2019 Tencent AI Lab Rhino-Bird Focused Research Program

Research Topics

1. Machine Learning Center (Shenzhen)

1.1 Automated Machine Learning (AutoML)

Automated Machine Learning (AutoML) attracts more and more attentions from the research community, since it can help machine learning researchers accelerate their research processes, and make it easier for the non-experts to apply machine learning to various real-world domains. From the perspective of technologies, AutoML consists of several stages, including: Automated Data Preparation, Automated Feature Engineering, Automated Model Selection, Hyper-parameter Optimization, Automated Metrics Selection, Automated Problem Checking, Automated Results Analysis, and so on. Among all these different stages, we would like to suggest the following research areas, while others are also welcome:

1.1.1 Hyper-parameter Optimization. It can be solved by various method, such as Bayesian Optimization.

1.1.2 Neural Architecture Search. New neural architecture search algorithms based on RL, Evolution, Network morphing, Back propagation, etc. Efficient and effect NAS models that can be used by both academic and industry.

1.1.3 Others. There are also some other techniques that probably can be used for AutoML, such as, Meta Learning, Transfer Learning, etc.

1.2 Automatic Model Compression (AutoMC)

Deep learning is widely used in various areas, such as computer vision, speech recognition, and natural language translation. However, deep learning models are often computational expensive, which limits further applications on mobile devices with limited computational resources. The next wave of ML applications will have significant processing on mobile and ambient devices. Recently, model compression for efficient training/inference with deep neural networks and other ML models is one of hottest research topics. Areas/topics of interests include, but not limited to:

1.2.1 Neural network compression techniques, such as pruning, quantization, low-rank factorization, etc.

1.2.2 Automatic model compression techniques: 1) Self-adaptive model compression with minimal human effort by adopting reinforcement learning and/or hyper-parameter optimization techniques. 2)Compact and efficient neural network design with neural architecture search.

1.2.3 Model compression and acceleration for speech and NLP tasks, including speech recognition/ language understanding and conversational assistants on mobile and IoT devices.

1.2.4 Efficient training procedure for model compression algorithm, which requires less computational resource and/or fewer training iterations and achieves higher training speed.

1.2.5 Model compression with no or limited number of training samples.

1.2.6 Model compression with hardware-software co-design: 1) Software libraries (including open-source) optimized for on-device ML. 2) Hardware design optimized for the data patterns with model compression, such as sparsity, low bit.

1.2.7 Video & media compression methods using DNNs.

2. **Reinforcement Learning Center** (Shenzhen)

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

2.1 Perceive-Plan-Control. Key algorithms for robot learning: perception (calibration, hand-eye coordination, 6D object pose estimation), planning (active, long-term), and control (flexible manipulation, sensorimotor control, optimal control).

2.2 Transfer Learning. Algorithms that can generalize to unseen domains and tasks, transferring what they've learned from simulation to real-world and from one task to new tasks.

2.3 Developmental Learning. Agents that learn on their own (by interacting with environments and themselves), and build upon what they've learned previously to acquire new capabilities.

2.4 Learning with Self-play. Large scale distributed RL training algorithms and systems for complex video games (e.g., MOBA, FPS), reward shaping, multi-agents coordination.

2.5 Game Development. Learning to generate game content (animations, rules, maps, levels), and evaluate the candidate content and game quality.

3. **Computer Vision Center** (Shenzhen)

3.1 Face AI. Video based face detection, alignment, and recognition; 3D face anti-spoofing; face recognition on low-resolution or occluded faces.

3.2 Video Understanding. Large-scale video classification, video object detection, video proposal/detection/grounding/captioning, video recommendation and retrieval.

3.3 Image/Video Generation. Theory and applications of GAN and VAE, such as cartoon painting, banner generation for AD promotion, image/video transformation, image/video generation from language.

3.4 3D Vision.
Reconstruction of 3D scenes/objects/human faces/human bodies from RGB or RGBD inputs, realtime 3D l ocalization and mapping (SLAM), sensor fusion, semantic understanding of 3D representations, image-based matching and correspondence problems, 3D rendering and animation, etc.

4. **Natural Language Processing Center** (Shenzhen & Seattle)

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 architecture design for NLU and unsupervised pre-training.
- Incorporating external commonsense and background knowledge in language understanding.
- Knowledge graph representation and reasoning, as well as their combinations with deep learning techniques.

4.2 Natural Language Generation

- Text summarization, including extractive summarization, compressive summarization, and abstractive summarization, especially the neural models for long document summarization and multi-sentence summary generation.
- Knowledge-based question answering in general and/or specific (e.g. financial) scenarios, incorporating inference into question answering process, question generation over various application scenarios (e.g. CQA and dialogue) and datasets (e.g. SQuAD and RACE), etc.
- Conditional sentence generation (conditioning on 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.

- Dialog response generation, including question-answer modeling, response generation, response quality assessment, etc.
- 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.
- Leveraging common sense knowledge in dialog system, including constructing dialog related common sense knowledge base and incorporating these knowledge into the model.
- Semantic parsing in single and multi-turn dialog including the logical form generation and deeper reasoning over them.

4.4 Machine Translation

Improving machine translation from amateur to professional.

- Adequacy-oriented NMT models that include various techniques such as advanced architectures and learning strategies, to alleviate the key problem of NMT – inadequate translation.
- Large-scale NMT system that build a high-performance system on a large-scale data, which consists of hundreds of millions of (possibly noisy) bilingual sentences from multiple domains.
- 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 Center** (Shenzhen & Seattle)

5.1 Far-field Signal Processing

In the far-field speech recognition task, the speech signal energy attenuation, the stationary and non-stationary noise, the reverberation, and the echo of the loudspeaker during the target sound propagation to the microphones will increase the difficulties of speech recognition and voice wake-up. Through the microphone array signal processing and deep learning speech noise reduction/separation technology, it could improve the speech quality for solving the problem of far field speech recognition. Suggested research area:

- Microphone array algorithm design to improve the speech recognition ability of multiple speakers and interference sources.
- Reverberation algorithm design, to enhance the ability of far-field speech recognition.
- Design of sound source localization algorithm to improve the accurate positioning ability under the far-field noisy environment.
- Echo cancellation, noise suppression and other algorithms designed to enhance the ability of speech recognition in the noisy environment.
- The design of neural network algorithm to enhance single-channel and multi-channel end-to-end far-field speech enhancement.
- Joint training and optimization of front-end speech processing and back-end speech recognition acoustic models to upgrade both systems.

Acoustic scene detection and determination aims to determine the current acoustic scene or event by acoustic features, such as stadiums, concert halls, rain, police car sound and so on.

- End-to-end neural network algorithm design.
- Accurate time positioning of acoustic scene/event.
- Accurate detection of multi-scene/event.

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. With the successful application of deep learning in the field of speech recognition, more and more new models and novel algorithms are proposed and continuously improve the recognition accuracy. Although a super-human performance has been achieved on a well-known conversational speech benchmark, in the real scenarios, robustness and naturalness of nowadays' speech recognition system are still far from satisfactory. In some difficult conditions, an industry-deployed speech recognition system can completely fail. Here are some unsolved or challenging problems in the field:

- End-to-end speech recognition.
- Multilingual speech recognition with special focus on Mandarin-English code-switching.
- Deep learning based acoustic model joint optimization with front-end speech processing.
- Robust speech recognition and far-field speech recognition.
- Cocktail party problem.
- Contextual speech recognition.
- Multi-modality speech recognition.

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 increases when generated voice is subjectively attractive to the listeners. Personalized expressive speech generation technology aims to build generated voice 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 a generated voiced with limited data has remain a challenging task. We encourage research directions including but not limited to the following:

- Multi-speaker speech synthesis.
- Speaker adaptation to the target voice characteristic and speaking style.
- Multi-lingual and cross-lingual speech synthesis.
- Expressive speech synthesis with controllable speaking styles.
- Speech synthesis with unlabeled data.
- New paradigm of speech synthesis.
- Singing synthesis.
- Multi-modal talking head synthesis.
- Personalized voice conversion

5.4 Speaker Recognition

Identifying a person by his or her voice is an important human trait most take for granted in natural human-to-human interaction/communication. 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. Future direction includes:

- Domain and environment mismatch. Systems often perform very well in the domain/environment for which they are trained. However, their performance suffers when the users use the system in other domain/environment. So how to adapt a system from a resource-rich domain/environment to a resource-limited domain/environment and how to make speaker recognition systems robust to domain/environment mismatch are great challenges.
- Short utterance in text-independent speaker recognition. Performance of i-vector/PLDA systems degrades rapidly in presence of short utterances or utterances with varying durations. The reason is that short utterance contains limited phonemic information and the i-vectors of short utterances have much bigger posterior covariance.
- Text-dependent speaker recognition using short utterances. It is more natural to use HMMs rather than GMMs for text-dependent tasks. But HMMs require local hidden variables, which are difficult to handle because of data fragmentation. More recently, using DNN/RNN to extract utterance-level features or building an end to end based on DNN/RNN is attracting more and more attention.

5.5 Audio-visual fusion for multi-speaker speech tracking

Multi-speaker speech tracking is important for many applications such as human-computer/android interaction, monitoring and surveillance systems, etc. In this topic, we address the challenging task of tracking multiple moving speakers 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 two complementary modalities is beneficial when the information is correctly processed and fused.

The research content of audio-visual fusion for multi-speaker speech tracking includes the follows:

- Carefully designed inference algorithms to exploit the inherent and hopefully complementary nature of the two modalities.

- Robust inference algorithms to acoustic variability and data corruption such as noise, reverberation, interference, outliers, etc. Despite the fact that the visual observations, e.g. face detection, is usually continuous for speakers looking towards the camera and within the field of view, however, in multi-person indoor environments, e.g. an android interacting with a group of persons holding a conversation, natural speech often happens intermittently with occasional overlaps between several speech signals. Multi-channel speech recordings can be used to identify which source emitted which part of the speech signal, i.e. to separate and diarize the sources, for instance by using beamforming techniques, especially for moving sources; on the other hand, using camera signals to obtain the knowledge about who is where and when in the scene could also help separating the sound sources.
- Robust inference algorithms to vision variability and vision data corruption. Vision modality suffers from such limitations as visual occlusions, limited field of view, lighting conditions, etc. Audio processing can help overcoming these limitations due to the complementary nature of the information encoded in the acoustic signals.