developed for multi-player stochastic game matching --- can be adopted to convert pairwise response comparisons to reward values and make predictions. Especially given the fact that only a limited number of prompt-response pairs are sparsely compared with others. In this paper, we first establish the convergence rate of BT reward models based on deep neural networks using embeddings, providing a theoretical foundation for their use.Despite theoretically sound, we argue that the BT model is not a necessary choice from the perspective of downstream optimization, this is because a reward model only needs to preserve the correct ranking predictions through a monotonic transformation of the true reward. We highlight the critical concept of *order consistency* in reward modeling and demonstrate that the BT model possesses this property.Moreover, we propose a simple and straightforward upper-bound algorithm, compatible with off-the-shelf binary classifiers, as an alternative order-consistent reward modeling objective. To offer practical insights, we empirically evaluate the performance of these different reward modeling approaches across more than 12,000 experimental setups, using 6 base LLMs, 2 datasets, and diverse annotation designs that vary in quantity, quality, and pairing choices in preference annotations.

920. Rethinking Artistic Copyright Infringements In the Era Of Text-to-Image Generative Models

链接：https://iclr.cc/virtual/2025/poster/31257 abstract： The advent of text-to-image generative models has led artists to worry that their individual styles may be copied, creating a pressing need to reconsider the lack of protection for artistic styles under copyright law. This requires answering challenging questions, like what defines style and what constitutes style infringment. In this work, we build on prior legal scholarship to develop an automatic and interpretable framework to \emph{quantitatively} assess style infringement. Our methods hinge on a simple logical argument: if an artist's works can consistently be recognized as their own, then they have a unique style. Based on this argument, we introduce ArtSavant, a practical (i.e., efficient and easy to understand) tool to (i) determine the unique style of an artist by comparing it to a reference corpus of works from hundreds of artists, and (ii) recognize if the identified style reappears in generated images. We then apply ArtSavant in an empirical study to quantify the prevalence of artistic style copying across 3 popular text-to-image generative models, finding that under simple prompting, 20% of 372 prolific artists studied appear to have their styles be at risk of copying by today's generative models. Our findings show that prior legal arguments can be operationalized in quantitative ways, towards more nuanced examination of the issue of artistic style infringements.

921. On Calibration of LLM-based Guard Models for Reliable Content Moderation

链接：https://iclr.cc/virtual/2025/poster/27843 abstract： Large language models (LLMs) pose significant risks due to the potential for generating harmful content or users attempting to evade guardrails. Existing studies have developed LLM-based guard models designed to moderate the input and output of threat LLMs, ensuring adherence to safety policies by blocking content that violates these protocols upon deployment. However, limited attention has been given to the reliability and calibration of such guard models. In this work, we empirically conduct comprehensive investigations of confidence calibration for 9 existing LLM-based guard models on 12 benchmarks in both user input and model output classification. Our findings reveal that current LLM-based guard models tend to 1) produce overconfident predictions, 2) exhibit significant miscalibration when subjected to jailbreak attacks, and 3) demonstrate limited robustness to the outputs generated by different types of response models. Additionally, we assess the effectiveness of post-hoc calibration methods to mitigate miscalibration. We demonstrate the efficacy of temperature scaling and, for the first time, highlight the benefits of contextual calibration for confidence calibration of guard models, particularly in the absence of validation sets. Our analysis and experiments underscore the limitations of current LLM-based guard models and provide valuable insights for the future development of well-calibrated guard models toward more reliable content moderation. We also advocate for incorporating reliability evaluation of confidence calibration when releasing future LLM-based guard models.

922. Exploiting Hankel-Toeplitz Structures for Fast Computation of Kernel Precision Matrices

链接：https://iclr.cc/virtual/2025/poster/31467 abstract： The Hilbert-space Gaussian process (HGP) approach offers a hyperparameter-independent basis function approximation for speeding up Gaussian process (GP) inference by projecting the GP onto M basis functions. These properties result in a favorable data-independent $\mathcal{O}(M^3)$ computational complexity during hyperparameter optimization but require a dominating one-time precomputation of the precision matrix costing $\mathcal{O}(NM^2)$ operations. In this paper, we lower this dominating computational complexity to $\mathcal{O}

(NM)$ with no additional approximations. We can do this because we realize that the precision matrix can be split into a sum of Hankel-Toeplitz matrices, each having $\mathcal{O}(M)$ unique entries. Based on this realization we propose computing only these unique entries at $\mathcal{O}(NM)$ costs. Further, we develop two theorems that prescribe sufficient conditions for the complexity reduction to hold generally for a wide range of other approximate GP models, such as the Variational Fourier features approach. The two theorems do this with no assumptions on the data and no additional approximations of the GP models themselves. Thus, our contribution provides a pure speed-up of several existing, widely used, GP approximations, without further approximations

923. MuHBoost: Multi-Label Boosting For Practical Longitudinal Human Behavior Modeling

链接：https://iclr.cc/virtual/2025/poster/30588 abstract： Longitudinal human behavior modeling has received increasing attention over the years due to its widespread applications to patient monitoring, dietary and lifestyle recommendations, and just-in-time intervention for at-risk individuals (e.g., problematic drug users and struggling students), to name a few. Using in-the-moment health data collected via ubiquitous devices (e.g., smartphones and smartwatches), this multidisciplinary field focuses on developing predictive models for certain health or well-being outcomes (e.g., depression and stress) in the short future given the time series of individual behaviors (e.g., resting heart rate, sleep quality, and current feelings). Yet, most existing models on these data, which we refer to as ubiquitous health data, do not achieve adequate accuracy. The latest works that yielded promising results have yet to consider realistic aspects of ubiquitous health data (e.g., containing features of different types and high rate of missing values) and the consumption of various resources (e.g., computing power, time, and cost). Given these two shortcomings, it is dubious whether these studies could translate to realistic settings. In this paper, we propose MuHBoost, a multi-label boosting method for addressing these shortcomings, by leveraging advanced methods in large language model (LLM) prompting and multi-label classification (MLC) to jointly predict multiple health or well-being outcomes. Because LLMs can hallucinate when tasked with answering multiple questions simultaneously, we also develop two variants of MuHBoost that alleviate this issue and thereby enhance its predictive performance. We conduct extensive experiments to evaluate MuHBoost and its variants on 13 health and well-being prediction tasks defined from four realistic ubiquitous health datasets. Our results show that our three developed methods outperform all considered baselines across three standard MLC metrics, demonstrating their effectiveness while ensuring resource efficiency.

924. PETRA: Parallel End-to-end Training with Reversible Architectures

链接：https://iclr.cc/virtual/2025/poster/31242 abstract： Reversible architectures have been shown to be capable of performing on par with their non-reversible architectures, being applied in deep learning for memory savings and generative modeling. In this work, we show how reversible architectures can solve challenges in parallelizing deep model training. We introduce PETRA, a novel alternative to backpropagation for parallelizing gradient computations. PETRA facilitates effective model parallelism by enabling stages (i.e., a set of layers) to compute independently on different devices, while only needing to communicate activations and gradients between each other. By decoupling the forward and backward passes and keeping a single updated version of the parameters, the need for weight stashing is also removed. We develop a custom autograd-like training framework for PETRA, and we demonstrate its effectiveness on standard computer vision benchmarks, achieving competitive accuracies comparable to backpropagation using ResNet-18, ResNet-34, and ResNet-50 models.

925. Diffusion On Syntax Trees For Program Synthesis

链接：https://iclr.cc/virtual/2025/poster/27850 abstract： Large language models generate code one token at a time. Their autoregressive generation process lacks the feedback of observing the program's output. Training LLMs to suggest edits directly can be challenging due to the scarcity of rich edit data. To address these problems, we propose neural diffusion models that operate on syntax trees of any context-free grammar. Similar to image diffusion models, our method also inverts "noise" applied to syntax trees. Rather than generating code sequentially, we iteratively edit it while preserving syntactic validity, which makes it easy to combine this neural model with search. We apply our approach to inverse graphics tasks, where our model learns to convert images into programs that produce those images. Combined with search, our model is able to write graphics programs, see the execution result, and debug them to meet the required specifications. We additionally show how our system can write graphics programs for hand-drawn sketches. Video results can be found at https://tree-diffusion.github.io.

926. What Makes a Good Diffusion Planner for Decision Making?

链接：https://iclr.cc/virtual/2025/poster/30839 abstract： Diffusion models have recently shown significant potential in solving decision-making problems, particularly in generating behavior plans -- also known as diffusion planning. While numerous studies have demonstrated the impressive performance of diffusion planning, the mechanisms behind the key components of a good diffusion planner remain unclear and the design choices are highly inconsistent in existing studies. In this work, we address this issue through systematic empirical experiments on diffusion planning in an offline reinforcement learning (RL) setting, providing practical insights into the essential components of diffusion planning. We trained and evaluated over 6,000 diffusion models, identifying the critical components such as guided sampling, network architecture, action generation and planning strategy. We revealed that some design choices opposite to the common practice in previous work in diffusion planning actually lead to better performance, e.g., unconditional sampling with selection can be better than guided sampling and Transformer outperforms U-Net as denoising network. Based on these insights, we suggest a simple yet strong diffusion planning baseline that achieves state-of-the-art results on standard offline RL benchmarks. Code: https://github.com/Josh00-Lu/DiffusionVeteran.

927. Test-time Adaptation for Regression by Subspace Alignment

链接：https://iclr.cc/virtual/2025/poster/29596 abstract： This paper investigates test-time adaptation (TTA) for regression, where a regression model pre-trained in a source domain is adapted to an unknown target distribution with unlabeled target data.Although regression is one of the fundamental tasks in machine learning, most of the existing TTA methods have classification-specific designs, which assume that models output class-categorical predictions, whereas regression models typically output only single scalar values.To enable TTA for regression, we adopt a feature alignment approach, which aligns the feature distributions between the source and target domains to mitigate the domain gap.However, we found that naive feature alignment employed in existing TTA methods for classification is ineffective or even worse for regression because the features are distributed in a small subspace and many of the raw feature dimensions have little significance to the output.For an effective feature alignment in TTA for regression, we propose Significant-subspace Alignment (SSA).SSA consists of two components: subspace detection and dimension weighting.Subspace detection finds the feature subspace that is representative and significant to the output.Then, the feature alignment is performed in the subspace during TTA.Meanwhile, dimension weighting raises the importance of the dimensions of the feature subspace that have greater significance to the output.We experimentally show that SSA outperforms various baselines on real-world datasets.The code is available at https://github.com/kzkadc/regression-tta.

928. Attributing Culture-Conditioned Generations to Pretraining Corpora

链接：https://iclr.cc/virtual/2025/poster/29287 abstract： In open-ended generative tasks like narrative writing or dialogue, large language models often exhibit cultural biases, showing limited knowledge and generating templated outputs for less prevalent cultures. Recent works show that these biases may stem from uneven cultural representation in pretraining corpora. This work investigates how pretraining leads to biased culture-conditioned generationsby analyzing how models associate entities with cultures based on pretraining data patterns. We propose the MEMOED framework (MEMOrization from prEtraining Document) to determine whether a generation for a culture arises from memorization. Using MEMOED on culture-conditioned generations about food and clothing for 110 cultures, we find that high-frequency cultures in pretraining data yield more generations with memorized symbols, while some low-frequency cultures produce none. Additionally, the model favors generating entities with extraordinarily high frequency regardless of the conditioned culture, reflecting biases toward frequent pretraining terms irrespective of relevance. We hope that the MEMOED framework and our insights will inspire more works on attributing model performance on pretraining data.

929. X-ALMA: Plug & Play Modules and Adaptive Rejection for Quality Translation at Scale

链接：https://iclr.cc/virtual/2025/poster/29024 abstract： Large language models (LLMs) have achieved remarkable success across various NLP tasks with a focus on English due to English-centric pre-training and limited multilingual data. In this work, we focus on the problem of translation, and while some multilingual LLMs claim to support for hundreds of languages, models often fail to provide high-quality responses for mid- and low-resource languages, leading to imbalanced performance heavily skewed in favor of high-resource languages. We introduce X-ALMA, a model designed to ensure top-tier performance across 50 diverse languages, regardless of their resource levels. X-ALMA surpasses state-of-the-art open-source multilingual LLMs, such as Aya-101 and Aya-23, in every single translation direction on the FLORES-200 and WMT'23 test datasets according to COMET-22. This is achieved by plug-and-play language-specific module architecture to prevent language conflicts during training and a carefully designed training regimen with novel optimization methods to maximize the translation performance. After the final stage of training regimen, our proposed Adaptive Rejection Preference Optimization (ARPO) surpasses existing preference optimization methods in translation tasks.

930. Semantix: An Energy-guided Sampler for Semantic Style Transfer

链接：https://iclr.cc/virtual/2025/poster/28103 abstract： Recent advances in style and appearance transfer are impressive, but most methods isolate global style and local appearance transfer, neglecting semantic correspondence. Additionally, image and video tasks are typically handled in isolation, with little focus on integrating them for video transfer. To address these limitations, we introduce a novel task, Semantic Style Transfer, which involves transferring style and appearance features from a reference image to a target visual content based on semantic correspondence. We subsequently propose a training-free method, Semantix, an energy-guided sampler designed for Semantic Style Transfer that simultaneously guides both style and appearance transfer based on semantic understanding capacity of pre-trained diffusion models. Additionally, as a sampler, Semantix can be seamlessly applied to both image and video models, enabling semantic style transfer to be generic across various visual media. Specifically, once inverting both reference and context images or videos to noise space by SDEs, Semantix utilizes a meticulously crafted energy function to guide the sampling process, including three key components: Style Feature Guidance, Spatial Feature Guidance and Semantic Distance as a regularisation term. Experimental results demonstrate that Semantix not only effectively accomplishes the task of semantic style transfer across images and videos, but also surpasses existing state-of-the-art solutions in both fields.

931. LANTERN: Accelerating Visual Autoregressive Models with Relaxed

Speculative Decoding

链接：https://iclr.cc/virtual/2025/poster/30718 abstract： Auto-Regressive (AR) models have recently gained prominence in image generation, often matching or even surpassing the performance of diffusion models. However, one major limitation of AR models is their sequential nature, which processes tokens one at a time, slowing down generation compared to models like GANs or diffusion-based methods that operate more efficiently. While speculative decoding has proven effective for accelerating LLMs by generating multiple tokens in a single forward, its application in visual AR models remains largely unexplored. In this work, we identify a challenge in this setting, which we term \textit{token selection ambiguity}, wherein visual AR models frequently assign uniformly low probabilities to tokens, hampering the performance of speculative decoding. To overcome this challenge, we propose a relaxed acceptance condition referred to as LANTERN that leverages the interchangeability of tokens in latent space. This relaxation restores the effectiveness of speculative decoding in visual AR models by enabling more flexible use of candidate tokens that would otherwise be prematurely rejected. Furthermore, by incorporating a total variation distance bound, we ensure that these speed gains are achieved without significantly compromising image quality or semantic coherence. Experimental results demonstrate the efficacy of our method in providing a substantial speed-up over speculative decoding. In specific, compared to a na\"ive application of the state-of-the-art speculative decoding, LANTERN increases speed-ups by $\mathbf{1.75}\times$ and $\mathbf{1.82}\times$, as compared to greedy decoding and random sampling, respectively, when applied to LlamaGen, a contemporary visual AR model. The code is publicly available at \url{https://github.com/jadohu/LANTERN}.

932. ConcreTizer: Model Inversion Attack via Occupancy Classification and Dispersion Control for 3D Point Cloud Restoration

链接：https://iclr.cc/virtual/2025/poster/32098 abstract： The growing use of 3D point cloud data in autonomous vehicles (AVs) has raised serious privacy concerns, particularly due to the sensitive information that can be extracted from 3D data. While model inversion attacks have been widely studied in the context of 2D data, their application to 3D point clouds remains largely unexplored. To fill this gap, we present the first in-depth study of model inversion attacks aimed at restoring 3D point cloud scenes. Our analysis reveals the unique challenges, the inherent sparsity of 3D point clouds and the ambiguity between empty and non-empty voxels after voxelization, which are further exacerbated by the dispersion of non-empty voxels across feature extractor layers. To address these challenges, we introduce ConcreTizer, a simple yet effective model inversion attack designed specifically for voxel-based 3D point cloud data. ConcreTizer incorporates Voxel Occupancy Classification to distinguish between empty and non-empty voxels and Dispersion-Controlled Supervision to mitigate non-empty voxel dispersion. Extensive experiments on widely used 3D feature extractors and benchmark datasets, such as KITTI and Waymo, demonstrate that ConcreTizer concretely restores the original 3D point cloud scene from disrupted 3D feature data. Our findings highlight both the vulnerability of 3D data to inversion attacks and the urgent need for robust defense strategies.

933. Rethinking LLM Unlearning Objectives: A Gradient Perspective and Go Beyond

链接：https://iclr.cc/virtual/2025/poster/28734 abstract： Large language models (LLMs) should undergo rigorous audits to identify potential risks, such as copyright and privacy infringements. Once these risks emerge, timely updates are crucial to remove undesirable responses, ensuring legal and safe model usage. It has spurred recent research into LLM unlearning, focusing on erasing targeted undesirable knowledge without compromising the integrity of other, non-targeted responses. Existing studies have introduced various unlearning objectives to pursue LLM unlearning without necessitating complete retraining. However, each of these objectives has unique properties, and no unified framework is currently available to comprehend them thoroughly. To fill the gap, we propose the metric of the G-effect, quantifying the impacts of unlearning objectives on model performance from a gradient lens. A significant advantage of our metric is its broad ability to detail the unlearning impacts from various aspects across instances, updating steps, and LLM layers. Accordingly, the G-effect offers new insights into identifying drawbacks of existing unlearning objectives, further motivating us to explore a series of candidate solutions for their mitigation and improvements. Finally, we outline promising directions that merit further studies, aiming at contributing to the community to advance this critical field.

934. Eliminating Oversaturation and Artifacts of High Guidance Scales in Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/28952 abstract： Classifier-free guidance (CFG) is crucial for improving both generation quality and alignment between the input condition and final output in diffusion models. While a high guidance scale is generally required to enhance these aspects, it also causes oversaturation and unrealistic artifacts. In this paper, we revisit the CFG update rule and introduce modifications to address this issue. We first decompose the update term in CFG into parallel and orthogonal components with respect to the conditional model prediction and observe that the parallel component primarily causes oversaturation, while the orthogonal component enhances image quality. Accordingly, we propose down-weighting the parallel component to achieve high-quality generations without oversaturation. Additionally, we draw a connection between CFG and gradient ascent and introduce a new rescaling and momentum method for the CFG update rule based on this insight. Our approach, termed adaptive projected guidance (APG), retains the quality-boosting advantages of CFG while enabling the use of higher guidance scales without oversaturation. APG is easy to implement and introduces practically no additional computational

overhead to the sampling process. Through extensive experiments, we demonstrate that APG is compatible with various conditional diffusion models and samplers, leading to improved FID, recall, and saturation scores while maintaining precision comparable to CFG, making our method a superior plug-and-play alternative to standard classifier-free guidance.

935. No Training, No Problem: Rethinking Classifier-Free Guidance for Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/29126 abstract： Classifier-free guidance (CFG) has become the standard method for enhancing the quality of conditional diffusion models. However, employing CFG requires either training an unconditional model alongside the main diffusion model or modifying the training procedure by periodically inserting a null condition. There is also no clear extension of CFG to unconditional models. In this paper, we revisit the core principles of CFG and introduce a new method, independent condition guidance (ICG), which provides the benefits of CFG without the need for any special training procedures. Our approach streamlines the training process of conditional diffusion models and can also be applied during inference on any pre-trained conditional model. Additionally, by leveraging the time-step information encoded in all diffusion networks, we propose an extension of CFG, called time-step guidance (TSG), which can be applied to any diffusion model, including unconditional ones. Our guidance techniques are easy to implement and have the same sampling cost as CFG. Through extensive experiments, we demonstrate that ICG matches the performance of standard CFG across various conditional diffusion models. Moreover, we show that TSG improves generation quality in a manner similar to CFG, without relying on any conditional information.

936. Efficient Off-Policy Learning for High-Dimensional Action Spaces

链接：https://iclr.cc/virtual/2025/poster/30118 abstract： Existing off-policy reinforcement learning algorithms often rely on an explicit state-action-value function representation, which can be problematic in high-dimensional action spaces due to the curse of dimensionality.This reliance results in data inefficiency as maintaining a state-action-value function in such spaces is challenging. We present an efficient approach that utilizes only a state-value function as the critic for off-policy deep reinforcement learning.This approach, which we refer to as Vlearn, effectively circumvents the limitations of existing methods by eliminating the necessity for an explicit state-action-value function. To this end, we leverage a weighted importance sampling loss for learning deep value functions from off-policy data. While this is common for linear methods, it has not been combined with deep value function networks. This transfer to deep methods is not straightforward and requires novel design choices such as robust policy updates, twin value function networks to avoid an optimization bias, and importance weight clipping.We also present a novel analysis of the variance of our estimate compared to commonly used importance sampling estimators such as V-trace. Our approach improves sample complexity as well as final performance and ensures consistent and robust performance across various benchmark tasks.Eliminating the state-action-value function in Vlearn facilitates a streamlined learning process, yielding high-return agents.

937. Towards Semantic Equivalence of Tokenization in Multimodal LLM

链接：https://iclr.cc/virtual/2025/poster/28428 abstract： Multimodal Large Language Models (MLLMs) have demonstrated exceptional capabilities in processing vision-language tasks. One of the crux of MLLMs lies in vision tokenization, which involves efficiently transforming input visual signals into feature representations that are most beneficial for LLMs. However, existing vision tokenizers, essential for semantic alignment between vision and language, remain problematic. Existing methods aggressively fragment visual input, corrupting the visual semantic integrity. To address this, this paper proposes a novel dynamic Semantic-Equivalent Vision Tokenizer (SeTok), which groups visual features into semantic units via a dynamic clustering algorithm, flexibly determining the number of tokens based on image complexity. The resulting vision tokens effectively preserve semantic integrity and capture both low-frequency and high-frequency visual features. The proposed MLLM (Setokim) equipped with SeTok significantly demonstrates superior performance across various tasks, as evidenced by our experimental results.

938. Searching for Optimal Solutions with LLMs via Bayesian Optimization

链接：https://iclr.cc/virtual/2025/poster/29164 abstract： Scaling test-time compute to search for optimal solutions is an important step towards building generally-capable language models that can reason. Recent work, however, shows that tasks of varying complexity require distinct search strategies to solve optimally, thus making it challenging to design a one-size-fits-all approach. Prior solutions either attempt to predict task difficulty to select the optimal search strategy, often infeasible in practice, or use a static, pre-defined strategy, e.g., repeated parallel sampling or greedy sequential search, which is sub-optimal. In this work, we argue for an alternative view using the probabilistic framework of Bayesian optimization (BO), where the search strategy is adapted dynamically based on the evolving uncertainty estimates of solutions as search progresses. To this end, we introduce Bayesian-OPRO (BOPRO)—a generalization of a recent method for in-context optimization, which iteratively samples from new proposal distributions by modifying the prompt to the LLM with a subset of its previous generations selected to explore or exploit different parts of the search space. We evaluate our method on word search, molecule optimization, and a joint hypothesis+program search task using a 1-D version of the challenging Abstraction and Reasoning Corpus (1D-ARC). Our results show that BOPRO outperforms all baselines in word search (≥10 points) and molecule optimization (higher quality and 17% fewer invalid molecules), but trails a best-k prompting strategy in program search. Our analysis reveals that despite the ability to balance exploration and exploitation using BOPRO, failure is likely due to the inability of code representation models in distinguishing sequences with low edit-distances.

939. RandLoRA: Full rank parameter-efficient fine-tuning of large models

链接：https://iclr.cc/virtual/2025/poster/30212 abstract： Low-Rank Adaptation (LoRA) and its variants have shown impressive results in reducing the number of trainable parameters and memory requirements of large transformer networks while maintaining fine-tuning performance. The low-rank nature of the weight update inherently limits the representation power of fine-tuned models, however, thus potentially compromising performance on complex tasks. This raises a critical question: when a performance gap between LoRA and standard fine-tuning is observed, is it due to the reduced number of train-able parameters or the rank deficiency? This paper aims to answer this question by introducing RandLoRA, a parameter-efficient method that performs full-rank updates using a learned linear combinations of low-rank, non-trainable random matrices. Our method limits the number of trainable parameters by restricting optimization to diagonal scaling matrices applied to the fixed random matrices. This allows us to effectively overcome the low-rank limitations while maintaining parameter and memory efficiency during training. Through extensive experimen-tation across vision, language, and vision-language benchmarks, we systematically evaluate the limitations of LoRA and existing random basis methods. Our findings reveal that full-rank updates are beneficial across vision and language tasks individually, and even more so for vision-language tasks, where RandLoRA significantly reduces—and sometimes eliminates—the performance gap between standard fine-tuning and LoRA, demonstrating its efficacy.

940. AlphaEdit: Null-Space Constrained Knowledge Editing for Language Models

链接：https://iclr.cc/virtual/2025/poster/30202 abstract： Large language models (LLMs) often exhibit hallucinations, producing incorrect or outdated knowledge. Hence, model editing methods have emerged to enable targeted knowledge updates. To achieve this, a prevailing paradigm is the locating-then-editing approach, which first locates influential parameters and then edits them by introducing a perturbation. While effective, current studies have demonstrated that this perturbation inevitably disrupt the originally preserved knowledge within LLMs, especially in sequential editing scenarios.To address this, we introduce AlphaEdit, a novel solution that projects perturbation onto the null space of the preserved knowledge before applying it to the parameters. We theoretically prove that this projection ensures the output of post-edited LLMs remains unchanged when queried about the preserved knowledge, thereby mitigating the issue of disruption. Extensive experiments on various LLMs, including LLaMA3, GPT2-XL, and GPT-J, show that AlphaEdit boosts the performance of most locating-then-editing methods by an average of 36.7% with a single line of additional code for projection solely.

941. Precise Parameter Localization for Textual Generation in Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/28808 abstract： Novel diffusion models can synthesize photo-realistic images with integrated high-quality text. Surprisingly, we demonstrate through attention activation patching that only less than 1\% of diffusion models' parameters, all contained in attention layers, influence the generation of textual content within the images. Building on this observation, we improve textual generation efficiency and performance by targeting cross and joint attention layers of diffusion models. We introduce several applications that benefit from localizing the layers responsible for textual content generation. We first show that a LoRA-based fine-tuning solely of the localized layers enhances, even more, the general text-generation capabilities of large diffusion models while preserving the quality and diversity of the diffusion models' generations. Then, we demonstrate how we can use the localized layers to edit textual content in generated images. Finally, we extend this idea to the practical use case of preventing the generation of toxic text in a cost-free manner. In contrast to prior work, our localization approach is broadly applicable across various diffusion model architectures, including U-Net (e.g., SDXL and DeepFloyd IF) and transformer-based (e.g., Stable Diffusion 3), utilizing diverse text encoders (e.g., from CLIP to the large language models like T5). Project page available at https://t2i-text-loc.github.io/.

942. Decoupled Graph Energy-based Model for Node Out-of-Distribution Detection on Heterophilic Graphs

链接：https://iclr.cc/virtual/2025/poster/29852 abstract： Despite extensive research efforts focused on Out-of-Distribution (OOD) detection on images, OOD detection on nodes in graph learning remains underexplored. The dependence among graph nodes hinders the trivial adaptation of existing approaches on images that assume inputs to be i.i.d. sampled, since many unique features and challenges specific to graphs are not considered, such as the heterophily issue. Recently, GNNSafe, which considers node dependence, adapted energy-based detection to the graph domain with state-of-the-art performance, however, it has two serious issues: 1) it derives node energy from classification logits without specifically tailored training for modeling data distribution, making it less effective at recognizing OOD data; 2) it highly relies on energy propagation, which is based on homophily assumption and will cause significant performance degradation on heterophilic graphs, where the node tends to have dissimilar distribution with its neighbors. To address the above issues, we suggest training Energy-based Models (EBMs) by Maximum Likelihood Estimation (MLE) to enhance data distribution modeling and removing energy propagation to overcome the heterophily issues. However, training EBMs via MLE requires performing Markov Chain Monte Carlo (MCMC) sampling on both node feature and node neighbors, which is challenging due to the node interdependence and discrete graph topology. To tackle the sampling challenge, we introduce Decoupled Graph Energy-based Model (DeGEM), which decomposes the learning process into two parts—a graph encoder that leverages topology information for node representations and an energy head that operates in latent space. Additionally, we propose a Multi-Hop Graph encoder (MH) and Energy Readout (ERo) to enhance

node representation learning, Conditional Energy (CE) for improved EBM training, and Recurrent Update for the graph encoder and energy head to promote each other. This approach avoids sampling adjacency matrices and removes the need for energy propagation to extract graph topology information. Extensive experiments validate that DeGEM, without OOD exposure during training, surpasses previous state-of-the-art methods, achieving an average AUROC improvement of 6.71% on homophilic graphs and 20.29% on heterophilic graphs, and even outperform methods trained with OOD exposure. Our code is available at: https://github.com/draym28/DeGEM.

943. Toward Generalizing Visual Brain Decoding to Unseen Subjects

链接：https://iclr.cc/virtual/2025/poster/30608 abstract： Visual brain decoding aims to decode visual information from human brain activities. Despite the great progress, one critical limitation of current brain decoding research lies in the lack of generalization capability to unseen subjects. Prior work typically focuses on decoding brain activity of individuals based on the observation that different subjects exhibit different brain activities, while it remains unclear whether brain decoding can be generalized to unseen subjects. This study aims to answer this question. We first consolidate an image-fMRI dataset consisting of stimulus-image and fMRI-response pairs, involving 177 subjects in the movie-viewing task of the Human Connectome Project (HCP). This dataset allows us to investigate the brain decoding performance with the increase of participants. We then present a learning paradigm that applies uniform processing across all subjects, instead of employing different network heads or tokenizers for individuals as in previous methods, so that we can accommodate a large number of subjects to explore the generalization capability across different subjects. A series of experiments are conducted and we have the following findings. First, the network exhibits clear generalization capabilities with the increase of training subjects. Second, the generalization capability is common to popular network architectures (MLP, CNN and Transformer). Third, the generalization performance is affected by the similarity between subjects. Our findings reveal the inherent similarities in brain activities across individuals. With the emergence of larger and more comprehensive datasets, it is possible to train a brain decoding foundation model in the future. Codes and models can be found at https://github.com/Xiangtaokong/TGBD}{https://github.com/Xiangtaokong/TGBD.

944. Debiasing Mini-Batch Quadratics for Applications in Deep Learning

链接：https://iclr.cc/virtual/2025/poster/29718 abstract： Quadratic approximations form a fundamental building block of machine learning methods. E.g., second-order optimizers try to find the Newton step into the minimum of a local quadratic proxy to the objective function; and the second-order approximation of a network's loss function can be used to quantify the uncertainty of its outputs via the Laplace approximation. When computations on the entire training set are intractable - typical for deep learning - the relevant quantities are computed on mini-batches. This, however, distorts and biases the shape of the associated stochastic quadratic approximations in an intricate way with detrimental effects on applications. In this paper, we (i) show that this bias introduces a systematic error, (ii) provide a theoretical explanation for it, (iii) explain its relevance for second-order optimization and uncertainty quantification via the Laplace approximation in deep learning, and (iv) develop and evaluate debiasing strategies.

945. Optimal Protocols for Continual Learning via Statistical Physics and Control Theory

链接：https://iclr.cc/virtual/2025/poster/28178 abstract： Artificial neural networks often struggle with catastrophic forgetting when learning multiple tasks sequentially, as training on new tasks degrades the performance on previously learned tasks. Recent theoretical work has addressed this issue by analysing learning curves in synthetic frameworks under predefined training protocols. However, these protocols relied on heuristics and lacked a solid theoretical foundation assessing their optimality. In this paper, we fill this gap by combining exact equations for training dynamics, derived using statistical physics techniques, with optimal control methods. We apply this approach to teacher-student models for continual learning and multi-task problems, obtaining a theory for task-selection protocols maximising performance while minimising forgetting. Our theoretical analysis offers non-trivial yet interpretable strategies for mitigating catastrophic forgetting, shedding light on how optimal learning protocols modulate established effects, such as the influence of task similarity on forgetting. Finally, we validate our theoretical findings with experiments on real-world data.

946. A Theory of Initialisation's Impact on Specialisation

链接：https://iclr.cc/virtual/2025/poster/29650 abstract： Prior work has demonstrated a consistent tendency in neural networks engaged in continual learning tasks, wherein intermediate task similarity results in the highest levels of catastrophic interference. This phenomenon is attributed to the network's tendency to reuse learned features across tasks. However, this explanation heavily relies on the premise that neuron specialisation occurs, i.e. the emergence of localised representations. Our investigation challenges the validity of this assumption. Using theoretical frameworks for the analysis of neural networks, we show a strong dependence of specialisation on the initial condition. More precisely, we show that weight imbalance and high weight entropy can favour specialised solutions. We then apply these insights in the context of continual learning, first showing the emergence of a monotonic relation between task-similarity and forgetting in non-specialised networks. Finally, we show that specialization by weight imbalance is beneficial on the commonly employed elastic weight consolidation regularisation technique.

947. Correlating instruction-tuning (in multimodal models) with vision-

language processing (in the brain)

链接：https://iclr.cc/virtual/2025/poster/27766 abstract： Transformer-based language models, though not explicitly trained to mimic brain recordings, have demonstrated surprising alignment with brain activity. Progress in these models—through increased size, instruction-tuning, and multimodality—has led to better representational alignment with neural data. Recently, a new class of instruction-tuned multimodal LLMs (MLLMs) have emerged, showing remarkable zero-shot capabilities in open-ended multimodal vision tasks. However, it is unknown whether MLLMs, when prompted with natural instructions, lead to better brain alignment and effectively capture instruction-specific representations. To address this, we first investigate the brain alignment, i.e., measuring the degree of predictivity of neural visual activity using text output response embeddings from MLLMs as participants engage in watching natural scenes. Experiments with 10 different instructions (like image captioning, visual question answering, etc.) show that MLLMs exhibit significantly better brain alignment than vision-only models and perform comparably to non-instruction-tuned multimodal models like CLIP. We also find that while these MLLMs are effective at generating high-quality responses suitable to the task-specific instructions, not all instructions are relevant for brain alignment. Further, by varying instructions, we make the MLLMs encode instruction-specific visual concepts related to the input image. This analysis shows that MLLMs effectively capture count-related and recognition-related concepts, demonstrating strong alignment with brain activity. Notably, the majority of the explained variance of the brain encoding models is shared between MLLM embeddings of image captioning and other instructions. These results indicate that enhancing MLLMs' ability to capture more task-specific information could allow for better differentiation between various types of instructions, and hence improve their precision in predicting brain responses.

948. Non-Equilibrium Dynamics of Hybrid Continuous-Discrete Ground-State Sampling

链接：https://iclr.cc/virtual/2025/poster/30550 abstract： We propose a general framework for a hybrid continuous-discrete algorithm that integrates continuous-time deterministic dynamics with Metropolis-Hastings (MH) steps to combine search dynamics that either preserve or break detailed balance. Our purpose is to study the non-equilibrium dynamics that leads to the ground state of rugged energy landscapes in this general setting. Our results show that MH-driven dynamics reach ``easy'' ground states more quickly, indicating a stronger bias toward these solutions in algorithms using reversible transition probabilities. To validate this, we construct a set of Ising problem instances with a controllable bias in the energy landscape that makes certain degenerate solutions more accessible than others. The constructed hybrid algorithm demonstrates significant improvements in convergence and ground-state sampling accuracy, achieving a 100x speedup on GPU compared to simulated annealing, making it well-suited for large-scale applications.

949. Do as I do (Safely): Mitigating Task-Specific Fine-tuning Risks in Large Language Models

链接：https://iclr.cc/virtual/2025/poster/28514 abstract： Recent research shows that fine-tuning on benign instruction-following data can inadvertently undo the safety alignment process and increase a model's propensity to comply with harmful queries. While instruction-following fine-tuning is important, task-specific fine-tuning-where models are trained on datasets with clear ground truth answers (e.g., multiple choice questions)-can enhance model performance on specialized downstream tasks. Understanding and mitigating safety risks in the task-specific setting remains distinct from the instruction-following context due to structural differences in the data. Our work demonstrates how malicious actors can subtly manipulate the structure of almost any task-specific dataset to foster significantly more dangerous model behaviors, while maintaining an appearance of innocuity and reasonable downstream task performance. To address this issue, we propose a novel mitigation strategy that mixes in safety data which mimics the task format and prompting style of the user data, showing this is significantly more effective and efficient than existing baselines at re-establishing safety alignment while maintaining similar task performance.

950. Disentangling 3D Animal Pose Dynamics with Scrubbed Conditional Latent Variables

链接：https://iclr.cc/virtual/2025/poster/32066 abstract： Methods for tracking lab animal movements in unconstrained environments have become increasingly common and powerful tools for neuroscience. The prevailing hypothesis is that animal behavior in these environments comprises sequences of discrete stereotyped body movements ("motifs" or "actions"). However, the same action can occur at different speeds or heading directions, and the same action may manifest slightly differently across subjects due to, for example, variation in body size. These and other forms of nuisance variability complicate attempts to quantify animal behavior in terms of discrete action sequences and draw meaningful comparisons across individual subjects. To address this, we present a framework for motion analysis that uses conditional variational autoencoders in conjunction with adversarial learning paradigms to disentangle behavioral factors. We demonstrate the utility of this approach in downstream tasks such as clustering, decodability, and motion synthesis. Further, we apply our technique to improve disease detection in a Parkinsonian mouse model.

951. Towards Interpreting Visual Information Processing in Vision-Language Models

链接：https://iclr.cc/virtual/2025/poster/29034 abstract： Vision-Language Models (VLMs) are powerful tools for processing and understanding text and images. We study the processing of visual tokens in the language model component of LLaVA, a prominent VLM. Our approach focuses on analyzing the localization of object information, the evolution of visual token representations across layers, and the mechanism of integrating visual information for predictions. Through ablation studies, we demonstrated that object identification accuracy drops by over 70\% when object-specific tokens are removed. We observed that visual token representations become increasingly interpretable in the vocabulary space across layers, suggesting an alignment with textual tokens corresponding to image content. Finally, we found that the model extracts object information from these refined representations at the last token position for prediction, mirroring the process in text-only language models for factual association tasks. These findings provide crucial insights into how VLMs process and integrate visual information, bridging the gap between our understanding of language and vision models, and paving the way for more interpretable and controllable multimodal systems.

952. L-WISE: Boosting Human Visual Category Learning Through Model-Based Image Selection and Enhancement

链接：https://iclr.cc/virtual/2025/poster/30613 abstract： The currently leading artificial neural network models of the visual ventral stream - which are derived from a combination of performance optimization and robustification methods - have demonstrated a remarkable degree of behavioral alignment with humans on visual categorization tasks. We show that image perturbations generated by these models can enhance the ability of humans to accurately report the ground truth class. Furthermore, we find that the same models can also be used out-of-the-box to predict the proportion of correct human responses to individual images, providing a simple, human-aligned estimator of the relative difficulty of each image. Motivated by these observations, we propose to augment visual learning in humans in a way that improves human categorization accuracy at test time. Our learning augmentation approach consists of (i) selecting images based on their model-estimated recognition difficulty, and (ii) applying image perturbations that aid recognition for novice learners. We find that combining these model-based strategies leads to categorization accuracy gains of 33-72% relative to control subjects without these interventions, on unmodified, randomly selected held-out test images. Beyond the accuracy gain, the training time for the augmented learning group was also shortened by 20-23%, despite both groups completing the same number of training trials. We demonstrate the efficacy of our approach in a fine-grained categorization task with natural images, as well as two tasks in clinically relevant image domains - histology and dermoscopy - where visual learning is notoriously challenging. To the best of our knowledge, our work is the first application of artificial neural networks to increase visual learning performance in humans by enhancing category-specific image features.

953. In-context Time Series Predictor

链接：https://iclr.cc/virtual/2025/poster/28999 abstract： Recent Transformer-based large language models (LLMs) demonstrate in-context learning ability to perform various functions based solely on the provided context, without updating model parameters. To fully utilize the in-context capabilities in time series forecasting (TSF) problems, unlike previous Transformer-based or LLM-based time series forecasting methods, we reformulate "time series forecasting tasks" as input tokens by constructing a series of (lookback, future) pairs within the tokens. This method aligns more closely with the inherent in-context mechanisms and is more parameter-efficient without the need of using pre-trained LLM parameters. Furthermore, it addresses issues such as overfitting in existing Transformer-based TSF models, consistently achieving better performance across full-data, few-shot, and zero-shot settings compared to previous architectures.

954. The Loss Landscape of Deep Linear Neural Networks: a Second-order Analysis

链接：https://iclr.cc/virtual/2025/poster/31378 abstract： We study the optimization landscape of deep linear neural networks with square loss. It is known that, under weak assumptions, there are no spurious local minima and no local maxima. However, the existence and diversity of non-strict saddle points, which can play a role in first-order algorithms' dynamics, have only been lightly studied. We go a step further with a complete analysis of the optimization landscape at order 2. Among all critical points, we characterize global minimizers, strict saddle points, and non-strict saddle points. We enumerate all the associated critical values. The characterization is simple, involves conditions on the ranks of partial matrix products, and sheds some light on global convergence or implicit regularization that has been proved or observed when optimizing linear neural networks. In passing, we provide an explicit parameterization of the set of all global minimizers and exhibit large sets of strict and non-strict saddle points.

955. Task-Adaptive Pretrained Language Models via Clustered-Importance Sampling

链接：https://iclr.cc/virtual/2025/poster/28318 abstract： Specialist language models (LMs) focus on a specific task or domain on which they often outperform generalist LMs of the same size. However, the specialist data needed to pretrain these models is only available in limited amount for most tasks. In this work, we build specialist models from large generalist training sets instead. We adjust the training distribution of the generalist data with guidance from the limited domain-specific data. We explore several approaches, with clustered importance sampling standing out. This method clusters the generalist dataset and

samples from these clusters based on their frequencies in the smaller specialist dataset. It is scalable, suitable for pretraining and continued pretraining, it works well in multi-task settings. Our findings demonstrate improvements across different domains in terms of language modeling perplexity and accuracy on multiple-choice question tasks. We also present ablation studies that examine the impact of dataset sizes, clustering configurations, and model sizes.

956. Feedback Favors the Generalization of Neural ODEs

链接：https://iclr.cc/virtual/2025/poster/29029 abstract： The well-known generalization problem hinders the application of artificial neural networks in continuous-time prediction tasks with varying latent dynamics. In sharp contrast, biological systems can neatly adapt to evolving environments benefiting from real-time feedback mechanisms. Inspired by the feedback philosophy, we present feedback neural networks, showing that a feedback loop can flexibly correct the learned latent dynamics of neural ordinary differential equations (neural ODEs), leading to a prominent generalization improvement. The feedback neural network is a novel two-DOF neural network, which possesses robust performance in unseen scenarios with no loss of accuracy performance on previous tasks. A linear feedback form is presented to correct the learned latent dynamics firstly, with a convergence guarantee. Then, domain randomization is utilized to learn a nonlinear neural feedback form. Finally, extensive tests including trajectory prediction of a real irregular object and model predictive control of a quadrotor with various uncertainties, are implemented, indicating significant improvements over state-of-the-art model-based and learning-based methods.

957. Zero-cost Proxy for Adversarial Robustness Evaluation

链接：https://iclr.cc/virtual/2025/poster/27675 abstract： Deep neural networks (DNNs) easily cause security issues due to the lack of adversarial robustness. An emerging research topic for this problem is to design adversarially robust architectures via neural architecture search (NAS), i.e., robust NAS. However, robust NAS needs to train numerous DNNs for robustness estimation, making the search process prohibitively expensive. In this paper, we propose a zero-cost proxy to evaluate the adversarial robustness without training. Specifically, the proposed zero-cost proxy formulates the upper bound of adversarial loss, which can directly reflect the adversarial robustness. The formulation involves only the initialized weights of DNNs, thus the training process is no longer needed. Moreover, we theoretically justify the validity of the proposed proxy based on the theory of neural tangent kernel and input loss landscape. Experimental results show that the proposed zero-cost proxy can bring more than $20\times$ speedup compared with the state-of-the-art robust NAS methods, while the searched architecture has superior robustness and transferability under white-box and black-box attacks. Furthermore, compared with the state-of-the-art zero-cost proxies, the calculation of the proposed method has the strongest correlation with adversarial robustness. Our source code is available at https://github.com/fyqsama/Robust_ZCP.

958. Safety Alignment Should be Made More Than Just a Few Tokens Deep

链接：https://iclr.cc/virtual/2025/poster/30893 abstract： The safety alignment of current Large Language Models (LLMs) is vulnerable. Simple attacks, or even benign fine-tuning, can jailbreak aligned models. We note that many of these vulnerabilities are related to a shared underlying issue: safety alignment can take shortcuts, wherein the alignment adapts a model's generative distribution primarily over only its very first few output tokens. We unifiedly refer to this issue as shallow safety alignment. In this paper, we present case studies to explain why shallow safety alignment can exist and show how this issue universally contributes to multiple recently discovered vulnerabilities in LLMs, including the susceptibility to adversarial suffix attacks, prefilling attacks, decoding parameter attacks, and fine-tuning attacks. The key contribution of this work is that we demonstrate how this consolidated notion of shallow safety alignment sheds light on promising research directions for mitigating these vulnerabilities. We show that deepening the safety alignment beyond the first few tokens can meaningfully improve robustness against some common exploits. We also design a regularized fine-tuning objective that makes the safety alignment more persistent against fine-tuning attacks by constraining updates on initial tokens. Overall, we advocate that future safety alignment should be made more than just a few tokens deep.

959. MaestroMotif: Skill Design from Artificial Intelligence Feedback

链接：https://iclr.cc/virtual/2025/poster/28331 abstract： Describing skills in natural language has the potential to provide an accessible way to inject human knowledge about decision-making into an AI system. We present MaestroMotif, a method for AI-assisted skill design, which yields high-performing and adaptable agents. MaestroMotif leverages the capabilities of Large Language Models (LLMs) to effectively create and reuse skills. It first uses an LLM's feedback to automatically design rewards corresponding to each skill, starting from their natural language description. Then, it employs an LLM's code generation abilities, together with reinforcement learning, for training the skills and combining them to implement complex behaviors specified in language. We evaluate MaestroMotif using a suite of complex tasks in the NetHack Learning Environment (NLE), demonstrating that it surpasses existing approaches in both performance and usability.

960. Optimal Learning of Kernel Logistic Regression for Complex Classification Scenarios

链接：https://iclr.cc/virtual/2025/poster/29346 abstract： Complex classification scenarios, including long-tailed learning, domain adaptation, and transfer learning, present substantial challenges for traditional algorithms. Conditional class probability

(CCP) predictions have recently become critical components of many state-of-the-art algorithms designed to address these challenging scenarios. Among kernel methods, kernel logistic regression (KLR) is distinguished by its effectiveness in predicting CCPs through the minimization of the cross-entropy (CE) loss. Despite the empirical success of CCP-based approaches, the theoretical understanding of their performance, particularly regarding the CE loss, remains limited. In this paper, we bridge this gap by demonstrating that KLR-based algorithms achieve minimax optimal convergence rates for the CE loss under mild assumptions in these complex tasks, thereby establishing their theoretical efficiency in such demanding contexts.

961. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models

链接：https://iclr.cc/virtual/2025/poster/30620 abstract： Recent advancements in Large Language Models (LLMs) have sparked interest in their mathematical reasoning capabilities. While performance on the widely popular GSM8K benchmark has improved, questions remain about whether reported evaluation metrics are reliable, and reasoning abilities of LLMs have advanced. To overcome the limitations of existing evaluations, we introduce GSM-Symbolic, an improved benchmark created from symbolic templates that allow for the generation of a diverse set of questions. GSM-Symbolic enables more controllable evaluations, providing key insights and more reliable metrics for measuring the reasoning capabilities of models. Our findings reveal that LLMs exhibit noticeable variance when responding to different instantiations of the same question. Specifically, the performance of models declines when only the numerical values in the question are altered in the GSM-Symbolic benchmark. Furthermore, we investigate the fragility of mathematical reasoning in these models and demonstrate that their performance significantly deteriorates as the number of clauses in a question increases. We hypothesize that this decline is due to the fact that current LLMs are not capable of genuine logical reasoning; instead, they attempt to replicate the reasoning steps observed in their training data. When we add a single clause that appears relevant to the question, we observe significant performance drops (up to 65%) across all state-of-the-art models, even though the added clause does not contribute to the reasoning chain needed to reach the final answer. Overall, our work provides a more nuanced understanding of LLMs' capabilities and limitations in mathematical reasoning.

962. Privately Counting Partially Ordered Data

链接：https://iclr.cc/virtual/2025/poster/28757 abstract： We consider differentially private counting when each data point consists of d bits satisfying a partial order. Our main technical contribution is a problem-specific K-norm mechanism that runs in time $O(d^2)$. Experiments show that, depending on the partial order in question, our solution dominates existing pure differentially private mechanisms and can reduce their error by an order of magnitude or more.

963. CONGO: Compressive Online Gradient Optimization

链接：https://iclr.cc/virtual/2025/poster/31029 abstract： We address the challenge of zeroth-order online convex optimization where the objective function's gradient exhibits sparsity, indicating that only a small number of dimensions possess non-zero gradients. Our aim is to leverage this sparsity to obtain useful estimates of the objective function's gradient even when the only information available is a limited number of function samples. Our motivation stems from the optimization of large-scale queueing networks that process time-sensitive jobs. Here, a job must be processed by potentially many queues in sequence to produce an output, and the service time at any queue is a function of the resources allocated to that queue. Since resources are costly, the end-to-end latency for jobs must be balanced with the overall cost of the resources used. While the number of queues is substantial, the latency function primarily reacts to resource changes in only a few, rendering the gradient sparse. We tackle this problem by introducing the Compressive Online Gradient Optimization framework which allows compressive sensing methods previously applied to stochastic optimization to achieve regret bounds with an optimal dependence on the time horizon without the full problem dimension appearing in the bound. For specific algorithms, we reduce the samples required per gradient estimate to scale with the gradient's sparsity factor rather than its full dimensionality. Numerical simulations and real-world microservices benchmarks demonstrate CONGO's superiority over gradient descent approaches that do not account for sparsity.

964. EIA: ENVIRONMENTAL INJECTION ATTACK ON GENERALIST WEB AGENTS FOR PRIVACY LEAKAGE

链接：https://iclr.cc/virtual/2025/poster/27789 abstract： Recently, generalist web agents have demonstrated remarkable potential in autonomously completing a wide range of tasks on real websites, significantly boosting human productivity. However, web tasks, such as booking flights, usually involve users' personally identifiable information (PII), which may be exposed to potential privacy risks if web agents accidentally interact with compromised websites—a scenario that remains largely unexplored in the literature.In this work, we narrow this gap by conducting the first study on the privacy risks of generalist web agents in adversarial environments. First, we present a realistic threat model for attacks on the website, where we consider two adversarial targets: stealing users' specific PII or the entire user request.Then, we propose a novel attack method, termed Environmental Injection Attack (EIA). EIA injects malicious content designed to adapt well to environments where the agents operate and our work instantiates EIA specifically for privacy scenarios in web environments.We collect 177 action steps that involve diverse PII categories on realistic websites from the Mind2Web dataset, and conduct experiments using one of the most capable generalist web agent frameworks to date. The results demonstrate that EIA achieves up to 70\% attack success rate (ASR) in stealing users' specific PII and 16\% ASR in stealing a full user request at an action step. Additionally, by evaluating the

detectability and testing defensive system prompts, we indicate that EIA is challenging to detect and mitigate.Notably, attacks that are not well adapted for a webpage can be detected through careful human inspection, leading to our discussion about the trade-off between security and autonomy. However, extra attackers' efforts can make EIA seamlessly adapted, rendering such human supervision ineffective. Thus, we further discuss the implications on defenses at the pre- and post-deployment stages of the websites without relying on human supervision and call for more advanced defense strategies.

965. Graph Neural Networks for Edge Signals: Orientation Equivariance and Invariance

链接：https://iclr.cc/virtual/2025/poster/29303 abstract： Many applications in traffic, civil engineering, or electrical engineering revolve around edge-level signals. Such signals can be categorized as inherently directed, for example, the water flow in a pipe network, and undirected, like the diameter of a pipe. Topological methods model edge signals with inherent direction by representing them relative to a so-called orientation assigned to each edge. They can neither model undirected edge signals nor distinguish if an edge itself is directed or undirected. We address these shortcomings by (i) revising the notion of orientation equivariance to enable edge direction-aware topological models, (ii) proposing orientation invariance as an additional requirement to describe signals without inherent direction, and (iii) developing EIGN, an architecture composed of novel direction-aware edge-level graph shift operators, that provably fulfils the aforementioned desiderata. It is the first work that discusses modeling directed and undirected signals while distinguishing between directed and undirected edges. A comprehensive evaluation shows that EIGN outperforms prior work in edge-level tasks, improving in RMSE on flow simulation tasks by up to 23.5%.

966. SAM 2: Segment Anything in Images and Videos

链接：https://iclr.cc/virtual/2025/poster/30219 abstract： We present Segment Anything Model 2 (SAM 2), a foundation model towards solving promptable visual segmentation in images and videos. We build a data engine, which improves model and data via user interaction, to collect the largest video segmentation dataset to date. Our model is a simple transformer architecture with streaming memory for real-time video processing. SAM 2 trained on our data provides strong performance across a wide range of tasks. In video segmentation, we observe better accuracy, using 3x fewer interactions than prior approaches. In image segmentation, our model is more accurate and 6x faster than the Segment Anything Model (SAM). We believe that our data, model, and insights will serve as a significant milestone for video segmentation and related perception tasks. We are releasing our main model, the dataset, an interactive demo and code.

967. Extreme Risk Mitigation in Reinforcement Learning using Extreme Value Theory

链接：https://iclr.cc/virtual/2025/poster/31490 abstract： Risk-sensitive reinforcement learning (RL) has garnered significant attention in recent years due to the growing interest in deploying RL agents in real-world scenarios. A critical aspect of risk awareness involves modelling highly rare risk events (rewards) that could potentially lead to catastrophic outcomes. These infrequent occurrences present a formidable challenge for data-driven methods aiming to capture such risky events accurately. While risk-aware RL techniques do exist, they suffer from high variance estimation due to the inherent data scarcity. Our work proposes to enhance the resilience of RL agents when faced with very rare and risky events by focusing on refining the predictions of the extreme values predicted by the state-action value distribution. To achieve this, we formulate the extreme values of the state-action value function distribution as parameterized distributions, drawing inspiration from the principles of extreme value theory (EVT). We propose an extreme value theory based actor-critic approach, namely, Extreme Valued Actor-Critic (EVAC) which effectively addresses the issue of infrequent occurrence by leveraging EVT-based parameterization. Importantly, we theoretically demonstrate the advantages of employing these parameterized distributions in contrast to other risk-averse algorithms. Our evaluations show that the proposed method outperforms other risk averse RL algorithms on a diverse range of benchmark tasks, each encompassing distinct risk scenarios.

968. Systems with Switching Causal Relations: A Meta-Causal Perspective

链接：https://iclr.cc/virtual/2025/poster/30125 abstract： Most work on causality in machine learning assumes that causal relationships are driven by a constant underlying process. However, the flexibility of agents' actions or tipping points in the environmental process can change the qualitative dynamics of the system. As a result, new causal relationships may emerge, while existing ones change or disappear, resulting in an altered causal graph. To analyze these qualitative changes on the causal graph, we propose the concept of meta-causal states, which groups classical causal models into clusters based on equivalent qualitative behavior and consolidates specific mechanism parameterizations. We demonstrate how meta-causal states can be inferred from observed agent behavior, and discuss potential methods for disentangling these states from unlabeled data. Finally, we direct our analysis towards the application of a dynamical system, showing that meta-causal states can also emerge from inherent system dynamics, and thus constitute more than a context-dependent framework in which mechanisms emerge only as a result of external factors.

969. Towards Scalable Topological Regularizers

链接：https://iclr.cc/virtual/2025/poster/30329 abstract： Latent space matching, which consists of matching distributions of features in latent space, is a crucial component for tasks such as adversarial attacks and defenses, domain adaptation, and generative modelling. Metrics for probability measures, such as Wasserstein and maximum mean discrepancy, are commonly used to quantify the differences between such distributions. However, these are often costly to compute, or do not appropriately take the geometric and topological features of the distributions into consideration. Persistent homology is a tool from topological data analysis which quantifies the multi-scale topological structure of point clouds, and has recently been used as a topological regularizer in learning tasks. However, computation costs preclude larger scale computations, and discontinuities in the gradient lead to unstable training behavior such as in adversarial tasks. We propose the use of principal persistence measures, based on computing the persistent homology of a large number of small subsamples, as a topological regularizer. We provide a parallelized GPU implementation of this regularizer, and prove that gradients are continuous for smooth densities. Furthermore, we demonstrate the efficacy of this regularizer on shape matching, image generation, and semi-supervised learning tasks, opening the door towards a scalable regularizer for topological features.

970. ToVE: Efficient Vision-Language Learning via Knowledge Transfer from Vision Experts

链接：https://iclr.cc/virtual/2025/poster/30410 abstract： Vision-language (VL) learning requires extensive visual perception capabilities, such as fine-grained object recognition and spatial perception. Recent works typically rely on training huge models on massive datasets to develop these capabilities. As a more efficient alternative, this paper proposes a new framework that Transfers the knowledge from a hub of Vision Experts (ToVE) for efficient VL learning, leveraging pre-trained vision expert models to promote visual perception capability. Specifically, building on a frozen CLIP image encoder that provides vision tokens for image-conditioned language generation, ToVE introduces a hub of multiple vision experts and a token-aware gating network that dynamically routes expert knowledge to vision tokens. In the transfer phase, we propose a "residual knowledge transfer" strategy, which not only preserves the generalizability of the vision tokens but also allows selective detachment of low-contributing experts to improve inference efficiency. Further, we explore to merge these expert knowledge to a single CLIP encoder, creating a knowledge-merged CLIP that produces more informative vision tokens without expert inference during deployment. Experiment results across various VL tasks demonstrate that the proposed ToVE achieves competitive performance with two orders of magnitude fewer training data.

971. ChartMoE: Mixture of Diversely Aligned Expert Connector for Chart Understanding

链接：https://iclr.cc/virtual/2025/poster/28378 abstract： Automatic chart understanding is crucial for content comprehension and document parsing. Multimodal Large Language Models (MLLMs) have demonstrated remarkable capabilities in chart understanding through domain-specific alignment and fine-tuning. However, current MLLMs still struggle to provide faithful data and reliable analysis only based on charts. To address it, we propose ChartMoE, which employs the Mixture of Expert (MoE) architecture to replace the traditional linear projector to bridge the modality gap. Specifically, we train several linear connectors through distinct alignment tasks, which are utilized as the foundational initialization parameters for different experts. Additionally, we introduce ChartMoE-Align, a dataset with nearly 1 million chart-table-JSON-code quadruples to conduct three alignment tasks (chart-table/JSON/code). Combined with the vanilla connector, we initialize different experts diversely and adopt high-quality knowledge learning to further refine the MoE connector and LLM parameters. Extensive experiments demonstrate the effectiveness of the MoE connector and our initialization strategy, e.g., ChartMoE improves the accuracy of the previous state-of-the-art from 80.48% to 84.64% on the ChartQA benchmark.

972. LeanAgent: Lifelong Learning for Formal Theorem Proving

链接：https://iclr.cc/virtual/2025/poster/29455 abstract： Large Language Models (LLMs) have been successful in mathematical reasoning tasks such as formal theorem proving when integrated with interactive proof assistants like Lean. Existing approaches involve training or fine-tuning an LLM on a specific dataset to perform well on particular domains, such as undergraduate-level mathematics. These methods struggle with generalizability to advanced mathematics. A fundamental limitation is that these approaches operate on static domains, failing to capture how mathematicians often work across multiple domains and projects simultaneously or cyclically. We present LeanAgent, a novel lifelong learning framework for formal theorem proving that continuously generalizes to and improves on ever-expanding mathematical knowledge without forgetting previously learned knowledge. LeanAgent introduces several key innovations, including a curriculum learning strategy that optimizes the learning trajectory in terms of mathematical difficulty, a dynamic database for efficient management of evolving mathematical knowledge, and progressive training to balance stability and plasticity. LeanAgent successfully generates formal proofs for 155 theorems across 23 diverse Lean repositories where formal proofs were previously missing, many from advanced mathematics. It performs significantly better than the static LLM baseline, proving challenging theorems in domains like abstract algebra and algebraic topology while showcasing a clear progression of learning from basic concepts to advanced topics. In addition, we analyze LeanAgent's superior performance on key lifelong learning metrics. LeanAgent achieves exceptional scores in stability and backward transfer, where learning new tasks improves performance on previously learned tasks. This emphasizes LeanAgent's continuous generalizability and improvement, explaining its superior theorem-proving performance.

973. CLOVER: Cross-Layer Orthogonal Vectors Pruning and Fine-Tuning

链接：https://iclr.cc/virtual/2025/poster/31331 abstract： The absorb operation utilized in DeepSeek, which merges Query-Key and Value-Output weight matrices during inference, significantly increases parameter count and computational overhead. We observe that these absorbed matrices inherently exhibit low-rank structures. Motivated by this insight, we introduce CLOVER (Cross-Layer Orthogonal Vectors), a method that factorizes these matrices into four head-wise orthogonal matrices and two sets of singular values without any loss of information. By eliminating redundant vectors, CLOVER reduces the encoder parameters in Whisper-large-v3 by 46.42% without requiring additional training. Moreover, by freezing singular vectors and fine-tuning only singular values, CLOVER enables efficient full-rank fine-tuning. When evaluated on eight commonsense reasoning tasks with LLaMA-2 7B, CLOVER surpasses existing SoTA methods—LoRA, DoRA, HiRA, and PiSSA—by 7.6%, 5.5%, 3.8%, and 0.7%, respectively.

974. Enhancing Graph Of Thought: Enhancing Prompts with LLM Rationales and Dynamic Temperature Control

链接：https://iclr.cc/virtual/2025/poster/28537 abstract： We introduce Enhancing Graph of Thoughts (EGoT), a method designed to enhance the performance of large language models (LLMs) on complex reasoning tasks. EGoT automates the process of generating accurate responses using given data and a base prompt. The process consists of several steps: It obtains an initial response from the answering node using the base prompt. Evaluation node evaluates the response and generates reasoning for it, utilizing the score's probabilities to enhance evaluation accuracy. The reasoning from both the answering node and the evaluation node is aggregated to identify the problem in the response. This aggregated reasoning is incorporated into the base prompt to obtain an enhanced response. These steps are organized in a graph architecture, where the final leaf nodes are merged to produce a final response. As the graph descends, the temperature is lowered using Cosine Annealing and scoring, to explore diverse responses with earlier nodes and to focus on precise responses with later nodes. The minimum temperature in Cosine Annealing is adjusted based on scoring, ensuring that nodes with low scores continue to explore diverse responses, while those with high scores confirm accurate responses. In sorting 256 elements using GPT-4o mini, EGoT performs 88.31\% accuracy, while GoT (Graph of Thoughts) achieves 84.37\% accuracy. In the frozen lake problem using GPT-4o, EGoT averages 0.55 jumps or falls into the hole, while ToT (Tree of Thoughts) averages 0.89.

975. GOFA: A Generative One-For-All Model for Joint Graph Language Modeling

链接：https://iclr.cc/virtual/2025/poster/28473 abstract：

976. Copyright-Protected Language Generation via Adaptive Model Fusion

链接：https://iclr.cc/virtual/2025/poster/28583 abstract： The risk of language models reproducing copyrighted material from their training data has led to the development of various protective measures. Among these, inference-time strategies that impose constraints via post-processing have shown promise in addressing the complexities of copyright regulation. However, they often incur prohibitive computational costs or suffer from performance trade-offs. To overcome these limitations, we introduce Copyright-Protecting Model Fusion (CP-Fuse), a novel approach that combines models trained on disjoint sets of copyrighted material during inference. In particular, CP-Fuse adaptively aggregates the model outputs to minimize the reproduction of copyrighted content, adhering to a crucial balancing property to prevent the regurgitation of memorized data. Through extensive experiments, we show that CP-Fuse significantly reduces the reproduction of protected material without compromising the quality of text and code generation. Moreover, its post-hoc nature allows seamless integration with other protective measures, further enhancing copyright safeguards. Lastly, we show that CP-Fuse is robust against common techniques for extracting training data.

977. Edge-aware Image Smoothing with Relative Wavelet Domain Representation

链接：https://iclr.cc/virtual/2025/poster/31252 abstract： Image smoothing is a fundamental technique in image processing, designed to eliminate perturbations and textures while preserving dominant structures. It plays a pivotal role in numerous high-level computer vision tasks. More recently, both traditional and deep learning-based smoothing methods have been developed. However, existing algorithms frequently encounter issues such as gradient reversals and halo artifacts. Furthermore, the smoothing strength of deep learning-based models, once trained, cannot be adjusted for adapting different complexity levels of textures. These limitations stem from the inability of previous approaches to achieve an optimal balance between smoothing intensity and edge preservation. Consequently, image smoothing while maintaining edge integrity remains a significant challenge. To address these challenges, we propose a novel edge-aware smoothing model that leverages a relative wavelet domain representation. Specifically, by employing wavelet transformation, we introduce a new measure, termed Relative Wavelet Domain Representation (RWDR), which effectively distinguishes between textures and structures. Additionally, we present an innovative edge-aware scale map that is incorporated into the adaptive bilateral filter, facilitating mutual guidance in the smoothing process. This paper provides complete theoretical derivations for solving the proposed non-convex optimization model. Extensive experiments substantiate that our method has a competitive superiority with previous algorithms in edge-preserving and artifact removal. Visual and numerical comparisons further validate the effectiveness and efficiency of our approach in several applications of image smoothing.

978. When does compositional structure yield compositional generalization? A kernel theory.

链接：https://iclr.cc/virtual/2025/poster/30345 abstract： Compositional generalization (the ability to respond correctly to novel combinations of familiar components) is thought to be a cornerstone of intelligent behavior. Compositionally structured (e.g. disentangled) representations support this ability; however, the conditions under which they are sufficient for the emergence of compositional generalization remain unclear. To address this gap, we present a theory of compositional generalization in kernel models with fixed, compositionally structured representations. This provides a tractable framework for characterizing the impact of training data statistics on generalization. We find that these models are limited to functions that assign values to each combination of components seen during training, and then sum up these values ("conjunction-wise additivity"). This imposes fundamental restrictions on the set of tasks compositionally structured kernel models can learn, in particular preventing them from transitively generalizing equivalence relations. Even for compositional tasks that they can learn in principle, we identify novel failure modes in compositional generalization (memorization leak and shortcut bias) that arise from biases in the training data. Finally, we empirically validate our theory, showing that it captures the behavior of deep neural networks (convolutional networks, residual networks, and Vision Transformers) trained on a set of compositional tasks with similarly structured data. Ultimately, this work examines how statistical structure in the training data can affect compositional generalization, with implications for how to identify and remedy failure modes in deep learning models.

979. When GNNs meet symmetry in ILPs: an orbit-based feature augmentation approach

链接：https://iclr.cc/virtual/2025/poster/27839 abstract： A common characteristic in integer linear programs (ILPs) is symmetry, allowing variables to be permuted without altering the underlying problem structure. Recently, GNNs have emerged as a promising approach for solving ILPs. However, a significant challenge arises when applying GNNs to ILPs with symmetry: classic GNN architectures struggle to differentiate between symmetric variables, which limits their predictive accuracy. In this work, we investigate the properties of permutation equivalence and invariance in GNNs, particularly in relation to the inherent symmetry of ILP formulations. We reveal that the interaction between these two factors contributes to the difficulty of distinguishing between symmetric variables.To address this challenge, we explore the potential of feature augmentation and propose several guiding principles for constructing augmented features. Building on these principles, we develop an orbit-based augmentation scheme that first groups symmetric variables and then samples augmented features for each group from a discrete uniform distribution. Empirical results demonstrate that our proposed approach significantly enhances both training efficiency and predictive performance.

980. Sail into the Headwind: Alignment via Robust Rewards and Dynamic Labels against Reward Hacking

链接：https://iclr.cc/virtual/2025/poster/30189 abstract： Aligning AI systems with human preferences typically suffers from the infamous reward hacking problem, where optimization of an imperfect reward model leads to undesired behaviors. In this paper, we investigate reward hacking in offline preference optimization, which aims to improve an initial model using a preference dataset. We identify two types of reward hacking stemming from statistical fluctuations in the dataset: Type I Reward Hacking due to subpar choices appearing more favorable, and Type II Reward Hacking due to decent choices appearing less desirable. We prove that many (mainstream or theoretical) preference optimization methods suffer from both types of reward hacking. To mitigate Type I Reward Hacking, we propose POWER, a new preference optimization method that combines Guiasu's weighted entropy with a robust reward maximization objective. POWER enjoys finite-sample guarantees under general function approximation, competing with the best covered policy in the data. To mitigate Type II Reward Hacking, we analyze the learning dynamics of preference optimization and develop a novel technique that dynamically updates preference labels toward certain "stationary labels", resulting in diminishing gradients for untrustworthy samples. Empirically, POWER with dynamic labels (DL) consistently outperforms state-of-the-art methods on alignment benchmarks, achieving improvements of up to 13.0 points on AlpacaEval 2 and 11.5 points on Arena-Hard over DPO, while also improving or maintaining performance on downstream tasks such as mathematical reasoning. Strong theoretical guarantees and empirical results demonstrate the promise of POWER-DL in mitigating reward hacking.

981. Programming Refusal with Conditional Activation Steering

链接：https://iclr.cc/virtual/2025/poster/29806 abstract： LLMs have shown remarkable capabilities, but precisely controlling their response behavior remains challenging.Existing activation steering methods alter LLM behavior indiscriminately, limiting their practical applicability in settings where selective responses are essential, such as content moderation or domain-specific assistants.In this paper, we propose Conditional Activation Steering (CAST), which analyzes LLM activation patterns during inference to selectively apply or withhold activation steering based on the input context.Our method is based on the observation that different categories of prompts activate distinct patterns in the model's hidden states.Using CAST, one can systematically control LLM behavior with rules like "if input is about hate speech or adult content, then refuse" or "if input is not about legal advice, then refuse."This allows for selective modification of responses to specific content while maintaining normal responses to other content, all without requiring weight optimization.We release an open-source implementation of our framework.

982. Mentored Learning: Improving Generalization and Convergence of Student Learner

链接：https://iclr.cc/virtual/2025/poster/31385 abstract： Student learners typically engage in an iterative process of actively updating its hypotheses, like active learning. While this behavior can be advantageous, there is an inherent risk of introducing mistakes through incremental updates including weak initialization, inaccurate or insignificant history states, resulting in expensive convergence cost. In this work, rather than solely monitoring the update of the learner's status, we propose monitoring the disagreement w.r.t. $\mathcal{F}^\mathcal{T}(\cdot)$ between the learner and teacher, and call this new paradigm `Mentored Learning'', which consists ofhow to teach' and `how to learn'. By actively incorporating feedback that deviates from the learner's current hypotheses, convergence will be much easier to analyze without strict assumptions on learner's historical status, then deriving tighter generalization bounds on error and label complexity. Formally, we introduce an approximately optimal teaching hypothesis, $h^\mathcal{T}$, incorporating a tighter slack term $\left(1+\mathcal{F}^{\mathcal{T}}(\widehat{h}_t)\right)\Delta_t$ to replace the typical $2\Delta_t$ used in hypothesis pruning. Theoretically, we demonstrate that, guided by this teaching hypothesis, the learner can converge to tighter generalization bounds on error and label complexity compared to non-educated learners who lack guidance from a teacher: 1) the generalization error upper bound can be reduced from $R(h^)+4\Delta_{T-1}$ to approximately $R(h^{\mathcal{T}})+2\Delta_{T-1}$, and 2) the label complexity upper bound can be decreased from $4 \theta\left(TR(h^{})+2O(\sqrt{T})\right)$ to approximately $2\theta\left(2TR(h^{\mathcal{T}})+3 O(\sqrt{T})\right)$. To adhere strictly to our assumption, self-improvement of teaching is proposed when $h^\mathcal{T}$ loosely approximates h^*. In the context of learning, we further consider two teaching scenarios: instructing a white-box and black-box learner. Experiments validate this teaching concept and demonstrate superior generalization performance compared to fundamental active learning strategies, such as IWAL, IWAL-D, etc.

983. Towards General-Purpose Model-Free Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/29666 abstract： Reinforcement learning (RL) promises a framework for near-universal problem-solving. In practice however, RL algorithms are often tailored to specific benchmarks, relying on carefully tuned hyperparameters and algorithmic choices. Recently, powerful model-based RL methods have shown impressive general results across benchmarks but come at the cost of increased complexity and slow run times, limiting their broader applicability. In this paper, we attempt to find a unifying model-free deep RL algorithm that can address a diverse class of domains and problem settings. To achieve this, we leverage model-based representations that approximately linearize the value function, taking advantage of the denser task objectives used by model-based RL while avoiding the costs associated with planning or simulated trajectories. We evaluate our algorithm, MR.Q, on a variety of common RL benchmarks with a single set of hyperparameters and show a competitive performance against domain-specific and general baselines, providing a concrete step towards building general-purpose model-free deep RL algorithms.

984. Dualformer: Controllable Fast and Slow Thinking by Learning with Randomized Reasoning Traces

链接：https://iclr.cc/virtual/2025/poster/29093 abstract： In cognition theory, human thinking is governed by two systems: the fast and intuitive System 1 and the slower but more deliberative System 2. Analogously, Large Language Models (LLMs) can operate in two reasoning modes: outputting only the solutions (\emph{fast mode}) or both the reasoning chain and the final solution (\emph{slow mode}). We present \dualformer, a single Transformer model that seamlessly integrates both the fast and slow reasoning modes by training on randomized reasoning traces, where different parts of the traces are strategically dropped during training. At inference time, \dualformer can be easily configured to execute in either fast or slow mode, or automatically decide which mode to engage (\emph{auto mode}). It outperforms baselines in both performance and computational efficiency across all three modes: \textbf{(1)} in slow mode, \dualformer achieves 97.6% optimal rate on unseen 30×30 maze tasks, surpassing the \searchformer baseline (93.3\%) trained on data with complete reasoning traces, with 45.5% fewer reasoning steps; \textbf{(2)} in fast mode, \dualformer achieves 80% optimal rate, significantly outperforming the Solution-Only model trained on solution-only data, which has an optimal rate of only 30\%; \textbf{(3)} in auto mode, \dualformer achieves 96.6% optimal rate with 59.9% fewer steps than \searchformer. For math reasoning problems, our techniques have also achieved improved performance with LLM fine-tuning, demonstrating its generalization beyond task-specific models. We open source our code at https://github.com/facebookresearch/dualformer.

985. Soft Merging of Experts with Adaptive Routing

链接：https://iclr.cc/virtual/2025/poster/31491 abstract： Neural networks that learn to route their inputs through different "expert" subnetworks provide a form of modularity that standard dense models lack. Despite their possible benefits, modular models with learned routing often underperform their parameter-matched dense counterparts as well as models that use non-learned heuristic routing strategies. In this paper, we hypothesize that these shortcomings stem from the gradient estimation techniques used to train modular models that use non-differentiable discrete routing decisions. To address this issue, we introduce $\textbf{S}$oft $\textbf{M}$erging of $\textbf{E}$xperts with $\textbf{A}$daptive $\textbf{R}$outing (SMEAR), which avoids discrete routing by using a single "merged" expert constructed via a weighted average of all of the experts' parameters. By routing activations through a single merged expert, SMEAR does not incur a significant increase in computational costs and enables standard gradient-based training. We empirically validate that models using SMEAR outperform models that route based on metadata or learn routing through gradient estimation. Furthermore, we provide qualitative analysis demonstrating that

the experts learned via SMEAR exhibit a significant amount of specialization.

986. Scaling Autonomous Agents via Automatic Reward Modeling And Planning

链接：https://iclr.cc/virtual/2025/poster/27820 abstract：

987. Scaling Laws for Adversarial Attacks on Language Model Activations and Tokens

链接：https://iclr.cc/virtual/2025/poster/29238 abstract： We explore a class of adversarial attacks targeting the activations of language models to derive upper-bound scaling laws on their attack susceptibility. By manipulating a relatively small subset of model activations, a, we demonstrate the ability to control the exact prediction of a significant number (in some cases up to 1000) of subsequent tokens t. We empirically verify a scaling law where the maximum number of target tokens predicted, t_max, depends linearly on the number of tokens a whose activations the attacker controls as $t_\mathrm{max} = \kappa a$. We find that the number of bits the attacker controls on the input to exert a single bit of control on the output (a property we call \textit{attack resistance χ}) is remarkably stable between ≈ 16 and ≈ 25 over orders of magnitude of model sizes and between model families. Compared to attacks directly on input tokens, attacks on activations are predictably much stronger, however, we identify a surprising regularity where one bit of input steered either via activations or via tokens is able to exert a surprisingly similar amount of control over the model predictions. This gives support for the hypothesis that adversarial attacks are a consequence of dimensionality mismatch between the input and output spaces. A practical implication of the ease of attacking language model activations instead of tokens is for multi-modal and selected retrieval models. By using language models as a controllable test-bed to study adversarial attacks, we explored input-output dimension regimes that are inaccessible in computer vision and greatly extended the empirical support for the dimensionality theory of adversarial attacks.

988. EC-DIT: Scaling Diffusion Transformers with Adaptive Expert-Choice Routing

链接：https://iclr.cc/virtual/2025/poster/29721 abstract： Diffusion transformers have been widely adopted for text-to-image synthesis. While scaling these models up to billions of parameters shows promise, the effectiveness of scaling beyond current sizes remains underexplored and challenging. By explicitly exploiting the computational heterogeneity of image generations, we develop a new family of Mixture-of-Experts (MoE) models (EC-DIT) for diffusion transformers with expert-choice routing. EC-DIT learns to adaptively optimize the compute allocated to understand the input texts and generate the respective image patches, enabling heterogeneous computation aligned with varying text-image complexities. This heterogeneity provides an efficient way of scaling EC-DIT up to 97 billion parameters and achieving significant improvements in training convergence, text-to-image alignment, and overall generation quality over dense models and conventional MoE models. Through extensive ablations, we show that EC-DIT demonstrates superior scalability and adaptive compute allocation by recognizing varying textual importance through end-to-end training. Notably, in text-to-image alignment evaluation, our largest models achieve a state-of-the-art GenEval score of 71.68% and still maintain competitive inference speed with intuitive interpretability.

989. TVNet: A Novel Time Series Analysis Method Based on Dynamic Convolution and 3D-Variation

链接：https://iclr.cc/virtual/2025/poster/29927 abstract： With the recent development and advancement of Transformer and MLP architectures, significant strides have been made in time series analysis. Conversely, the performance of Convolutional Neural Networks (CNNs) in time series analysis has fallen short of expectations, diminishing their potential for future applications. Our research aims to enhance the representational capacity of Convolutional Neural Networks (CNNs) in time series analysis by introducing novel perspectives and design innovations. To be specific, We introduce a novel time series reshaping technique that considers the inter-patch, intra-patch, and cross-variable dimensions. Consequently, we propose TVNet, a dynamic convolutional network leveraging a 3D perspective to employ time series analysis. TVNet retains the computational efficiency of CNNs and achieves state-of-the-art results in five key time series analysis tasks, offering a superior balance of efficiency and performance over the state-of-the-art Transformer-based and MLP-based models. Additionally, our findings suggest that TVNet exhibits enhanced transferability and robustness. Therefore, it provides a new perspective for applying CNN in advanced time series analysis tasks.

990. DiffGAD: A Diffusion-based Unsupervised Graph Anomaly Detector

链接：https://iclr.cc/virtual/2025/poster/30622 abstract： Graph Anomaly Detection (GAD) is crucial for identifying abnormal entities within networks, garnering significant attention across various fields. Traditional unsupervised methods, which decode encoded latent representations of unlabeled data with a reconstruction focus, often fail to capture critical discriminative content, leading to suboptimal anomaly detection.To address these challenges, we present a Diffusion-based Graph Anomaly Detector (DiffGAD). At the heart of DiffGAD is a novel latent space learning paradigm, meticulously designed to enhance the model's

proficiency by guiding it with discriminative content. This innovative approach leverages diffusion sampling to infuse the latent space with discriminative content and introduces a content-preservation mechanism that retains valuable information across different scales, significantly improving the model's adeptness at identifying anomalies with limited time and space complexity. Our comprehensive evaluation of DiffGAD, conducted on six real-world and large-scale datasets with various metrics, demonstrated its exceptional performance. Our code is available at https://github.com/fortunato-all/DiffGAD

991. The Hyperfitting Phenomenon: Sharpening and Stabilizing LLMs for Open-Ended Text Generation

链接：https://iclr.cc/virtual/2025/poster/30156 abstract： This paper introduces the counter-intuitive generalization results of overfitting pre-trained large language models (LLMs) on very small datasets. In the setting of open-ended text generation, it is well-documented that LLMs tend to generate repetitive and dull sequences, a phenomenon that is especially apparent when generating using greedy decoding. This issue persists even with state-of-the-art LLMs containing billions of parameters, trained via next-token prediction on large datasets. We find that by further fine-tuning these models to achieve a near-zero training loss on a small set of samples -- a process we refer to as hyperfitting -- the long-sequence generative capabilities are greatly enhanced.Greedy decoding with these Hyperfitted models even outperform Top-P sampling over long-sequences, both in terms of diversity and human preferences. This phenomenon extends to LLMs of various sizes, different domains, and even autoregressive image generation. We further find this phenomena to be distinctly different from that of Grokking and double descent. Surprisingly, our experiments indicate that hyperfitted models rarely fall into repeating sequences they were trained on, and even explicitly blocking these sequences results in high-quality output. All hyperfitted models produce extremely low-entropy predictions, often allocating nearly all probability to a single token.

992. Generative Adversarial Ranking Nets

链接：https://iclr.cc/virtual/2025/poster/31380 abstract： We propose a new adversarial training framework -- generative adversarial ranking networks (GARNet) to learn from user preferences among a list of samples so as to generate data meeting user-specific criteria. Verbosely, GARNet consists of two modules: a ranker and a generator. The generator fools the ranker to raise generated samples to the top; while the ranker learns to rank generated samples at the bottom. Meanwhile, the ranker learns to rank samples regarding the interested property by training with preferences collected on real samples. The adversarial ranking game between the ranker and the generator enables an alignment between the generated data distribution and the user-preferred data distribution with theoretical guarantees and empirical verification. Specifically, we first prove that when training with full preferences on a discrete property, the learned distribution of GARNet rigorously coincides with the distribution specified by the given score vector based on user preferences. The theoretical results are then extended to partial preferences on a discrete property and further generalized to preferences on a continuous property. Meanwhile, numerous experiments show that GARNet can retrieve the distribution of user-desired data based on full/partial preferences in terms of various interested properties (i.e., discrete/continuous property, single/multiple properties). Code is available at https://github.com/EvaFlower/GARNet.

993. ADMM for Nonconvex Optimization under Minimal Continuity Assumption

链接：https://iclr.cc/virtual/2025/poster/30287 abstract： This paper introduces a novel approach to solving multi-block nonconvex composite optimization problems through a proximal linearized Alternating Direction Method of Multipliers (ADMM). This method incorporates an Increasing Penalization and Decreasing Smoothing (IPDS) strategy. Distinguishing itself from existing ADMM-style algorithms, our approach (denoted IPDS-ADMM) imposes a less stringent condition, specifically requiring continuity in just one block of the objective function. IPDS-ADMM requires that the penalty increases and the smoothing parameter decreases, both at a controlled pace. When the associated linear operator is bijective, IPDS-ADMM uses an over-relaxation stepsize for faster convergence; however, when the linear operator is surjective, IPDS-ADMM uses an under-relaxation stepsize for global convergence. We devise a novel potential function to facilitate our convergence analysis and prove an oracle complexity $\mathcal{O}(\epsilon^{-3})$ to achieve an ϵ-approximate critical point. To the best of our knowledge, this is the first complexity result for using ADMM to solve this class of nonsmooth nonconvex problems. Finally, some experiments on the sparse PCA problem are conducted to demonstrate the effectiveness of our approach.

994. Leveraging Sub-Optimal Data for Human-in-the-Loop Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/30457 abstract： To create useful reinforcement learning (RL) agents, step zero is to design a suitable reward function that captures the nuances of the task. However, reward engineering can be a difficult and time-consuming process. Instead, human-in-the-loop RL methods hold the promise of learning reward functions from human feedback. Despite recent successes, many of the human-in-the-loop RL methods still require numerous human interactions to learn successful reward functions.To improve the feedback efficiency of human-in-the-loop RL methods (i.e., require less human interaction), this paper introduces Sub-optimal Data Pre-training, SDP, an approach that leverages reward-free, sub-optimal data to improve scalar- and preference-based RL algorithms. In SDP, we start by pseudo-labeling all low-quality data with the minimum environment reward. Through this process, we obtain reward labels to pre-train our reward model without requiring

human labeling or preferences. This pre-training phase provides the reward model a head start in learning, enabling it to recognize that low-quality transitions should be assigned low rewards. Through extensive experiments with both simulated and human teachers, we find that SDP can at least meet, but often significantly improve, state of the art human-in-the-loop RL performance across a variety of simulated robotic tasks.

995. Root Cause Analysis of Anomalies in Multivariate Time Series through Granger Causal Discovery

链接：https://iclr.cc/virtual/2025/poster/28602 abstract： Identifying the root causes of anomalies in multivariate time series is challenging due to the complex dependencies among the series. In this paper, we propose a comprehensive approach called AERCA that inherently integrates Granger causal discovery with root cause analysis. By defining anomalies as interventions on the exogenous variables of time series, AERCA not only learns the Granger causality among time series but also explicitly models the distributions of exogenous variables under normal conditions. AERCA then identifies the root causes of anomalies by highlighting exogenous variables that significantly deviate from their normal states. Experiments on multiple synthetic and real-world datasets demonstrate that AERCA can accurately capture the causal relationships among time series and effectively identify the root causes of anomalies.

996. UniWav: Towards Unified Pre-training for Speech Representation Learning and Generation

链接：https://iclr.cc/virtual/2025/poster/27705 abstract： Pre-training and representation learning have been playing an increasingly important role in modern speech processing. Nevertheless, different applications have been relying on different foundation models, since predominant pre-training techniques are either designed for discriminative tasks or generative tasks. In this work, we make the first attempt at building a unified pre-training framework for both types of tasks in speech. We show that with the appropriate design choices for pre-training, one can jointly learn a representation encoder and generative audio decoder that can be applied to both types of tasks. We propose UniWav, an encoder-decoder framework designed to unify pre-training representation learning and generative tasks. On speech recognition, text-to-speech, and speech tokenization, UniWav achieves comparable performance to different existing foundation models, each trained on a specific task. Our findings suggest that a single general-purpose foundation model for speech can be built to replace different foundation models, reducing the overhead and cost of pre-training.

997. Provable unlearning in topic modeling and downstream tasks

链接：https://iclr.cc/virtual/2025/poster/28973 abstract： Machine unlearning algorithms are increasingly important as legal concerns arise around the provenance of training data, but verifying the success of unlearning is often difficult. Provable guarantees for unlearning are often limited to supervised learning settings. In this paper, we provide the first theoretical guarantees for unlearning in the pre-training and fine-tuning paradigm by studying topic models, simple bag-of-words language models that can be adapted to solve downstream tasks like retrieval and classification. First, we design a provably effective unlearning algorithm for topic models that incurs a computational overhead independent of the size of the original dataset. Our analysis additionally quantifies the deletion capacity of the model -- i.e., the number of examples that can be unlearned without incurring a significant cost in model performance. Finally, we formally extend our analyses to account for adaptation to a given downstream task. In particular, we design an efficient algorithm to perform unlearning after fine-tuning the topic model via a linear head. Notably, we show that it is easier to unlearn pre-training data from models that have been fine-tuned to a particular task, and one can unlearn this data without modifying the base model.

998. Object-Centric Pretraining via Target Encoder Bootstrapping

链接：https://iclr.cc/virtual/2025/poster/30811 abstract： Object-centric representation learning has recently been successfully applied to real-world datasets. This success can be attributed to pretrained non-object-centric foundation models, whose features serve as reconstruction targets for slot attention. However, targets must remain frozen throughout the training, which sets an upper bound on the performance object-centric models can attain. Attempts to update the target encoder by bootstrapping result in large performance drops, which can be attributed to its lack of object-centric inductive biases, causing the object-centric model’s encoder to drift away from representations useful as reconstruction targets. To address these limitations, we propose Object-CEntric Pretraining by Target Encoder BOotstrapping, a self-distillation setup for training object-centric models from scratch, on real-world data, for the first time ever. In OCEBO, the target encoder is updated as an exponential moving average of the object-centric model, thus explicitly being enriched with object-centric inductive biases introduced by slot attention while removing the upper bound on performance present in other models. We mitigate the slot collapse caused by random initialization of the target encoder by introducing a novel cross-view patch filtering approach that limits the supervision to sufficiently informative patches. When pretrained on 241k images from COCO, OCEBO achieves unsupervised object discovery performance comparable to that of object-centric models with frozen non-object-centric target encoders pretrained on hundreds of millions of images. The code and pretrained models are publicly available at https://github.com/djukicn/ocebo.

999. A Simple Framework for Open-Vocabulary Zero-Shot Segmentation

链接：https://iclr.cc/virtual/2025/poster/29668 abstract： Zero-shot classification capabilities naturally arise in models trained within a vision-language contrastive framework. Despite their classification prowess, these models struggle in dense tasks like zero-shot open-vocabulary segmentation. This deficiency is often attributed to the absence of localization cues in captions and the intertwined nature of the learning process, which encompasses both image/text representation learning and cross-modality alignment. To tackle these issues, we propose SimZSS, a $\textbf{Sim}$ple framework for open-vocabulary $\textbf{Z}$ero-$\textbf{S}$hot $\textbf{S}$egmentation. The method is founded on two key principles: i) leveraging frozen vision-only models that exhibit spatial awareness while exclusively aligning the text encoder and ii) exploiting the discrete nature of text and linguistic knowledge to pinpoint local concepts within captions. By capitalizing on the quality of the visual representations, our method requires only image-caption pair datasets and adapts to both small curated and large-scale noisy datasets. When trained on COCO Captions across 8 GPUs, SimZSS achieves state-of-the-art results on 7 out of 8 benchmark datasets in less than 15 minutes. Our code and pretrained models are publicly available at https://github.com/tileb1/simzss.

1000. LLaVA-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models

链接：https://iclr.cc/virtual/2025/poster/28357 abstract： Visual instruction tuning has made considerable strides in enhancing the capabilities of Large Multimodal Models (LMMs). However, existing open LMMs largely focus on single-image tasks, their applications to multi-image scenarios remains less explored. Additionally, prior LMM research separately tackles different scenarios, leaving it impossible to generalize cross scenarios with newemerging capabilities. To this end, we introduce LLaVA-Interleave, which simultaneously tackles Multi-image, Multi-frame (video), Multi-view (3D), and Multi-patch (single-image) scenarios in LMMs. To enable these capabilities, we regard the interleaved data format as a general template and compile the M4-Instruct dataset with 1,177.6k samples, spanning 4 primary domains with 14tasks and 41 datasets. We also curate the LLaVA-Interleave Bench to comprehensively evaluate the multi-image performance of LMMs. Through extensiveexperiments, LLaVA-Interleave achieves leading results in multi-image, video,and 3D benchmarks, while maintaining the performance of single-image tasks.Besides, our model also exhibits several emerging capabilities, e.g., transferring tasks across different settings and modalities.