2201. A Rainbow in Deep Network Black Boxes

链接：https://iclr.cc/virtual/2025/poster/31381 abstract： A central question in deep learning is to understand the functions learned by deep networks. What is their approximation class? Do the learned weights and representations depend on initialization? Previous empirical work has evidenced that kernels defined by network activations are similar across initializations. For shallow networks, this has been theoretically studied with random feature models, but an extension to deep networks has remained elusive. Here, we provide a deep extension of such random feature models, which we call the rainbow model. We prove that rainbow networks define deterministic (hierarchical) kernels in the infinite-width limit. The resulting functions thus belong to a data-dependent RKHS which does not depend on the weight randomness. We also verify numerically our modeling assumptions on deep CNNs trained on image classification tasks, and show that the trained networks approximately satisfy the rainbow hypothesis. In particular, rainbow networks sampled from the corresponding random feature model achieve similar performance as the trained networks. Our results highlight the central role played by the covariances of network weights at each layer, which are observed to be low-rank as a result of feature learning.

2202. Large Scale Knowledge Washing

链接：https://iclr.cc/virtual/2025/poster/28979 abstract： Large language models show impressive abilities in memorizing world knowledge, which leads to concerns regarding memorization of private information, toxic or sensitive knowledge, and copyrighted content. We introduce the problem of Large Scale Knowledge Washing, focusing on unlearning an extensive amount of factual knowledge. Previous unlearning methods usually define the reverse loss and update the model via backpropagation, which may affect the model's fluency and reasoning ability or even destroy the model due to extensive training with the reverse loss. Existing works introduce additional data from downstream tasks to prevent the model from losing capabilities, which requires downstream task awareness. Controlling the tradeoff of unlearning existing knowledge while maintaining existing capabilities is also challenging. To this end, we propose LaW (Large Scale Washing), where we update the MLP layers in decoder-only large language models to perform knowledge washing, as inspired by model editing methods. We derive a new objective with the knowledge to be unlearned to update the weights of certain MLP layers. Experimental results demonstrate the effectiveness of LaW in forgetting target knowledge while maximally maintaining reasoning ability. The code will be open-sourced.

2203. The Effectiveness of Curvature-Based Rewiring and the Role of Hyperparameters in GNNs Revisited

链接：https://iclr.cc/virtual/2025/poster/30396 abstract： Message passing is the dominant paradigm in Graph Neural Networks (GNNs). The efficiency of message passing, however, can be limited by the topology of the graph. This happens when information is lost during propagation due to being oversquashed when travelling through bottlenecks. To remedy this, recent efforts have focused on graph rewiring techniques, which disconnect the input graph originating from the data and the computational graph, on which message passing is performed. A prominent approach for this is to use discrete graph curvature measures, of which several variants have been proposed, to identify and rewire around bottlenecks, facilitating information propagation. While oversquashing has been demonstrated in synthetic datasets, in this work we reevaluate the performance gains that curvature-based rewiring brings to real-world datasets. We show that in these datasets, edges selected during the rewiring process are not in line with theoretical criteria identifying bottlenecks. This implies they do not necessarily oversquash information during message passing. Subsequently, we demonstrate that SOTA accuracies on these datasets are outliers originating from sweeps of hyperparameters—both the ones for training and dedicated ones related to the rewiring algorithm—instead of consistent performance gains. In conclusion, our analysis nuances the effectiveness of curvature-based rewiring in real-world datasets and brings a new perspective on the methods to evaluate GNN accuracy improvements.

2204. Building Interactable Replicas of Complex Articulated Objects via Gaussian Splatting

链接：https://iclr.cc/virtual/2025/poster/28664 abstract： Building interactable replicas of articulated objects is a key challenge in computer vision. Existing methods often fail to effectively integrate information across different object states, limiting the accuracy of part-mesh reconstruction and part dynamics modeling, particularly for complex multi-part articulated objects. We introduce ArtGS, a novel approach that leverages 3D Gaussians as a flexible and efficient representation to address these issues. Our method incorporates canonical Gaussians with coarse-to-fine initialization and updates for aligning articulated part information across different object states, and employs a skinning-inspired part dynamics modeling module to improve both part-mesh reconstruction and articulation learning. Extensive experiments on both synthetic and real-world datasets, including a new benchmark for complex multi-part objects, demonstrate that ArtGS achieves state-of-the-art performance in joint parameter estimation and part mesh reconstruction. Our approach significantly improves reconstruction quality and efficiency, especially for multi-part articulated objects. Additionally, we provide comprehensive analyses of our design choices, validating the effectiveness of each component to highlight potential areas for future improvement.

2205. NUDGE: Lightweight Non-Parametric Fine-Tuning of Embeddings for Retrieval

链接：https://iclr.cc/virtual/2025/poster/29928 abstract： k-Nearest Neighbor search on dense vector embeddings (k-NN retrieval) from pre-trained embedding models is the predominant retrieval method for text and images, as well as Retrieval-Augmented Generation (RAG) pipelines. In practice, application developers often fine-tune the embeddings to improve their accuracy on the dataset and query workload in hand. Existing approaches either fine-tune the pre-trained model itself or, more efficiently, but at the cost of accuracy, train adaptor models to transform the output of the pre-trained model. We present NUDGE, a family of novel *non-parametric* embedding fine-tuning approaches that are significantly more accurate and efficient than both sets of existing approaches. NUDGE directly modifies the embeddings of data records to maximize the accuracy of k-NN retrieval. We present a thorough theoretical and experimental study of NUDGE's non-parametric approach. We show that even though the underlying problem is NP-Hard, constrained variations can be solved efficiently. These constraints additionally ensure that the changes to the embeddings are modest, avoiding large distortions to the semantics learned during pre-training. In experiments across five pre-trained models and nine standard text and image retrieval datasets, *NUDGE runs in minutes and often improves NDCG@10 by more than 10\% over existing fine-tuning methods. On average, NUDGE provides 3.3$\times$ and 4.3$\times$ higher increase in accuracy and runs 200$\times$ and 3$\times$ faster, respectively, over fine-tuning the pre-trained model and training adaptors.*

2206. ConFIG: Towards Conflict-free Training of Physics Informed Neural Networks

链接：https://iclr.cc/virtual/2025/poster/30634 abstract： The loss functions of many learning problems contain multiple additive terms that can disagree and yield conflicting update directions. For Physics-Informed Neural Networks (PINNs), loss terms on initial/boundary conditions and physics equations are particularly interesting as they are well-established as highly difficult tasks. To improve learning the challenging multi-objective task posed by PINNs, we propose the ConFIG method, which provides conflict-free updates by ensuring a positive dot product between the final update and each loss-specific gradient. It also maintains consistent optimization rates for all loss terms and dynamically adjusts gradient magnitudes based on conflict levels. We additionally leverage momentum to accelerate optimizations by alternating the back-propagation of different loss terms. We provide a mathematical proof showing the convergence of the ConFIG method, and it is evaluated across a range of challenging PINN scenarios. ConFIG consistently shows superior performance and runtime compared to baseline methods. We also test the proposed method in a classic multi-task benchmark, where the ConFIG method likewise exhibits a highly promising performance. Source code is available at https://tum-pbs.github.io/ConFIG

2207. Multi-Dimensional Conformal Prediction

链接：https://iclr.cc/virtual/2025/poster/28498 abstract： Conformal prediction has attracted significant attention as a distribution-free method for uncertainty quantification in black-box models, providing prediction sets with guaranteed coverage. However, its practical utility is often limited when these prediction sets become excessively large, reducing its overall effectiveness. In this paper, we introduce a novel approach to conformal prediction for classification problems, which leverages a multi-dimensional nonconformity score. By extending standard conformal prediction to higher dimensions, we achieve better separation between correct and incorrect labels. Utilizing this we can focus on regions with low concentrations of incorrect labels, leading to smaller, more informative prediction sets. To efficiently generate the multi-dimensional score, we employ a self-ensembling technique that trains multiple diverse classification heads on top of a backbone model. We demonstrate the advantage of our approach compared to baselines across different benchmarks.

2208. MorphoDiff: Cellular Morphology Painting with Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/29726 abstract： Understanding cellular responses to external stimuli is critical for parsing biological mechanisms and advancing therapeutic development. High-content image-based assays provide a cost-effective approach to examine cellular phenotypes induced by diverse interventions, which offers valuable insights into biological processes and cellular states. We introduce MorphoDiff, a generative pipeline to predict high-resolution cell morphological responses under different conditions based on perturbation encoding. To the best of our knowledge, MorphoDiff is the first framework capable of producing guided, high-resolution predictions of cell morphology that generalize across both chemical and genetic interventions. The model integrates perturbation embeddings as guiding signals within a 2D latent diffusion model. The comprehensive computational, biological, and visual validations across three open-source Cell Painting datasets show that MorphoDiff can generate high-fidelity images and produce meaningful biology signals under various interventions. We envision the model will facilitate efficient in silico exploration of perturbational landscapes towards more effective drug discovery studies.

2209. Generalization v.s. Memorization: Tracing Language Models' Capabilities Back to Pretraining Data

链接：https://iclr.cc/virtual/2025/poster/30169 abstract： The impressive capabilities of large language models (LLMs) have sparked debate over whether these models genuinely generalize to unseen tasks or predominantly rely on memorizing vast amounts of pretraining data. To explore this issue, we introduce an extended concept of memorization, distributional memorization, which measures the correlation between the LLM output probabilities and the pretraining data frequency. To effectively capture task-specific pretraining data frequency, we propose a novel task-gram language model, which is built by counting the co-occurrence of semantically related n-gram pairs from task inputs and outputs in the pretraining corpus. Using the Pythia models trained on the Pile dataset, we evaluate four distinct tasks: machine translation, factual question answering, world knowledge understanding, and math reasoning. Our findings reveal varying levels of memorization, with the strongest effect observed in factual question answering. Furthermore, while model performance improves across all tasks as LLM size increases, only factual question answering shows an increase in memorization, whereas machine translation and reasoning tasks exhibit greater generalization, producing more novel outputs. This study demonstrates that memorization plays a larger role in simpler, knowledge-intensive tasks, while generalization is the key for harder, reasoning-based tasks, providing a scalable method for analyzing large pretraining corpora in greater depth.

2210. On Linear Representations and Pretraining Data Frequency in Language Models

链接：https://iclr.cc/virtual/2025/poster/30418 abstract： Pretraining data has a direct impact on the behaviors and quality of language models (LMs), but we only understand the most basic principles of this relationship. While most work focuses on pretraining data's effect on downstream task behavior, we investigate its relationship to LM representations. Previous work has discovered that, in language models, some concepts are encoded "linearly" in the representations, but what factors cause these representations to form (or not)? We study the connection between pretraining data frequency and models' linear representations of factual relations (e.g., mapping France to Paris in a capital prediction task). We find evidence that the formation of linear representations is strongly connected to pretraining term frequencies; specifically for subject-relation-object fact triplets, both subject-object co-occurrence frequency and in-context learning accuracy for the relation are highly correlated with linear representations. This is the case across all phases of pretraining, i.e., it is not affected by the model's underlying capability. In OLMo-7B and GPT-J (6B), we discover that a linear representation consistently (but not exclusively) forms when the subjects and objects within a relation co-occur at least 1k and 2k times, respectively, regardless of when these occurrences happen during pretraining (and around 4k times for OLMo-1B). Finally, we train a regression model on measurements of linear representation quality in fully-trained LMs that can predict how often a term was seen in pretraining. Our model achieves low error even on inputs from a different model with a different pretraining dataset, providing a new method for estimating properties of the otherwise-unknown training data of closed-data models. We conclude that the strength of linear representations in LMs contains signal about the models' pretraining corpora that may provide new avenues for controlling and improving model behavior: particularly, manipulating the models' training data to meet specific frequency thresholds. We release our code to support future work.

2211. PN-GAIL: Leveraging Non-optimal Information from Imperfect Demonstrations

链接：https://iclr.cc/virtual/2025/poster/31246 abstract： Imitation learning aims at constructing an optimal policy by emulating expert demonstrations. However, the prevailing approaches in this domain typically presume that the demonstrations are optimal, an assumption that seldom holds true in the complexities of real-world applications. The data collected in practical scenarios often contains imperfections, encompassing both optimal and non-optimal examples. In this study, we propose Positive-Negative Generative Adversarial Imitation Learning (PN-GAIL), a novel approach that falls within the framework of Generative Adversarial Imitation Learning (GAIL). PN-GAIL innovatively leverages non-optimal information from imperfect demonstrations, allowing the discriminator to comprehensively assess the positive and negative risks associated with these demonstrations. Furthermore, it requires only a small subset of labeled confidence scores. Theoretical analysis indicates that PN-GAIL deviates from the non-optimal data while mimicking imperfect demonstrations. Experimental results demonstrate that PN-GAIL surpasses conventional baseline methods in dealing with imperfect demonstrations, thereby significantly augmenting the practical utility of imitation learning in real-world contexts. Our codes are available at https://github.com/QiangLiuT/PN-GAIL.

2212. Exploring channel distinguishability in local neighborhoods of the model space in quantum neural networks

链接：https://iclr.cc/virtual/2025/poster/28831 abstract： With the increasing interest in Quantum Machine Learning, Quantum Neural Networks (QNNs) have emerged and gained significant attention. These models have, however, been shown to be notoriously difficult to train, which we hypothesize is partially due to the architectures, called ansatzes, that are hardly studied at this point. Therefore, in this paper, we take a step back and analyze ansatzes. We initially consider their expressivity, i.e., the space of operations they are able to express, and show that the closeness to being a 2-design, the primarily used measure, fails at capturing this property. Hence, we look for alternative ways to characterize ansatzes, unrelated to expressivity, by considering the local neighborhood of the model space, in particular, analyzing model distinguishability upon small perturbation of parameters. We derive an upper bound on their distinguishability, showcasing that QNNs using the Hardware Efficient Ansatz with few parameters are hardly discriminable upon update. Our numerical experiments support our bounds and further indicate that there is a significant degree of variability, which stresses the need for warm-starting or clever initialization. Altogether, our work provides an ansatz-centric perspective on training dynamics and difficulties in QNNs, ultimately suggesting that iterative training of small quantum models may not be effective, which contrasts their initial motivation.

2213. Rethinking Graph Neural Networks From A Geometric Perspective Of Node Features

链接：https://iclr.cc/virtual/2025/poster/28532 abstract： Many works on graph neural networks (GNNs) focus on graph topologies and analyze graph-related operations to enhance performance on tasks such as node classification. In this paper, we propose to understand GNNs based on a feature-centric approach. Our main idea is to treat the features of nodes from each label class as a whole, from which we can identify the centroid. The convex hull of these centroids forms a simplex called the feature centroid simplex, where a simplex is a high-dimensional generalization of a triangle. We borrow ideas from coarse geometry to analyze the geometric properties of the feature centroid simplex by comparing them with basic geometric models, such as regular simplexes and degenerate simplexes. Such a simplex provides a simple platform to understand graph-based feature aggregation, including phenomena such as heterophily, oversmoothing, and feature re-shuffling. Based on the theory, we also identify simple and useful tricks for the node classification task.

2214. InfoGS: Efficient Structure-Aware 3D Gaussians via Lightweight Information Shaping

链接：https://iclr.cc/virtual/2025/poster/29735 abstract： 3D Gaussians, as an explicit scene representation, typically involve thousands to millions of elements per scene. This makes it challenging to control the scene in ways that reflect the underlying semantics, where the number of independent entities is typically much smaller. Especially, if one wants to animate or edit objects in the scene, as this requires coordination among the many Gaussians involved in representing each object. To address this issue, we develop a mutual information shaping technique that enforces resonance and coordination between correlated Gaussians via a Gaussian attribute decoding network. Such correlations can be learned from putative 2D object masks in different views. By approximating the mutual information with the gradients concerning the network parameters, our method ensures consistency between scene elements and enables efficient scene editing by operating on network parameters rather than massive Gaussians. In particular, we develop an effective learning pipeline named InfoGS with lightweight optimization to shape the attribute decoding network ,while ensuring that the shaping (consistency) is maintained during continuous edits, avoiding re-shaping after parameter changes. Notably, our training only touches a small fraction of all Gaussians in the scene yet attains the desired correlated behavior according to the underlying scene structure. The proposed technique is evaluated on challenging scenes and demonstrates significant performance improvements in 3D object segmentation and promoting scene interactions, while inducing low computation and memory requirements. Our code is available at: https://github.com/StylesZhang/InfoGS.

2215. GI-GS: Global Illumination Decomposition on Gaussian Splatting for Inverse Rendering

链接：https://iclr.cc/virtual/2025/poster/28771 abstract： We present GI-GS, a novel inverse rendering framework that leverages 3D Gaussian Splatting (3DGS) and deferred shading to achieve photo-realistic novel view synthesis and relighting. In inverse rendering, accurately modeling the shading processes of objects is essential for achieving high-fidelity results. Therefore, it is critical to incorporate global illumination to account for indirect lighting that reaches an object after multiple bounces across the scene. Previous 3DGS-based methods have attempted to model indirect lighting by characterizing indirect illumination as learnable lighting volumes or additional attributes of each Gaussian, while using baked occlusion to represent shadow effects. These methods, however, fail to accurately model the complex physical interactions between light and objects, making it impossible to construct realistic indirect illumination during relighting. To address this limitation, we propose to calculate indirect lighting using efficient path tracing with deferred shading. In our framework, we first render a G-buffer to capture the detailed geometry and material properties of the scene. Then, we perform physically-based rendering (PBR) only for direct lighting. With the G-buffer and previous rendering results, the indirect lighting can be calculated through a lightweight path tracing. Our method effectively models indirect lighting under any given lighting conditions, thereby achieving better novel view synthesis and competitive relighting. Quantitative and qualitative results show that our GI-GS outperforms existing baselines in both rendering quality and efficiency. Project page: https://stopaimme.github.io/GI-GS-site/.

2216. IV-mixed Sampler: Leveraging Image Diffusion Models for Enhanced Video Synthesis

链接：https://iclr.cc/virtual/2025/poster/30150 abstract： Exploring suitable solutions to improve performance by increasing the computational cost of inference in visual diffusion models is a highly promising direction. Sufficient prior studies have demonstrated that correctly scaling up computation in the sampling process can successfully lead to improved generation quality, enhanced image editing, and compositional generalization. While there have been rapid advancements in developing inference-heavy algorithms for improved image generation, relatively little work has explored inference scaling laws in video diffusion models (VDMs). Furthermore, existing research shows only minimal performance gains that are perceptible to the naked eye. To address this, we design a novel training-free algorithm IV-Mixed Sampler that leverages the strengths of image diffusion models (IDMs) to assist VDMs surpass their current capabilities. The core of IV-Mixed Sampler is to use IDMs to significantly enhance the quality of each video frame and VDMs ensure the temporal coherence of the video during the sampling process. Our experiments have demonstrated that IV-Mixed Sampler achieves state-of-the-art performance on 4 benchmarks including UCF-101-FVD, MSR-VTT-FVD, Chronomagic-Bench-150/1649, and VBench. For example, the open-source Animatediff with IV-Mixed Sampler reduces the UMT-FVD score from 275.2 to 228.6, closing to 223.1 from the closed-source Pika-2.0.

2217. Animate Your Thoughts: Reconstruction of Dynamic Natural Vision

from Human Brain Activity

链接：https://iclr.cc/virtual/2025/poster/30544 abstract： Reconstructing human dynamic vision from brain activity is a challenging task with great scientific significance. Although prior video reconstruction methods have made substantial progress, they still suffer from several limitations, including: (1) difficulty in simultaneously reconciling semantic (e.g. categorical descriptions), structure (e.g. size and color), and consistent motion information (e.g. order of frames); (2) low temporal resolution of fMRI, which poses a challenge in decoding multiple frames of video dynamics from a single fMRI frame; (3) reliance on video generation models, which introduces ambiguity regarding whether the dynamics observed in the reconstructed videos are genuinely derived from fMRI data or are hallucinations from generative model. To overcome these limitations, we propose a two-stage model named Mind-Animator. During the fMRI-to-feature stage, we decouple semantic, structure, and motion features from fMRI. Specifically, we employ fMRI-vision-language tri-modal contrastive learning to decode semantic feature from fMRI and design a sparse causal attention mechanism for decoding multi-frame video motion features through a next-frame-prediction task. In the feature-to-video stage, these features are integrated into videos using an inflated Stable Diffusion, effectively eliminating external video data interference. Extensive experiments on multiple video-fMRI datasets demonstrate that our model achieves state-of-the-art performance. Comprehensive visualization analyses further elucidate the interpretability of our model from a neurobiological perspective. Project page: https://mind-animator-design.github.io/.

2218. Unified Parameter-Efficient Unlearning for LLMs

链接：https://iclr.cc/virtual/2025/poster/27670 abstract： The advent of Large Language Models (LLMs) has revolutionized natural language processing, enabling advanced understanding and reasoning capabilities across a variety of tasks. Fine-tuning these models for specific domains, particularly through Parameter-Efficient Fine-Tuning (PEFT) strategies like LoRA, has become a prevalent practice due to its efficiency. However, this raises significant privacy and security concerns, as models may inadvertently retain and disseminate sensitive or undesirable information. To address these issues, we introduce a novel instance-wise unlearning framework, LLMEraser, which systematically categorizes unlearning tasks and applies precise parameter adjustments using influence functions. Unlike traditional unlearning techniques that are often limited in scope and require extensive retraining, LLMEraser is designed to handle a broad spectrum of unlearning tasks without compromising model performance. Extensive experiments on benchmark datasets demonstrate that LLMEraser excels in efficiently managing various unlearning scenarios while maintaining the overall integrity and efficacy of the models.

2219. Breaking Mental Set to Improve Reasoning through Diverse Multi-Agent Debate

链接：https://iclr.cc/virtual/2025/poster/28079 abstract： Large Language Models (LLMs) have seen significant progress but continue to struggle with persistent reasoning mistakes.Previous methods of self-reflection have been proven limited due to the models' inherent fixed thinking patterns. While Multi-Agent Debate (MAD) attempts to mitigate this by incorporating multiple agents, it often employs the same reasoning methods, even though assigning different personas to models. This leads to a "fixed mental set", where models rely on homogeneous thought processes without exploring alternative perspectives.In this paper, we introduce Diverse Multi-Agent Debate (DMAD), a method that encourages agents to think with distinct reasoning approaches. By leveraging diverse problem-solving strategies, each agent can gain insights from different perspectives, refining its responses through discussion and collectively arriving at the optimal solution. DMAD effectively breaks the limitations of fixed mental sets. We evaluate DMAD against various prompting techniques, including self-reflection and traditional MAD, across multiple benchmarks using both LLMs and Multimodal LLMs. Our experiments show that DMAD consistently outperforms other methods, delivering better results than MAD in fewer rounds. Code is available at https://github.com/MraDonkey/DMAD.

2220. Consistency Checks for Language Model Forecasters

链接：https://iclr.cc/virtual/2025/poster/28220 abstract： Forecasting is a task that is difficult to evaluate: the ground truth can only be known in the future. Recent work showing LLM forecasters rapidly approaching human-level performance begs the question: how can we benchmark and evaluate these forecasters instantaneously? Following the consistency check framework, we measure the performance of forecasters in terms of the consistency of their predictions on different logically-related questions. We propose a new, general consistency metric based on arbitrage: for example, if a forecasting AI illogically predicts that both the Democratic and Republican parties have 60\% probability of winning the 2024 US presidential election, an arbitrageur could trade against the forecaster's predictions and make a profit. We build an automated evaluation system that generates a set of base questions, instantiates consistency checks from these questions, elicits the predictions of the forecaster, and measures the consistency of the predictions. We then build a standard, proper-scoring-rule forecasting benchmark, and show that our (instantaneous) consistency metrics correlate strongly with LLM forecasters' ground truth Brier scores (which are only known in the future). We also release a consistency benchmark that resolves in 2028, providing a long-term evaluation tool for forecasting.

2221. Multi-objective Differentiable Neural Architecture Search

链接：https://iclr.cc/virtual/2025/poster/30667 abstract： Pareto front profiling in multi-objective optimization (MOO), i.e., finding a diverse set of Pareto optimal solutions, is challenging, especially with expensive objectives that require training a neural network. Typically, in MOO for neural architecture search (NAS), we aim to balance performance and hardware metrics across devices. Prior NAS approaches simplify this task by incorporating hardware constraints into the objective function, but profiling the Pareto front necessitates a computationally expensive search for each constraint. In this work, we propose a novel NAS algorithm that encodes user preferences to trade-off performance and hardware metrics, yielding representative and diverse architectures across multiple devices in just a single search run. To this end, we parameterize the joint architectural distribution across devices and multiple objectives via a hypernetwork that can be conditioned on hardware features and preference vectors, enabling zero-shot transferability to new devices. Extensive experiments involving up to 19 hardware devices and 3 different objectives demonstrate the effectiveness and scalability of our method. Finally, we show that, without any additional costs, our method outperforms existing MOO NAS methods across a broad range of qualitatively different search spaces and datasets, including MobileNetV3 on ImageNet-1k, an encoder-decoder transformer space for machine translation and a decoder-only space for language modelling.

2222. Unlocking State-Tracking in Linear RNNs Through Negative Eigenvalues

链接：https://iclr.cc/virtual/2025/poster/29442 abstract：

2223. uniINF: Best-of-Both-Worlds Algorithm for Parameter-Free Heavy-Tailed MABs

链接：https://iclr.cc/virtual/2025/poster/31113 abstract： In this paper, we present a novel algorithm, `uniINF`, for the Heavy-Tailed Multi-Armed Bandits (HTMAB) problem, demonstrating robustness and adaptability in both stochastic and adversarial environments. Unlike the stochastic MAB setting where loss distributions are stationary with time, our study extends to the adversarial setup, where losses are generated from heavy-tailed distributions that depend on both arms and time. Our novel algorithm `uniINF` enjoys the so-called Best-of-Both-Worlds (BoBW) property, performing optimally in both stochastic and adversarial environments *without* knowing the exact environment type. Moreover, our algorithm also possesses a Parameter-Free feature, *i.e.*, it operates *without* the need of knowing the heavy-tail parameters (σ, α) a-priori.To be precise, `uniINF` ensures nearly-optimal regret in both stochastic and adversarial environments, matching the corresponding lower bounds when (σ, α) is known (up to logarithmic factors). To our knowledge, `uniINF` is the first parameter-free algorithm to achieve the BoBW property for the heavy-tailed MAB problem. Technically, we develop innovative techniques to achieve BoBW guarantees for Parameter-Free HTMABs, including a refined analysis for the dynamics of log-barrier, an auto-balancing learning rate scheduling scheme, an adaptive skipping-clipping loss tuning technique, and a stopping-time analysis for logarithmic regret.

2224. Quantitative Approximation for Neural Operators in Nonlinear Parabolic Equations

链接：https://iclr.cc/virtual/2025/poster/27724 abstract： Neural operators serve as universal approximators for general continuous operators. In this paper, we derive the approximation rate of solution operators for the nonlinear parabolic partial differential equations (PDEs), contributing to the quantitative approximation theorem for solution operators of nonlinear PDEs. Our results show that neural operators can efficiently approximate these solution operators without the exponential growth in model complexity, thus strengthening the theoretical foundation of neural operators. A key insight in our proof is to transfer PDEs into the corresponding integral equations via Duahamel's principle, and to leverage the similarity between neural operators and Picard’s iteration—a classical algorithm for solving PDEs. This approach is potentially generalizable beyond parabolic PDEs to a class of PDEs which can be solved by Picard's iteration.

2225. Neural Interactive Proofs

链接：https://iclr.cc/virtual/2025/poster/29664 abstract： We consider the problem of how a trusted, but computationally bounded agent (a 'verifier') can learn to interact with one or more powerful but untrusted agents ('provers') in order to solve a given task. More specifically, we study the case in which agents are represented using neural networks and refer to solutions of this problem as neural interactive proofs. First we introduce a unifying framework based on prover-verifier games (Anil et al., 2021), which generalises previously proposed interaction protocols. We then describe several new protocols for generating neural interactive proofs, and provide a theoretical comparison of both new and existing approaches. Finally, we support this theory with experiments in two domains: a toy graph isomorphism problem that illustrates the key ideas, and a code validation task using large language models. In so doing, we aim to create a foundation for future work on neural interactive proofs and their application in building safer AI systems.

2226. RAPID: Retrieval Augmented Training of Differentially Private Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/28006 abstract： Differentially private diffusion models (DPDMs) harness the remarkable generative capabilities of diffusion models while enforcing differential privacy (DP) for sensitive data. However, existing DPDM training approaches often suffer from significant utility loss, large memory footprint, and expensive inference cost, impeding their practical uses. To overcome such limitations, we present RAPID: Retrieval Augmented PrIvate Diffusion model, a novel approach that integrates retrieval augmented generation (RAG) into DPDM training. Specifically, RAPID leverages available public data to build a knowledge base of sample trajectories; when training the diffusion model on private data, RAPID computes the early sampling steps as queries, retrieves similar trajectories from the knowledge base as surrogates, and focuses on training the later sampling steps in a differentially private manner. Extensive evaluation using benchmark datasets and models demonstrates that, with the same privacy guarantee, RAPID significantly outperforms state-of-the-art approaches by large margins in generative quality, memory footprint, and inference cost, suggesting that retrieval-augmented DP training represents a promising direction for developing future privacy-preserving generative models. The code is available at: https://github.com/TanqiuJiang/RAPID

2227. OMNI-EPIC: Open-endedness via Models of human Notions of Interestingness with Environments Programmed in Code

链接：https://iclr.cc/virtual/2025/poster/29279 abstract： Open-ended and AI-generating algorithms aim to continuously generate and solve increasingly complex tasks indefinitely, offering a promising path toward more general intelligence. To accomplish this grand vision, learning must occur within a vast array of potential tasks. Existing approaches to automatically generating environments are constrained within manually predefined, often narrow distributions of environments, limiting their ability to create any learning environment. To address this limitation, we introduce a novel framework, OMNI-EPIC, that augments previous work in Open-endedness via Models of human Notions of Interestingness (OMNI) with Environments Programmed in Code (EPIC). OMNI-EPIC leverages foundation models to autonomously generate code specifying the next learnable (i.e., not too easy or difficult for the agent’s current skill set) and interesting (e.g., worthwhile and novel) tasks. OMNI-EPIC generates both environments (e.g., an obstacle course) and reward functions (e.g., progress through the obstacle course quickly without touching red objects), enabling it, in principle, to create any simulatable learning task. We showcase the explosive creativity of OMNI-EPIC, which continuously innovates to suggest new, interesting learning challenges. We also highlight how OMNI-EPIC can adapt to reinforcement learning agents’ learning progress, generating tasks that are of suitable difficulty. Overall, OMNI-EPIC has the potential to endlessly create learnable and interesting environments, further propelling the development of self-improving AI systems and AI-Generating Algorithms.

2228. CAX: Cellular Automata Accelerated in JAX

链接：https://iclr.cc/virtual/2025/poster/28380 abstract： Cellular automata have become a cornerstone for investigating emergence and self-organization across diverse scientific disciplines. However, the absence of a hardware-accelerated cellular automata library limits the exploration of new research directions, hinders collaboration, and impedes reproducibility. In this work, we introduce CAX (Cellular Automata Accelerated in JAX), a high-performance and flexible open-source library designed to accelerate cellular automata research. CAX delivers cutting-edge performance through hardware acceleration while maintaining flexibility through its modular architecture, intuitive API, and support for both discrete and continuous cellular automata in arbitrary dimensions. We demonstrate CAX's performance and flexibility through a wide range of benchmarks and applications. From classic models like elementary cellular automata and Conway's Game of Life to advanced applications such as growing neural cellular automata and self-classifying MNIST digits, CAX speeds up simulations up to 2,000 times faster. Furthermore, we demonstrate CAX's potential to accelerate research by presenting a collection of three novel cellular automata experiments, each implemented in just a few lines of code thanks to the library's modular architecture. Notably, we show that a simple one-dimensional cellular automaton can outperform GPT-4 on the 1D-ARC challenge.

2229. Sparse Autoencoders Reveal Temporal Difference Learning in Large Language Models

链接：https://iclr.cc/virtual/2025/poster/31107 abstract： In-context learning, the ability to adapt based on a few examples in the input prompt, is a ubiquitous feature of large language models (LLMs). However, as LLMs' in-context learning abilities continue to improve, understanding this phenomenon mechanistically becomes increasingly important. In particular, it is not well-understood how LLMs learn to solve specific classes of problems, such as reinforcement learning (RL) problems, in-context. Through three different tasks, we first show that Llama 3 70B can solve simple RL problems in-context. We then analyze the residual stream of Llama using Sparse Autoencoders (SAEs) and find representations that closely match temporal difference (TD) errors. Notably, these representations emerge despite the model only being trained to predict the next token. We verify that these representations are indeed causally involved in the computation of TD errors and Q-values by performing carefully designed interventions on them. Taken together, our work establishes a methodology for studying and manipulating in-context learning with SAEs, paving the way for a more mechanistic understanding.

2230. Modeling dynamic social vision highlights gaps between deep learning and humans

链接：https://iclr.cc/virtual/2025/poster/27867 abstract： Deep learning models trained on computer vision tasks are widely considered the most successful models of human vision to date. The majority of work that supports this idea evaluates how accurately these models predict behavior and brain responses to static images of objects and scenes. Real-world vision, however, is highly dynamic, and far less work has evaluated deep learning models on human responses to moving stimuli, especially those that involve more complicated, higher-order phenomena like social interactions. Here, we extend a dataset of natural videos depicting complex multi-agent interactions by collecting human-annotated sentence captions for each video, and we benchmark 350+ image, video, and language models on behavior and neural responses to the videos. As in prior work, we find that many vision models reach the noise ceiling in predicting visual scene features and responses along the ventral visual stream (often considered the primary neural substrate of object and scene recognition). In contrast, vision models poorly predict human action and social interaction ratings and neural responses in the lateral stream (a neural pathway theorized to specialize in dynamic, social vision), though video models show a striking advantage in predicting mid-level lateral stream regions. Language models (given human sentence captions of the videos) predict action and social ratings better than image and video models, but perform poorly at predicting neural responses in the lateral stream. Together, these results identify a major gap in AI's ability to match human social vision and provide insights to guide future model development for dynamic, natural contexts.

2231. LLM-based Typed Hyperresolution for Commonsense Reasoning with Knowledge Bases

链接：https://iclr.cc/virtual/2025/poster/27849 abstract： Large language models (LLM) are being increasingly applied to tasks requiring commonsense reasoning. Despite their outstanding potential, the reasoning process of LLMs is prone to errors and hallucinations that hinder their applicability, especially in high-stakes scenarios. Several works have attempted to enhance commonsense reasoning performance of LLMs by (i) using prompting styles that elicit more accurate reasoning, (ii) utilizing the LLM as a semantic parser for a symbolic reasoner, or (iii) enforcing the LLM to simulate a logical inference rule. However, all these solutions have critical limitations: they are unable to leverage the internal commonsense knowledge of the LLM in tandem with an axiomatic knowledge base, they lack a mechanism to reliably repair erroneous inference steps, and their application is restricted to small knowledge bases that fit the context limit of the LLM. In this work, we present LLM-based Typed Hyperresolution (LLM-TH), a logical commonsense reasoning framework that leverages "theory resolution", a concept from classical logical inference which enables integrating LLMs into the "resolution" inference rule, thus mitigating reasoning errors and hallucinations and enabling verification of the reasoning procedure. LLM-TH is also equipped with a mechanism for repairing erroneous inference steps supported by theoretical guarantees. Using "Hyperresolution" and "Typed inference" schemes, we show that LLM-TH can efficiently reason over large knowledge bases consisting of tens of thousands of rules with arbitrary predicate arities. Our experiments on three diverse language-based reasoning tasks—preference reasoning, multi-domain deductive reasoning, and geographical question answering—showcase that LLM-TH, using merely a BART 406M parameter NLI entailment model, significantly reduces reasoning errors compared to baselines using Llama3-70B, Gemini1.5-Flash, GPT-3.5-Turbo, and Mixtral-46.7B.

2232. MoLEx: Mixture of Layer Experts for Fine-tuning with Sparse Upcycling

链接：https://iclr.cc/virtual/2025/poster/28189 abstract： Large-scale pre-training of deep models, followed by fine-tuning them to adapt to downstream tasks, has become the cornerstone of natural language processing (NLP). The prevalence of vast corpses of data coupled with computational resources has led to large models with a considerable number of parameters. While the massive size of these models has led to remarkable success in many NLP tasks, a detriment is the expense required to retrain all the base model's parameters for the adaptation to each task or domain. Parameter Efficient Fine-Tuning (PEFT) provides a highly effective solution for this challenge by minimizing the number of parameters required to be trained in adjusting to the new task while maintaining the quality of the model. While existing methods have achieved impressive results, they mainly focus on adapting a subset of parameters using adapters, weight reparameterization, and prompt engineering. In this paper, we study layers as extractors of different types of linguistic information that are valuable when used in conjunction with each other. We then propose the Mixture of Layer Experts (MoLEx), a novel Sparse Mixture of Experts (SMoE) whose experts are layers in the pre-trained model. In particular, MoLEx is applied at each layer of the pre-trained model. It performs a conditional computation of a mixture of layers during fine-tuning to provide the model with more structural knowledge about the data. By providing an avenue for information exchange between layers, MoLEx enables the model to make a more well-informed prediction for the downstream task, leading to better fine-tuning results with the same number of effective parameters. As experts can be processed in parallel, MoLEx introduces minimal additional computational overhead. We empirically corroborate the advantages of MoLEx when combined with popular PEFT baseline methods on a variety of downstream fine-tuning tasks, including the popular GLUE benchmark for natural language understanding (NLU) as well as the natural language generation (NLG) End-to-End Challenge (E2E).

2233. Generalization in VAE and Diffusion Models: A Unified Information-Theoretic Analysis

链接：https://iclr.cc/virtual/2025/poster/29890 abstract： Despite the empirical success of Diffusion Models (DMs) and Variational Autoencoders (VAEs), their generalization performance remains theoretically underexplored, especially lacking a full consideration of the shared encoder-generator structure. Leveraging recent information-theoretic tools, we propose a unified theoretical framework that provides guarantees for the generalization of both the encoder and generator by treating them as randomized mappings. This framework further enables (1) a refined analysis for VAEs, accounting for the generator's generalization, which was previously overlooked; (2) illustrating an explicit trade-off in generalization terms for DMs that depends on the diffusion time T; and (3) providing computable bounds for DMs based solely on the training data, allowing the selection of the optimal T and the integration of such bounds into the optimization process to improve model performance. Empirical results on both synthetic and real datasets illustrate the validity of the proposed theory.

2234. Transformer Meets Twicing: Harnessing Unattended Residual Information

链接：https://iclr.cc/virtual/2025/poster/31222 abstract： Transformer-based deep learning models have achieved state-of-the-art performance across numerous language and vision tasks. While the self-attention mechanism, a core component of transformers, has proven capable of handling complex data patterns, it has been observed that the representational capacity of the attention matrix degrades significantly across transformer layers, thereby hurting its overall performance. In this work, we leverage the connection between self-attention computations and low-pass non-local means (NLM) smoothing filters and propose the Twicing Attention, a novel attention mechanism that uses kernel twicing procedure in nonparametric regression to alleviate the low-pass behavior of associated NLM smoothing with compelling theoretical guarantees. This approach enables the extraction and reuse of meaningful information retained in the residuals following the imperfect smoothing operation at each layer. Our proposed method offers two key advantages over standard self-attention: 1) a provably slower decay of representational capacity and 2) improved accuracy across various data modalities and tasks. We empirically demonstrate the performance gains of our model over baseline transformers on multiple tasks and benchmarks, including image classification and language modeling, on both clean and corrupted data.

2235. RAG-SR: Retrieval-Augmented Generation for Neural Symbolic Regression

链接：https://iclr.cc/virtual/2025/poster/29868 abstract： Symbolic regression is a key task in machine learning, aiming to discover mathematical expressions that best describe a dataset. While deep learning has increased interest in using neural networks for symbolic regression, many existing approaches rely on pre-trained models. These models require significant computational resources and struggle with regression tasks involving unseen functions and variables. A pre-training-free paradigm is needed to better integrate with search-based symbolic regression algorithms. To address these limitations, we propose a novel framework for symbolic regression that integrates evolutionary feature construction with a neural network, without the need for pre-training. Our approach adaptively generates symbolic trees that align with the desired semantics in real-time using a language model trained via online supervised learning, providing effective building blocks for feature construction. To mitigate hallucinations from the language model, we design a retrieval-augmented generation mechanism that explicitly leverages searched symbolic expressions. Additionally, we introduce a scale-invariant data augmentation technique that further improves the robustness and generalization of the model. Experimental results demonstrate that our framework achieves state-of-the-art accuracy across 25 regression algorithms and 120 regression tasks.

2236. Unlocking Guidance for Discrete State-Space Diffusion and Flow Models

链接：https://iclr.cc/virtual/2025/poster/29286 abstract： Generative models on discrete state-spaces have a wide range of potential applications, particularly in the domain of natural sciences. In continuous state-spaces, controllable and flexible generation of samples with desired properties has been realized using guidance on diffusion and flow models. However, these guidance approaches are not readily amenable to discrete state-space models. Consequently, we introduce a general and principled method for applying guidance on such models. Our method depends on leveraging continuous-time Markov processes on discrete state-spaces, which unlocks computational tractability for sampling from a desired guided distribution. We demonstrate the utility of our approach, Discrete Guidance, on a range of applications including guided generation of small-molecules, DNA sequences and protein sequences.

2237. Causal Order: The Key to Leveraging Imperfect Experts in Causal Inference

链接：https://iclr.cc/virtual/2025/poster/30671 abstract： Large Language Models (LLMs) have recently been used as experts to infer causal graphs, often by repeatedly applying a pairwise prompt that asks about the causal relationship of each variable pair. However, such experts, including human domain experts, cannot distinguish between direct and indirect effects given a pairwise prompt. Therefore, instead of the graph, we propose that causal order be used as a more stable output interface for utilizing expert knowledge. When querying a perfect expert with a pairwise prompt, we show that the inferred graph can have significant errors whereas the causal order is always correct. In practice, however, LLMs are imperfect experts and we find that pairwise prompts lead to multiple cycles and do not yield a valid order. Hence, we propose a prompting strategy that introduces an auxiliary variable for every variable pair and instructs the LLM to avoid cycles within this triplet. We show, both theoretically and empirically, that such a triplet prompt leads to fewer cycles than the pairwise prompt. Across multiple real-world graphs, the triplet prompt yields a more accurate order using both LLMs and human annotators as experts. By querying the expert with different auxiliary variables for the same variable pair, it also increases robustness---triplet method with much smaller models such as Phi-3 and Llama-3 8B outperforms a pairwise prompt with GPT-4. For practical usage, we show how the estimated causal order from the triplet method can be used to reduce error in downstream discovery and effect inference tasks.

2238. INCLUDE: Evaluating Multilingual Language Understanding with Regional Knowledge

链接：https://iclr.cc/virtual/2025/poster/28601 abstract： The performance differential of large language models (LLM) between languages hinders their effective deployment in many regions, inhibiting the potential economic and societal value of generative AI tools in many communities. However, the development of functional LLMs in many languages (i.e., multilingual LLMs) is bottlenecked by the lack of high-quality evaluation resources in languages other than English. Moreover, current practices in multilingual benchmark construction often translate English resources, ignoring the regional and cultural knowledge of the environments in which multilingual systems would be used. In this work, we construct an evaluation suite of 197,243 QA pairs from local exam sources to measure the capabilities of multilingual LLMs in a variety of regional contexts.Our novel resource, INCLUDE, is a comprehensive knowledge- and reasoning-centric benchmark across 44 written languages that evaluates multilingual LLMs for performance in the actual language environments where they would be deployed.

2239. Tight Clusters Make Specialized Experts

链接：https://iclr.cc/virtual/2025/poster/29725 abstract： Sparse Mixture-of-Experts (MoE) architectures have emerged as a promising approach to decoupling model capacity from computational cost. At the core of the MoE model is the router, which learns the underlying clustering structure of the input distribution in order to send input tokens to appropriate experts. However, latent clusters may be unidentifiable in high dimension, which causes slow convergence, susceptibility to data contamination, and overall degraded representations as the router is unable to perform appropriate token-expert matching. We examine the router through the lens of clustering optimization and derive optimal feature weights that maximally identify the latent clusters. We use these weights to compute the token-expert routing assignments in an adaptively transformed space that promotes well-separated clusters, which helps identify the best-matched expert for each token. In particular, for each expert cluster, we compute a set of weights that scales features according to whether that expert clusters tightly along that feature. We term this novel router the Adaptive Clustering (AC) router. Our AC router enables the MoE model to obtain three connected benefits: 1) faster convergence, 2) better robustness to data corruption, and 3) overall performance improvement, as experts are specialized in semantically distinct regions of the input space. We empirically demonstrate the advantages of our AC router over baseline routing methods when applied on a variety of MoE backbones for language modeling and image recognition tasks in both clean and corrupted settings.

2240. Facilitating Multi-turn Function Calling for LLMs via Compositional Instruction Tuning

链接：https://iclr.cc/virtual/2025/poster/28328 abstract： Large Language Models (LLMs) have exhibited significant potential in performing diverse tasks, including the ability to call functions or use external tools to enhance their performance. While current research on function calling by LLMs primarily focuses on single-turn interactions, this paper addresses the overlooked necessity for LLMs to engage in multi-turn function calling—critical for handling compositional, real-world queries that require planning with functions but not only use functions. To facilitate this, we introduce an approach, BUTTON, which generates synthetic compositional instruction tuning data via bottom-up instruction construction and top-down trajectory generation. In the bottom-up phase, we generate simple atomic tasks based on real-world scenarios and build compositional tasks using heuristic strategies based on atomic tasks. Corresponding function definitions are then synthesized for these compositional tasks. The top-down phase features a multi-agent environment where interactions among simulated humans, assistants, and tools are utilized to gather multi-turn function calling trajectories. This approach ensures task compositionality and allows for effective function and trajectory generation by examining atomic tasks within compositional tasks. We produce a dataset BUTTONInstruct comprising 8k data points and demonstrate its effectiveness through extensive experiments across various LLMs.

2241. CAMEx: Curvature-aware Merging of Experts

链接：https://iclr.cc/virtual/2025/poster/28408 abstract： Existing methods for merging experts during model training and fine-tuning predominantly rely on Euclidean geometry, which assumes a flat parameter space. This assumption can limit the model's generalization ability, especially during the pre-training phase, where the parameter manifold might exhibit more complex curvature. Curvature-aware merging methods typically require additional information and computational resources to approximate the Fisher Information Matrix, adding memory overhead. In this paper, we introduce CAMEx (Curvature-Aware Merging of Experts), a novel expert merging protocol that incorporates natural gradients to account for the non-Euclidean curvature of the parameter manifold. By leveraging natural gradients, CAMEx adapts more effectively to the structure of the parameter space, improving alignment between model updates and the manifold's geometry. This approach enhances both pre-training and fine-tuning, resulting in better optimization trajectories and improved generalization without the substantial memory overhead typically associated with curvature-aware methods. Our contributions are threefold: (1) CAMEx significantly outperforms traditional Euclidean-based expert merging techniques across various natural language processing tasks, leading to enhanced performance during pre-training and fine-tuning; (2) we introduce a dynamic merging architecture that optimizes resource utilization, achieving high performance while reducing computational costs, facilitating efficient scaling of large language models; and (3) we provide both theoretical and empirical evidence to demonstrate the efficiency of our proposed method. The code is publicly available at: https://github.com/kpup1710/CAMEx.

2242. ClawMachine: Learning to Fetch Visual Tokens for Referential Comprehension

链接：https://iclr.cc/virtual/2025/poster/29545 abstract： Aligning vision and language concepts at a finer level remains an essential topic of multimodal large language models (MLLMs), particularly for tasks such as referring and grounding. Existing methods, such as proxy encoding and geometry encoding genres, incorporate additional syntax to encode spatial information, imposing extra burdens when communicating between language with vision modules. In this study, we propose ClawMachine, offering a new methodology that explicitly notates each entity using token collectives—groups of visual tokens that collaboratively represent higher-level semantics. A hybrid perception mechanism is also explored to perceive and understand scenes from both discrete and continuous spaces. Our method unifies the prompt and answer of visual referential tasks without using additional syntax. By leveraging a joint vision-language vocabulary, ClawMachine integrates referring and grounding in an auto-regressive manner, demonstrating great potential with scaled up pre-training data. Experiments show that ClawMachine achieves superior performance on scene-level and referential understanding tasks with higher efficiency. It also exhibits the potential to integrate multi-source information for complex visual reasoning, which is beyond the capability of many MLLMs. Our code is available at https://github.com/martian422/ClawMachine.

2243. Distribution-Specific Agnostic Conditional Classification With Halfspaces

链接：https://iclr.cc/virtual/2025/poster/30049 abstract： We study "selective" or "conditional" classification problems under an agnostic setting. Classification tasks commonly focus on modeling the relationship between features and categories that captures the vast majority of data. In contrast to common machine learning frameworks, conditional classification intends to model such relationships only on a subset of the data defined by some selection rule. Most work on conditional classification either solves the problem in a realizable setting or does not guarantee the error is bounded compared to an optimal solution. In this work, we consider selective/conditional classification by sparse linear classifiers for subsets defined by halfspaces, and give both positive as well as negative results for Gaussian feature distributions. On the positive side, we present the first PAC-learning algorithm for homogeneous halfspace selectors with error guarantee $\tilde{O}(\sqrt{\mathrm{opt}})$, where opt is the smallest conditional classification error over the given class of classifiers and homogeneous halfspaces. On the negative side, we find that, under cryptographic assumptions, approximating the conditional classification loss within a small additive error is computationally hard even under Gaussian distribution. We prove that approximating conditional classification is at least as hard as approximating agnostic classification in both additive and multiplicative form.

2244. Last-Iterate Convergence Properties of Regret-Matching Algorithms in Games

链接：https://iclr.cc/virtual/2025/poster/29992 abstract： We study last-iterate convergence properties of algorithms for solving two-player zero-sum games based on Regret Matching$^+$ (RM$^+$). Despite their widespread use for solving real games, virtually nothing is known about their last-iterate convergence. A major obstacle to analyzing RM-type dynamics is that their regret operators lack Lipschitzness and (pseudo)monotonicity.We start by showing numerically that several variants used in practice, such as RM$^+$, predictive RM$^+$ and alternating RM$^+$, all lack last-iterate convergence guarantees even on a simple 3×3 matrix game.We then prove that recent variants of these algorithms based on a smoothing technique, extragradient RM^{+} and smooth Predictive RM$^+$, enjoy asymptotic last-iterate convergence (without a rate), $1/\sqrt{t}$ best-iterate convergence, and when combined with restarting, linear-rate last-iterate convergence. Our analysis builds on a new characterization of the geometric structure of the limit points of our algorithms, marking a significant departure from most of the literature on last-iterate convergence. We believe that our analysis may be of independent interest and offers a fresh perspective for studying last-iterate convergence in algorithms based on non-monotone operators.

2245. Joint Reward and Policy Learning with Demonstrations and Human Feedback Improves Alignment

链接：https://iclr.cc/virtual/2025/poster/29428 abstract： Aligning to human preferences and/or intentions is an important requirement for contemporary foundation models. To ensure alignment, popular approaches such as reinforcement learning with human feedback (RLHF) break down the task into three stages: (i) a model is computed with supervised fine-tuning (SFT) based upon large demonstrations data, (ii) a reward model (RM) is estimated based upon human feedback data, and (iii) reinforcement learning (RL) is used to further refine the SFT model by optimizing the estimated reward model. Demonstrations and human feedback data reflect human user preferences in different ways. As a result, the reward model estimate obtained from only human feedback data is likely not as accurate as a reward model estimate obtained from both demonstration and human feedback data. A policy model that optimizes the reward model estimate obtained from both demonstration and human feedback data will likely exhibit better alignment performance. We introduce a tractable algorithm for finding the reward and policy models and provide a finite-time performance guarantee. Additionally, we demonstrate the efficiency of the proposed solution with extensive experiments including alignment problems in LLMs and robotic control problems in MuJoCo. We observe that the proposed solutions outperform the existing alignment algorithm by large margins, especially when the amounts of demonstration and preference data are unbalanced.

2246. Learning to Discretize Denoising Diffusion ODEs

链接：https://iclr.cc/virtual/2025/poster/27797 abstract： Diffusion Probabilistic Models (DPMs) are generative models showing competitive performance in various domains, including image synthesis and 3D point cloud generation. Sampling from pre-trained DPMs involves multiple neural function evaluations (NFEs) to transform Gaussian noise samples into images, resulting in higher computational costs compared to single-step generative models such as GANs or VAEs. Therefore, reducing the number of NFEs while preserving generation quality is crucial. To address this, we propose LD3, a lightweight framework designed to learn the optimal time discretization for sampling. LD3 can be combined with various samplers and consistently improves generation quality without having to retrain resource-intensive neural networks. We demonstrate analytically and empirically that LD3 improves sampling efficiency with much less computational overhead. We evaluate our method with extensive experiments on 7 pre-trained models, covering unconditional and conditional sampling in both pixel-space and latent-space DPMs. We achieve FIDs of 2.38 (10 NFE), and 2.27 (10 NFE) on unconditional CIFAR10 and AFHQv2 in 5-10 minutes of training. LD3 offers an efficient approach to sampling from pre-trained diffusion models. Code is available at https://github.com/vinhsuhi/LD3.

2247. Divide and Translate: Compositional First-Order Logic Translation and Verification for Complex Logical Reasoning

链接：https://iclr.cc/virtual/2025/poster/31268 abstract： Complex logical reasoning tasks require a long sequence of reasoning, which a large language model (LLM) with chain-of-thought prompting still falls short. To alleviate this issue, neurosymbolic approaches incorporate a symbolic solver. Specifically, an LLM only translates a natural language problem into a satisfiability (SAT) problem that consists of first-order logic formulas, and a sound symbolic solver returns a mathematically correct solution. However, we discover that LLMs have difficulties to capture complex logical semantics hidden in the natural language during translation. To resolve this limitation, we propose a Compositional First-Order Logic Translation. An LLM first parses a natural language sentence into newly defined logical dependency structures that consist of an atomic subsentence and its dependents, then sequentially translate the parsed subsentences. Since multiple logical dependency structures and sequential translations are possible for a single sentence, we also introduce two Verification algorithms to ensure more reliable results. We utilize an SAT solver to rigorously compare semantics of generated first-order logic formulas and select the most probable one. We evaluate the proposed method, dubbed CLOVER, on seven logical reasoning benchmarks and show that it outperforms the previous neurosymbolic approaches and achieves new state-of-the-art results.

2248. REBIND: Enhancing Ground-state Molecular Conformation Prediction via Force-Based Graph Rewiring

链接：https://iclr.cc/virtual/2025/poster/29367 abstract： Predicting the ground-state 3D molecular conformations from 2D molecular graphs is critical in computational chemistry due to its profound impact on molecular properties. Deep learning (DL) approaches have recently emerged as promising alternatives to computationally-heavy classical methods such as density functional theory (DFT). However, we discover that existing DL methods inadequately model inter-atomic forces, particularly for non-bonded atomic pairs, due to their naive usage of bonds and pairwise distances. Consequently, significant prediction errors occur for atoms with low degree (ie., low coordination numbers) whose conformations are primarily influenced by non-bonded interactions. To address this, we propose ReBIND, a novel framework that rewires molecular graphs by adding edges based on the Lennard-Jones potential to capture non-bonded interactions for low-degree atoms. Experimental results demonstrate that ReBIND significantly outperforms state-of-the-art methods across various molecular sizes, achieving up to a 20% reduction in prediction error. The code is available in: https://github.com/holymollyhao/ReBIND

2249. Token-Supervised Value Models for Enhancing Mathematical Problem-Solving Capabilities of Large Language Models

链接：https://iclr.cc/virtual/2025/poster/30898 abstract： With the rapid advancement of test-time compute search strategies to improve the mathematical problem-solving capabilities of large language models (LLMs), the need for building robust verifiers has become increasingly important. However, all these inference strategies rely on existing verifiers originally designed for Best-of-N search, which makes them sub-optimal for tree search techniques at test time. During tree search, existing verifiers can only offer indirect and implicit assessments of partial solutions or under-value prospective intermediate steps, thus resulting in the premature pruning of promising intermediate steps. To overcome these limitations, we propose token-supervised value models (TVMs) -- a new class of verifiers that assign each token a probability that reflects the likelihood of reaching the correct final answer. This new token-level supervision enables TVMs to directly and explicitly evaluate partial solutions, effectively distinguishing between promising and incorrect intermediate steps during tree search at test time. Experimental results demonstrate that combining tree-search-based inference strategies with TVMs significantly improves the accuracy of LLMs in mathematical problem-solving tasks, surpassing the performance of existing verifiers.

2250. Test-Time Ensemble via Linear Mode Connectivity: A Path to Better Adaptation

链接：https://iclr.cc/virtual/2025/poster/30982 abstract： Test-time adaptation updates pretrained models on the fly to handle distribution shifts in test data. While existing research has focused on stable optimization during adaptation, less attention has been given to enhancing model representations for adaptation capability. To address this gap, we propose Test-Time Ensemble (TTE) grounded in the intriguing property of linear mode connectivity. TTE leverages ensemble strategies during adaptation: 1) adaptively averaging the parameter weights of assorted test-time adapted models and 2) incorporating dropout to further promote representation diversity. These strategies encapsulate model diversity into a single model, avoiding computational burden associated with managing multiple models. Besides, we propose a robust knowledge distillation scheme to prevent model collapse, ensuring stable optimization and preserving the ensemble benefits during adaptation. Notably, TTE integrates seamlessly with existing TTA approaches, advancing their adaptation capabilities. In extensive experiments, integration with TTE consistently outperformed baseline models across various challenging scenarios, demonstrating its effectiveness and general applicability.

2251. Towards Principled Evaluations of Sparse Autoencoders for Interpretability and Control

链接：https://iclr.cc/virtual/2025/poster/31206 abstract： Disentangling model activations into human-interpretable features is a centralproblem in interpretability. Sparse autoencoders (SAEs) have recently attractedmuch attention as a scalable unsupervised approach to this problem. However, ourimprecise understanding of ground-truth features in realistic scenarios makes itdifficult to measure the success of SAEs. To address this challenge, we proposeto evaluate SAEs on specific tasks by comparing them to supervisedfeature dictionaries computed with knowledge of the concepts relevant to thetask. Specifically, we suggest that it is possible to (1) compute supervised sparsefeature dictionaries that disentangle model computations for a specific task;(2) use them to evaluate and contextualize the degree of disentanglement andcontrol offered by SAE latents on this task. Importantly, we can do this in away that is agnostic to whether the SAEs have learned the exact ground-truthfeatures or a different but similarly useful representation.As a case study, we apply this framework to the indirect object identification(IOI) task using GPT-2 Small, with SAEs trained on either the IOI or OpenWebTextdatasets. We find that SAEs capture interpretable features for the IOI task, andthat more recent SAE variants such as Gated SAEs and Top-K SAEs are competitivewith supervised features in terms of disentanglement and control over the model.We also exhibit, through this setup and toy models, some qualitative phenomenain SAE training illustrating feature splitting and the role of featuremagnitudes in solutions preferred by SAEs.

2252. Solving Token Gradient Conflict in Mixture-of-Experts for Large Vision-Language Model

链接：https://iclr.cc/virtual/2025/poster/29390 abstract： The Mixture-of-Experts (MoE) has gained increasing attention in studying Large Vision-Language Models (LVLMs). It uses a sparse model to replace the dense model, achieving comparable performance while activating fewer parameters during inference, thus significantly reducing the inference cost. Existing MoE methods in LVLM encourage different experts to specialize in different tokens, and they usually employ a router to predict the routing of each token. However, the router is not optimized concerning distinct parameter optimization directions generated from tokens within an expert. This may lead to severe interference between tokens within an expert. To address this problem, we propose to use the token-level gradient analysis to Solving Token Gradient Conflict (STGC) in this paper. Specifically, we first use token-level gradients to identify conflicting tokens in experts. After that, we add a regularization loss tailored to encourage conflicting tokens routing from their current experts to other experts, for reducing interference between tokens within an expert. Our method can serve as a plug-in for diverse LVLM methods, and extensive experimental results demonstrate its effectiveness. demonstrate its effectiveness. The code will be publicly available at https://github.com/longrongyang/STGC.

2253. TopoGaussian: Inferring Internal Topology Structures from Visual Clues

链接：https://iclr.cc/virtual/2025/poster/30595 abstract： We present TopoGaussian, a holistic, particle-based pipeline for inferring the interior structure of an opaque object from easily accessible photos and videos as input. Traditional mesh-based approaches require tedious and error-prone mesh filling and fixing process, while typically output rough boundary surface. Our pipeline combines Gaussian Splatting with a novel, versatile particle-based differentiable simulator that simultaneously accommodates constitutive model, actuator, and collision, without interference with mesh. Based on the gradients from this simulator, we provide flexible choice of topology representation for optimization, including particle, neural implicit surface, and quadratic surface. The resultant pipeline takes easily accessible photos and videos as input and outputs the topology that matches the physical characteristics of the input. We demonstrate the efficacy of our pipeline on a synthetic dataset and four real-world tasks with 3D-printed prototypes. Compared with existing mesh-based method, our pipeline is 5.26x faster on average with improved shape quality. These results highlight the potential of our pipeline in 3D vision, soft robotics, and manufacturing applications.

2254. SonicSim: A customizable simulation platform for speech processing in moving sound source scenarios

链接：https://iclr.cc/virtual/2025/poster/30201 abstract： Systematic evaluation of speech separation and enhancement models under moving sound source conditions requires extensive and diverse data. However, real-world datasets often lack sufficient data for training and evaluation, and synthetic datasets, while larger, lack acoustic realism. Consequently, neither effectively meets practical needs. To address this issue, we introduce SonicSim, a synthetic toolkit based on the embodied AI simulation platform Habitat-sim, designed to generate highly customizable data for moving sound sources. SonicSim supports multi-level adjustments—including scene-level, microphone-level, and source-level—enabling the creation of more diverse synthetic data. Leveraging SonicSim, we constructed a benchmark dataset called SonicSet, utilizing LibriSpeech, Freesound Dataset 50k

(FSD50K), Free Music Archive (FMA), and 90 scenes from Matterport3D to evaluate speech separation and enhancement models. Additionally, to investigate the differences between synthetic and real-world data, we selected 5 hours of raw, non-reverberant data from the SonicSet validation set and recorded a real-world speech separation dataset, providing a reference for comparing SonicSet with other synthetic datasets. For speech enhancement, we utilized the real-world dataset RealMAN to validate the acoustic gap between SonicSet and existing synthetic datasets. The results indicate that models trained on SonicSet generalize better to real-world scenarios compared to other synthetic datasets. Code is publicly available at https://cslikai.cn/SonicSim/.

2255. Inverse Attention Agents for Multi-Agent Systems

链接：https://iclr.cc/virtual/2025/poster/29819 abstract： A major challenge for Multi-Agent Systems (MAS) is enabling agents to adapt dynamically to diverse environments in which opponents and teammates may continually change. Agents trained using conventional methods tend to excel only within the confines of their training cohorts; their performance drops significantly when confronting unfamiliar agents. To address this shortcoming, we introduce Inverse Attention Agents that adopt concepts from the Theory of Mind (ToM) implemented algorithmically using an attention mechanism trained in an end-to-end manner. Crucial to determining the final actions of these agents, the weights in their attention model explicitly represent attention to different goals. We furthermore propose an inverse attention network that deduces the ToM of agents based on observations and prior actions. The network infers the attentional states of other agents, thereby refining the attention weights to adjust the agent's final action. We conduct experiments in a continuous environment, tackling demanding tasks encompassing cooperation, competition, and a blend of both. They demonstrate that the inverse attention network successfully infers the attention of other agents, and that this information improves agent performance. Additional human experiments show that, compared to baseline agent models, our inverse attention agents exhibit superior cooperation with humans and better emulate human behaviors.

2256. ToddlerDiffusion: Interactive Structured Image Generation with Cascaded Schrödinger Bridge

链接：https://iclr.cc/virtual/2025/poster/30087 abstract： Diffusion models break down the challenging task of generating data from high-dimensional distributions into a series of easier denoising steps. Inspired by this paradigm, we propose a novel approach that extends the diffusion framework into modality space, decomposing the complex task of RGB image generation into simpler, interpretable stages. Our method, termed {\papernameAbbrev}, cascades modality-specific models, each responsible for generating an intermediate representation, such as contours, palettes, and detailed textures, ultimately culminating in a high-quality RGB image.Instead of relying on the naive LDM concatenation conditioning mechanism to connect the different stages together, we employ Schr\"odinger Bridge to determine the optimal transport between different modalities.Although employing a cascaded pipeline introduces more stages, which could lead to a more complex architecture, each stage is meticulously formulated for efficiency and accuracy, surpassing Stable-Diffusion (LDM) performance.Modality composition not only enhances overall performance but enables emerging proprieties such as consistent editing, interaction capabilities, high-level interpretability, and faster convergence and sampling rate. Extensive experiments on diverse datasets, including LSUN-Churches, ImageNet, CelebHQ, and LAION-Art, demonstrate the efficacy of our approach, consistently outperforming state-of-the-art methods.For instance, {\papernameAbbrev} achieves notable efficiency, matching LDM performance on LSUN-Churches while operating 2$\times$ faster with a 3$\times$ smaller architecture.The project website is available at:\href{https://toddlerdiffusion.github.io/website/}{$https://toddlerdiffusion.github.io/website/$}

2257. SelectFormer in Data Markets: Privacy-Preserving and Efficient Data Selection for Transformers with Multi-Party Computation

链接：https://iclr.cc/virtual/2025/poster/31130 abstract： Critical to a free data market is $ \textit{private data selection}$, i.e. the model owner selects and then appraises training data from the data owner before both parties commit to a transaction. To keep the data and model private, this process shall evaluate the target model to be trained over Multi-Party Computation (MPC). While prior work suggests that evaluating Transformer-based models over MPC is prohibitively expensive, this paper makes it practical for the purpose of data selection. Our contributions are three: (1) a new pipeline for private data selection over MPC; (2) emulating high-dimensional nonlinear operators with low-dimension MLPs, which are trained on a small sample of the data of interest; (3) scheduling MPC in a parallel, multiphase fashion. We evaluate our method on diverse Transformer models and NLP/CV benchmarks. Compared to directly evaluating the target model over MPC, our method reduces the delay from thousands of hours to tens of hours, while only seeing around 0.20% accuracy degradation from training with the selected data.

2258. Zeroth-Order Policy Gradient for Reinforcement Learning from Human Feedback without Reward Inference

链接：https://iclr.cc/virtual/2025/poster/29030 abstract： Reward inference (learning a reward model from human preferences) is a critical intermediate step in the Reinforcement Learning from Human Feedback (RLHF) pipeline for fine-tuning Large Language Models (LLMs). In practice, RLHF faces fundamental challenges such as distribution shift, reward model overfitting, and problem misspecification. An alternative approach is direct policy optimization without reward inference, such as Direct Preference Optimization (DPO), which provides a much simpler pipeline and has shown empirical success in LLM applications. However, DPO utilizes the closed-form expression between the optimal policy and the reward function, which is only suitable under the bandit setting or deterministic MDPs. This paper develops two RLHF algorithms without reward inference for general RL problems beyond bandits and deterministic MDPs, and general preference models beyond the Bradley-Terry model. The key idea is to estimate the local value function difference from human preferences and then approximate the policy gradient with a zeroth-order gradient approximator. For both algorithms, we establish polynomial convergence rates in terms of the number of policy gradient iterations, the number of trajectory samples, and human preference queries per iteration. Numerical experiments in stochastic environments validate the performance of our proposed algorithms, outperforming popular RLHF baselines such as DPO and PPO. Our paper shows there exist provably efficient methods to solve general RLHF problems without reward inference.

2259. Differentiable Rule Induction from Raw Sequence Inputs

链接：https://iclr.cc/virtual/2025/poster/27680 abstract： Rule learning-based models are widely used in highly interpretable scenarios due to their transparent structures. Inductive logic programming (ILP), a form of machine learning, induces rules from facts while maintaining interpretability. Differentiable ILP models enhance this process by leveraging neural networks to improve robustness and scalability. However, most differentiable ILP methods rely on symbolic datasets, facing challenges when learning directly from raw data. Specifically, they struggle with explicit label leakage: The inability to map continuous inputs to symbolic variables without explicit supervision of input feature labels. In this work, we address this issue by integrating a self-supervised differentiable clustering model with a novel differentiable ILP model, enabling rule learning from raw data without explicit label leakage. The learned rules effectively describe raw data through its features. We demonstrate that our method intuitively and precisely learns generalized rules from time series and image data.

2260. Atlas Gaussians Diffusion for 3D Generation

链接：https://iclr.cc/virtual/2025/poster/30246 abstract： Using the latent diffusion model has proven effective in developing novel 3D generation techniques. To harness the latent diffusion model, a key challenge is designing a high-fidelity and efficient representation that links the latent space and the 3D space. In this paper, we introduce Atlas Gaussians, a novel representation for feed-forward native 3D generation. Atlas Gaussians represent a shape as the union of local patches, and each patch can decode 3D Gaussians. We parameterize a patch as a sequence of feature vectors and design a learnable function to decode 3D Gaussians from the feature vectors. In this process, we incorporate UV-based sampling, enabling the generation of a sufficiently large, and theoretically infinite, number of 3D Gaussian points. The large amount of 3D Gaussians enables the generation of high-quality details. Moreover, due to local awareness of the representation, the transformer-based decoding procedure operates on a patch level, ensuring efficiency. We train a variational autoencoder to learn the Atlas Gaussians representation, and then apply a latent diffusion model on its latent space for learning 3D Generation. Experiments show that our approach outperforms the prior arts of feed-forward native 3D generation. Project page: https://yanghtr.github.io/projects/atlas_gaussians.

2261. Generalizing Reasoning Problems to Longer Lengths

链接：https://iclr.cc/virtual/2025/poster/27647 abstract： Length generalization (LG) is a challenging problem in learning to reason. It refers to the phenomenon that when trained on reasoning problems of smaller lengths/sizes, the model struggles with problems of larger sizes or lengths. Although it has been proven that reasoning can be learned if the intermediate reasoning steps (also known as chain-of-thought (CoT)) are given in the training data, existing studies only apply to within a given length (interpolation), while LG is about extrapolation beyond the given length. This paper begins by presenting a theorem that identifies the root cause of the LG problem. It then defines a class of reasoning problems for which achieving LG with Transformers can be theoretically guaranteed, provided the CoT schemes are constructed to meet a proposed condition called (n,r)-consistency.

2262. Cybench: A Framework for Evaluating Cybersecurity Capabilities and Risks of Language Models

链接：https://iclr.cc/virtual/2025/poster/28041 abstract： Language Model (LM) agents for cybersecurity that are capable of autonomously identifying vulnerabilities and executing exploits have potential to cause real-world impact. Policymakers, model providers, and researchers in the AI and cybersecurity communities are interested in quantifying the capabilities of such agents to help mitigate cyberrisk and investigate opportunities for penetration testing. Toward that end, we introduce Cybench, a framework for specifying cybersecurity tasks and evaluating agents on those tasks. We include 40 professional-level Capture the Flag (CTF) tasks from 4 distinct CTF competitions, chosen to be recent, meaningful, and spanning a wide range of difficulties. Each task includes its own description, starter files, and is initialized in an environment where an agent can execute commands and observe outputs. Since many tasks are beyond the capabilities of existing LM agents, we introduce subtasks for each task, which break down a task into intermediary steps for a more detailed evaluation. To evaluate agent capabilities, we construct a cybersecurity agent and evaluate 8 models: GPT-4o, OpenAI o1-preview, Claude 3 Opus, Claude 3.5 Sonnet, Mixtral 8x22b Instruct, Gemini 1.5 Pro, Llama 3 70B Chat, and Llama 3.1 405B Instruct. For the top performing models (GPT-4o and Claude 3.5 Sonnet), we further investigate performance across 4 agent scaffolds (structured bash, action-only, pseudoterminal, and web search). Without subtask guidance, agents leveraging Claude 3.5 Sonnet, GPT-4o, OpenAI o1-preview, and Claude 3 Opus successfully solved complete tasks that took human teams up to 11 minutes to solve. In comparison, the most difficult task took human teams 24 hours and 54 minutes to solve. Anonymized code and data are available at https://drive.google.com/file/d/1kp3H0pw1WMAH-Qyyn9WA0ZKmEa7Cr4D4 and https://drive.google.com/file/d/1BcTQ02BBR0m5LYTiK-tQmlK17_Txijly.

2263. Learn-by-interact: A Data-Centric Framework For Self-Adaptive Agents in Realistic Environments

链接：https://iclr.cc/virtual/2025/poster/31073 abstract： Autonomous agents powered by large language models (LLMs) have the potential to enhance human capabilities, assisting with digital tasks from sending emails to performing data analysis. The abilities of existing LLMs at such tasks are often hindered by the lack of high-quality agent data from the corresponding environments they interact with. We propose LEARN-BY-INTERACT, a data-centric framework to adapt LLM agents to any given environments without human annotations. LEARN-BY-INTERACT synthesizes trajectories of agent-environment interactions based on documentations, and constructs instructions by summarizing or abstracting the interaction histories, a process called backward construction. We assess the quality of our synthetic data by using them in both training-based scenarios and training-free in-context learning (ICL), where we craft innovative retrieval approaches optimized for agents. Extensive experiments on SWE-bench, WebArena, OSWorld, and Spider2-V spanning across realistic coding, web, and desktop environments show the effectiveness of LEARN-BY-INTERACT in various downstream agentic tasks — baseline results are improved up to 11.1% for ICL with Claude-3.5 and 23.1% for training with Codestral-22B. We further demonstrate the critical role of backward construction, which provides up to 10.6% improvement for training. Our ablation studies demonstrate the efficiency provided by our synthesized data in ICL and the superiority of our retrieval pipeline over alternative approaches like conventional retrieval-augmented generation (RAG). We expect that LEARN-BY-INTERACT will serve as a foundation for agent data synthesis as LLMs are increasingly deployed at real-world environments.

2264. Intrinsic User-Centric Interpretability through Global Mixture of Experts

链接：https://iclr.cc/virtual/2025/poster/27864 abstract： In human-centric settings like education or healthcare, model accuracy and model explainability are key factors for user adoption. Towards these two goals, intrinsically interpretable deep learning models have gained popularity, focusing on accurate predictions alongside faithful explanations. However, there exists a gap in the human-centeredness of these approaches, which often produce nuanced and complex explanations that are not easily actionable for downstream users. We present InterpretCC (interpretable conditional computation), a family of intrinsically interpretable neural networks at a unique point in the design space that optimizes for ease of human understanding and explanation faithfulness, while maintaining comparable performance to state-of-the-art models. InterpretCC achieves this through adaptive sparse activation of features before prediction, allowing the model to use a different, minimal set of features for each instance. We extend this idea into an interpretable, global mixture-of-experts (MoE) model that allows users to specify topics of interest, discretely separates the feature space for each data point into topical subnetworks, and adaptively and sparsely activates these topical subnetworks for prediction. We apply InterpretCC for text, time series and tabular data across several real-world datasets, demonstrating comparable performance with non-interpretable baselines and outperforming intrinsically interpretable baselines. Through a user study involving 56 teachers, InterpretCC explanations are found to have higher actionability and usefulness over other intrinsically interpretable approaches.

2265. RegMix: Data Mixture as Regression for Language Model Pre-training

链接：https://iclr.cc/virtual/2025/poster/30960 abstract： The data mixture for large language model pre-training significantly impacts performance, yet how to determine an effective mixture remains unclear. We propose RegMix to automatically identify a high-performing data mixture by formulating it as a regression task. RegMix trains many small models on diverse data mixtures, uses regression to predict performance of unseen mixtures, and applies the best predicted mixture to train a large-scale model with orders of magnitude more compute. To empirically validate RegMix, we train 512 models with 1M parameters for 1B tokens to fit the regression model and predict the best data mixture. Using this mixture we train a 1B parameter model for 25B tokens (i.e. 1000× larger and 25× longer) which we find performs best among 64 candidate 1B parameter models with other mixtures. Furthermore, RegMix consistently outperforms human selection in experiments involving models up to 7B models trained on 100B tokens, while matching or exceeding DoReMi using just 10% of the computational resources. Our experiments also show that (1) Data mixtures significantly impact performance; (2) Web corpora rather than data perceived as high-quality like Wikipedia have the strongest positive correlation with downstream performance; (3) Domains interact in complex ways often contradicting common sense, thus automatic approaches like RegMix are needed; (4) Data mixture effects transcend scaling laws. Our code is available at https://github.com/sail-sg/regmix.

2266. Inner Information Analysis Algorithm for Deep Neural Network based on Community

链接：https://iclr.cc/virtual/2025/poster/29134 abstract： Deep learning has achieved advancements across a variety of forefront fields. However, its inherent 'black box' characteristic poses challenges to the comprehension and trustworthiness of the decision-making processes within neural networks. To mitigate these challenges, we introduce InnerSightNet, an inner information analysis algorithm designed to illuminate the inner workings of deep neural networks through the perspectives of community. This approach is aimed at deciphering the intricate patterns of neurons within deep neural networks, thereby shedding light on the networks' information processing and decision-making pathways. InnerSightNet operates in three primary phases, 'neuronization-aggregation-evaluation'. Initially, it transforms learnable units into a structured network of neurons. Subsequently, these neurons are aggregated into distinct communities according to representation attributes. The final phase involves the evaluation of these communities' roles and functionalities, to unpick the information flow and decision-making. By transcending focus on single-layer or individual neuron, InnerSightNet broadens the horizon for deep neural network interpretation. InnerSightNet offers a unique vantage point, enabling insights into the collective behavior of communities within the overarching architecture, thereby enhancing transparency and trust in deep learning systems.

2267. Verifying Properties of Binary Neural Networks Using Sparse Polynomial Optimization

链接：https://iclr.cc/virtual/2025/poster/30678 abstract： This paper explores methods for verifying the properties of Binary Neural Networks (BNNs), focusing on robustness against adversarial attacks. Despite their lower computational and memory needs, BNNs, like their full-precision counterparts, are also sensitive to input perturbations. Established methods for solving this problem are predominantly based on Satisfiability Modulo Theories and Mixed-Integer Linear Programming techniques, which are characterized by NP complexity and often face scalability issues.We introduce an alternative approach using Semidefinite Programming relaxations derived from sparse Polynomial Optimization. Our approach, compatible with continuous input space, not only mitigates numerical issues associated with floating-point calculations but also enhances verification scalability through the strategic use of tighter first-order semidefinite relaxations. We demonstrate the effectiveness of our method in verifying robustness against both $\||.|\|_\infty$ and $\||.|\|_2$-based adversarial attacks.

2268. Scaling up Masked Diffusion Models on Text

链接：https://iclr.cc/virtual/2025/poster/29366 abstract： Masked diffusion models (MDMs) have shown promise in language modeling, yet their scalability and effectiveness in core language tasks, such as text generation and language understanding, remain underexplored. This paper establishes the first scaling law for MDMs, demonstrating a scaling rate comparable to autoregressive models (ARMs) and a relatively small compute gap. Motivated by their scalability, we train a family of MDMs with up to 1.1 billion (B) parameters to systematically evaluate their performance against ARMs of comparable or larger sizes. Fully leveraging the probabilistic formulation of MDMs, we propose a simple yet effective unsupervised classifier-free guidance that effectively exploits large-scale unpaired data, boosting performance for conditional inference. In language understanding, the 1.1B MDM outperforms the 1.1B TinyLlama model trained on the same data across four of eight zero-shot benchmarks. Notably, it achieves competitive math reasoning ability with the 7B Llama-2 model on the GSM8K dataset. In text generation, MDMs with 16 times more pre-training time offer a flexible trade-off against ARMs with the accelerated sampling technique KV-Cache: MDMs match ARMs in performance while being 1.4 times faster during sampling.Moreover, MDMs address challenging tasks for ARMs by effectively handling bidirectional reasoning and adapting to temporal shifts in data. Notably, a 1.1B MDM breaks the reverse curse encountered by much larger ARMs with significantly more data and computation, such as 13B Llama-2 and 175B GPT-3. Our code is available at https://github.com/ML-GSAI/SMDM.

2269. GPUDrive: Data-driven, multi-agent driving simulation at 1 million FPS

链接：https://iclr.cc/virtual/2025/poster/30404 abstract： Multi-agent learning algorithms have been successful at generating superhuman planning in various games but have had limited impact on the design of deployed multi-agent planners. A key bottleneck in applying these techniques to multi-agent planning is that they require billions of steps of experience. To enable the study of multi-agent planning at scale, we present GPUDrive, a GPU-accelerated, multi-agent simulator built on top of the Madrona Game Engine capable of generating over a million simulation steps per second. Observation, reward, and dynamics functions are written directly in C++, allowing users to define complex, heterogeneous agent behaviors that are lowered to high-performance CUDA. Despite these low-level optimizations, GPUDrive is fully accessible through Python, offering a seamless and efficient workflow for multi-agent, closed-loop simulation. Using GPUDrive, we train reinforcement learning agents on the Waymo Open Motion Dataset, achieving efficient goal-reaching in minutes and scaling to thousands of scenarios in hours. We open-source the code and pre-trained agents at \url{www.github.com/Emerge-Lab/gpudrive}.

2270. On the Learn-to-Optimize Capabilities of Transformers in In-Context Sparse Recovery

链接：https://iclr.cc/virtual/2025/poster/29887 abstract： An intriguing property of the Transformer is its ability to perform in-context learning (ICL), where the Transformer can solve different inference tasks without parameter updating based on the contextual information provided by the corresponding input-output demonstration pairs. It has been theoretically proved that ICL is enabled by the capability of Transformers to perform gradient-descent algorithms (Von Oswald et al., 2023a; Bai et al., 2024). This work takes a step further and shows that Transformers can perform learning-to-optimize (L2O) algorithms. Specifically, for the ICL sparse recovery (formulated as LASSO) tasks, we show that a K-layer Transformer can perform an L2O algorithm with a provable convergence rate linear in K. This provides a new perspective explaining the superior ICL capability of Transformers, even with only a few layers, which cannot be achieved by the standard gradient-descent algorithms. Moreover, unlike the conventional L2O algorithms that require the measurement matrix involved in training to match that in testing, the trained Transformer is able to solve sparse recovery problems generated with different measurement matrices. Besides, Transformers as an L2O algorithm can leverage structural information embedded in the training tasks to accelerate its convergence during ICL, and generalize across different lengths of demonstration pairs, where conventional L2O algorithms typically struggle or fail.

Such theoretical findings are supported by our experimental results.

2271. Data-adaptive Differentially Private Prompt Synthesis for In-Context Learning

链接：https://iclr.cc/virtual/2025/poster/28125 abstract： Large Language Models (LLMs) rely on the contextual information embedded in examples/demonstrations to perform in-context learning (ICL). To mitigate the risk of LLMs potentially leaking private information contained in examples in the prompt, we introduce a novel data-adaptive differentially private algorithm called AdaDPSyn to generate synthetic examples from the private dataset and then use these synthetic examples to perform ICL. The objective of AdaDPSyn is to adaptively adjust the noise level in the data synthesis mechanism according to the inherent statistical properties of the data, thereby preserving high ICL accuracy while maintaining formal differential privacy guarantees. A key innovation in AdaDPSyn is the Precision-Focused Iterative Radius Reduction technique, which dynamically refines the aggregation radius - the scope of data grouping for noise addition - based on patterns observed in data clustering, thereby minimizing the amount of additive noise. We conduct extensive experiments on standard benchmarks and compare AdaDPSyn with DP few-shot generation algorithm (Tang et al., 2023). The experiments demonstrate that AdaDPSyn not only outperforms DP few-shot generation, but also maintains high accuracy levels close to those of non-private baselines, providing an effective solution for ICL with privacy protection.

2272. Optimal Flow Transport and its Entropic Regularization: a GPU-friendly Matrix Iterative Algorithm for Flow Balance Satisfaction

链接：https://iclr.cc/virtual/2025/poster/29855 abstract： The Sinkhorn algorithm, based on Entropic Regularized Optimal Transport (OT), has garnered significant attention due to its computational efficiency enabled by GPU-friendly matrix-vector multiplications. However, vanilla OT primarily deals with computations between the source and target nodes in a bipartite graph, limiting its practical application in real-world transportation scenarios.In this paper, we introduce the concept of Optimal Flow Transport (OFT) as an extension, where we consider a more general graph setting and the marginal constraints in vanilla OT are replaced by flow balance constraints. To obtain solutions, we incorporate entropic regularization into the OFT and introduce virtual flows for individual nodes to tackle the issue of potentially numerous isolated nodes lacking flow passages. Our proposition, the OFT-Sinkhorn algorithm, utilizes GPU-friendly matrix iterations to maintain flow balance constraints and minimize the objective function, and theoretical results for global convergence is also proposed in this paper.Furthermore, we enhance OFT by introducing capacity constraints on nodes and edges, transforming the OFT problem into a minimum-cost flow problem. We then present the Capacity-Constrained EOFT-Sinkhorn algorithm and compare it with the traditional Minimum cost flow (MCF) algorithm, showing that our algorithm is quite efficient for calculation. In particular, our EOFT-Sinkhorn is evaluated on high-precision and integer-precision MCF problems with different scales from one hundred to five thousand size, exhibiting significant time efficiency and the ability to approximate optimal solutions.

2273. Semantics-Adaptive Activation Intervention for LLMs via Dynamic Steering Vectors

链接：https://iclr.cc/virtual/2025/poster/30762 abstract： Large language models (LLMs) have achieved remarkable performance across many tasks, yet aligning them with desired behaviors remains challenging. Activation intervention has emerged as an effective and economical method to modify the behavior of LLMs. Despite considerable interest in this area, current intervention methods exclusively employ a fixed steering vector to modify model activations, lacking adaptability to diverse input semantics. To address this limitation, we propose Semantics-Adaptive Dynamic Intervention (SADI), a novel method that constructs a dynamic steering vector to intervene model activations at inference time. More specifically, SADI utilizes activation differences in contrastive pairs to precisely identify critical elements of an LLM (i.e., attention heads, hidden states, and neurons) for targeted intervention. During inference, SADI dynamically steers model behavior by scaling element-wise activations based on the directions of input semantics. Experimental results show that SADI outperforms established baselines by substantial margins, improving task performance without training. SADI's cost-effectiveness and generalizability across various LLM backbones and tasks highlight its potential as a versatile alignment technique. We will release the code to foster research in this area.

2274. IDInit: A Universal and Stable Initialization Method for Neural Network Training

链接：https://iclr.cc/virtual/2025/poster/30009 abstract： Deep neural networks have achieved remarkable accomplishments in practice. The success of these networks hinges on effective initialization methods, which are vital for ensuring stable and rapid convergence during training. Recently, initialization methods that maintain identity transition within layers have shown good efficiency in network training. These techniques (e.g., Fixup) set specific weights to zero to achieve identity control. However, settings of remaining weight (e.g., Fixup uses random values to initialize non-zero weights) will affect the inductive bias that is achieved only by a zero weight, which may be harmful to training. Addressing this concern, we introduce fully identical initialization (IDInit), a novel method that preserves identity in both the main and sub-stem layers of residual networks. IDInit employs a padded identity-like matrix to overcome rank constraints in non-square weight matrices. Furthermore, we show the convergence problem of an identity matrix can be solved by stochastic gradient descent. Additionally, we enhance the universality of IDInit by processing higher-order weights and addressing dead neuron problems. IDInit is a straightforward yet effective initialization method, with improved convergence, stability, and performance across various settings, including large-scale datasets and deep models.

2275. Mitigating the Backdoor Effect for Multi-Task Model Merging via Safety-Aware Subspace

链接：https://iclr.cc/virtual/2025/poster/28963 abstract： Model merging has gained significant attention as a cost-effective approach to integrate multiple single-task fine-tuned models into a unified one that can perform well on multiple tasks. However, existing model merging techniques primarily focus on resolving conflicts between task-specific models, they often overlook potential security threats, particularly the risk of backdoor attacks in the open-source model ecosystem. In this paper, we first investigate the vulnerabilities of existing model merging methods to backdoor attacks, identifying two critical challenges: backdoor succession and backdoor transfer. To address these issues, we propose a novel Defense-Aware Merging (DAM) approach that simultaneously mitigates task interference and backdoor vulnerabilities. Specifically, DAM employs a meta-learning-based optimization method with dual masks to identify a shared and safety-aware subspace for model merging. These masks are alternately optimized: the Task-Shared mask identifies common beneficial parameters across tasks, aiming to preserve task-specific knowledge while reducing interference, while the Backdoor-Detection mask isolates potentially harmful parameters to neutralize security threats. This dual-mask design allows us to carefully balance the preservation of useful knowledge and the removal of potential vulnerabilities. Compared to existing merging methods, DAM achieves a more favorable balance between performance and security, reducing the attack success rate by 2-10 percentage points while sacrificing only about 1\% in accuracy. Furthermore, DAM exhibits robust performance and broad applicability across various types of backdoor attacks and the number of compromised models involved in the merging process. Our codes and models can be accessed through https://github.com/Yangjinluan/DAM.

2276. EdgeRunner: Auto-regressive Auto-encoder for Artistic Mesh Generation

链接：https://iclr.cc/virtual/2025/poster/30789 abstract： Current auto-regressive mesh generation methods suffer from issues such as incompleteness, insufficient detail, and poor generalization. In this paper, we propose an Auto-regressive Auto-encoder (ArAE) model capable of generating high-quality 3D meshes with up to 4,000 faces at a spatial resolution of 512^3.We introduce a novel mesh tokenization algorithm that efficiently compresses triangular meshes into 1D token sequences, significantly enhancing training efficiency. Furthermore, our model compresses variable-length triangular meshes into a fixed-length latent space, enabling training latent diffusion models for better generalization. Extensive experiments demonstrate the superior quality, diversity, and generalization capabilities of our model in both point cloud and image-conditioned mesh generation tasks.

2277. Action Sequence Augmentation for Action Anticipation

链接：https://iclr.cc/virtual/2025/poster/28900 abstract： Action anticipation models require an understanding of temporal action patterns and dependencies to predict future actions from previous events. The key challenges arise from the vast number of possible action sequences, given the flexibility in action ordering and the interleaving of multiple goals. Since only a subset of such action sequences are present in action anticipation datasets, there is an inherent ordering bias in them. Another challenge is the presence of noisy input to the models due to erroneous action recognition or other upstream tasks. This paper addresses these challenges by introducing a novel data augmentation strategy that separately augments observed action sequences and next actions. To address biased action ordering, we introduce a grammar induction algorithm that derives a powerful context-free grammar from action sequence data. We also develop an efficient parser to generate plausible next-action candidates beyond the ground truth. For noisy input, we enhance model robustness by randomly deleting or replacing actions in observed sequences. Our experiments on the 50Salads, EGTEA Gaze+, and Epic-Kitchens-100 datasets demonstrate significant performance improvements over existing state-of-the-art methods.

2278. IterGen: Iterative Semantic-aware Structured LLM Generation with Backtracking

链接：https://iclr.cc/virtual/2025/poster/29157 abstract： Large Language Models (LLMs) are widely used for tasks such as natural language and code generation, but their outputs often suffer from issues like hallucination, toxicity, and incorrect results. Current libraries for structured LLM generation rely on left-to-right decoding without support for backtracking, limiting the ability to correct or refine outputs mid-generation. To address this, we introduce IterGen, a user-friendly library for iterative, grammar-guided LLM generation that enables users to move both forward and backward within the generated output based on grammar symbols. By leveraging a symbol-to-position mapping and maintaining the key-value (KV) cache state, IterGen ensures efficient and structured generation while allowing for corrections during the process. We demonstrate IterGen's effectiveness in two important applications: reducing privacy leakage in LLM outputs, improving the accuracy of LLM-generated SQL and Vega-Lite queries.Our code and additional resources are available at https://structuredllm.com.

2279. ZETA: Leveraging Z-order Curves for Efficient Top-k Attention

链接：https://iclr.cc/virtual/2025/poster/28649 abstract： Over recent years, the Transformer has become a fundamental building block for sequence modeling architectures. Yet at its core is the use of self-attention, whose memory and computational cost grow quadratically with the sequence length N, rendering it prohibitively expensive for long sequences. A promising approach is top-k attention, which selects only the k most relevant tokens and achieves performance comparable to vanilla self-attention while significantly reducing space and computational demands. However, causal masks require the current query token to only attend to past tokens, preventing existing top-k attention methods from efficiently searching for the most relevant tokens in parallel, thereby limiting training efficiency. In this work, we propose ZETA, leveraging Z-Order Curves for Efficient Top-k Attention, to enable parallel querying of past tokens for entire sequences. We first theoretically show that the choice of key and query dimensions involves a trade-off between the curse of dimensionality and the preservation of relative distances after projection. In light of this insight, we propose reducing the dimensionality of keys and queries in contrast to values and further leveraging Z-order curves to map low-dimensional keys and queries into one-dimensional space, which permits parallel sorting, thereby largely improving the efficiency for top-k token selection. Experimental results demonstrate that ZETA~matches the performance of standard attention on synthetic tasks Associative Recall and outperforms attention and its variants on Long-Range Arena and WikiText-103 language modeling.

2280. AnalogGenie: A Generative Engine for Automatic Discovery of Analog Circuit Topologies

链接：https://iclr.cc/virtual/2025/poster/28646 abstract： The massive and large-scale design of foundational semiconductor integrated circuits (ICs) is crucial to sustaining the advancement of many emerging and future technologies, such as generative AI, 5G/6G, and quantum computing.Excitingly, recent studies have shown the great capabilities of foundational models in expediting the design of digital ICs.Yet, applying generative AI techniques to accelerate the design of analog ICs remains a significant challenge due to critical domain-specific issues, such as the lack of a comprehensive dataset and effective representation methods for analog circuits.This paper proposes, $\textbf{AnalogGenie}$, a $\underline{\textbf{Gen}}$erat$\underline{\textbf{i}}$ve $\underline{\textbf{e}}$ngine for automatic design/discovery of $\underline{\textbf{Analog}}$ circuit topologies--the most challenging and creative task in the conventional manual design flow of analog ICs.AnalogGenie addresses two key gaps in the field: building a foundational comprehensive dataset of analog circuit topology and developing a scalable sequence-based graph representation universal to analog circuits.Experimental results show the remarkable generation performance of AnalogGenie in broadening the variety of analog ICs, increasing the number of devices within a single design, and discovering unseen circuit topologies far beyond any prior arts.Our work paves the way to transform the longstanding time-consuming manual design flow of analog ICs to an automatic and massive manner powered by generative AI.Our source code is available at https://github.com/xz-group/AnalogGenie.

2281. The Last Iterate Advantage: Empirical Auditing and Principled Heuristic Analysis of Differentially Private SGD

链接：https://iclr.cc/virtual/2025/poster/30434 abstract： We propose a simple heuristic privacy analysis of noisy clipped stochastic gradient descent (DP-SGD) in the setting where only the last iterate is released and the intermediate iterates remain hidden. Namely, our heuristic assumes a linear structure for the model.We show experimentally that our heuristic is predictive of the outcome of privacy auditing applied to various training procedures. Thus it can be used prior to training as a rough estimate of the final privacy leakage. We also probe the limitations of our heuristic by providing some artificial counterexamples where it underestimates the privacy leakage.The standard composition-based privacy analysis of DP-SGD effectively assumes that the adversary has access to all intermediate iterates, which is often unrealistic. However, this analysis remains the state of the art in practice. While our heuristic does not replace a rigorous privacy analysis, it illustrates the large gap between the best theoretical upper bounds and the privacy auditing lower bounds and sets a target for further work to improve the theoretical privacy analyses.

2282. Unified Convergence Analysis for Score-Based Diffusion Models with Deterministic Samplers

链接：https://iclr.cc/virtual/2025/poster/30206 abstract： Score-based diffusion models have emerged as powerful techniques for generating samples from high-dimensional data distributions. These models involve a two-phase process: first, injecting noise to transform the data distribution into a known prior distribution, and second, sampling to recover the original data distribution from noise. Among the various sampling methods, deterministic samplers stand out for their enhanced efficiency. However, analyzing these deterministic samplers presents unique challenges, as they preclude the use of established techniques such as Girsanov's theorem, which are only applicable to stochastic samplers. Furthermore, existing analysis for deterministic samplers usually focuses on specific examples, lacking a generalized approach for general forward processes and various deterministic samplers. Our paper addresses these limitations by introducing a unified convergence analysis framework. To demonstrate the power of our framework, we analyze the variance-preserving (VP) forward process with the exponential integrator (EI) scheme, achieving iteration complexity of $\tilde{O}(d^2/\epsilon)$.Additionally, we provide a detailed analysis of Denoising Diffusion Implicit Models (DDIM)-type samplers, which have been underexplored in previous research, achieving polynomial iteration complexity.

2283. Near-Exact Privacy Amplification for Matrix Mechanisms

链接：https://iclr.cc/virtual/2025/poster/28007 abstract： We study the problem of computing the privacy parameters for DP machine learning when using privacy amplification via random batching and noise correlated across rounds via a correlation matrix $\textbf{C}$ (i.e., the matrix mechanism). Past work on this problem either only applied to banded $\textbf{C}$, or gave loose privacy parameters. In this work, we give a framework for computing near-exact privacy parameters for any lower-triangular, non-negative $\textbf{C}$. Our framework allows us to optimize the correlation matrix $\textbf{C}$ while accounting for amplification, whereas past work could not. Empirically, we show this lets us achieve smaller RMSE on prefix sums than the previous state-of-the-art (SOTA). We also show that we can improve on the SOTA performance on deep learning tasks. Our two main technical tools are (i) using Monte Carlo accounting to bypass composition, which was the main technical challenge for past work, and (ii) a ``balls-in-bins'' batching scheme that enables easy privacy analysis and is closer to practical random batching than Poisson sampling.

2284. Selective Aggregation for Low-Rank Adaptation in Federated Learning

链接：https://iclr.cc/virtual/2025/poster/28693 abstract： We investigate LoRA in federated learning through the lens of the asymmetry analysis of the learned A and B matrices. In doing so, we uncover that A matrices are responsible for learning general knowledge, while B matrices focus on capturing client-specific knowledge. Based on this finding, we introduce Federated Share-A Low-Rank Adaptation (FedSA-LoRA), which employs two low-rank trainable matrices A and B to model the weight update, but only A matrices are shared with the server for aggregation. Moreover, we delve into the relationship between the learned A and B matrices in other LoRA variants, such as rsLoRA and VeRA, revealing a consistent pattern. Consequently, we extend our FedSA-LoRA method to these LoRA variants, resulting in FedSA-rsLoRA and FedSA-VeRA. In this way, we establish a general paradigm for integrating LoRA with FL, offering guidance for future work on subsequent LoRA variants combined with FL. Extensive experimental results on natural language understanding and generation tasks demonstrate the effectiveness of the proposed method. Our code is available at https://github.com/Pengxin-Guo/FedSA-LoRA.

2285. Adaptive Shrinkage Estimation for Personalized Deep Kernel Regression in Modeling Brain Trajectories

链接：https://iclr.cc/virtual/2025/poster/28284 abstract： Longitudinal biomedical studies monitor individuals over time to capture dynamics in brain development, disease progression, and treatment effects. However, estimating trajectories of brain biomarkers is challenging due to biological variability, inconsistencies in measurement protocols (e.g., differences in MRI scanners) as well as scarcity and irregularity in longitudinal measurements. Herein,we introduce a novel personalized deep kernel regression framework for forecasting brain biomarkers, with application to regional volumetric measurements. Our approach integrates two key components: a population model that captures brain trajectories from a large and diverse cohort, and a subject-specific model that captures individual trajectories. To optimally combine these, we propose Adaptive Shrinkage Estimation, which effectively balances population and subject-specific models. We assess our model’s performance through predictive accuracy metrics, uncertainty quantification, and validation against external clinical studies. Benchmarking against state-of-the-art statistical and machine learning models—including linear mixed effects models, generalized additive models, and deep learning methods—demonstrates the superior predictive performance of our approach. Additionally, we apply our method to predict trajectories of composite neuroimaging biomarkers, which highlights the versatility of our approach in modeling the progression of longitudinal neuroimaging biomarkers. Furthermore, validation on three external neuroimaging studies confirms the robustness of our method across different clinical contexts. We make the code available at https://github.com/vatass/AdaptiveShrinkageDKGP.

2286. VTDexManip: A Dataset and Benchmark for Visual-tactile Pretraining and Dexterous Manipulation with Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/28627 abstract： Vision and touch are the most commonly used senses in human manipulation. While leveraging human manipulation videos for robotic task pretraining has shown promise in prior works, it is limited to image and language modalities and deployment to simple parallel grippers. In this paper, aiming to address the limitations, we collect a vision-tactile dataset by humans manipulating 10 daily tasks and 182 objects. In contrast with the existing datasets, our dataset is the first visual-tactile dataset for complex robotic manipulation skill learning. Also, we introduce a novel benchmark, featuring six complex dexterous manipulation tasks and a reinforcement learning-based vision-tactile skill learning framework. 18 non-pretraining and pretraining methods within the framework are designed and compared to investigate the effectiveness of different modalities and pertaining strategies. Key findings based on our benchmark results and analyses experiments include: 1) Despite the tactile modality used in our experiments being binary and sparse, including it directly in the policy training boosts the success rate by about 20\% and joint pretraining it with vision gains a further 20\%. 2) Joint pretraining visual-tactile modalities exhibits strong adaptability in unknown tasks and achieves robust performance among all tasks. 3) Using binary tactile signals with vision is robust to viewpoint setting, tactile noise, and the binarization threshold, which facilitates to the visual-tactile policy to be deployed in reality. The dataset and benchmark are available at \url{https://github.com/LQTS/VTDexManip}.

2287. GameArena: Evaluating LLM Reasoning through Live Computer Games

链接：https://iclr.cc/virtual/2025/poster/29588 abstract： Evaluating the reasoning abilities of large language models (LLMs) is challenging. Existing benchmarks often depend on static datasets, which are vulnerable to data contamination and may get saturated over time, or on binary live human feedback that conflates reasoning with other abilities. As the most prominent dynamic benchmark, Chatbot Arena evaluates open-ended questions in real-world settings, but lacks the granularity in assessing specific reasoning capabilities. We introduce GameArena, a dynamic benchmark designed to evaluate LLM reasoning capabilities through interactive gameplay with humans. GameArena consists of three games designed to test specific reasoning capabilities (e.g., deductive and inductive reasoning), while keeping participants entertained and engaged. We analyze the gaming data retrospectively to uncover the underlying reasoning processes of LLMs and measure their fine-grained reasoning capabilities. We collect over 2000 game sessions and provide detailed assessments of various reasoning capabilities for five state-of-the-art LLMs. Our user study with 100 participants suggests that GameArena improves user engagement compared to Chatbot Arena. For the first time, GameArena enables the collection of step-by-step LLM reasoning data in the wild.

2288. Learning to Select Nodes in Branch and Bound with Sufficient Tree Representation

链接：https://iclr.cc/virtual/2025/poster/28790 abstract： Branch-and-bound methods are pivotal in solving Mixed Integer Linear Programming (MILP), where the challenge of node selection arises, necessitating the prioritization of different regions of the space for subsequent exploration. While machine learning techniques have been proposed to address this, two crucial problems concerning \textbf{(P1)} how to sufficiently extract features from the branch-and-bound tree, and \textbf{(P2)} how to assess the node quality comprehensively based on the features remain open. To tackle these challenges, we propose to tackle the node selection problem employing a novel Tripartite graph representation and Reinforcement learning with a Graph Neural Network model (TRGNN). The tripartite graph is theoretically proved to encompass sufficient information for tree representation in information theory. We learn node selection via reinforcement learning for learning delay rewards and give more comprehensive node metrics. Experiments show that TRGNN significantly improves the efficiency of solving MILPs compared to human-designed and learning-based node selection methods on both synthetic and large-scale real-world MILPs. Moreover, experiments demonstrate that TRGNN well generalizes to MILPs that are significantly larger than those seen during training.

2289. Looking into User’s Long-term Interests through the Lens of Conservative Evidential Learning

链接：https://iclr.cc/virtual/2025/poster/28375 abstract： Reinforcement learning (RL) provides an effective means to capture users' evolving preferences, leading to improved recommendation performance over time. However, existing RL approaches primarily rely on standard exploration strategies, which are less effective for a large item space with sparse reward signals given the limited interactions for most users. Therefore, they may not be able to learn the optimal policy that effectively captures user's evolving preferences and achieves the maximum expected reward over the long term. In this paper, we propose a novel evidential conservative Q-learning framework (ECQL) that learns an effective and conservative recommendation policy by integrating evidence-based uncertainty and conservative learning. ECQL conducts evidence-aware explorations to discover items that are located beyond current observations but reflect users' long-term interests. It offers an uncertainty-aware conservative view on policy evaluation to discourage deviating too much from users' current interests. Two central components of ECQL include a uniquely designed sequential state encoder and a novel conservative evidential-actor-critic (CEAC) module. The former generates the current state of the environment by aggregating historical information and a sliding window that contains the current user interactions as well as newly recommended items from RL exploration that may represent short and long-term interests respectively. The latter performs an evidence-based rating prediction by maximizing the conservative evidential Q-value and leverages an uncertainty-aware ranking score to explore the item space for a more diverse and valuable recommendation. Experiments on multiple real-world dynamic datasets demonstrate the state-of-the-art performance of ECQL and its capability to capture users' long-term interests.

2290. SuperCorrect: Advancing Small LLM Reasoning with Thought Template Distillation and Self-Correction

链接：https://iclr.cc/virtual/2025/poster/29720 abstract： Large language models (LLMs) like GPT-4, DeepSeek-R1, and ReasonFlux have shown significant improvements in various reasoning tasks. However, smaller LLMs still struggle with complex mathematical reasoning because they fail to effectively identify and correct reasoning errors. Recent reflection-based methods aim to address these issues by enabling self-reflection and self-correction, but they still face challenges in independently detecting errors in their reasoning steps. To overcome these limitations, we propose SuperCorrect, a novel two-stage framework that uses a large teacher model to supervise and correct both the reasoning and reflection processes of a smaller student model. In the first stage, we extract hierarchical high-level and detailed thought templates from the teacher model to guide the student model in eliciting more fine-grained reasoning thoughts. In the second stage, we introduce cross-model collaborative direct preference optimization (DPO) to enhance the self-correction abilities of the student model by following the teacher's correction traces during training. This cross-model DPO approach teaches the student model to effectively locate and resolve erroneous thoughts with error-driven insights from the teacher model, breaking the bottleneck of its thoughts and acquiring new skills and knowledge to tackle challenging problems. Extensive experiments consistently demonstrate our superiority over previous methods. Notably, our SuperCorrect-7B model significantly surpasses powerful DeepSeekMath-7B by 7.8\%/5.3\% and Qwen2.5-Math-7B by 15.1\%/6.3\% on MATH/GSM8K benchmarks, achieving new SOTA performance among all 7B models. Code is available at: https://github.com/YangLing0818/SuperCorrect-llm

2291. Large-scale and Fine-grained Vision-language Pre-training for Enhanced CT Image Understanding

链接：https://iclr.cc/virtual/2025/poster/28403 abstract： Artificial intelligence (AI) shows great potential in assisting radiologists to improve the efficiency and accuracy of medical image interpretation and diagnosis. However, a versatile AI model requires large-scale data and comprehensive annotations, which are often impractical in medical settings. Recent studies leverage radiology reports as a naturally high-quality supervision for medical images, using contrastive language-image pre-training (CLIP) to develop language-informed models for radiological image interpretation. Nonetheless, these approaches typically contrast entire images with reports, neglecting the local associations between imaging regions and report sentences, which may undermine model performance and interoperability. In this paper, we propose a fine-grained vision-language model (fVLM) for anatomy-level CT image interpretation. Specifically, we explicitly match anatomical regions of CT images with corresponding descriptions in radiology reports and perform contrastive pre-training for each anatomy individually. Fine-grained alignment, however, faces considerable false-negative challenges, mainly from the abundance of anatomy-level healthy samples and similarly diseased abnormalities, leading to ambiguous patient-level pairings. To tackle this issue, we propose identifying false negatives of both normal and abnormal samples and calibrating contrastive learning from patient-level to disease-aware pairing. We curated the largest CT dataset to date, comprising imaging and report data from 69,086 patients, and conducted a comprehensive evaluation of 54 major and important disease (including several most deadly cancers) diagnosis tasks across 15 main anatomies. Experimental results demonstrate the substantial potential of fVLM in versatile medical image interpretation. In the zero-shot classification task, we achieved an average AUC of 81.3% on 54 diagnosis tasks, surpassing CLIP and supervised methods by 12.9% and 8.0%, respectively. Additionally, on the publicly available CT-RATE and Rad-ChestCT benchmarks, our fVLM outperformed the current state-of-the-art methods with absolute AUC gains of 7.4% and 4.8%, respectively.

2292. MarS: a Financial Market Simulation Engine Powered by Generative Foundation Model

链接：https://iclr.cc/virtual/2025/poster/29246 abstract： Generative models aim to simulate realistic effects of various actions across different contexts, from text generation to visual effects. Despite significant efforts to build real-world simulators, the application of generative models to virtual worlds, like financial markets, remains under-explored. In financial markets, generative models can simulate complex market effects of participants with various behaviors, enabling interaction under different market conditions, and training strategies without financial risk. This simulation relies on the finest structured data in financial market like orders thus building the finest realistic simulation. We propose Large Market Model (LMM), an order-level generative foundation model, for financial market simulation, akin to language modeling in the digital world. Our financial Market Simulation engine (MarS), powered by LMM, addresses the domain-specific need for realistic, interactive and controllable order generation. Key observations include LMM's strong scalability across data size and model complexity, and MarS's robust and practicable realism in controlled generation with market impact. We showcase MarS as a forecast tool, detection system, analysis platform, and agent training environment, thus demonstrating MarS's ``paradigm shift'' potential for a variety of financial applications. We release the code of MarS at https://github.com/microsoft/MarS/.

2293. Computational Explorations of Total Variation Distance

链接：https://iclr.cc/virtual/2025/poster/27775 abstract： We investigate some previously unexplored (or underexplored) computational aspects of total variation (TV) distance.First, we give a simple deterministic polynomial-time algorithm for checking equivalence between mixtures of product distributions, over arbitrary alphabets.This corresponds to a special case, whereby the TV distance between the two distributions is zero.Second, we prove that unless $\mathsf{NP} \subseteq \mathsf{RP}$ it is impossible to efficiently estimate the TV distance between arbitrary Ising models, even in a bounded-error randomized setting.

2294. Presto! Distilling Steps and Layers for Accelerating Music Generation

链接：https://iclr.cc/virtual/2025/poster/30264 abstract： Despite advances in diffusion-based text-to-music (TTM) methods, efficient, high-quality generation remains a challenge. We introduce Presto!, an approach to inference acceleration for score-based diffusion transformers via reducing both sampling steps and cost per step. To reduce steps, we develop a new score-based distribution matching distillation (DMD) method for the EDM-family of diffusion models, the first GAN-based distillation method for TTM. To reduce the cost per step, we develop a simple, but powerful improvement to a recent layer distillation method that improves learning via better preserving hidden state variance. Finally, we combine our step and layer distillation methods together for a dual-faceted approach. We evaluate our step and layer distillation methods independently and show each yield best-in-class performance. Our combined distillation method can generate high-quality outputs with improved diversity, accelerating our base model by 10-18x (230/435ms latency for 32 second mono/stereo 44.1kHz, 15x faster than the comparable SOTA model) — the fastest TTM to our knowledge.

2295. TopoLM: brain-like spatio-functional organization in a topographic

language model

链接：https://iclr.cc/virtual/2025/poster/29162 abstract： Neurons in the brain are spatially organized such that neighbors on tissue often exhibit similar response profiles. In the human language system, experimental studies have observed clusters for syntactic and semantic categories, but the mechanisms underlying this functional organization remain unclear. Here, building on work from the vision literature, we develop TopoLM, a transformer language model with an explicit two-dimensional spatial representation of model units. By combining a next-token prediction objective with a spatial smoothness loss, representations in this model assemble into clusters that correspond to semantically interpretable groupings of text and closely match the functional organization in the brain's language system. TopoLM successfully predicts the emergence of a spatially organized cortical language system as well as the organization of functional clusters selective for fine-grained linguistic features empirically observed in human cortex. Our results suggest that the functional organization of the human language system is driven by a unified spatial objective, and provide a functionally and spatially aligned model of language processing in the brain.Neurons in the brain are spatially organized such that neighbors on tissue often exhibit similar response profiles. In the human language system, experimental studies have observed clusters for syntactic and semantic categories, but the mechanisms underlying this functional organization remain unclear. Here, building on work from the vision literature, we develop TopoLM, a transformer language model with an explicit two-dimensional spatial representation of model units. By combining a next-token prediction objective with a spatial smoothness loss, representations in this model assemble into clusters that correspond to semantically interpretable groupings of text and closely match the functional organization in the brain's language system. TopoLM successfully predicts the emergence of a spatially organized cortical language system as well as the organization of functional clusters selective for fine-grained linguistic features empirically observed in human cortex. Our results suggest that the functional organization of the human language system is driven by a unified spatial objective, and provide a functionally and spatially aligned model of language processing in the brain.

2296. Do LLMs Recognize Your Preferences? Evaluating Personalized Preference Following in LLMs

链接：https://iclr.cc/virtual/2025/poster/29686 abstract： Large Language Models (LLMs) are increasingly deployed as chatbots, yet their ability to personalize responses to user preferences remains limited. We introduce PrefEval, a benchmark for evaluating LLMs' ability to infer, memorize and adhere to user preferences in long-context conversational setting.PrefEval comprises 3,000 manually curated user preference and query pairs spanning 20 topics. PrefEval contains user personalization or preference information in both explicit and implicit preference forms, and evaluates LLM performance using a generation and a classification task. With PrefEval, we have evaluated 10 open-sourced andproprietary LLMs in multi-session conversations with varying context lengths up to 100k tokens. We benchmark with various prompting, iterative feedback, and retrieval-augmented generation methods. Our benchmarking effort reveals that state-of-the-art LLMs face significant challenges in following users' preference during conversations. In particular, in zero-shot settings, preference following accuracy falls below 10\% at merely 10 turns (~3k tokens) across most evaluated models. Even with advanced prompting and retrieval methods, preference following still deteriorates in long-context conversations. Furthermore, we show that fine-tuning on PrefEval significantly improves performance. We believe PrefEval serves as a valuable resource for measuring, understanding, and enhancing LLMs' proactive preference following abilities, paving the way for personalized conversational agents.

2297. Convex Formulations for Training Two-Layer ReLU Neural Networks

链接：https://iclr.cc/virtual/2025/poster/28956 abstract： Solving non-convex, NP-hard optimization problems is crucial for training machine learning models, including neural networks. However, non-convexity often leads to black-box machine learning models with unclear inner workings. While convex formulations have been used for verifying neural network robustness, their application to training neural networks remains less explored. In response to this challenge, we reformulate the problem of training infinite-width two-layer ReLU networks as a convex completely positive program in a finite-dimensional (lifted) space. Despite the convexity, solving this problem remains NP-hard due to the complete positivity constraint. To overcome this challenge, we introduce a semidefinite relaxation that can be solved in polynomial time. We then experimentally evaluate the tightness of this relaxation, demonstrating its competitive performance in test accuracy across a range of classification tasks.

2298. Improved Techniques for Optimization-Based Jailbreaking on Large Language Models

链接：https://iclr.cc/virtual/2025/poster/28946 abstract： Large language models (LLMs) are being rapidly developed, and a key component of their widespread deployment is their safety-related alignment. Many red-teaming efforts aim to jailbreak LLMs, where among these efforts, the Greedy Coordinate Gradient (GCG) attack's success has led to a growing interest in the study of optimization-based jailbreaking techniques. Although GCG is a significant milestone, its attacking efficiency remains unsatisfactory. In this paper, we present several improved (empirical) techniques for optimization-based jailbreaks like GCG. We first observe that the single target template of "Sure" largely limits the attacking performance of GCG; given this, we propose to apply diverse target templates containing harmful self-suggestion and/or guidance to mislead LLMs. Besides, from the optimization aspects, we propose an automatic multi-coordinate updating strategy in GCG (i.e., adaptively deciding how many tokens to replace in each step) to accelerate convergence, as well as tricks like easy-to-hard initialization. Then, we combine these improved technologies to develop an efficient jailbreak method, dubbed $\mathcal{I}$-GCG. In our experiments, we evaluate our $\mathcal{I}$-GCG on a series of benchmarks (such as NeurIPS 2023 Red Teaming Track). The results demonstrate that our improved techniques can help GCG outperform state-of-the-art jailbreaking attacks and achieve a nearly 100\% attack success rate.The code is released at https://github.com/jiaxiaojunQAQ/I-GCG.

2299. Grokking at the Edge of Numerical Stability

链接：https://iclr.cc/virtual/2025/poster/29501 abstract： Grokking, or sudden generalization that occurs after prolonged overfitting, is a surprising phenomenon that has challenged our understanding of deep learning. While a lot of progress has been made in understanding grokking, it is still not clear why generalization is delayed and why grokking often does not happen without regularization. In this work we argue that without regularization, grokking tasks push models to the edge of numerical stability, introducing floating point errors in the Softmax that we refer to as *Softmax Collapse* (SC). We show that SC prevents grokking and that mitigating SC leads to grokking *without* regularization. Investigating the root cause of SC, we find that beyond the point of overfitting, the gradients strongly align with what we call the *naïve loss minimization* (NLM) direction. This component of the gradient does not change the predictions of the model but decreases the loss by scaling the logits, usually through the scaling of the weights along their current direction. We show that this scaling of the logits explains the delay in generalization characteristic of grokking, and eventually leads to SC, stopping learning altogether. To validate these hypotheses, we introduce two key contributions that mitigate the issues faced in grokking tasks: (i) $\mathrm{StableMax}$, a new activation function that prevents SC and enables grokking without regularization, and (ii) $\perp\mathrm{Grad}$, a training algorithm that leads to quick generalization in grokking tasks by preventing NLM altogether. These contributions provide new insights into grokking, shedding light on its delayed generalization, reliance on regularization, and the effectiveness of known grokking-inducing methods.

2300. Accelerating Task Generalisation with Multi-Level Skill Hierarchies

链接：https://iclr.cc/virtual/2025/poster/30040 abstract： Developing reinforcement learning agents that can generalise effectively to new tasks is one of the main challenges in AI research. This paper introduces Fracture Cluster Options (FraCOs), a multi-level hierarchical reinforcement learning method designed to improve generalisation performance. FraCOs identifies patterns in agent behaviour and forms temporally-extended actions (options) based on the expected future usefulness of those patterns, enabling rapid adaptation to new tasks. In tabular settings, FraCOs demonstrates effective transfer and improves performance as the depth of the hierarchy increases. In several complex procedurally-generated environments, FraCOs consistently outperforms state-of-the-art deep reinforcement learning algorithms, achieving superior results in both in-distribution and out-of-distribution scenarios.

2301. SSOLE: Rethinking Orthogonal Low-rank Embedding for Self-Supervised Learning

链接：https://iclr.cc/virtual/2025/poster/27685 abstract： Self-supervised learning (SSL) aims to learn meaningful representations from unlabeled data. Orthogonal Low-rank Embedding (OLE) shows promise for SSL by enhancing intra-class similarity in a low-rank subspace and promoting inter-class dissimilarity in a high-rank subspace, making it particularly suitable for multi-view learning tasks. However, directly applying OLE to SSL poses significant challenges: (1) the virtually infinite number of "classes" in SSL makes achieving the OLE objective impractical, leading to representational collapse; and (2) low-rank constraints may fail to distinguish between positively and negatively correlated features, further undermining learning. To address these issues, we propose SSOLE (Self-Supervised Orthogonal Low-rank Embedding), a novel framework that integrates OLE principles into SSL by (1) decoupling the low-rank and high-rank enforcement to align with SSL objectives; and (2) applying low-rank constraints to feature deviations from their mean, ensuring better alignment of positive pairs by accounting for the signs of cosine similarities. Our theoretical analysis and empirical results demonstrate that these adaptations are crucial to SSOLE’s effectiveness. Moreover, SSOLE achieves competitive performance across SSL benchmarks without relying on large batch sizes, memory banks, or dual-encoder architectures, making it an efficient and scalable solution for self-supervised tasks. Code is available at https://github.com/husthuaan/ssole.

2302. Regretful Decisions under Label Noise

链接：https://iclr.cc/virtual/2025/poster/30842 abstract： Machine learning models are routinely used to support decisions that affect individuals – be it to screen a patient for a serious illness or to gauge their response to treatment. In these tasks, we are limited to learning models from datasets with noisy labels. In this paper, we study the instance-level impact of learning under label noise. We introduce a notion of regret for this regime which measures the number of unforeseen mistakes due to noisy labels. We show that standard approaches to learning under label noise can return models that perform well at a population level while subjecting individuals to a lottery of mistakes. We present a versatile approach to estimate the likelihood of mistakes at the individual level from a noisy dataset by training models over plausible realizations of datasets without label noise. This is supported by a comprehensive empirical study of label noise in clinical prediction tasks. Our results reveal how failure to anticipate mistakes can compromise model reliability and adoption, and demonstrate how we can address these challenges by anticipating and avoiding regretful decisions.

2303. Large Convolutional Model Tuning via Filter Subspace

链接：https://iclr.cc/virtual/2025/poster/30422 abstract： Efficient fine-tuning methods are critical to address the high computational and parameter complexity while adapting large pre-trained models to downstream tasks.Our study is inspired by prior research that represents each convolution filter as a linear combination of a small set of filter subspace elements, referred to as filter atoms. In this paper, we propose to fine-tune pre-trained models by adjusting only filter atoms, which are responsible for spatial-only convolution, while preserving spatially-invariant channel combination knowledge in atom coefficients.In this way, we bring a new filter subspace view for model tuning. Furthermore, each filter atom can be recursively decomposed as a combination of another set of atoms, which naturally expands the number of tunable parameters in the filter subspace.By only adapting filter atoms constructed by a small number of parameters, while maintaining the rest of model parameters constant, the proposed approach is highly parameter-efficient. It effectively preserves the capabilities of pre-trained models and prevents overfitting to downstream tasks. Extensive experiments show that such a simple scheme surpasses previous tuning baselines for both discriminate and generative tasks.

2304. ACC-Collab: An Actor-Critic Approach to Multi-Agent LLM Collaboration

链接：https://iclr.cc/virtual/2025/poster/28401 abstract： Large language models (LLMs) have demonstrated a remarkable ability to serve as general-purpose tools for various language-based tasks. Recent works have demonstrated that the efficacy of such models can be improved through iterative dialog between multiple models. While these paradigms show promise in improving model efficacy, most works in this area treat collaboration as an emergent behavior, rather than a learned behavior. In doing so, current multi-agent frameworks rely on collaborative behaviors to have been sufficiently trained into off-the-shelf models. To address this limitation, we propose ACC-Collab, an Actor-Critic based learning framework to produce a two-agent team (an actor-agent and a critic-agent) specialized in collaboration. We demonstrate that ACC-Collab outperforms SotA multi-agent techniques on a wide array of benchmarks.

2305. Improving Language Model Distillation through Hidden State Matching

链接：https://iclr.cc/virtual/2025/poster/30163 abstract： Hidden State Matching is shown to improve knowledge distillation of language models by encouraging similarity between a student and its teacher's hidden states since DistilBERT. This typically uses a cosine loss, which restricts the dimensionality of the student to the teacher's, severely limiting the compression ratio. We present an alternative technique using Centered Kernel Alignment (CKA) to match hidden states of different dimensionality, allowing for smaller students and higher compression ratios. We show the efficacy of our method using encoder--decoder (BART, mBART \& T5) and encoder-only (BERT) architectures across a range of tasks from classification to summarization and translation. Our technique is competitive with the current state-of-the-art distillation methods at comparable compression rates and does not require already pretrained student models. It can scale to students smaller than the current methods, is no slower in training and inference, and is considerably more flexible. The code is available on github.

2306. Bridging and Modeling Correlations in Pairwise Data for Direct Preference Optimization

链接：https://iclr.cc/virtual/2025/poster/28762 abstract： Direct preference optimization (DPO), a widely adopted offline preference optimization algorithm, aims to align large language models (LLMs) with human-desired behaviors using pairwise preference data. However, the generation of the winning response and the losing response within pairwise data are typically isolated, leading to weak correlations between them as well as suboptimal alignment performance. To address this issue, we propose an effective framework for Bridging and Modeling Correlations in pairwise data, named BMC. Firstly, we increase the consistency and informativeness of the pairwise preference signals through targeted modifications, synthesizing a pseudo-winning response by improving the losing response with the winning response as a reference. Secondly, we identify that DPO alone is insufficient to model these correlations and capture nuanced variations. Therefore, we propose learning token-level correlations by dynamically leveraging the policy model's confidence during training. Comprehensive experiments on QA, math, and instruction-following tasks demonstrate the effectiveness of our approach, significantly surpassing competitive baselines, including DPO. Additionally, our in-depth quantitative analysis reveals the reasons behind our method's superior performance over DPO and showcases its versatility to other DPO variants.

2307. Lightweight Neural App Control

链接：https://iclr.cc/virtual/2025/poster/30577 abstract： This paper introduces a novel mobile phone control architecture, Lightweight Multi-modal App Control (LiMAC), for efficient interactions and control across various Android apps. LiMAC takes as input a textual goal and a sequence of past mobile observations, such as screenshots and corresponding UI trees, to generate precise actions. To address the computational constraints inherent to smartphones, we introduce a small Action Transformer (AcT) integrated with a fine-tuned vision-language model (VLM) for real-time decision-making and task execution. We evaluate LiMAC on two open-source mobile control datasets, demonstrating the superior performance of our small-form-factor approach against fine-tuned versions of open-source VLMs, such as Florence2 and Qwen2-VL. It also significantly outperforms prompt engineering baselines utilising closed-source foundation models like GPT-4o. More specifically, LiMAC increases the overall action accuracy by up to 19% compared to fine-tuned VLMs, and up to 42% compared to prompt-engineering baselines.

2308. Plastic Learning with Deep Fourier Features

链接：https://iclr.cc/virtual/2025/poster/29885 abstract： Deep neural networks can struggle to learn continually in the face of non-stationarity, a phenomenon known as loss of plasticity. In this paper, we identify underlying principles that lead to plastic algorithms. We provide theoretical results showing that linear function approximation, as well as a special case of deep linear networks, do not suffer from loss of plasticity. We then propose deep Fourier features, which are the concatenation of a sine and cosine in every layer, and we show that this combination provides a dynamic balance between the trainability obtained through linearity and the effectiveness obtained through the nonlinearity of neural networks. Deep networks composed entirely of deep Fourier features are highly trainable and sustain their trainability over the course of learning. Our empirical results show that continual learning performance can be improved by replacing ReLU activations with deep Fourier features combined with regularization. These results hold for different continual learning scenarios (e.g., label noise, class incremental learning, pixel permutations) on all major supervised learning datasets used for continual learning research, such as CIFAR10, CIFAR100, and tiny-ImageNet.

2309. A Common Pitfall of Margin-based Language Model Alignment: Gradient Entanglement

链接：https://iclr.cc/virtual/2025/poster/29259 abstract： Reinforcement Learning from Human Feedback (RLHF) has become the predominant approach for aligning language models (LMs) to be more helpful and less harmful. At its core, RLHF uses a margin-based loss for preference optimization, which specifies the ideal LM behavior only in terms of the difference between preferred and dispreferred responses. In this paper, we identify a common pitfall of margin-based methods---the under-specification of ideal LM behavior on preferred and dispreferred responses individually, which results in two unintended consequences as the margin increases:(1) The probability of dispreferred (e.g., unsafe) responses may increase, resulting in potential safety alignment failures.(2) The probability of preferred responses may decrease, even when those responses are ideal.We demystify the reasons behind these problematic behaviors: margin-based losses couple the change in the preferred probability with the gradient of the dispreferred one, and vice versa, often preventing the preferred probability from increasing while the dispreferred one decreases, and thus causing a synchronized increase or decrease in both probabilities. We term this effect, inherent in margin-based objectives, gradient entanglement. Formally, we derive conditions for general margin-based alignment objectives under which gradient entanglement becomes concerning: the inner product between the gradient of preferred log-probability and the gradient of dispreferred log-probability is large relative to the individual gradient norms. Furthermore, we theoretically investigate why such inner products can be large when aligning language models and empirically validate our findings. Empirical implications of our framework further extend to explaining important differences in the training dynamics of various preference optimization algorithms and suggesting future directions for improvement.

2310. DPaI: Differentiable Pruning at Initialization with Node-Path Balance Principle

链接：https://iclr.cc/virtual/2025/poster/28732 abstract： Pruning at Initialization (PaI) is a technique in neural network optimization characterized by the proactive elimination of weights before the network's training on designated tasks. This innovative strategy potentially reduces the costs for training and inference, significantly advancing computational efficiency. A key factor leading to PaI's effectiveness is that it considers the saliency of weights in an untrained network, and prioritizes the trainability and optimization potential of the pruned subnetworks. Recent methods can effectively prevent the formation of hard-to-optimize networks, e.g. through iterative adjustments at each network layer. However, this way often results in large-scale discrete optimization problems, which could make PaI further challenging. This paper introduces a novel method, called DPaI, that involves a differentiable optimization of the pruning mask. DPaI adopts a dynamic and adaptable pruning process, allowing easier optimization processes and better solutions. More importantly, our differentiable formulation enables readily use of the existing rich body of efficient gradient-based methods for PaI. Our empirical results demonstrate that DPaI significantly outperforms current state-of-the-art PaI methods on various architectures, such as Convolutional Neural Networks and Vision-Transformers. Code is available at https://github.com/QuanNguyen-Tri/DPaI.git

2311. Erasing Concept Combination from Text-to-Image Diffusion Model

链接：https://iclr.cc/virtual/2025/poster/29839 abstract： Advancements in the text-to-image diffusion model have raised security concerns due to their potential to generate images with inappropriate themes such as societal biases and copyright infringements. Current studies have made notable progress in preventing the model from generating images containing specific high-risk visual concepts. However, these methods neglect the issue that inappropriate themes may also arise from the combination of benign visual concepts. A crucial challenge arises because the same image theme can be represented through multiple distinct visual concept combinations, and the model's ability to generate individual concepts may become distorted when processing these combinations. Consequently, effectively erasing such visual concept combinations from the diffusion model remains a formidable challenge. To tackle this problem, we formalize the problem as the Concept Combination Erasing (CCE) problem and propose a Concept Graph-based high-level Feature Decoupling framework (CoGFD) to address CCE. CoGFD identifies and decomposes visual concept combinations with a consistent image theme from an LLM-induced concept logic graph, and erases these combinations through decoupling co-occurrent high-level features. These techniques enable CoGFD to eliminate undesirable visual concept combinations while minimizing adverse effects on the generative fidelity of related individual concepts, outperforming state-of-the-art baselines. Extensive experiments across diverse visual concept combination scenarios verify the effectiveness of CoGFD.

2312. Test-time Adaptation for Cross-modal Retrieval with Query Shift

链接：https://iclr.cc/virtual/2025/poster/30549 abstract： The success of most existing cross-modal retrieval methods heavily relies on the assumption that the given queries follow the same distribution of the source domain. However, such an assumption is easily violated in real-world scenarios due to the complexity and diversity of queries, thus leading to the query shift problem.Specifically, query shift refers to the online query stream originating from the domain that follows a different distribution with the source one.In this paper, we observe that query shift would not only diminish the uniformity (namely, within-modality scatter) of the query modality but also amplify the gap between query and gallery modalities. Based on the observations, we propose a novel method dubbed Test-time adaptation for Cross-modal Retrieval (TCR). In brief, TCR employs a novel module to refine the query predictions (namely, retrieval results of the query) and a joint objective to prevent query shift from disturbing the common space, thus achieving online adaptation for the cross-modal retrieval models with query shift.Expensive experiments demonstrate the effectiveness of the proposed TCR against query shift. Code is available at https://github.com/XLearning-SCU/2025-ICLR-TCR.

2313. What's the Move? Hybrid Imitation Learning via Salient Points

链接：https://iclr.cc/virtual/2025/poster/28225 abstract： While imitation learning (IL) offers a promising framework for teaching robots various behaviors, learning complex tasks remains challenging. Existing IL policies struggle to generalize effectively across visual and spatial variations even for simple tasks. In this work, we introduce SPHINX: Salient Point-based Hybrid ImitatioN and eXecution, a flexible IL policy that leverages multimodal observations (point clouds and wrist images), along with a hybrid action space of low-frequency, sparse waypoints and high-frequency, dense end effector movements. Given 3D point cloud observations, SPHINX learns to infer task-relevant points within a point cloud, or salient points, which support spatial generalization by focusing on semantically meaningful features. These salient points serve as anchor points to predict waypoints for long-range movement, such as reaching target poses in free-space. Once near a salient point, SPHINX learns to switch to predicting dense end-effector movements given close-up wrist images for precise phases of a task. By exploiting the strengths of different input modalities and action representations for different manipulation phases, SPHINX tackles complex tasks in a sample-efficient, generalizable manner. Our method achieves 86.7% success across 4 real-world and 2 simulated tasks, outperforming the next best state-of-the-art IL baseline by 41.1% on average across 440 real world trials. SPHINX additionally generalizes to novel viewpoints, visual distractors, spatial arrangements, and execution speeds with a 1.7x speedup over the most competitive baseline. Our website (http://sphinx-manip.github.io) provides open-sourced code for data collection, training, and evaluation, along with supplementary videos.

2314. Metalic: Meta-Learning In-Context with Protein Language Models

链接：https://iclr.cc/virtual/2025/poster/29540 abstract： Predicting the biophysical and functional properties of proteins is essential for in silico protein design. Machine learning has emerged as a promising technique for such prediction tasks. However, the relative scarcity of in vitro annotations means that these models often have little, or no, specific data on the desired fitness prediction task. As a result of limited data, protein language models (PLMs) are typically trained on general protein sequence modeling tasks, and then fine-tuned, or applied zero-shot, to protein fitness prediction. When no task data is available, the models make strong assumptions about the correlation between the protein sequence likelihood and fitness scores. In contrast, we propose meta-learning over a distribution of standard fitness prediction tasks, and demonstrate positive transfer to unseen fitness prediction tasks. Our method, called Metalic (Meta-Learning In-Context), uses in-context learning and fine-tuning, when data is available, to adapt to new tasks. Crucially, fine-tuning enables considerable generalization, even though it is not accounted for during meta-training. Our fine-tuned models achieve strong results with 18 times fewer parameters than state-of-the-art models. Moreover, our method sets a new state-of-the-art in low-data settings on ProteinGym, an established fitness-prediction benchmark. Due to data scarcity, we believe meta-learning will play a pivotal role in advancing protein engineering.

2315. Unlocking Global Optimality in Bilevel Optimization: A Pilot Study

链接：https://iclr.cc/virtual/2025/poster/31102 abstract： Bilevel optimization has witnessed a resurgence of interest, driven by its critical role in trustworthy and efficient AI applications. Recent focus has been on finding efficient methods with provable convergence guarantees. However, while many prior works have established convergence to stationary points or local minima, obtaining the global optimum of bilevel optimization remains an important yet open problem. The difficulty lies in the fact that unlike many prior non-convex single-level problems, bilevel problems often do not admit a ``benign" landscape, and may indeed have multiple spurious local solutions. Nevertheless, attaining the global optimality is indispensable for ensuring reliability, safety, and cost-effectiveness, particularly in high-stakes engineering applications that rely on bilevel optimization. In this paper, we first explore the challenges of establishing a global convergence theory for bilevel optimization, and present two sufficient conditions for global convergence. We provide {\em algorithm-dependent} proofs to rigorously substantiate these sufficient conditions on two specific bilevel learning scenarios: representation learning and data hypercleaning (a.k.a. reweighting). Experiments corroborate the theoretical findings, demonstrating convergence to global minimum in both cases.

2316. GOPlan: Goal-conditioned Offline Reinforcement Learning by Planning with Learned Models

链接：https://iclr.cc/virtual/2025/poster/31488 abstract： Offline Goal-Conditioned RL (GCRL) offers a feasible paradigm for learning general-purpose policies from diverse and multi-task offline datasets. Despite notable recent progress, the predominant offline GCRL methods, mainly model-free, face constraints in handling limited data and generalizing to unseen goals. In this work, we propose Goal-conditioned Offline Planning (GOPlan), a novel model-based framework that contains two key phases: (1) pretraining a prior policy capable of capturing multi-modal action distribution within the multi-goal dataset; (2) employing the reanalysis method with planning to generate imagined trajectories for funetuning policies. Specifically, we base the prior policy on an advantage-weighted conditioned generative adversarial network, which facilitates distinct mode separation, mitigating the pitfalls of out-of-distribution (OOD) actions. For further policy optimization, the reanalysis method generates high-quality imaginary data by planning with learned models for both intra-trajectory and inter-trajectory goals. With thorough experimental evaluations, we demonstrate that GOPlan achieves state-of-the-art performance on various offline multi-goal navigation and manipulation tasks. Moreover, our results highlight the superior ability of GOPlan to handle small data budgets and generalize to OOD goals.

2317. Balancing Act: Diversity and Consistency in Large Language Model Ensembles

链接：https://iclr.cc/virtual/2025/poster/30442 abstract： Ensembling strategies for Large Language Models (LLMs) have demonstrated significant potential in improving performance across various tasks by combining the strengths of individual models. However, identifying the most effective ensembling method remains an open challenge, as neither maximizing output consistency through self-consistency decoding nor enhancing model diversity via frameworks like "Mixture of Agents" has proven universally optimal. Motivated by this, we propose a unified framework to examine the trade-offs between task performance, model diversity, and output consistency in ensembles. More specifically, we introduce a consistency score that defines a gating mechanism for mixtures of agents and an algorithm for mixture refinement to investigate these trade-offs at the semantic and model levels, respectively. We incorporate our insights into a novel inference-time LLM ensembling strategy called the Dynamic Mixture of Agents (DMoA) and demonstrate that it achieves a new state-of-the-art result in the challenging Big Bench Hard mixed evaluations benchmark. Our analysis reveals that cross-validation bias can enhance performance, contingent on the expertise of the constituent models. We further demonstrate that distinct reasoning tasks—such as arithmetic reasoning, commonsense reasoning, and instruction following—require different model capabilities, leading to inherent task-dependent trade-offs that DMoA balances effectively.

2318. What should a neuron aim for? Designing local objective functions based on information theory

链接：https://iclr.cc/virtual/2025/poster/30518 abstract： In modern deep neural networks, the learning dynamics of individual neurons are often obscure, as the networks are trained via global optimization. Conversely, biological systems build on self-organized, local learning, achieving robustness and efficiency with limited global information. Here, we show how self-organization between individual artificial neurons can be achieved by designing abstract bio-inspired local learning goals. These goals are parameterized using a recent extension of information theory, Partial Information Decomposition (PID), which decomposes the information that a set of information sources holds about an outcome into unique, redundant and synergistic contributions. Our framework enables neurons to locally shape the integration of information from various input classes, i.e., feedforward, feedback, and lateral, by selecting which of the three inputs should contribute uniquely, redundantly or synergistically to the output. This selection is expressed as a weighted sum of PID terms, which, for a given problem, can be directly derived from intuitive reasoning or via numerical optimization, offering a window into understanding task-relevant local information processing. Achieving neuron-level interpretability while enabling strong performance using local learning, our work advances a principled information-theoretic foundation for local learning strategies.

2319. Tackling Data Corruption in Offline Reinforcement Learning via Sequence Modeling

链接：https://iclr.cc/virtual/2025/poster/28283 abstract： Learning policy from offline datasets through offline reinforcement learning (RL) holds promise for scaling data-driven decision-making while avoiding unsafe and costly online interactions. However, real-world data collected from sensors or humans often contains noise and errors, posing a significant challenge for existing offline RL methods, particularly when the real-world data is limited. Our study reveals that prior research focusing on adapting predominant offline RL methods based on temporal difference learning still falls short under data corruption when the dataset is limited. In contrast, we discover that vanilla sequence modeling methods, such as Decision Transformer, exhibit robustness against data corruption, even without specialized modifications. To unlock the full potential of sequence modeling, we propose Robust Decision Transformer (RDT) by incorporating three simple yet effective robust techniques: embedding dropout to improve the model's robustness against erroneous inputs, Gaussian weighted learning to mitigate the effects of corrupted labels, and iterative data correction to eliminate corrupted data from the source. Extensive experiments on MuJoCo, Kitchen, and Adroit tasks demonstrate RDT's superior performance under various data corruption scenarios compared to prior methods. Furthermore, RDT exhibits remarkable robustness in a more challenging setting that combines training-time data corruption with test-time observation perturbations. These results highlight the potential of sequence modeling for learning from noisy or corrupted offline datasets, thereby promoting the reliable application of offline RL in real-world scenarios.Our code is available at https://github.com/jiawei415/RobustDecisionTransformer。

2320. DynaMath: A Dynamic Visual Benchmark for Evaluating Mathematical Reasoning Robustness of Vision Language Models

链接：https://iclr.cc/virtual/2025/poster/29415 abstract： The rapid advancements in Vision-Language Models (VLMs) have shown great potential in tackling mathematical reasoning tasks that involve visual context. Unlike humans who can reliably apply solution steps to similar problems with minor modifications, we found that state-of-the-art VLMs like GPT-4o can consistently fail in these scenarios, revealing limitations in their mathematical reasoning capabilities. In this paper, we investigate the mathematical reasoning robustness in VLMs and evaluate how well these models perform under different variants of the same question, such as changes in visual numerical values or function graphs.While several vision-based math benchmarks have been developed to assess VLMs' problem-solving capabilities, these benchmarks contain only static sets of problems and cannot easily evaluate mathematical reasoning robustness.To fill this gap, we introduce DynaMath, a dynamic visual math benchmark designed for in-depth assessment of VLMs. DynaMath includes 501 high-quality, multi-topic seed questions, each represented as a Python program. Those programs are carefully designed and annotated to enable the automatic generation of a much larger set of concrete questions, including many different types of visual and textual variations. DynaMath allows us to evaluate the generalization ability of VLMs, by assessing their performance under varying input conditions of a seed question. We evaluated 14 state-of-the-art VLMs with 5,010 generated concrete questions (10 per seed question). Our results show that the worst-case model accuracy, defined as the percentage of correctly answered seed questions in all 10 variants, is significantly lower than the average-case accuracy. In addition, many models show high consistency in answering these questions -- the incorrectness of a certain variant of a seed question is not only due to inherent randomness. Our analysis emphasizes the need to study the robustness of VLMs' reasoning abilities, and DynaMath provides valuable insights to guide the development of more reliable models for mathematical reasoning.

2321. Fast Summation of Radial Kernels via QMC Slicing

链接：https://iclr.cc/virtual/2025/poster/28699 abstract： The fast computation of large kernel sums is a challenging task, which arises as a subproblem in any kernel method. We approach the problem by slicing, which relies on random projections to one-dimensional subspaces and fast Fourier summation. We prove bounds for the slicing error and propose a quasi-Monte Carlo (QMC) approach for selecting the projections based on spherical quadrature rules. Numerical examples demonstrate that our QMC-slicing approach significantly outperforms existing methods like (QMC-)random Fourier features, orthogonal Fourier features or non-QMC slicing on standard test datasets.

2322. DELTA: DENSE EFFICIENT LONG-RANGE 3D TRACKING FOR ANY VIDEO

链接：https://iclr.cc/virtual/2025/poster/29002 abstract： Tracking dense 3D motion from monocular videos remains challenging, particularly when aiming for pixel-level precision over long sequences. We introduce DELTA, a novel method that efficiently tracks every pixel in 3D space, enabling accurate motion estimation across entire videos. Our approach leverages a joint global-local attention mechanism for reduced-resolution tracking, followed by a transformer-based upsampler to achieve high-resolution predictions. Unlike existing methods, which are limited by computational inefficiency or sparse tracking, DELTA delivers dense 3D tracking at scale, running over 8x faster than previous methods while achieving state-of-the-art accuracy. Furthermore, we explore the impact of depth representation on tracking performance and identify log-depth as the optimal choice. Extensive experiments demonstrate the superiority of DELTA on multiple benchmarks, achieving new state-of-the-art results in both 2D and 3D dense tracking tasks. Our method provides a robust solution for applications requiring fine-grained, long-term motion tracking in 3D space.

2323. Tailoring Mixup to Data for Calibration

链接：https://iclr.cc/virtual/2025/poster/31046 abstract： Among all data augmentation techniques proposed so far, linear interpolation of training samples, also called Mixup, has found to be effective for a large panel of applications. Along with improved predictive performance, Mixup is also a good technique for improving calibration. However, mixing data carelessly can lead to manifold mismatch, i.e., synthetic data lying outside original class manifolds, which can deteriorate calibration. In this work, we show that the likelihood of assigning a wrong label with mixup increases with the distance between data to mix. To this end, we propose to dynamically change the underlying distributions of interpolation coefficients depending on the similarity between samples to mix, and define a flexible framework to do so without losing in diversity. We provide extensive experiments for classification and regression tasks, showing that our proposed method improves predictive performance and calibration of models, while being much more efficient.

2324. Restyling Unsupervised Concept Based Interpretable Networks with Generative Models

链接：https://iclr.cc/virtual/2025/poster/30505 abstract： Developing inherently interpretable models for prediction has gained prominence in recent years. A subclass of these models, wherein the interpretable network relies on learning high-level concepts, are valued because of closeness of concept representations to human communication. However, the visualization and understanding of the learnt unsupervised dictionary of concepts encounters major limitations, especially for large-scale images.

We propose here a novel method that relies on mapping the concept features to the latent space of a pretrained generative model. The use of a generative model enables high quality visualization, and lays out an intuitive and interactive procedure for better interpretation of the learnt concepts by imputing concept activations and visualizing generated modifications. Furthermore, leveraging pretrained generative models has the additional advantage of making the training of the system more efficient. We quantitatively ascertain the efficacy of our method in terms of accuracy of the interpretable prediction network, fidelity of reconstruction, as well as faithfulness and consistency of learnt concepts. The experiments are conducted on multiple image recognition benchmarks for large-scale images. Project page available at https://jayneelparekh.github.io/VisCoINprojectpage/

2325. BlendRL: A Framework for Merging Symbolic and Neural Policy Learning

链接：https://iclr.cc/virtual/2025/poster/30913 abstract： Humans can leverage both symbolic reasoning and intuitive responses. In contrast, reinforcement learning policies are typically encoded in either opaque systems like neural networks or symbolic systems that rely on predefined symbols and rules. This disjointed approach severely limits the agents' capabilities, as they often lack either the flexible low-level reaction characteristic of neural agents or the interpretable reasoning of symbolic agents. To overcome this challenge, we introduce BlendRL, a neuro-symbolic RL framework that harmoniously integrates both paradigms. We empirically demonstrate that BlendRL agents outperform both neural and symbolic baselines in standard Atari environments, and showcase their robustness to environmental changes. Additionally, we analyze the interaction between neural and symbolic policies, illustrating how their hybrid use helps agents overcome each other's limitations.

2326. MGMapNet: Multi-Granularity Representation Learning for End-to-End Vectorized HD Map Construction

链接：https://iclr.cc/virtual/2025/poster/30420 abstract： The construction of vectorized high-definition map typically requires capturing both category and geometry information of map elements. Current state-of-the-art methods often adopt solely either point-level or instance-level representation, overlooking the strong intrinsic relationship between points and instances. In this work, we propose a simple yet efficient framework named MGMapNet (multi-granularity map network) to model map elements with multi-granularity representation, integrating both coarse-grained instance-level and fine-grained point-level queries. Specifically, these two granularities of queries are generated from the multi-scale bird's eye view features using a proposed multi-granularity aggregator. In this module, instance-level query aggregates features over the entire scope covered by an instance, and the point-level query aggregates features locally. Furthermore, a point-instance interaction module is designed to encourage information exchange between instance-level and point-level queries. Experimental results demonstrate that the proposed MGMapNet achieves state-of-the-art performances, surpassing MapTRv2 by 5.3 mAP on the nuScenes dataset and 4.4 mAP on the Argoverse2 dataset, respectively.

2327. MUSE: Machine Unlearning Six-Way Evaluation for Language Models

链接：https://iclr.cc/virtual/2025/poster/29559 abstract： Language models (LMs) are trained on vast amounts of text data, which may include private and copyrighted content. Data owners may request the removal of their data from a trained model due to privacy or copyright concerns. However, exactly unlearning only these datapoints (i.e., retraining with the data removed) is intractable in modern-day models. This has led to the development of many approximate unlearning algorithms. The evaluation of the efficacy of these algorithms has traditionally been narrow in scope, failing to precisely quantify the success and practicality of the algorithm from the perspectives of both the model deployers and the data owners. We address this issue by proposing MUSE, a comprehensive machine unlearning evaluation benchmark that enumerates six diverse desirable properties for unlearned models: (1) no verbatim memorization, (2) no knowledge memorization, (3) no privacy leakage, (4) utility preservation on data not intended for removal, (5) scalability with respect to the size of removal requests, and (6) sustainability over sequential unlearning requests. Using these criteria, we benchmark how effectively eight popular unlearning algorithms on 7B-parameter LMs can unlearn Harry Potter books and news articles. Our results demonstrate that most algorithms can prevent verbatim memorization and knowledge memorization to varying degrees, but only one algorithm does not lead to severe privacy leakage. Furthermore, existing algorithms fail to meet deployer's expectations because they often degrade general model utility and also cannot sustainably accommodate successive unlearning requests or large-scale content removal. Our findings identify key issues with the practicality of existing unlearning algorithms on language models.

2328. GMValuator: Similarity-based Data Valuation for Generative Models

链接：https://iclr.cc/virtual/2025/poster/29345 abstract： Data valuation plays a crucial role in machine learning. Existing data valuation methods, mainly focused on discriminative models, overlook generative models that have gained attention recently. In generative models, data valuation measures the impact of training data on generated datasets. Very few existing attempts at data valuation methods designed for deep generative models either concentrate on specific models or lack robustness in their outcomes. Moreover, efficiency still reveals vulnerable shortcomings. We formulate the data valuation problem in generative models from a similarity matching perspective to bridge the gaps. Specifically, we introduce Generative Model Valuator (GMValuator), the first training-free and model-agnostic approach to providing data valuation for image generation tasks. It empowers efficient data valuation through our innovative similarity matching module, calibrates biased contributions by incorporating image quality assessment, and attributes credits to all training samples based on their contributions to the generated samples. Additionally, we introduce four evaluation criteria for assessing data valuation methods in generative models. GMValuator is extensively evaluated on benchmark and high-resolution datasets and various mainstream generative architectures to demonstrate its effectiveness. Our code is available at: https://github.com/ubc-tea/GMValuator.

2329. On Evaluating the Durability of Safeguards for Open-Weight LLMs

链接：https://iclr.cc/virtual/2025/poster/28875 abstract： Many stakeholders---from model developers to policymakers---seek to minimize the risks of large language models (LLMs). Key to this goal is whether technical safeguards can impede the misuse of LLMs, even when models are customizable via fine-tuning or when model weights are openly available. Several recent studies have proposed methods to produce durable LLM safeguards for open-weight LLMs that can withstand adversarial modifications of the model's weights via fine-tuning. This holds the promise of raising adversaries' costs even under strong threat models where adversaries can directly fine-tune parameters. However, we caution against over-reliance on such methods in their current state. Through several case studies, we demonstrate that even the evaluation of these defenses is exceedingly difficult and can easily mislead audiences into thinking that safeguards are more durable than they really are. We draw lessons from the failure modes that we identify and suggest that future research carefully cabin claims to more constrained, well-defined, and rigorously examined threat models, which can provide useful and candid assessments to stakeholders.

2330. Not All LLM-Generated Data Are Equal: Rethinking Data Weighting in Text Classification

链接：https://iclr.cc/virtual/2025/poster/28366 abstract： Synthetic data augmentation via Large Language Models (LLMs) allows researchers to leverage additional training data, thus enhancing the performance of downstream tasks, especially when real-world data is scarce. However, the generated data can deviate from the real-world data, and this misalignment can bring about deficient results while applying the trained model to applications. Therefore, we proposed efficient weighted-loss approaches to align synthetic data with real-world distribution by emphasizing high-quality and diversified data generated by LLMs using merely a tiny amount of real-world data. We empirically assessed the effectiveness of our methods on multiple text classification tasks, and the results showed that leveraging our approaches on a BERT-level model robustly outperformed standard cross-entropy and other data weighting approaches, providing potential solutions to effectively leveraging synthetic data from any suitable data generator.

2331. LeanVec: Searching vectors faster by making them fit

链接：https://iclr.cc/virtual/2025/poster/31495 abstract： Modern deep learning models have the ability to generate high-dimensional vectors whose similarity reflects semantic resemblance. Thus, similarity search, i.e., the operation of retrieving those vectors in a large collection that are similar to a given query, has become a critical component of a wide range of applications that demand highly accurate and timely answers. In this setting, the high vector dimensionality puts similarity search systems under compute and memory pressure, leading to subpar performance. Additionally, cross-modal retrieval tasks have become increasingly common, e.g., where a user inputs a text query to find the most relevant images for that query. However, these queries often have different distributions than the database embeddings, making it challenging to achieve high accuracy. In this work, we present LeanVec, a framework that combines linear dimensionality reduction with vector quantization to accelerate similarity search on high-dimensional vectors while maintaining accuracy. We present LeanVec variants for in-distribution (ID) and out-of-distribution (OOD) queries. LeanVec-ID yields accuracies on par with those from recently introduced deep learning alternatives whose computational overhead precludes their usage in practice. LeanVec-OOD uses a novel technique for dimensionality reduction that considers the query and database distributions to simultaneously boost the accuracy and the performance of the framework even further (even presenting competitive results when the query and database distributions match). All in all, our extensive and varied experimental results show that LeanVec produces state-of-the-art results, with up to 3.7x improvement in search throughput and up to 4.9x faster index build time over the state of the art.

2332. HiBug2: Efficient and Interpretable Error Slice Discovery for Comprehensive Model Debugging

链接：https://iclr.cc/virtual/2025/poster/28538 abstract： Despite the significant success of deep learning models in computer vision, they often exhibit systematic failures on specific data subsets, known as error slices. Identifying and mitigating these error slices is crucial to enhancing model robustness and reliability in real-world scenarios. In this paper, we introduce HiBug2, an automated framework for error slice discovery and model repair. HiBug2 first generates task-specific visual attributes to highlight instances prone to errors through an interpretable and structured process. It then employs an efficient slice enumeration algorithm to systematically identify error slices, overcoming the combinatorial challenges that arise during slice exploration. Additionally, HiBug2 extends its capabilities by predicting error slices beyond the validation set, addressing a key limitation of prior approaches. Extensive experiments across multiple domains — including image classification, pose estimation, and object detection — show that HiBug2 not only improves the coherence and precision of identified error slices but also significantly enhances the model repair capabilities.

2333. Gated Delta Networks: Improving Mamba2 with Delta Rule

链接：https://iclr.cc/virtual/2025/poster/28219 abstract： Linear Transformers have gained attention as efficient alternatives to standard Transformers, but their performance in retrieval and long-context tasks has been limited. To address these limitations, recent work has explored two distinct mechanisms: gating for adaptive memory control and the delta update rule for precise memory modifications. We observe that these mechanisms are complementary—gating enables rapid memory erasure while the delta rule facilitates targeted updates. Building on this insight, we introduce the gated delta rule and develop a parallel training algorithm optimized for modern hardware. Our proposed architecture, Gated DeltaNet, consistently surpasses existing models like Mamba2 and DeltaNet across multiple benchmarks, including language modeling, common-sense reasoning, in-context retrieval, length extrapolation, and long-context understanding. We further enhance performance by developing hybrid architectures that combine Gated DeltaNet layers with sliding window attention or Mamba2 layers, achieving both improved training efficiency and superior task performance.

2334. No Free Lunch: Fundamental Limits of Learning Non-Hallucinating Generative Models

链接：https://iclr.cc/virtual/2025/poster/29791 abstract： Generative models have shown impressive capabilities in synthesizing high-quality outputs across various domains. However, a persistent challenge is the occurrence of "hallucinations," where the model produces outputs that are not grounded in the underlying facts. While empirical strategies have been explored to mitigate this issue, a rigorous theoretical understanding remains elusive. In this paper, we develop a theoretical framework to analyze the learnability of non-hallucinating generative models from a learning-theoretic perspective. Our results reveal that non-hallucinating learning is statistically impossible when relying solely on the training dataset, even for a hypothesis class of size two and when the entire training set is truthful. To overcome these limitations, we show that incorporating inductive biases aligned with the actual facts into the learning process is essential. We provide a systematic approach to achieve this by restricting the fact set to a concept class of finite VC-dimension and demonstrate its effectiveness under various learning paradigms. Although our findings are primarily conceptual, they represent a first step towards a principled approach to addressing hallucinations in learning generative models.

2335. Efficient Dictionary Learning with Switch Sparse Autoencoders

链接：https://iclr.cc/virtual/2025/poster/28604 abstract： Sparse autoencoders (SAEs) are a recent technique for decomposing neural network activations into human-interpretable features. However, in order for SAEs to identify all features represented in frontier models, it will be necessary to scale them up to very high width, posing a computational challenge. In this work, we introduce Switch Sparse Autoencoders, a novel SAE architecture aimed at reducing the compute cost of training SAEs. Inspired by sparse mixture of experts models, Switch SAEs route activation vectors between smaller "expert" SAEs, enabling SAEs to efficiently scale to many more features. We present experiments comparing Switch SAEs with other SAE architectures, and find that Switch SAEs deliver a substantial Pareto improvement in the reconstruction vs. sparsity frontier for a given fixed training compute budget. We also study the geometry of features across experts, analyze features duplicated across experts, and verify that Switch SAE features are as interpretable as features found by other SAE architectures.

2336. Curriculum-aware Training for Discriminating Molecular Property Prediction Models

链接：https://iclr.cc/virtual/2025/poster/30904 abstract： Despite their wide application across various fields, current molecular property prediction models struggle with the challenge of activity cliff, which refers to the situation where molecules with similar chemical structures display remarkable different properties. This phenomenon hinders existing models' ability to learn distinctive representations for molecules with similar chemical structures, and results in inaccurate predictions on molecules with activity cliff. To address this limitation, we first present empirical evidence demonstrating the ineffectiveness of standard training pipelines on molecules with activity cliff. We propose a novel approach that reformulates molecular property prediction as a node classification problem, introducing two innovative tasks at both the node and edge levels to improve learning outcomes for these challenging molecules with activity cliff. Our method is versatile, allowing seamless integration with a variety of base models, whether pre-trained or randomly initialized. Extensive evaluation across different molecular property prediction datasets validate the effectiveness of our approach.

2337. MMIU: Multimodal Multi-image Understanding for Evaluating Large Vision-Language Models

链接：https://iclr.cc/virtual/2025/poster/29339 abstract： The capability to process multiple images is crucial for Large Vision-Language Models (LVLMs) to develop a more thorough and nuanced understanding of a scene. Recent multi-image LVLMs have begun to address this need. However, their evaluation has not kept pace with their development. To fill this gap, we introduce the Multimodal Multi-image Understanding (MMIU) benchmark, a comprehensive evaluation suite designed to assess LVLMs across a wide range of multi-image tasks. MMIU encompasses 7 types of multi-image relationships, 52 tasks, 77K images, and 11K meticulously curated multiple-choice questions, making it the most extensive benchmark of its kind. Our evaluation of nearly 30 popular LVLMs, including both open-source and proprietary models, reveals significant challenges in multi-image comprehension, particularly in tasks involving spatial understanding. Even the most advanced models, such as GPT-4o, achieve only 55.7\% accuracy on MMIU. Through multi-faceted analytical experiments, we identify key performance gaps and limitations, providing valuable insights for future model and data improvements. We aim for MMIU to advance the frontier of LVLM research and development. We release the data and code at https://github.com/MMIUBenchmark/MMIU.

2338. Trajectory attention for fine-grained video motion control

链接：https://iclr.cc/virtual/2025/poster/31100 abstract： Recent advancements in video generation have been greatly driven by video diffusion models, with camera motion control emerging as a crucial challenge in creating view-customized visual content. This paper introduces trajectory attention, a novel approach that performs attention along available pixel trajectories for fine-grained camera motion control. Unlike existing methods that often yield imprecise outputs or neglect temporal correlations, our approach possesses a stronger inductive bias that seamlessly injects trajectory information into the video generation process. Importantly, our approach models trajectory attention as an auxiliary branch alongside traditional temporal attention. This design enables the original temporal attention and the trajectory attention to work in synergy, ensuring bothprecise motion control and new content generation capability, which is critical when the trajectory is only partially available. Experiments on camera motion control for images and videos demonstrate significant improvements in precision and long-range consistency while maintaining high-quality generation. Furthermore, we show that our approach can be extended to other video motion control tasks, such as first-frame-guided video editing, where it excels in maintaining content consistency over large spatial and temporal ranges.

2339. MLE-bench: Evaluating Machine Learning Agents on Machine Learning Engineering

链接：https://iclr.cc/virtual/2025/poster/30860 abstract： We introduce MLE-bench, a benchmark for measuring how well AI agents perform at machine learning engineering. To this end, we curate 75 ML engineering-related competitions from Kaggle, creating a diverse set of challenging tasks that test real-world ML engineering skills such as training models, preparing datasets, and running experiments. We establish human baselines for each competition using Kaggle's publicly available leaderboards. We use open-source agent scaffolds to evaluate several frontier language models on our benchmark, finding that the best-performing setup — OpenAI's o1-preview with AIDE scaffolding — achieves at least the level of a Kaggle bronze medal in 16.9% of competitions. In addition to our main results, we investigate various forms of resource-scaling for AI agents and the impact of contamination from pre-training. We open-source our benchmark code https://github.com/openai/mle-bench to facilitate future research in understanding the ML engineering capabilities of AI agents.

2340. Surprising Effectiveness of pretraining Ternary Language Model at Scale

链接：https://iclr.cc/virtual/2025/poster/29550 abstract： Rapid advancements in GPU computational power has outpaced memory capacity and bandwidth growth, creating bottlenecks in Large Language Model (LLM) inference. Post-training quantization is the leading method for addressing memory-related bottlenecks in LLM inference, but it suffers from significant performance degradation below 4-bit precision. This paper addresses these challenges by investigating the pretraining of low-bitwidth models specifically Ternary Language Models (TriLMs) as an alternative to traditional floating-point models (FloatLMs) and their post-training quantized versions (QuantLMs). We present Spectra LLM suite, the first open suite of LLMs spanning multiple bit-widths, including FloatLMs, QuantLMs, and TriLMs, ranging from 99M to 3.9B parameters trained on 300B tokens. Our comprehensive evaluation demonstrates that TriLMs offer superior scaling behavior in terms of model size (in bits). Surprisingly, at scales exceeding one billion parameters, TriLMs consistently outperform their QuantLM and FloatLM counterparts for a given bit size across various benchmarks. Notably, the 3.9B parameter TriLM matches the performance of the FloatLM 3.9B across all benchmarks, despite having fewer bits than FloatLM 830M. Overall, this research provides valuable insights into the feasibility and scalability of low-bitwidth language models, paving the way for the development of more efficient LLMs.

2341. PARTNR: A Benchmark for Planning and Reasoning in Embodied Multi-agent Tasks

链接：https://iclr.cc/virtual/2025/poster/29562 abstract： We present a benchmark for Planning And Reasoning Tasks in humaN-Robot collaboration (PARTNR) designed to study human-robot coordination in household activities. PARTNR tasks exhibit characteristics of everyday tasks, such as spatial, temporal, and heterogeneous agent capability constraints. We employ a semi-automated task generation pipeline using Large Language Models (LLMs), incorporating simulation-in-the-loop for the grounding and verification. PARTNR stands as the largest benchmark of its kind, comprising 100,000 natural language tasks, spanning 60 houses and 5,819 unique objects. We analyze state-of-the-art LLMs on PARTNR tasks, across the axes of planning, perception and skill execution. The analysis reveals significant limitations in SoTA models, such as poor coordination and failures in task tracking and recovery from errors. When LLMs are paired with 'real' humans, they require 1.5x as many steps as two humans collaborating and 1.1x more steps than a single human, underscoring the potential for improvement in these models. We further show that fine-tuning smaller LLMs with planning data can achieve performance on par with models 9 times larger, while being 8.6x faster at inference. Overall, PARTNR highlights significant challenges facing collaborative embodied agents and aims to drive research in this direction.

2342. Rationalizing and Augmenting Dynamic Graph Neural Networks

链接：https://iclr.cc/virtual/2025/poster/28032 abstract： Graph data augmentation (GDA) has shown significant promise in enhancing the performance, generalization, and robustness of graph neural networks (GNNs). However, contemporary methodologies are often limited to static graphs, whose applicability on dynamic graphs—more prevalent in real-world applications—remains unexamined. In this paper, we empirically highlight the challenges faced by static GDA methods when applied to dynamic graphs, particularly their inability to maintain temporal consistency. In light of this limitation, we propose a dedicated augmentation framework for dynamic graphs, termed $\texttt{DyAug}$, which adaptively augments the evolving graph structure with temporal consistency awareness. Specifically, we introduce the paradigm of graph rationalization for dynamic GNNs, progressively distinguishing between causal subgraphs (\textit{rationale}) and the non-causal complement (\textit{environment}) across snapshots. We develop three types of environment replacement, including, spatial, temporal, and spatial-temporal, to facilitate data augmentation in the latent representation space, thereby improving the performance, generalization, and robustness of dynamic GNNs. Extensive experiments on six benchmarks and three GNN backbones demonstrate that $\texttt{DyAug}$ can \textbf{(I)} improve the performance of dynamic GNNs by $0.89\%\sim3.13\%\uparrow$; \textbf{(II)} effectively counter targeted and non-targeted adversarial attacks with $6.2\%\sim12.2\%\uparrow$ performance boost; \textbf{(III)} make stable predictions under temporal distribution shifts.

2343. Large Language Models can Become Strong Self-Detoxifiers

链接：https://iclr.cc/virtual/2025/poster/28633 abstract： Reducing the likelihood of generating harmful and toxic output is an essential task when aligning large language models (LLMs). Existing methods mainly rely on training an external reward model (i.e., another language model) or fine-tuning the LLM using self-generated data to influence the outcome. In this paper, we show that LLMs have the capability of self-detoxification without external reward model learning or retraining of the LM. We propose \textit{Self-disciplined Autoregressive Sampling (SASA)}, a lightweight controlled decoding algorithm for toxicity reduction of LLMs. SASA leverages the contextual representations from an LLM to learn linear subspaces from labeled data characterizing toxic v.s. non-toxic output in analytical forms. When auto-completing a response token-by-token, SASA dynamically tracks the margin of the current output to steer the generation away from the toxic subspace, by adjusting the autoregressive sampling strategy. Evaluated on LLMs of different scale and nature, namely Llama-3.1-Instruct (8B), Llama-2 (7B), and GPT2-L models with the RealToxicityPrompts, BOLD, and AttaQ benchmarks, SASA markedly enhances the quality of the generated sentences relative to the original models and attains comparable performance to state-of-the-art detoxification techniques, significantly reducing the toxicity level by only using the LLM's internal representations.

2344. From Promise to Practice: Realizing High-performance Decentralized Training

链接：https://iclr.cc/virtual/2025/poster/28499 abstract： Decentralized training of deep neural networks has attracted significant attention for its theoretically superior scalability compared to synchronous data-parallel methods like All-Reduce. However, realizing this potential in multi-node training is challenging due to the complex design space that involves communication topologies, computation patterns, and optimization algorithms. This paper identifies three key factors that can lead to speedups over All-Reduce training and constructs a runtime model to determine when and how decentralization can shorten the per-iteration runtimes. To support the decentralized training of transformer-based models, we introduce a decentralized Adam algorithm that overlaps communications with computations, prove its convergence, and propose an accumulation technique to mitigate the high variance caused by small local batch sizes. We deploy our solution in clusters with up to 64 GPUs, demonstrating its practical advantages in both runtime and generalization performance under a fixed iteration budget.The experiment code is open-source at https://github.com/WangZesen/Decentralized-Training-Exp, and the extension code is open-source at https://github.com/WangZesen/Decent-DP.

2345. Which Tasks Should Be Compressed Together? A Causal Discovery Approach for Efficient Multi-Task Representation Compression

链接：https://iclr.cc/virtual/2025/poster/27804 abstract： Conventional image compression methods are inadequate for intelligent analysis, as they overemphasize pixel-level precision while neglecting semantic significance and the interaction among multiple tasks. This paper introduces a Taskonomy-Aware Multi-Task Compression framework comprising (1) inter-coherent task grouping, which organizes synergistic tasks into shared representations to improve multi-task accuracy and reduce encoding volume, and (2) a conditional entropy-based directed acyclic graph (DAG) that captures causal dependencies among grouped representations. By leveraging parent representations as contextual priors for child representations, the framework effectively utilizes cross-task information to improve entropy model accuracy. Experiments on diverse vision tasks, including Keypoint 2D, Depth Z-buffer, Semantic Segmentation, Surface Normal, Edge Texture, and Autoencoder, demonstrate significant bitrate-performance gains, validating the method’s capability to reduce system entropy uncertainty. These findings underscore the potential of leveraging representation disentanglement, synergy, and causal modeling to learn compact representations, which enable efficient multi-task compression in intelligent systems.

2346. High-Quality Joint Image and Video Tokenization with Causal VAE

链接：https://iclr.cc/virtual/2025/poster/29168 abstract： Generative modeling has seen significant advancements in image and video synthesis. However, the curse of dimensionality remains a significant obstacle, especially for video generation, given its inherently complex and high-dimensional nature. Many existing works rely on low-dimensional latent spaces from pretrained image autoencoders. However, this approach overlooks temporal redundancy in videos and often leads to temporally incoherent decoding. To address this issue, we propose a video compression network that reduces the dimensionality of visual data both spatially and temporally. Our model, based on a variational autoencoder, employs causal 3D convolution to handle images and videos jointly. The key contributions of our work include a scale-agnostic encoder for preserving video fidelity, a novel spatio-temporal down/upsampling block for robust long-sequence modeling, and a flow regularization loss for accurate motion decoding. Our approach outperforms competitors in video quality and compression rates across various datasets. Experimental analyses also highlight its potential as a robust autoencoder for video generation training.

2347. ReGenesis: LLMs can Grow into Reasoning Generalists via Self-Improvement

链接：https://iclr.cc/virtual/2025/poster/29261 abstract： Post-training Large Language Models (LLMs) with explicit reasoning trajectories can enhance their reasoning abilities. However, acquiring such high-quality trajectory data typically demands meticulous supervision from humans or superior models, which can be either expensive or license-constrained. In this paper, we explore how far an LLM can improve its reasoning by self-synthesizing reasoning paths as training data without any additional supervision. Existing self-synthesizing methods, such as STaR, suffer from poor generalization to out-of-domain (OOD) reasoning tasks. We hypothesize it is due to that their self-synthesized reasoning paths are too task-specific, lacking general task-agnostic reasoning guidance. To address this, we propose Reasoning Generalist via Self-Improvement (ReGenesis), a method to self-synthesize reasoning paths as post-training data by progressing from abstract to concrete. More specifically, ReGenesis self-synthesizes reasoning paths by converting general reasoning guidelines into task-specific ones, generating reasoning structures, and subsequently transforming these structures into reasoning paths, without the need for human-designed task-specific examples used in existing methods. We show that ReGenesis achieves superior performance on all in-domain and OOD settings tested compared to existing methods. For six OOD tasks specifically, while previous methods exhibited an average performance decrease of approximately 4.6% after post training, ReGenesis delivers around 6.1% performance improvement. We also conduct an in-depth analysis of our framework and show ReGenesis is effective across various language models and design choices.

2348. Attribute-based Visual Reprogramming for Vision-Language Models

链接：https://iclr.cc/virtual/2025/poster/28650 abstract： *Visual reprogramming* (VR) reuses pre-trained vision models for downstream image classification tasks by adding trainable noise patterns to inputs. When applied to vision-language models (e.g., CLIP), existing VR approaches follow the same pipeline used in vision models (e.g., ResNet, ViT), where ground-truth class labels are inserted into fixed text templates to guide the optimization of VR patterns. This label-based approach, however, overlooks the rich information and diverse attribute-guided textual representations that CLIP can exploit, which may lead to the misclassification of samples. In this paper, we propose **_Attr_*ibute-based** V**isual** R**eprogramming** ***(AttrVR) for CLIP, utilizing*** des**criptive** attr**ibutes** ***(DesAttrs) and*** dist**inctive** attr*_ibutes_ (DistAttrs), which respectively represent common and unique feature descriptions for different classes. Besides, as images of the same class may reflect different attributes after VR, AttrVR iteratively refines patterns using the k-nearest DesAttrs and DistAttrs for each image sample, enabling more dynamic and sample-specific optimization. Theoretically, AttrVR is shown to reduce intra-class variance and increase inter-class separation. Empirically, it achieves superior performance in 12 downstream tasks for both ViT-based and ResNet-based CLIP. The success of AttrVR facilitates more effective integration of VR from unimodal vision models into vision-language models. Our code is available at https://github.com/tmlr-group/AttrVR.

2349. Maintaining Structural Integrity in Parameter Spaces for Parameter Efficient Fine-tuning

链接：https://iclr.cc/virtual/2025/poster/29842 abstract： Adapting pre-trained foundation models for various downstream tasks has been prevalent in artificial intelligence. Due to the vast number of tasks and high costs, adjusting all parameters becomes unfeasible. To mitigate this, several fine-tuning techniques have been developed to update the pre-trained model weights in a more resource-efficient manner, such as through low-rank adjustments. Yet, almost all of these methods focus on linear weights, neglecting the intricacies of parameter spaces in higher dimensions like 4D. Alternatively, some methods can be adapted for high-dimensional parameter space by compressing changes in the original space into two dimensions and then employing low-rank matrix adaptations. However, these approaches destructs the structural integrity of the involved high-dimensional spaces. To tackle the diversity of dimensional spaces across different foundation models and provide a more precise representation of the changes within these spaces, this paper introduces a generalized parameter-efficient fine-tuning framework, designed for various dimensional parameter space. Specifically, our method asserts that changes in each dimensional parameter space are based on a low-rank core space which maintains the consistent topological structure with the original space. It then models the changes through this core space alongside corresponding weights to reconstruct alterations in the original space. It effectively preserves the structural integrity of the change of original N-dimensional parameter space, meanwhile models it via low-rank tensor adaptation. Extensive experiments on computer vision, natural language processing and multi-modal tasks validate the effectiveness of our method.

2350. VideoPhy: Evaluating Physical Commonsense for Video Generation

链接：https://iclr.cc/virtual/2025/poster/30714 abstract： Recent advances in internet-scale video data pretraining have led to the development of text-to-video generative models that can create high-quality videos across a broad range of visual concepts, synthesize realistic motions and render complex objects. Hence, these generative models have the potential to become general-purpose simulators of the physical world. However, it is unclear how far we are from this goal with the existing text-to-video generative models. To this end, we present VideoPhy, a benchmark designed to assess whether the generated videos follow physical commonsense for real-world activities (e.g. marbles will roll down when placed on a slanted surface). Specifically, we curate diverse prompts that involve interactions between various material types in the physical world (e.g., solid-solid, solid-fluid, fluid-fluid). We then generate videos conditioned on these captions from diverse state-of-the-art text-to-video generative models, including open models (e.g., CogVideoX) and closed models (e.g., Lumiere, Dream Machine). Our human evaluation reveals that the existing models severely lack the ability to generate videos adhering to the given text prompts, while also lack physical commonsense. Specifically, the best performing model, CogVideoX-5B, generates videos that adhere to the caption and physical laws for 39.6% of the instances. VideoPhy thus highlights that the video generative models are far from accurately simulating the physical world. Finally, we propose an auto-evaluator, VideoCon-Physics, to assess the performance reliably for the newly released models. The code is available here: https://github.com/Hritikbansal/videophy.

2351. Text-to-Image Rectified Flow as Plug-and-Play Priors

链接：https://iclr.cc/virtual/2025/poster/29568 abstract： Large-scale diffusion models have achieved remarkable performance in generative tasks. Beyond their initial training applications, these models have proven their ability to function as versatile plug-and-play priors. For instance, 2D diffusion models can serve as loss functions to optimize 3D implicit models. Rectified Flow, a novel class of generative models, has demonstrated superior performance across various domains. Compared to diffusion-based methods, rectified flow approaches surpass them in terms of generation quality and efficiency. In this work, we present theoretical and experimental evidence demonstrating that rectified flow based methods offer similar functionalities to diffusion models — they can also serve as effective priors. Besides the generative capabilities of diffusion priors, motivated by the unique time-symmetry properties of rectified flow models, a variant of our method can additionally perform image inversion. Experimentally, our rectified flow based priors outperform their diffusion counterparts — the SDS and VSD losses — in text-to-3D generation. Our method also displays competitive performance in image inversion and editing. Code is available at: https://github.com/yangxiaofeng/rectifiedflowprior.

2352. On the Adversarial Risk of Test Time Adaptation: An Investigation into Realistic Test-Time Data Poisoning

链接：https://iclr.cc/virtual/2025/poster/30848 abstract： Test-time adaptation (TTA) updates the model weights during the inference stage using testing data to enhance generalization. However, this practice exposes TTA to adversarial risks. Existing studies have shown that when TTA is updated with crafted adversarial test samples, also known as test-time poisoned data, the performance on benign samples can deteriorate. Nonetheless, the perceived adversarial risk may be overstated if the poisoned data is generated under overly strong assumptions. In this work, we first review realistic assumptions for test-time data poisoning, including white-box versus grey-box attacks, access to benign data, attack order, and more. We then propose an effective and realistic attack method that better produces poisoned samples without access to benign samples, and derive an effective in-distribution attack objective. We also design two TTA-aware attack objectives. Our benchmarks of existing attack methods reveal that the TTA methods are more robust than previously believed. In addition, we analyze effective defense strategies to help develop adversarially robust TTA methods. The source code is available at https://github.com/Gorilla-Lab-SCUT/RTTDP.

2353. Evidential Learning-based Certainty Estimation for Robust Dense Feature Matching

链接：https://iclr.cc/virtual/2025/poster/31015 abstract： Dense feature matching methods aim to estimate a dense correspondence field between images. Inaccurate correspondence can occur due to the presence of unmatchable region, necessitating the need for certainty measurement. This is typically addressed by training a binary classifier to decide whether each predicted correspondence is reliable. However, deep neural network-based classifiers can be vulnerable to image corruptions or perturbations, making it difficult to obtain reliable matching pairs in corrupted scenario. In this work, we propose an evidential deep learning framework to enhance the robustness of dense matching against corruptions. We modify the certainty prediction branch in dense matching models to generate appropriate belief masses and compute the certainty score by taking expectation over the resulting Dirichlet distribution. We evaluate our method on a wide range of benchmarks and show that our method leads to improved robustness against common corruptions and adversarial attacks, achieving up to 10.1\% improvement under severe corruptions.

2354. Long-Sequence Recommendation Models Need Decoupled Embeddings

链接：https://iclr.cc/virtual/2025/poster/28620 abstract： Lifelong user behavior sequences are crucial for capturing user interests and predicting user responses in modern recommendation systems. A two-stage paradigm is typically adopted to handle these long sequences: a subset of relevant behaviors is first searched from the original long sequences via an attention mechanism in the first stage and then aggregated with the target item to construct a discriminative representation for prediction in the second stage. In this work, we identify and characterize, for the first time, a neglected deficiency in existing long-sequence recommendation models: a single set of embeddings struggles with learning both attention and representation, leading to interference between these two processes. Initial attempts to address this issue with some common methods (e.g., linear projections---a technique borrowed from language processing) proved ineffective, shedding light on the unique challenges of recommendation models. To overcome this, we propose the Decoupled Attention and Representation Embeddings (DARE) model, where two distinct embedding tables are initialized and learned separately to fully decouple attention and representation. Extensive experiments and analysis demonstrate that DARE provides more accurate searches of correlated behaviors and outperforms baselines with AUC gains up to 9‰ on public datasets and notable improvements on Tencent's advertising platform. Furthermore, decoupling embedding spaces allows us to reduce the attention embedding dimension and accelerate the search procedure by 50\% without significant performance impact, enabling more efficient, high-performance online serving. Code in PyTorch for experiments, including model analysis, is available at https://github.com/thuml/DARE.

2355. Outlier Synthesis via Hamiltonian Monte Carlo for Out-of-Distribution Detection

链接：https://iclr.cc/virtual/2025/poster/29899 abstract： Out-of-distribution (OOD) detection is crucial for developing trustworthy and reliable machine learning systems. Recent advances in training with auxiliary OOD data demonstrate efficacy in enhancing detection capabilities. Nonetheless, these methods heavily rely on acquiring a large pool of high-quality natural outliers. Some prior methods try to alleviate this problem by synthesizing virtual outliers but suffer from either poor quality or high cost due to the monotonous sampling strategy and the heavy-parameterized generative models. In this paper, we overcome all these problems by proposing the Hamiltonian Monte Carlo Outlier Synthesis (HamOS) framework, which views the synthesis process as sampling from Markov chains. Based solely on the in-distribution data, the Markov chains can extensively traverse the feature space and generate diverse and representative outliers, hence exposing the model to miscellaneous potential OOD scenarios. The Hamiltonian Monte Carlo with sampling acceptance rate almost close to 1 also makes our framework enjoy great efficiency. By empirically competing with SOTA baselines on both standard and large-scale benchmarks, we verify the efficacy and efficiency of our proposed HamOS.

2356. Style Outweighs Substance: Failure Modes of LLM Judges in Alignment Benchmarking

链接：https://iclr.cc/virtual/2025/poster/29906 abstract： The release of ChatGPT in November 2022 sparked an explosion of interest in post-training and an avalanche of new preference optimization (PO) methods. These methods claim superior alignment by virtue of better correspondence with human pairwise preferences, often measured by LLM-judges. In this work, we attempt to answer the following question -- do LLM-judge preferences translate to progress on other, more concrete metrics for alignment, and if not, why not? We define a concrete metric for alignment, and introduce SOS-Bench (Substance Outweighs Style Benchmark), the largest standardized, reproducible LLM meta-benchmark to date. We find that (1) LLM-judge preferences do not correlate with concrete measures of safety, world knowledge, and instruction following; (2) LLM-judges have powerful implicit biases, prioritizing style over factuality and safety; and (3) the supervised fine-tuning (SFT) stage of post-training has a large impact on alignment, with data scaling and prompt diversity as the driving factors.

2357. AtomSurf: Surface Representation for Learning on Protein Structures

链接：https://iclr.cc/virtual/2025/poster/30633 abstract： While there has been significant progress in evaluating and comparing different representations for learning on protein data, the role of surface-based learning approaches remains not well-understood. In particular, there is a lack of direct and fair benchmark comparison between the best available surface-based learning methods against alternative representations such as graphs. Moreover, the few existing surface-based approaches either use surface information in isolation or, at best, perform global pooling between surface and graph-based architectures. In this work, we fill this gap by first adapting a state-of-the-art surface encoder for protein learning tasks. We then perform a direct and fair comparison of the resulting method against alternative approaches within the Atom3D benchmark, highlighting the limitations of pure surface-based learning. Finally, we propose an integrated approach, which allows learned feature sharing between graphs and surface representations on the level of nodes and vertices \textit{across all layers}. We demonstrate that the resulting architecture achieves state-of-the-art results on all tasks in the Atom3D benchmark, while adhering to the strict benchmark protocol, as well as more broadly on binding site identification and binding pocket classification. Furthermore, we use coarsened surfaces and optimize our approach for efficiency, making our tool competitive in training and inference time with existing techniques.Code can be found online: https://github.com/Vincentx15/atomsurf

2358. Policy Design in Long-run Welfare Dynamics

链接：https://iclr.cc/virtual/2025/poster/29004 abstract： Improving social welfare is a complex challenge requiring policymakers to optimize objectives across multiple time horizons. Evaluating the impact of such policies presents a fundamental challenge, as those that appear suboptimal in the short run may yield significant long-term benefits. We tackle this challenge by analyzing the long-term dynamics of two prominent policy frameworks: Rawlsian policies, which prioritize those with the greatest need, and utilitarian policies, which maximize immediate welfare gains. Conventional wisdom suggests these policies are at odds, as Rawlsian policies are assumed to come at the cost of reducing the average social welfare, which their utilitarian counterparts directly optimize. We challenge this assumption by analyzing these policies in a sequential decision-making framework where individuals' welfare levels stochastically decay over time, and policymakers can intervene to prevent this decay. Under reasonable assumptions, we prove that interventions following Rawlsian policies can outperform utilitarian policies in the long run, even when the latter dominate in the short run. We characterize the exact conditions under which Rawlsian policies can outperform utilitarian policies. We further illustrate our theoretical findings using simulations, which highlight the risks of evaluating policies based solely on their short-term effects. Our results underscore the necessity of considering long-term horizons in designing and evaluating welfare policies; the true efficacy of even well-established policies may only emerge over time.

2359. Controllable Satellite-to-Street-View Synthesis with Precise Pose Alignment and Zero-Shot Environmental Control

链接：https://iclr.cc/virtual/2025/poster/28891 abstract： Generating street-view images from satellite imagery is a challenging task, particularly in maintaining accurate pose alignment and incorporating diverse environmental conditions. While diffusion models have shown promise in generative tasks, their ability to maintain strict pose alignment throughout the diffusion process is limited. In this paper, we propose a novel Iterative Homography Adjustment (IHA) scheme applied during the denoising process, which effectively addresses pose misalignment and ensures spatial consistency in the generated street-view images. Additionally, currently, available datasets for satellite-to-street-view generation are limited in their diversity of illumination and weather conditions, thereby restricting the generalizability of the generated outputs. To mitigate this, we introduce a text-guided illumination and weather-controlled sampling strategy that enables fine-grained control over the environmental factors. Extensive quantitative and qualitative evaluations demonstrate that our approach significantly improves pose accuracy and enhances the diversity and realism of generated street-view images, setting a new benchmark for satellite-to-street-view generation tasks.

2360. ImProver: Agent-Based Automated Proof Optimization

链接：https://iclr.cc/virtual/2025/poster/28980 abstract：

2361. KAA: Kolmogorov-Arnold Attention for Enhancing Attentive Graph Neural Networks

链接：https://iclr.cc/virtual/2025/poster/29141 abstract： Graph neural networks (GNNs) with attention mechanisms, often referred to as attentive GNNs, have emerged as a prominent paradigm in advanced GNN models in recent years. However, our understanding of the critical process of scoring neighbor nodes remains limited, leading to the underperformance of many existing attentive GNNs. In this paper, we unify the scoring functions of current attentive GNNs and propose Kolmogorov-Arnold Attention (KAA), which integrates the Kolmogorov-Arnold Network (KAN) architecture into the scoring process. KAA enhances the performance of scoring functions across the board and can be applied to nearly all existing attentive GNNs. To compare the expressive power of KAA with other scoring functions, we introduce Maximum Ranking Distance (MRD) to quantitatively estimate their upper bounds in ranking errors for node importance. Our analysis reveals that, under limited parameters and constraints on width and depth, both linear transformation-based and MLP-based scoring functions exhibit finite expressive power. In contrast, our proposed KAA, even with a single-layer KAN parameterized by zero-order B-spline functions, demonstrates nearly infinite expressive power. Extensive experiments on both node-level and graph-level tasks using various backbone models show that KAA-enhanced scoring functions consistently outperform their original counterparts, achieving performance improvements of over 20% in some cases.

2362. Policy Gradient with Kernel Quadrature

链接：https://iclr.cc/virtual/2025/poster/31498 abstract： Reward evaluation of episodes becomes a bottleneck in a broad range of reinforcement learning tasks. Our aim in this paper is to select a small but representative subset of a large batch of episodes, only on which we actually compute rewards for more efficient policy gradient iterations. We build a Gaussian process modeling of discounted returns or rewards to derive a positive definite kernel on the space of episodes, run an ``episodic'' kernel quadrature method to compress the information of sample episodes, and pass the reduced episodes to the policy network for gradient updates. We present the theoretical background of this procedure as well as its numerical illustrations in MuJoCo tasks.

2363. Hydra-SGG: Hybrid Relation Assignment for One-stage Scene Graph Generation

链接：https://iclr.cc/virtual/2025/poster/28023 abstract： DETR introduces a simplified one-stage framework for scene graph generation (SGG) but faces challenges of sparse supervision and false negative samples. The former occurs because each image typically contains fewer than 10 relation annotations, while DETR-based SGG models employ over 100 relation queries. Each ground truth relation is assigned to only one query during training. The latter arises when one ground truth relation may have multiple queries with similar matching scores, leading to suboptimally matched queries being treated as negative samples. To address these, we propose Hydra-SGG, a one-stage SGG method featuring a Hybrid Relation Assignment. This approach combines a One-to-One Relation Assignment with an IoU-based One-to-Many Relation Assignment, increasing positive training samples and mitigating sparse supervision. In addition, we empirically demonstrate that removing self-attention between relation queries leads to duplicate predictions, which actually benefits the proposed One-to-Many Relation Assignment. With this insight, we introduce Hydra Branch, an auxiliary decoder without self-attention layers, to further enhance One-to-Many Relation Assignment by promoting different queries to make the same relation prediction. Hydra-SGG achieves state-of-the-art performance on multiple datasets, including VG150 (16.0 mR@50), Open Images V6 (50.1 weighted score), and GQA (12.7 mR@50). Our code and pre-trained models will be released on Hydra-SGG.

2364. DECO: Unleashing the Potential of ConvNets for Query-based Detection and Segmentation

链接：https://iclr.cc/virtual/2025/poster/29537 abstract： Transformer and its variants have shown great potential for various vision tasks in recent years, including image classification, object detection and segmentation. Meanwhile, recent studies also reveal that with proper architecture design, convolutional networks (ConvNets) also achieve competitive performance with transformers. However, no prior methods have explored to utilize pure convolution to build a Transformer-style Decoder module, which is essential for Encoder-Decoder architecture like Detection Transformer (DETR).To this end, in this paper we explore whether we could build query-based detection and segmentation framework with ConvNets instead of sophisticated transformer architecture.We propose a novel mechanism dubbed InterConv to perform interaction between object queries and image features via convolutional layers. Equipped with the proposed InterConv, we build Detection ConvNet (DECO), which is composed of a backbone and convolutional encoder-decoder architecture. We compare the proposed DECO against prior detectors on the challenging COCO benchmark.Despite its simplicity, our DECO achieves competitive performance in terms of detection accuracy and running speed. Specifically, with the ResNet-18 and ResNet-50 backbone, our DECO achieves 40.5% and 47.8% AP with 66 and 34 FPS, respectively. The proposed method is also evaluated on the segment anything task, demonstrating similar performance and higher efficiency.We hope the proposed method brings another perspective for designing architectures for vision tasks.Codes are available at \url{https://github.com/xinghaochen/DECO} and \url{https://github.com/mindspore-lab/models/tree/master/research/huawei-noah/DECO}.

2365. A Theoretical Framework for Partially-Observed Reward States in RLHF

链接：https://iclr.cc/virtual/2025/poster/29804 abstract： The growing deployment of reinforcement learning from human feedback (RLHF) calls for a deeper theoretical investigation of its underlying models. The prevalent models of RLHF do not account for neuroscience-backed, partially-observed "internal states" that can affect human feedback, nor do they accommodate intermediate feedback during an interaction. Both of these can be instrumental in speeding up learning and improving alignment. To address these limitations, we model RLHF as reinforcement learning with partially observed reward-states (PORRL). We accommodate two kinds of feedback — cardinal and dueling feedback. We first demonstrate that PORRL subsumes a wide class of RL problems, including traditional RL, RLHF, and reward machines. For cardinal feedback, we present two model-based methods (POR-UCRL, POR-UCBVI). We give both cardinal regret and sample complexity guarantees for the methods, showing that they improve over naive history-summarization. We then discuss the benefits of a model-free method like GOLF with naive history-summarization in settings with recursive internal states and dense intermediate feedback. For this purpose, we define a new history aware version of the Bellman-eluder dimension and give a new guarantee for GOLF in our setting, which can be exponentially sharper in illustrative examples. For dueling feedback, we show that a naive reduction to cardinal feedback fails to achieve sublinear dueling regret. We then present the first explicit reduction that converts guarantees for cardinal regret to dueling regret. In both feedback settings, we show that our models and guarantees generalize and extend existing ones.

2366. LVSM: A Large View Synthesis Model with Minimal 3D Inductive Bias

链接：https://iclr.cc/virtual/2025/poster/29689 abstract： We propose the Large View Synthesis Model (LVSM), a novel transformer-based approach for scalable and generalizable novel view synthesis from sparse-view inputs. We introduce two architectures: (1) an encoder-decoder LVSM, which encodes input image tokens into a fixed number of 1D latent tokens, functioning as a fully learned scene representation, and decodes novel-view images from them; and (2) a decoder-only LVSM, which directly maps input images to novel-view outputs, completely eliminating intermediate scene representations. Both models bypass the 3D inductive biases used in previous methods---from 3D representations (e.g., NeRF, 3DGS) to network designs (e.g., epipolar projections, plane sweeps)---addressing novel view synthesis with a fully data-driven approach. While the encoder-decoder model offers faster inference due to its independent latent representation, the decoder-only LVSM achieves superior quality, scalability, and zero-shot generalization, outperforming previous state-of-the-art methods by 1.5 to 3.5 dB PSNR. Comprehensive evaluations across multiple datasets demonstrate that both LVSM variants achieve state-of-the-art novel view synthesis quality, delivering superior performance even with reduced computational resources (1-2 GPUs). Please see our anonymous website for more details: https://haian-jin.github.io/projects/LVSM/

2367. Global Convergence in Neural ODEs: Impact of Activation Functions

链接：https://iclr.cc/virtual/2025/poster/30612 abstract： Neural Ordinary Differential Equations (ODEs) have been successful in various applications due to their continuous nature and parameter-sharing efficiency. However, these unique characteristics also introduce challenges in training, particularly with respect to gradient computation accuracy and convergence analysis. In this paper, we address these challenges by investigating the impact of activation functions. We demonstrate that the properties of activation functions—specifically smoothness and nonlinearity—are critical to the training dynamics. Smooth activation functions guarantee globally unique solutions for both forward and backward ODEs, while sufficient nonlinearity is essential for maintaining the spectral properties of the Neural Tangent Kernel (NTK) during training. Together, these properties enable us to establish the global convergence of Neural ODEs under gradient descent in overparameterized regimes. Our theoretical findings are validated by numerical experiments, which not only support our analysis but also provide practical guidelines for scaling Neural ODEs, potentially leading to faster training and improved performance in real-world applications.

2368. InstaTrain: Adaptive Training via Ultra-Fast Natural Annealing within Dynamical Systems

链接：https://iclr.cc/virtual/2025/poster/29682 abstract： Time-series modeling is broadly adopted to capture underlying patterns present in historical data, allowing prediction of future values. However, one crucial aspect of such modeling is often overlooked: in highly dynamic environments, data distributions can shift drastically within a second or less. Under these circumstances, traditional predictive models, and even online learning methods, struggle to adapt to the ultra-fast and complex distribution shifts present in highly dynamic scenarios. To address this, we propose InstaTrain, a novel learning approach that enables ultra-fast model updates for real-world prediction tasks, thereby keeping pace with rapidly evolving data distributions. In this work, (1) we transform the slow and expensive training process into an ultra-fast natural annealing process within a dynamical system. (2) Leveraging a recently proposed electronic dynamical system, we augment the system with parameter update modules, extending its capabilities to encompass both rapid training and inference. Experimental results on highly dynamic datasets demonstrate that our method achieves orders-of-magnitude improvements in training speed and energy efficiency while delivering superior accuracy compared to baselines running on GPUs.

2369. Repetition Improves Language Model Embeddings

链接：https://iclr.cc/virtual/2025/poster/30621 abstract： Bidirectional models are considered essential for strong text embeddings. Recent approaches to adapt autoregressive language models (LMs) into strong text embedding models have largely had the requirement to modify the LM architecture to be bidirectional. We challenge this premise by introducing ``echo embeddings'' which converts autoregressive LMs into high quality text embedding models \emph{without} changing the architecture or requiring fine-tuning. By repeating the input and extracting embeddings from the repeated tokens—which have access to all original tokens—echo embeddings improve over classical LM embeddings by over 5\% in zero-shot settings. Our zero-shot embeddings nearly match those obtained by bidirectionally-converted LMs that undergo additional masked-language modeling training. Echo embeddings are also compatible with supervised fine-tuning, matching or outperforming bidirectionally-converted LMs in an apples-to-apples comparison, even with an identical compute budget during training and inference. Overall, repetition is a simple and effective strategy to circumvent the need for bidirectional attention in embedding models, paving the way towards a unified architecture for all NLP tasks.

2370. SPA-BENCH: A COMPREHENSIVE BENCHMARK FOR SMARTPHONE AGENT EVALUATION

链接：https://iclr.cc/virtual/2025/poster/29820 abstract： Smartphone agents are increasingly important for helping users control devices efficiently, with (Multimodal) Large Language Model (MLLM)-based approaches emerging as key contenders. Fairly comparing these agents is essential but challenging, requiring a varied task scope, the integration of agents with different implementations, and a generalisable evaluation pipeline to assess their strengths and weaknesses. In this paper, we present SPA-Bench, a comprehensive SmartPhone Agent Benchmark designed to evaluate (M)LLM-based agents in an interactive environment that simulates real-world conditions. SPA-Bench offers three key contributions: (1) A diverse set of tasks covering system and third-party apps in both English and Chinese, focusing on features commonly used in daily routines; (2) A plug-and-play framework enabling real-time agent interaction with Android devices, integrating over ten agents with the flexibility to add more; (3) A novel evaluation pipeline that automatically assesses agent performance across multiple dimensions, encompassing seven metrics related to task completion and resource consumption. Our extensive experiments across tasks and agents reveal challenges like interpreting mobile user interfaces, action grounding, memory retention, and execution costs. We propose future research directions to ease these difficulties, moving closer to real-world smartphone agent applications.

2371. h4rm3l: A Language for Composable Jailbreak Attack Synthesis

链接：https://iclr.cc/virtual/2025/poster/27663 abstract：

2372. Black-Box Detection of Language Model Watermarks

链接：https://iclr.cc/virtual/2025/poster/30423 abstract：

2373. Merging LoRAs like Playing LEGO: Pushing the Modularity of LoRA to Extremes Through Rank-Wise Clustering

链接：https://iclr.cc/virtual/2025/poster/28655 abstract：

2374. Optimizing Backward Policies in GFlowNets via Trajectory Likelihood

Maximization

链接：https://iclr.cc/virtual/2025/poster/29295 abstract： Generative Flow Networks (GFlowNets) are a family of generative models that learn to sample objects with probabilities proportional to a given reward function. The key concept behind GFlowNets is the use of two stochastic policies: a forward policy, which incrementally constructs compositional objects, and a backward policy, which sequentially deconstructs them. Recent results show a close relationship between GFlowNet training and entropy-regularized reinforcement learning (RL) problems with a particular reward design. However, this connection applies only in the setting of a fixed backward policy, which might be a significant limitation. As a remedy to this problem, we introduce a simple backward policy optimization algorithm that involves direct maximization of the value function in an entropy-regularized Markov Decision Process (MDP) over intermediate rewards. We provide an extensive experimental evaluation of the proposed approach across various benchmarks in combination with both RL and GFlowNet algorithms and demonstrate its faster convergence and mode discovery in complex environments.

2375. Judge Decoding: Faster Speculative Sampling Requires Going Beyond Model Alignment

链接：https://iclr.cc/virtual/2025/poster/28440 abstract： The performance of large language models (LLMs) is closely linked to their underlying size, leading to ever-growing networks and hence slower inference. Speculative decoding has been proposed as a technique to accelerate autoregressive generation, leveraging a fast draft model to propose candidate tokens, which are then verified in parallel based on their likelihood under the target model. While this approach guarantees to reproduce the target output, it incurs a substantial penalty: many high-quality draft tokens are rejected, even when they represent objectively valid continuations. Indeed, we show that even powerful draft models such as GPT-4o, as well as human text cannot achieve high acceptance rates under the standard verification scheme. This severely limits the speedup potential of current speculative decoding methods, as an early rejection becomes overwhelmingly likely when solely relying on alignment of draft and target.We thus ask the following question: Can we adapt verification to recognize correct, but non-aligned replies? To this end, we draw inspiration from the LLM-as-a-judge framework, which demonstrated that LLMs are able to rate answers in a versatile way. We carefully design a dataset coined TokenCourt to elicit the same capability in the target model by training a compact module on top of the embeddings to produce ``judgements" of the current continuation. We showcase our strategy on the Llama-3.1 family, where our 8B/405B-Judge achieves a speedup of $9\times$ over Llama-405B, while maintaining its quality on a large range of benchmarks. These benefits remain present even in optimized inference frameworks, where our method reaches up to $141$ tokens/s for 8B/70B-Judge and $129$ tokens/s for 8B/405B on $2$ and $8$ H100s respectively.

2376. Regulatory DNA Sequence Design with Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/30369 abstract： $\textit{Cis}$-regulatory elements (CREs), such as promoters and enhancers, are relatively short DNA sequences that directly regulate gene expression. The fitness of CREs, measured by their ability to modulate gene expression, highly depends on the nucleotide sequences, especially specific motifs known as transcription factor binding sites (TFBSs). Designing high-fitness CREs is crucial for therapeutic and bioengineering applications. Current CRE design methods are limited by two major drawbacks: (1) they typically rely on iterative optimization strategies that modify existing sequences and are prone to local optima, and (2) they lack the guidance of biological prior knowledge in sequence optimization. In this paper, we address these limitations by proposing a generative approach that leverages reinforcement learning (RL) to fine-tune a pre-trained autoregressive (AR) model. Our method incorporates data-driven biological priors by deriving computational inference-based rewards that simulate the addition of activator TFBSs and removal of repressor TFBSs, which are then integrated into the RL process. We evaluate our method on promoter design tasks in two yeast media conditions and enhancer design tasks for three human cell types, demonstrating its ability to generate high-fitness CREs while maintaining sequence diversity. The code is available at https://github.com/yangzhao1230/TACO.

2377. Learning Transformer-based World Models with Contrastive Predictive Coding

链接：https://iclr.cc/virtual/2025/poster/29267 abstract： The DreamerV3 algorithm recently obtained remarkable performance across diverse environment domains by learning an accurate world model based on Recurrent Neural Networks (RNNs). Following the success of model-based reinforcement learning algorithms and the rapid adoption of the Transformer architecture for its superior training efficiency and favorable scaling properties, recent works such as STORM have proposed replacing RNN-based world models with Transformer-based world models using masked self-attention. However, despite the improved training efficiency of these methods, their impact on performance remains limited compared to the Dreamer algorithm, struggling to learn competitive Transformer-based world models. In this work, we show that the next state prediction objective adopted in previous approaches is insufficient to fully exploit the representation capabilities of Transformers. We propose to extend world model predictions to longer time horizons by introducing TWISTER (Transformer-based World model wIth contraSTivE Representations), a world model using action-conditioned Contrastive Predictive Coding to learn high-level temporal feature representations and improve the agent performance. TWISTER achieves a human-normalized mean score of 162% on the Atari 100k benchmark, setting a new record among state-of-the-art methods that do not employ look-ahead search. We release our code at https://github.com/burchim/TWISTER.

2378. Decoupling Angles and Strength in Low-rank Adaptation

链接：https://iclr.cc/virtual/2025/poster/29327 abstract： Parameter-Efficient FineTuning (PEFT) methods have recently gained significant popularity thanks to the widespread availability of large-scale pretrained models. These methods allow for quick adaptation to downstream tasks with minimal computational cost. However, popular finetuning methods such as LoRA exhibit limited robustness when it comes to hyperparameter choices or extended training regimes, preventing optimal out-of-the-box performance. In contrast, bounded approaches, such as ETHER, provide greater robustness but are limited to extremely low-rank adaptations and fixed-strength transformations, reducing their adaptation expressive power. In this work, we propose Decoupled Low-rank Adaptation (DeLoRA), a novel finetuning method that normalizes and scales learnable low-rank matrices. By bounding the distance of the transformation, DeLoRA effectively decouples the angular learning from the adaptation strength, enhancing robustness without compromising performance. Through evaluations on subject-driven image generation, natural language understanding, and instruction tuning, we show that DeLoRA matches or surpasses performance of competing PEFT methods, while exhibiting stronger robustness. Code is available at https://github.com/ExplainableML/DeLoRA.

2379. Variational Search Distributions

链接：https://iclr.cc/virtual/2025/poster/31174 abstract： We develop VSD, a method for conditioning a generative model of discrete, combinatorial designs on a rare desired class by efficiently evaluating a black-box (e.g. experiment, simulation) in a batch sequential manner. We call this task active generation; we formalize active generation's requirements and desiderata, and formulate a solution via variational inference. VSD uses off-the-shelf gradient based optimization routines, can learn powerful generative models for desirable designs, and can take advantage of scalable predictive models. We derive asymptotic convergence rates for learning the true conditional generative distribution of designs with certain configurations of our method. After illustrating the generative model on images, we empirically demonstrate that VSD can outperform existing baseline methods on a set of real sequence-design problems in various protein and DNA/RNA engineering tasks.

2380. Interpretable Bilingual Multimodal Large Language Model for Diverse Biomedical Tasks

链接：https://iclr.cc/virtual/2025/poster/32076 abstract： Several medical Multimodal Large Languange Models (MLLMs) have been developed to address tasks involving visual images with textual instructions across various medical modalities, achieving impressive results. Most current medical generalist models are region-agnostic, treating the entire image as a holistic representation. However, they struggle to identify which specific regions they are focusing on when generating a sentence.To mimic the behavior of doctors, who typically begin by reviewing the entire image before concentrating on specific regions for a thorough evaluation, we aim to enhance the capability of medical MLLMs in understanding anatomical regions within entire medical scans.To achieve it, we first formulate \textbf{Region-Centric tasks} and construct a \textbf{large-scale dataset, MedRegInstruct,} to incorporate regional information into training. Combining our collected dataset with other medical multimodal corpora for training, we propose a \textbf{Region-Aware medical MLLM, MedRegA}, which is the first bilingual generalist medical AI system to simultaneously handle image-level and region-level medical vision-language tasks across a broad range of modalities. Our MedRegA not only enables three region-centric tasks, but also achieves the best performance for visual question answering, report generation and medical image classification over 8 modalities, showcasing significant versatility. Experiments demonstrate that our model can not only accomplish powerful performance across various medical vision-language tasks in bilingual settings, but also recognize and detect structures in multimodal medical scans, boosting the interpretability and user interactivity of medical MLLMs. The codes and model will be made publicly available.

2381. Stable Hadamard Memory: Revitalizing Memory-Augmented Agents for Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/29355 abstract： Effective decision-making in partially observable environments demands robust memory management. Despite their success in supervised learning, current deep-learning memory models struggle in reinforcement learning environments that are partially observable and long-term. They fail to efficiently capture relevant past information, adapt flexibly to changing observations, and maintain stable updates over long episodes. We theoretically analyze the limitations of existing memory models within a unified framework and introduce the Stable Hadamard Memory, a novel memory model for reinforcement learning agents. Our model dynamically adjusts memory by erasing no longer needed experiences and reinforcing crucial ones computationally efficiently. To this end, we leverage the Hadamard product for calibrating and updating memory, specifically designed to enhance memory capacity while mitigating numerical and learning challenges. Our approach significantly outperforms state-of-the-art memory-based methods on challenging partially observable benchmarks, such as meta-reinforcement learning, long-horizon credit assignment, and POPGym, demonstrating superior performance in handling long-term and evolving contexts.

2382. Building, Reusing, and Generalizing Abstract Representations from Concrete Sequences

链接：https://iclr.cc/virtual/2025/poster/27794 abstract： Humans excel at learning abstract patterns across different sequences, filtering outirrelevant details, and transferring these generalized concepts to new sequences.In contrast, many sequence learning models lack the ability to abstract, whichleads to memory inefficiency and poor transfer. We introduce a non-parametrichierarchical variable learning model (HVM) that learns chunks from sequencesand abstracts contextually similar chunks as variables. HVM efficiently organizesmemory while uncovering abstractions, leading to compact sequence representations.When learning on language datasets such as babyLM, HVM learns a more efficientdictionary than standard compression algorithms such as Lempel-Ziv. In a sequencerecall task requiring the acquisition and transfer of variables embedded in sequences,we demonstrate HVM's sequence likelihood correlates with human recall times. Incontrast, large language models (LLMs) struggle to transfer abstract variables aseffectively as humans. From HVM's adjustable layer of abstraction, we demonstratethat the model realizes a precise trade-off between compression and generalization.Our work offers a cognitive model that captures the learning and transfer of abstractrepresentations in human cognition and differentiates itself from LLMs.

2383. DeeperForward: Enhanced Forward-Forward Training for Deeper and Better Performance

链接：https://iclr.cc/virtual/2025/poster/28588 abstract： While backpropagation effectively trains models, it presents challenges related to bio-plausibility, resulting in high memory demands and limited parallelism. Recently, Hinton (2022) proposed the Forward-Forward (FF) algorithm for high-parallel local updates. FF leverages squared sums as the local update target, termed goodness, and decouples goodness by normalizing the vector length to extract new features. However, this design encounters issues with feature scaling and deactivated neurons, limiting its application mainly to shallow networks. This paper proposes a novel goodness design utilizing layer normalization and mean goodness to overcome these challenges, demonstrating performance improvements even in 17-layer CNNs. Experiments on CIFAR-10, MNIST, and Fashion-MNIST show significant advantages over existing FF-based algorithms, highlighting the potential of FF in deep models. Furthermore, the model parallel strategy is proposed to achieve highly efficient training based on the property of local updates.

2384. ELBOing Stein: Variational Bayes with Stein Mixture Inference

链接：https://iclr.cc/virtual/2025/poster/31111 abstract： Stein variational gradient descent (SVGD) (Liu & Wang, 2016) performs approximate Bayesian inference by representing the posterior with a set of particles.However, SVGD suffers from variance collapse, i.e. poor predictions due to underestimating uncertainty (Ba et al., 2021), even for moderately-dimensional modelssuch as small Bayesian neural networks (BNNs). To address this issue, we generalize SVGD by letting each particle parameterize a component distribution ina mixture model. Our method, Stein Mixture Inference (SMI), optimizes a lowerbound to the evidence (ELBO) and introduces user-specified guides parameterizedby particles. SMI extends the Nonlinear SVGD framework (Wang & Liu, 2019) tothe case of variational Bayes. SMI effectively avoids variance collapse, judging bya previously described test developed for this purpose, and performs well on standard data sets. In addition, SMI requires considerably fewer particles than SVGDto accurately estimate uncertainty for small BNNs. The synergistic combination ofNSVGD, ELBO optimization and user-specified guides establishes a promisingapproach towards variational Bayesian inference in the case of tall and wide data.

2385. Revealing and Reducing Gender Biases in Vision and Language Assistants (VLAs)

链接：https://iclr.cc/virtual/2025/poster/28356 abstract： Pre-trained large language models (LLMs) have been reliably integrated with visual input for multimodal tasks. The widespread adoption of instruction-tuned image-to-text vision-language assistants (VLAs) like LLaVA and InternVL necessitates evaluating gender biases. We study gender bias in 22 popular open-source VLAs with respect to personality traits, skills, and occupations. Our results show that VLAs replicate human biases likely present in the data, such as real-world occupational imbalances. Similarly, they tend to attribute more skills and positive personality traits to women than to men, and we see a consistent tendency to associate negative personality traits with men. To eliminate the gender bias in these models, we find that fine-tuning-based debiasing methods achieve the best trade-off between debiasing and retaining performance on downstream tasks. We argue for pre-deploying gender bias assessment in VLAs and motivate further development of debiasing strategies to ensure equitable societal outcomes. Code is available at https://github.com/ExplainableML/vla-gender-bias.

2386. SAMRefiner: Taming Segment Anything Model for Universal Mask Refinement

链接：https://iclr.cc/virtual/2025/poster/30090 abstract： In this paper, we explore a principal way to enhance the quality of widely pre-existing coarse masks, enabling them to serve as reliable training data for segmentation models to reduce the annotation cost. In contrast to prior refinement techniques that are tailored to specific models or tasks in a close-world manner, we propose SAMRefiner, a universal and efficient approach by adapting SAM to the mask refinement task. The core technique of our model is the noise-tolerant prompting scheme. Specifically, we introduce a multi-prompt excavation strategy to mine diverse input prompts for SAM (\ie, distance-guided points, context-aware elastic bounding boxes, and Gaussian-style masks) from initial coarse masks. These prompts can collaborate with each other to mitigate the effect of defects in coarse masks. In particular, considering the difficulty of SAM to handle the multi-object case in semantic segmentation, we introduce a split-then-merge (STM) pipeline. Additionally, we extend our method to SAMRefiner++ by introducing an additional IoU adaption step to further boost the performance of the generic SAMRefiner on the target dataset. This step is self-boosted and requires no additional annotation. The proposed framework is versatile and can flexibly cooperate with existing segmentation methods. We evaluate our mask framework on a wide range of benchmarks under different settings, demonstrating better accuracy and efficiency. SAMRefiner holds significant potential to expedite the evolution of refinement tools. Our code is available at https://github.com/linyq2117/SAMRefiner.

2387. SVBench: A Benchmark with Temporal Multi-Turn Dialogues for Streaming Video Understanding

链接：https://iclr.cc/virtual/2025/poster/30198 abstract： Despite the significant advancements of Large Vision-Language Models (LVLMs) on established benchmarks, there remains a notable gap in suitable evaluation regarding their applicability in the emerging domain of long-context streaming video understanding. Current benchmarks for video understanding typically emphasize isolated single-instance text inputs and fail to evaluate the capacity to sustain temporal reasoning throughout the entire duration of video streams. To address these limitations, we introduce SVBench, a pioneering benchmark with temporal multi-turn question-answering chains specifically designed to thoroughly assess the capabilities of streaming video understanding of current LVLMs. We design a semi-automated annotation pipeline to obtain 49,979 Question-Answer (QA) pairs of 1,353 streaming videos, which includes generating QA chains that represent a series of consecutive multi-turn dialogues over video segments and constructing temporal linkages between successive QA chains. Our experimental results, obtained from 14 models in dialogue and streaming evaluations, reveal that while the closed-source GPT-4o outperforms others, most open-source LVLMs struggle with long-context streaming video understanding. We also construct a StreamingChat model, which significantly outperforms open-source LVLMs on our SVBench and achieves comparable performance on diverse vision-language benchmarks. We expect SVBench to advance the research of streaming video understanding by providing a comprehensive and in-depth analysis of current LVLMs. Our benchmark and model can be accessed at https://yzy-bupt.github.io/SVBench.

2388. Spherical Tree-Sliced Wasserstein Distance

链接：https://iclr.cc/virtual/2025/poster/30344 abstract： Sliced Optimal Transport (OT) simplifies the OT problem in high-dimensional spaces by projecting supports of input measures onto one-dimensional lines, then exploiting the closed-form expression of the univariate OT to reduce the computational burden of OT. Recently, the Tree-Sliced method has been introduced to replace these lines with more intricate structures, known as tree systems. This approach enhances the ability to capture topological information of integration domains in Sliced OT while maintaining low computational cost. Inspired by this approach, in this paper, we present an adaptation of tree systems on OT problem for measures supported on a sphere. As counterpart to the Radon transform variant on tree systems, we propose a novel spherical Radon transform, with a new integration domain called spherical trees. By leveraging this transform and exploiting the spherical tree structures, we derive closed-form expressions for OT problems on the sphere. Consequently, we obtain an efficient metric for measures on the sphere, named Spherical Tree-Sliced Wasserstein (STSW) distance. We provide an extensive theoretical analysis to demonstrate the topology of spherical trees, the well-definedness and injectivity of our Radon transform variant, which leads to an orthogonally invariant distance between spherical measures. Finally, we conduct a wide range of numerical experiments, including gradient flows and self-supervised learning, to assess the performance of our proposed metric, comparing it to recent benchmarks.

2389. Distance-Based Tree-Sliced Wasserstein Distance

链接：https://iclr.cc/virtual/2025/poster/29805 abstract： To overcome computational challenges of Optimal Transport (OT), several variants of Sliced Wasserstein (SW) has been developed in the literature. These approaches exploit the closed-form expression of the univariate OT by projecting measures onto one-dimensional lines. However, projecting measures onto low-dimensional spaces can lead to a loss of topological information. Tree-Sliced Wasserstein distance on Systems of Lines (TSW-SL) has emerged as a promising alternative that replaces these lines with a more intricate structure called tree systems. The tree structures enhance the ability to capture topological information of the metric while preserving computational efficiency. However, at the core of TSW-SL, the splitting maps, which serve as the mechanism for pushing forward measures onto tree systems, focus solely on the position of the measure supports while disregarding the projecting domains. Moreover, the specific splitting map used in TSW-SL leads to a metric that is not invariant under Euclidean transformations, a typically expected property for OT on Euclidean space. In this work, we propose a novel class of splitting maps that generalizes the existing one studied in TSW-SL enabling the use of all positional information from input measures, resulting in a novel Distance-based Tree-Sliced Wasserstein (Db-TSW) distance. In addition, we introduce a simple tree sampling process better suited for Db-TSW, leading to an efficient GPU-friendly implementation for tree systems, similar to the original SW. We also provide a comprehensive theoretical analysis of proposed class of splitting maps to verify the injectivity of the corresponding Radon Transform, and demonstrate that Db-TSW is an Euclidean invariant metric. We empirically show that Db-TSW significantly improves accuracy compared to recent SW variants while maintaining low computational cost via a wide range of experiments on gradient flows, image style transfer, and generative models.

2390. MAI: A Multi-turn Aggregation-Iteration Model for Composed Image Retrieval

链接：https://iclr.cc/virtual/2025/poster/28813 abstract： Multi-Turn Composed Image Retrieval (MTCIR) addresses a real-world scenario where users iteratively refine retrieval results by providing additional information until a target meeting all their requirements is found. Existing methods primarily achieve MTCIR through a "multiple single-turn" paradigm, wherein methods incorrectly converge on shortcuts that only utilize the most recent turn's image, ignoring attributes from historical turns. Consequently, retrieval failures occur when modification requests involve historical information. We argue that explicitly incorporating historical information into the modified text is crucial to addressing this issue. To this end, we build a new retrospective-based MTCIR dataset, FashionMT, wherein modification demands are highly associated with historical turns. We also propose a Multi-turn Aggregation-Iteration (MAI) model, emphasizing efficient aggregation of multimodal semantics and optimization of information propagation in multi-turn retrieval. Specifically, we propose a new Two-stage Semantic Aggregation (TSA) paradigm coupled with a Cyclic Combination Loss (CCL), achieving improved semantic consistency and modality alignment by progressively interacting the reference image with its caption and the modified text. In addition, we design a Multi-turn Iterative Optimization (MIO) mechanism that dynamically selects representative tokens and reduces redundancy during multi-turn iterations. Extensive experiments demonstrate that the proposed MAI model achieves substantial improvements over state-of-the-art methods.

2391. Disentangled Representation Learning with the Gromov-Monge Gap

链接：https://iclr.cc/virtual/2025/poster/28910 abstract： Learning disentangled representations from unlabelled data is a fundamental challenge in machine learning. Solving it may unlock other problems, such as generalization, interpretability, or fairness. Although remarkably challenging to solve in theory, disentanglement is often achieved in practice through prior matching. Furthermore, recent works have shown that prior matching approaches can be enhanced by leveraging geometrical considerations, e.g., by learning representations that preserve geometric features of the data, such as distances or angles between points. However, matching the prior while preserving geometric features is challenging, as a mapping that fully preserves these features while aligning the data distribution with the prior does not exist in general. To address these challenges, we introduce a novel approach to disentangled representation learning based on quadratic optimal transport. We formulate the problem using Gromov-Monge maps that transport one distribution onto another with minimal distortion of predefined geometric features, preserving them as much as can be achieved. To compute such maps, we propose the Gromov-Monge-Gap (GMG), a regularizer quantifying whether a map moves a reference distribution with minimal geometry distortion. We demonstrate the effectiveness of our approach for disentanglement across four standard benchmarks, outperforming other methods leveraging geometric considerations.

2392. Towards Robust Alignment of Language Models: Distributionally Robustifying Direct Preference Optimization

链接：https://iclr.cc/virtual/2025/poster/30508 abstract： This study addresses the challenge of noise in training datasets for Direct Preference Optimization (DPO), a method for aligning Large Language Models (LLMs) with human preferences. We categorize noise into pointwise noise, which includes low-quality data points, and pairwise noise, which encompasses erroneous data pair associations that affect preference rankings. Utilizing Distributionally Robust Optimization (DRO), we enhance DPO's resilience to these types of noise. Our theoretical insights reveal that DPO inherently embeds DRO principles, conferring robustness to pointwise noise, with the regularization coefficient $\beta$ playing a critical role in its noise resistance. Extending this framework, we introduce Distributionally Robustifying DPO (Dr. DPO), which integrates pairwise robustness by optimizing against worst-case pairwise scenarios. The novel hyperparameter $\beta'$ in Dr. DPO allows for fine-tuned control over data pair reliability, providing a strategic balance between exploration and exploitation in noisy training environments. Empirical evaluations demonstrate that Dr. DPO substantially improves the quality of generated text and response accuracy in preference datasets, showcasing enhanced performance in both noisy and noise-free settings.

2393. kNN Attention Demystified: A Theoretical Exploration for Scalable Transformers

链接：https://iclr.cc/virtual/2025/poster/31032 abstract： Despite their power, Transformers face challenges with long sequences due to the quadratic complexity of self-attention. To address this limitation, methods like $k$-Nearest-Neighbor ($k$NN) attention have been introduced [Roy et al., 2017], enabling each token to attend to only its $k$ closest tokens. While $k$NN attention has shown empirical success in making Transformers more efficient, its exact approximation guarantees have not been theoretically analyzed. In this work, we establish a theoretical framework for $k$NN attention, reformulating self-attention as expectations over softmax distributions and leveraging lazy Gumbel sampling [Mussmann et al., 2017] with $k$NN indices for efficient approximation. Building on this framework, we also propose novel sub-quadratic algorithms that approximate self-attention gradients by leveraging efficient sampling techniques, such as Markov Chain-based estimation. Finally, we demonstrate the practical effectiveness of these algorithms through empirical experiments, showcasing their benefits in both training and inference.

2394. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers

链接：https://iclr.cc/virtual/2025/poster/29961 abstract： Recent advancements in large language models (LLMs) have sparked optimism about their potential to accelerate scientific discovery, with a growing number of works proposing research agents that autonomously generate and validate new ideas. Despite this, no evaluations have shown that LLM systems can take the very first step of producing novel, expert-level ideas, let alone perform the entire research process. We address this by establishing an experimental design that evaluates research idea generation while controlling for confounders and performs the first comparison between expert NLP researchers and an LLM ideation agent. By recruiting over 100 NLP researchers to write novel ideas and blind reviews of both LLM and human ideas, we obtain the first statistically significant conclusion on current LLM capabilities for research ideation: we find LLM-generated ideas are judged as more novel ($p < 0.05$) than human expert ideas while being judged slightly weaker on feasibility. Studying our agent baselines closely, we identify open problems in building and evaluating research agents, including failures of LLM self-evaluation and their lack of diversity in generation.

2395. Chunk-Distilled Language Modeling

链接：https://iclr.cc/virtual/2025/poster/28391 abstract： We introduce Chunk-Distilled Language Modeling (CD-LM), an approach to text generation that addresses two challenges in current large language models (LLMs): the inefficiency of token-level generation, and the difficulty of adapting to new data and knowledge. Our method combines deep network-based LLMs with a straightforward retrieval module, which allows the generation of multi-token text chunks at a single decoding step. Our retrieval framework enables flexible construction of model- or domain-specific datastores, either leveraging the internal knowledge of existing models, or incorporating expert insights from human-annotated corpora. This adaptability allows for enhanced control over the language model's distribution without necessitating additional training. We present the CD-LM formulation along with performance metrics demonstrating its ability to improve language model performance and efficiency across a diverse set of downstream applications. Code and data will be made publicly available.

2396. u-μP: The Unit-Scaled Maximal Update Parametrization

链接：https://iclr.cc/virtual/2025/poster/29774 abstract： The Maximal Update Parametrization (μP) aims to make the optimal hyperparameters (HPs) of a model independent of its size, allowing them to be swept using a cheap proxy model rather than the full-size target model. We present a new scheme, u-μP, which improves upon μP by combining it with Unit Scaling, a method for designing models that makes them easy to train in low-precision. The two techniques have a natural affinity: μP ensures that the scale of activations is independent of model size, and Unit Scaling ensures that activations, weights and gradients begin training with a scale of one. This synthesis opens the door to a simpler scheme, whose default values are near-optimal. This in turn facilitates a more efficient sweeping strategy, with u-μP models reaching a lower loss than comparable μP models and working out-of-the-box in FP8.

2397. Diffusion Models Are Real-Time Game Engines

链接：https://iclr.cc/virtual/2025/poster/29770 abstract： We present GameNGen, the first game engine powered entirely by a neural model that also enables real-time interaction with a complex environment over long trajectories at high quality. When trained on the classic game DOOM, GameNGen extracts gameplay and uses it to generate a playable environment that can interactively simulate new trajectories. GameNGen runs at 20 frames per second on a single TPU and remains stable over extended multi-minute play sessions. Next frame prediction achieves a PSNR of 29.4, comparable to lossy JPEG compression. Human raters are only slightly better than random chance at distinguishing short clips of the game from clips of the simulation, even after 5 minutes of auto-regressive generation. GameNGen is trained in two phases: (1) an RL-agent learns to play the game and the training sessions are recorded, and (2) a diffusion model is trained to produce the next frame, conditioned on the sequence of past frames and actions. Conditioning augmentations help ensure stable auto-regressive generation over long trajectories, and decoder fine-tuning improves the fidelity of visual details and text.

2398. Arithmetic Without Algorithms: Language Models Solve Math with a Bag of Heuristics

链接：https://iclr.cc/virtual/2025/poster/29843 abstract： Do large language models (LLMs) solve reasoning tasks by learning robust generalizable algorithms, or do they memorize training data? To investigate this question, we use arithmetic reasoning as a representative task. Using causal analysis, we identify a subset of the model (a circuit) that explains most of the model's behavior for basic arithmetic logic and examine its functionality. By zooming in on the level of individual circuit neurons, we discover a sparse set of important neurons that implement simple heuristics. Each heuristic identifies a numerical input pattern and outputs corresponding answers. We hypothesize that the combination of these heuristic neurons is the mechanism used to produce correct arithmetic answers. To test this, we categorize each neuron into several heuristic types---such as neurons that activate when an operand falls within a certain range---and find that the unordered combination of these heuristic types is the mechanism that explains most of the model's accuracy on arithmetic prompts. Finally, we demonstrate that this mechanism appears as the main source of arithmetic accuracy early in training. Overall, our experimental results across several LLMs show that LLMs perform arithmetic using neither robust algorithms nor memorization; rather, they rely on a ``bag of heuristics''.

2399. UV-Attack: Physical-World Adversarial Attacks on Person Detection via Dynamic-NeRF-based UV Mapping

链接：https://iclr.cc/virtual/2025/poster/28278 abstract： Recent works have attacked person detectors using adversarial patches or static-3D-model-based texture modifications. However, these methods suffer from low attack success rates when faced with significant human movements. The primary challenge stems from the highly non-rigid nature of the human body and clothing. Current attacks fail to model these 3D non-rigid deformations caused by varied actions.Fortunately, recent research has shown significant progress in using NeRF for dynamic human modeling. In this paper, we introduce \texttt{UV-Attack}, a novel physical adversarial attack achieving high attack success rates in scenarios involving extensive and unseen actions. We address the challenges above by leveraging dynamic-NeRF-based UV mapping. Our method can generate human images across diverse actions and viewpoints and even create novel unseen actions by sampling from the SMPL parameter space. While dynamic NeRF models are capable of modeling human bodies, modifying their clothing textures is challenging due to the texture being embedded within neural network parameters.To overcome this, \texttt{UV-Attack} generates UV maps instead of RGB images and modifies the texture stacks. This approach enables real-time texture edits and makes attacks more practical. Finally, we propose a novel Expectation over Pose Transformation loss (EoPT) to improve the evasion success rate on unseen poses and views.Our experiments show that \texttt{UV-Attack} achieves a 92.7\% attack success rate against the FastRCNN model across varied poses in dynamic video settings, significantly outperforming the state-of-the-art AdvCaT attack, which only had a 28.5\% ASR. Moreover, we achieve 49.5\% ASR on the latest YOLOv8 detector in black-box settings. The code is available at https://github.com/PolyLiYJ/UV-Attack

2400. Towards Unified Human Motion-Language Understanding via Sparse Interpretable Characterization

链接：https://iclr.cc/virtual/2025/poster/29810 abstract： Recently, the comprehensive understanding of human motion has been a prominent area of research due to its critical importance in many fields. However, existing methods often prioritize specific downstream tasks and roughly align text and motion features within a CLIP-like framework. This results in a lack of rich semantic information which restricts a more profound comprehension of human motions, ultimately leading to unsatisfactory performance.Therefore, we propose a novel motion-language representation paradigm to enhance the interpretability of motion representations by constructing a universal motion-language space, where both motion and text features are concretely lexicalized, ensuring that each element of features carries specific semantic meaning.Specifically, we introduce a multi-phase strategy mainly comprising Lexical Bottlenecked Masked Language Modeling to enhance the language model's focus on high-entropy words crucial for motion semantics, Contrastive Masked Motion Modeling to strengthen motion feature extraction by capturing spatiotemporal dynamics directly from skeletal motion, Lexical Bottlenecked Masked Motion Modeling to enable the motion model to capture the underlying semantic features of motion for improved cross-modal understanding, and Lexical Contrastive Motion-Language Pretraining to align motion and text lexicon representations, thereby ensuring enhanced cross-modal coherence.Comprehensive analyses and extensive experiments across multiple public datasets demonstrate that our model achieves state-of-the-art performance across various tasks and scenarios.