2021 Tencent AI Lab Rhino-Bird Focused Research Program

Research Topics

1. Machine Learning for Life Science

Machine Learning, especially, Deep Learning has been successfully applied to fields including computer vision, speech recognition, natural language processing, and boardgame programs. At the same time, Life science is also becoming a very important application for Machine Learning. The goal of this project is to develop machine learning algorithms for life science. Some possible directions include:

1.1 Machine Learning for Retrosynthesis. Retrosynthesis is the process of recursively decomposing target molecules into available building blocks. It plays an important role in solving problems in organic synthesis planning. To automate or assist in the retrosynthesis analysis, various retrosynthesis prediction algorithms have been proposed. Most of them are based on Graph Neural Networks(GNN) and Sequence to Sequence models.

1.2 Machine Learning for Virtual Screening. Virtual screening is a computational technique of identifying a small set from a big library of molecules, where the selected molecules are most likely to bind to a drug target, typically a protein receptor or enzyme. For this task, most existing algorithms are based on Deep Graph Learning, Meta Learning, etc.

1.3 Machine Learning for Scaffold Hopping. Scaffold hopping is to discover structurally novel compounds starting from known active compounds by modifying the central core structure of the molecule. It has led to several molecules with chemically completely different core structures, and yet binding to the same receptor. Some existing methods are based on Variational Autoencoder (VAE), Generative Adversarial Network (GAN), Graph Neural Networks (GNN) and Recurrent Neural Network (RNN).

1.4 Machine Learning for ADMET Prediction. ADMET is an abbreviation in pharmacology for "absorption, distribution, metabolism, excretion, and toxicity", and describes the disposition of a pharmaceutical compound within an organism. The five criteria all influence the performance of the compound as a drug. To predict ADMET of each molecules, many machine learning methods are proposed based on Graph Neural Networks (GNN), Transformer and so on. In addition, the expressiveness and interpret-ability of the model are also important directions of ADMET prediction.

1.5 Others. There are also some other application directions or machine learning techniques that are related to life science, such as transfer learning, contrastive learning, etc.

2. Deep Reinforcement Learning for Robotics

We are interested in pushing the state of the art in deep reinforcement learning, with applications in robotics and video game AI. Areas including but not limited to:

2.1 Imitation Learning for Robotics

- Modern approaches to Inverse Reinforcement Learning by recovering dynamics and/or reward.
- Partial expert trajectories (e.g., a trajectory of only states, no actions; Deep Mimic).
- Combining Imitation Learning and Reinforcement Learning (e.g., Guided Policy Search).
- Off-policy RL as Imitation Learning.

2.2 Multi-agent Reinforcement Learning

- Cooperative-Competitive Multi-Agent Reinforcement Learning in robotics or video games.

- Centralized or decentralized control of robot swarm.
- MARL methods for algorithmic trading/quantitative trading.

2.3 Non-Stationary Environment

- Agent that can learn and act when the environment dynamics is non-stationary.
- Theoretic connection to Game Theory, Multi-Agent Reinforcement Learning, Meta Reinforcement Learning, etc.

3. Computer Vision & Graphics

3.1 Face Recognition and Person Re-Identification

Theory, models, and algorithms for versatile and robust face representation and person re-identification. We primarily focus on two aspects: one is to design more advanced learning models to obtain video-based face representation towards higher recognition accuracy, the other is to develop versatile person re-identification models for tracking and identifying pedestrians from videos.

Recommended topics:

- Video-based face recognition with large pose variation and blur faces.
- Real-time multi-subject pedestrian detection.
- Person Re-Identification.

3.2 Video Understanding

Theories, algorithms and models for (robust) large-scale video (multi-label) classification, large-scale unsupervised (multi-modal)video representation learning, and spatio-temporal reasoning. We mainly focus on designing novel models and algorithms for higher classification accuracy, more general and robust video representations, and reasoning in the temporal and space domains for real-world applications.

Recommended topics:

- Large-scale video classification, multi-label classification, and robust models against to noise data.
- Unsupervised large-scale video representation learning, including novel models, new regularizations, new theoretical analysis, etc.
- Video representation learning under the multi-model setting, including learning with text, audio, knowledge.
- Reasoning, decision making, and planning related to videos in real-world applications.

3.3 Image/Video Generation

The image and video generation have been at the forefront of research on generative models in the past few years. With the development of the generative adversarial network, tremendous research investigations have been put on generating user-intended image and video contents. These content generation works mainly focus on both natural visual content and human face content.

Recommended topics:

- Image-to-image translation. Image generation and photo manipulation. Image generation from textual descriptions. Low-level and middle-level vision: super-resolution, denoising, inpainting, etc.
- Face related content generation. Face hallucination and deblurring. Face stylization and synthesis. Face restoration and facial expression generation.

- Video stylization and future prediction. Video forensics and compression. Joint video content analysis and compression. Video based biometrics.
- Network compression for GANs. Network architecture search. Knowledge distillation.

3.4 Adversarial Machine Learning

Theory, models and algorithms for adversarial machine learning. We primarily focus on the learning theory of adversarial examples and exploring and designing advanced adversarial attack and defense techniques for safety-critical applications, such as face recognition.

Recommended topics:

- Learning theory of adversarial examples, such as adversarially robust generalization, model interpretability and vulnerability, etc.
- Advanced adversarial attacks for real-world applications, such as transfer attacks, black-box attacks and physical attacks, etc.
- Advanced adversarial defense techniques for robust deep neural networks.

3.5 AutoML in Vision

Methodologies for automatic learning framework for computer vision tasks. We focus on designing new methods, frameworks for automating the pipeline of computer vision tasks, including automatic data augmentation, hyper-parameter search, architecture search, even optimization algorithm design.

Recommended topics:

- Automatic data augmentations: Learning the best augmentation operations for image classification or video classification.
- Automatic network design: Theory on the design space for NAS and find better design space in supervised learning setting and unsupervised learning setting, and better search algorithm for image classification or video classification.
- Automatic optimization algorithm design. Learning optimization algorithms from a large number of tasks.

3.6 Digital Human Generation

Theory and applications of deep generative models, such as GAN and VAE. We primarily focus on photo-realistic generations of human faces and bodies with well-trained deep neural networks, where we encourage to combine traditional computer graphics techniques with the latest deep generative models.

Recommended topics:

- Performance capture with the aid of deep neural networks.
- High-resolution photo-realistic motion transfer of face and body for a certain person.
- Real-time human video editing based on mobile camera input.

4. Natural Language Processing

4.1 Natural Language Understanding

NLU is to process, interpret and analyze both formal and social texts with necessary techniques that can help human or downstream systems understand them.

- Novel model frameworks for morphology, syntax and semantics.
- Large-scale pretrained language models for downstream text understanding tasks.

- Ultra-fine grained entity typing with types up to 10000.
- Exploring new tasks related to text understanding.
- Representation, construction and reasoning of knowledge graphs.
- Novel model architectures for unsupervised pre-training.
- Novel model architectures for neuro-symbolic reasoning in NLU.
- Incorporating external background knowledge in language understanding.
- Leverage multiple sources of teaching or educational materials for improving subject-area/examination-style question answering.
- Improve the robustness of machine reading comprehension (especially extractive) models in real-world settings.
- Theoretical understanding of self-supervised learning / pretraining.

4.2 Natural Language Generation

- Abstractive long text summarization, including long text modeling and multi-sentence summaries generation.
- Long text generation, such as stories and news.
- Controllable text generation, including generating text consistent with the given conditions such as the emotions and characters as well as some other specific styles.
- Model analysis for generation models, including interpretability analysis, robustness analysis, attack and defense analysis.
- Text generation (conditioning on prompts, tables, retrieved sentences, images, or videos) and language grounding in images and videos.

4.3 Dialogs

Dialog research has been a long-time hot spot for years since conversation systems are the key ability for artificial intelligence for enabling backend system to interact with people through language to assist, enable, or entertain.

- Leveraging commonsense knowledge in dialogue system, including constructing dialog related commonsense knowledge base, incorporating common sense knowledge into conversation generation, knowledge-grounded conversation generation.
- Multi-turn dialog systems, including multi-turn corpus construction, retrieval and/or generation based response prediction, intent classification and slot filling, topic sticking and recommendation, task-oriented dialog systems, etc.
- Personalized dialog models, including constructing personalized dialog corpus and incorporating personalized information into models.
- Automatic evaluation methods for open-domain chi-chat dialogue system.
- Semantic parsing in single and multi-turn dialog including the logical form generation and deeper reasoning over them.

4.4 Machine Translation

Machine translation research focuses on improving machine translation from amateur to professional, by conducting fundamental research on machine translation, as well as bridging the gap between machine translation systems and human translators.

- Adequacy-oriented NMT models that include various techniques such as advanced architectures and learning strategies, to alleviate the key problem of NMT – inadequate translation.
- Pre-Training for NMT: exploit existing pre-trained model or design specific pre-training algorithm for NMT models to improve translation performance.
- Interpretability of NMT: how information is transformed from source side to target side in NMT models.
- Interactive translation that bridges the gap between machine translation systems and human translators, including designing new human-machine interactive actions and evaluation on efficiency for interactive machine translation.

5. Speech Technology

5.1 Far-field Signal Processing

In far-field microphone situations, the speech signal energy attenuation, the stationary and non-stationary noise, the reverberation, and the echo of the loudspeaker during the target sound propagation increase the difficulties of speech recognition and voice wake-up. Through the microphone array signal processing, noise reduction and speech/noise separation technologies, we could improve both speech quality and speech recognition performance.

Suggested research topics:

- Microphone array algorithm design to improve speech recognition with multiple speakers and interference sources.
- Ad-hoc, distributed microphones and array independent multi-channel speech enhancement and separation.
- Joint training and optimization of front-end speech processing and back-end speech recognition acoustic models to upgrade both systems.
- Self-supervise/weakly-supervise and other learning methods for better utilizing large amount of unlabeled/weakly-labeled data and reducing domain/environment mismatch from training to testing.

5.2 Speech Recognition

Speech recognition, as one of the most natural way of human-computer interaction, plays a vital role in the AI era. Although human parity results have been reported on a clean benchmark dataset by models trained on affluent in-domain corpus, general domain ASR models still face challenges such as noise robustness, domain adaptation and long tail problems. We are interested in developing novel algorithms and models to address the above issues.

Suggested research topics:

- Research new neural network structures for SOTA End-to-end ASR.
- Research robust speech recognition using context information, multi-domain training, and adaptation.
- Multilingual speech recognition focusing on Mandarin-English code-switching.
- Improve E2E ASR end-pointing and latency.

5.3 Speech Generation

Speech generation technology, including both speech synthesis and voice conversion, is a key part of human-computer speech interaction. The user experience improves when generated voice is subjectively attractive to them. Personalized expressive speech generation technology aims to build generated voices that sounds familiar to the listeners, such as public figures, famous stars, friends and family members. However, the labeled data of the desired voices recorded in a clean environment is usually difficult to collect. Building high quality models with limited data remains a challenging task. We encourage research directions including but not limited to the following:

- Multi-speaker multi-style controllable expressive speech synthesis.
- Multi-lingual and cross-lingual speech synthesis.
- Speech synthesis that exploits low-quality/ASR data.
- Singing voice synthesis/conversion.
- Personalized voice conversion with limited data.
- Multi-modal (speech/face/gesture) synthesis.

5.4 Speaker Recognition

Identifying a person by his or her voice is an important human trait usually taken for granted in natural human-to-human interaction/communication, yet it remains as a challenging task for computers. Recently, automatic speaker-recognition systems have emerged as an important means of verifying identity in many e-commerce applications as well as in general business interactions, intelligent housing system, forensics, and law enforcement. We are interested in building high-quality systems through advanced paradigms with both labeled and unlabeled data.

Suggested Research topics:

- Domain and environment robust speaker recognition.
- Speaker recognition with short utterances.
- Self-supervised/weakly supervised learning for robust speaker representation.
- Joint training and optimization of single-channel/multi-channel front-end speech processing and back-end speaker recognition models.

5.5 Multi-model Speaker Tracking and Diarization

Speaker tracking and diarization are important for many applications such as human-computer/android interaction, meeting transcription and surveillance systems, etc. We attempt to address the challenging task of tracking multiple moving speakers and identify who is speaking with auditory and visual information. Proper fusion of multi-modal information is required in order to deal with corruption from audio data or visual data or both. The use of multiple complementary modalities is beneficial when the information is correctly processed and fused.

Suggested research topics:

- Investigating spectrum, spatial, voiceprint and visual modalities fusion techniques for speaker tracking, diarization, separation and recognition.
- Investigating new paradigm for robust speaker tracking based on multi-modal information.
- Investigating new paradigm for robust diarization based on multi-modal information.