Tencent AI Lab Rhino-Bird Joint Research Program

Research Topics

1. Computer Vision Center

Interested in multimedia (both image and video) AI, including:

1.1 **Generation**: theory and applications (e.g., cartoon painting, banner generation for AD promotion) of GAN.

1.2 **Editing**: image & video low-level and mid-level vision, such as image/video super resolution, enhancement, denoising, deblurring, harmonization, and so on.

1.3 **Analysis & Understanding**: large-scale image/video classification, video semantic segmentation, video localization, image & video captioning, and so on.

1.4 **Recommendation**: image & video recommendation and retrieval.

1.5 **Vision-driven RL**: vision based RL task (e.g., visual object tracking, in-door robot navigation) and its deployment in real world robot.

2. Speech Processing Center

2.1 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.

2.2 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 new models and new algorithms are proposed and continuously improve the recognition accuracy. In some test sets, speech recognition systems perform better than humans. These are constantly promoting voice recognition in the field of AI applications.

In spite of this, there are still many problems to be solved in the field of research, including

- End-to-end speech recognition.
- Multilingual speech recognition.
- Deep learning based model joint optimization.
- Robust speech recognition and far-field speech recognition.
- Cocktail party problem.

2.3 Speech Synthesis

Speech synthesis technology is a key part of human-computer speech interaction. The user experience decreases when synthesized voice is not subjectively attractive to the listeners. Personalized expressive speech synthesis technology aims to build synthesized 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 synthesized 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.

2.4 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.

3. Natural Language Processing Center

3.1 Natural Language Understanding (NLU)

NLU is to process, interpret and analyze natural languages with necessary techniques that can help human or downstream systems understand them. NLU is the core of NLP for years regarding to its fundamental role at the first step of processing natural languages. There are many aspects included in NLU research. In Tencent AI Lab, we focus on (include but not limited in) the following topics, which are also the suggested areas for applying the research funds:

- Fundamental NLP, including word segmentation, part-of-speech tagging, constituent and dependency parsing, named entity recognition, sentiment analysis, key-phrase extraction, etc.
- Semantics, including multi-granularity (word, phrase, sentence, document) embeddings, meaning representation and semantic tagging, etc.
- Knowledge representation and inference and its combination with deep learning techniques.
- Reading comprehension and causal-relationship extraction.

3.2 Natural Language Generation (NLG)

- Automatic summarization
- Automatic article writing

3.3 Dialogs

Dialog research has been a long-time hot spot for years since conversation systems are the key part for artificial intelligence for enabling backend system to interact with people through language to assist, enable, or entertain. In Tencent AI Lab, we focus on (include but not limited in) the following topics, which are also the suggested areas for applying the research funds:

- Extractive dialog system, including system construction, question understanding, answer ranking and re-ranking, slot tagging and intent classification, dialog management, etc.
- Dialog response generation, including question-answer modeling, response generation, response quality assessment, etc.
- Dialog management, including learning dialog manager, multi-turn conversation modeling, etc.
- Multi-user, multi-turn, multi-modality dialog system.

3.4 Machine Translation (MT)

We have two major MT areas, namely, neural machine translation (NMT) and interactive machine translation (IMT). NMT has advanced state of the art for MT in recent years. However, there are still a lot of remaining problems unsolved. IMT is a rising field which is more applicable to industry with the interwoven of MT and human-computer interactions. In Tencent AI Lab, we focus on (include but not limited in) the following topics, which are also the suggested areas for applying the research funds:

- Adequacy-oriented NMT, including various techniques of improving the adequacy of translations generated by NMT models.
- NMT visualization and interpretability, including visualizing and interpret the internal structure and composition of NMT models.
- Interactive MT with NMT, including interactive translation system on top of NMT models.
- Multi-domain NMT, including building a practical NMT system on a large-scale data, which consists of bilingual sentences from multiple domains.
- NMT with novel architectures, including building NMT models beyond the standard encoder-decoder framework and/or with novel networks such as capsule networks.

4. Machine Learning Center

4.1 **Deep learning theory and framework**. Theoretical understanding of deep learning or replacement framework of deep learning.

4.2 **Machine learning models and applications**. Machine learning models for different applications such as bandits problem, transfer learning, reinforcement learning, neural memory mechanism.

4.3 **Unsupervised learning with deep neural networks**. The potential and limitation of neural network based unsupervised learning methods, new unsupervised generative methods with deep neural networks, multi-modal learning.

4.4 **Large scale deep graph learning**. Node embedding for large scale social networks, discovering the communities in graphs, applying deep learning techniques for graph learning.

4.5 **Distributed optimization algorithm**. Design and develop more efficient distributed optimization algorithms with theoretical guarantee and/or more outstanding performance in practical applications.

5. Reinforcement Learning Center

5.1 Bridging between simulation and the physical world

Within the past decade, simulation has fostered tremendous progresses in modern machine learning, especially in reinforcement learning (e.g., AlphaGO, OpenAI Universe). This is mainly due to three main advantages of simulation: a) it can run much faster than real-time; b) the cost of simulation is much lower than collecting real data (e.g., accidents in autonomous driving); c) it is convenient to conduct controlled experiments for almost all cases, and repeat them. However, it is also extremely challenging to transfer the models learned from simulation, to the physical world. In this call for proposals, we hope to develop technologies to bridge between simulation and the physical world:

- Realistic simulation for the physical world.
- Photorealistic content generation for games.
- Transfer learning and domain adaptation.

5.2 Mastering StarCraft

Despite the promising performance of conventional reinforcement learning algorithms, learning to play real-time multiplayer strategy game (e.g., StarCraft) has remained an important, yet challenging task. Compared with chess and Go, StarCraft is orders of magnitude more complex, and you cannot see all of your opponents' troop deployments or construction projects. This forces you to use what you've seen, which is always imperfect, to predict what they may be planning, which can come from a huge action space. In this call for proposals, we encourage researchers to push the state-of-the-art game AI in mastering StarCraft, including but not limited to the following areas:

- API to train self-playing StarCraft bots.
- Learning to play StarCraft using game replays.
- Learning to act with imperfect information.
- Learning to coordinate multiple agents in StarCraft.
- Memory and planning in StarCraft.

5.3 Conversational AI

During the past half decade, we've seen an increasing number of so-called "intelligent" digital assistants (e.g., Alexa, Siri, Google Assistant, Cortana) being introduced on various devices. Although the conversational AI technology behind these applications keeps getting better, the expectation of human "intelligence" is far from being met. Here, we call for proposals on bridging the conversational gap between humans and AI bots. We encourage research directions including but not limited to the following:

- Performance evaluation for conversational AI.
- Natural language understanding for conversational AI.
- Speech understanding for conversational AI.
- Learning to understand human intentions beyond language.
- Dialog planning and management.