Introduction to the Special Issue on Deep Learning for Multi-Modal Intelligence Across Speech, Language, Vision, and Heterogeneous Signals

THANKS to the disruptive advances in deep learning, significant progress has been made in artificial intelligence (AI) applications with single modality, such as speech recognition, speech synthesis, image classification, object detection, as well as machine translation and reading comprehension, *etc*. However, many AI problems require more than one modality, and techniques developed for different modalities can often be successfully cross-fertilized. Therefore, the studies on the modeling and learning approaches across multiple modalities are of great interest. This special issue brings together a diverse but complementary set of contributions on emerging deep learning methods for problems based on multiple modalities including speech, text, image and video.

Following an open call for papers, we received a total number of 52 submissions for this special issue. After an extensive and competitive review process, 10 papers were selected for final publication. These papers span a range of diversified topics. Both audio and visual information are used for speech synchronization, enhancement, separation, recognition, and translation, as well as speaker verification and social role detection. One publicly available large dataset for audio-visual speech recognition and translation is introduced. The tasks of image captioning, text-to-image generation, visual question answering (VQA), visual reasoning and image classification with natural language explanations are based on image and text modalities. A new task that generates images directly based on speech without involving any text is proposed. Moreover, spiking neural networks, which more closely mimic natural neural networks compared to the widely used neural networks for deep learning, are also studied in one of the papers.

The issue begins with a review article, "Multimodal Intelligence: Representation Learning, Information Fusion, and Applications" by *Zhang et al.*, which presents a comprehensive analysis of some recent multimodal work related to computer vision and natural language processing. Three key topics are used to organise the paper to provide a more structured perspective. The representation learning topic presents the methods for training unimodal and multimodal embeddings. The information fusion topic reviews simple operation based, attention mechanism-based, and bilinear pooling based approaches to fuse embeddings of different modalities. The last key topic covers illustrative applications – image captioning, text-to-image generation, VQA, and visual reasoning tasks – in order to give a brief introduction of these rapidly developing fields to our community.

Common VQA systems obtain all visual concepts, which are used as the answers to the questions, equally as the softmax function labels from the same output layer, which ignores the rich structural-semantic meanings of the concepts. "Learning to Recognize Visual Concepts for Visual Question Answering With Structural Label Space" by *Gao et al.* proposes an approach to utilize such information by first clustering semantically related concepts into groups and measure the similarity between relevant groups. For each group of concepts, an independent output layer is constructed to classify the answer to the question as one of its belonging concepts. Some parameters can be shared adaptively across different output layers based on the similarity between their relevant groups. Experimental results of multiple benchmarks show the proposed method can improve the performance of VQA models.

Xu et al. proposes an explainable attention mechanism for image classification, to improve its interpretability in "Where is the Model Looking At? – Concentrate and Explain the Network Attention." A multitask training framework predicting both class category and the category dependent attributes is used in the method, from which the embeddings are integrated to derive the final image classification result. The introduction of the prediction to the attributes not only helps the model to concentrate attention on the foreground objects but also enables the generation of attribute-based textual and visual explanations. The multi-modal explanation improves users' trust of the model and can help to point out the weakness of the method and data.

Instead of using text descriptions, *Li et al.* proposes to generate images from speech in "Direct Speech-to-Image Translation." The raw speech signal is first converted into a representation using an audio encoder, which is trained as a student model via teacher-student training based on an image encoder teacher. The derived audio representation is then combined with Gaussian noise to use as the input to a set of stacked generative adversarial networks to generate high-resolution images progressively, which translates speech directly to image without any text.

For tasks across audio and video modalities, two papers of this special issue are devoted to audio-visual speech separation and enhancement. In "Multi-Modal Multi-Channel Target Speech Separation," *Gu et al.* proposes a general multimodal framework to separate the target talker's speech from overlapped speech segments with simultaneous talkers, which leverages different information embedded in the audio and video streams

Digital Object Identifier 10.1109/JSTSP.2020.2989852

including spatial location, lip movements, and voice characteristics. The attention mechanism is used to fuse the audio embeddings from different acoustic sub-spaces based on the video embeddings. *Tan et al.* proposes a two-stage multimodal network for audio-visual speech separation and enhancement in "Audio-Visual Speech Separation and Dereverberation With a Two-Stage Multimodal Network." The first-stage network fuses the audio and video embeddings to separate the target talker's speech from the interfering speech and the background noise. The second-stage network suppresses room reverberation to enhance the speech derived from the first-stage network. These stages are first separately trained as an initialization, then jointly trained by fine-tuning.

The paper, "Multi-Stream Recurrent Neural Network for Social Role Detection in Multiparty Interactions" from Zhang and Radke proposes a framework that integrates synchronized video, audio, and text streams from group people to capture the interaction dynamics in natural group meetings. Using multi-modality information, the authors estimate the dynamic social role of the meeting participants, *i.e.*, Protagonist, Neutral, Supporter, or Gatekeeper. On top of a multi-stream recurrent neural network, the authors incorporate both co-occurrence features and successive occurrence features in thin time windows to better describe the behavior of a target participant and his/her responses from others. The proposed method was evaluated on the widely-used AMI corpus and state-of-the-art accuracy was reported. Moreover, the authors analyze the importance of different video and audio features for estimating social roles, which presents a deeper understanding of this multi-modality task.

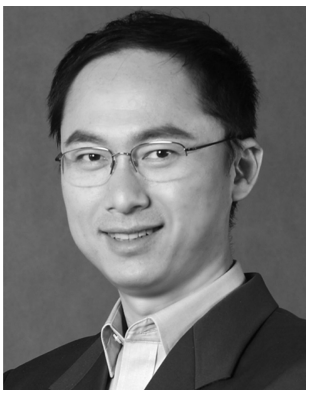
Cross-modal retrieval is a task to find the most relevant data in one modality given the input in another modality. In "Perfect Match: Self-Supervised Embeddings for Cross-Modal Retrieval," *Chung et al.* proposes a multi-way cross-entropy objective function for cross-modal retrieval, which computes the relative similarity of the matching data pair over non-matching data pairs. The method is validated using an audio-visual synchronization task that synchronizes a video clip with a separately recorded audio segment, as well as the audio-visual biometrics task based on 10-way cross-modal forced matching that retrieves the speaker's face image given the speech or vise versa. The proposed method is observed to improve both classification accuracy and convergence speed.

In "Grounded Sequence to Sequence Transduction," *Specia et al.* proposes a large multimodal dataset called How2, which consists of a 2,000-hour open-domain collection of instructional Youtube videos with 22.5 million-word manually derived English transcriptions and Portuguese translations. The dataset is used for two vision grounded sequence to sequence classification tasks in the paper: multimodal speech recognition and machine translation. Audio and visual features for speech, action, object, and scene are integrated into the standard approaches as well as some newly proposed ones, to validate the importance of joint use of the multimodal information in these tasks. This dataset has been used by multiple evaluations recently, showing how challenging it is to build multimodal systems that can perform better than unimodal systems.

Different from the other papers in this special issue, the work from *Zhang et al.*, "An Efficient Threshold-Driven Aggregate-Label Learning Algorithm for Multimodal Information Processing," is closely related to not only machine learning and deep learning, but neuroscience as well. It targets on the temporal-credit assignment problem, a fundamental problem about how to learn useful multi-sensory clues, such as acoustic and vision, from the delayed feedback signals, which can be resolved by matching the out spike count of spiking neurons with the magnitude of the delayed feedback signals using the aggregate-label learning algorithm. A threshold-driven algorithm is presented in this paper to improve the efficiency of aggregate-label learning, enabling the use of multi-layer spiking neural networks with many neurons for hand-written digits and spoken digits recognition. This results in significant improvements in classification accuracy when compared with the existing single-layer aggregate-label learning algorithms.

Despite the significant progress on many topics, such as shown in the aforementioned articles, research on multimodal intelligence is still in its infancy and has great potential in the future. We hope this special issue could become a stepping stone for future developments and advancements in multimodal intelligence towards building agents with the capabilities of multimodal perception and using the connection between different modalities.

In the end, the guest editorial team wants to thank all the authors and reviewers whose contributions have made this special issue possible. We would also like to thank Prof. Lina Karam, Editor-in-Chief, for her kind support and suggestions. Warmest thanks also go to Mikaela Langdon and Rebecca Wollman from the IEEE publication office for keeping this special issue on track at different stages of the process.



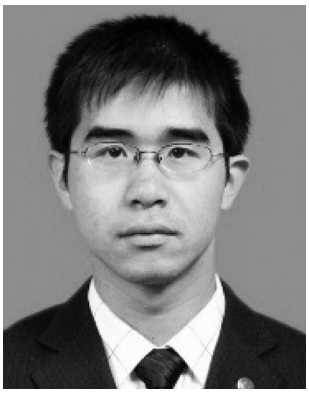
Xiaodong He (Fellow, IEEE) received the B.S. degree form Tsinhua University, M.S. degree form the Chinese Academy of Science, and the Ph.D. degree from the University of Missouri, Columbia. He is the Deputy Managing Director of JD AI Research and the Head of the Deep Learning, NLP, and Speech Lab. He is also an Affiliate Professor of ECE at the University of Washington, Seattle. His current research interests mainly focus on deep learning, natural language processing, speech recognition, computer vision, information retrieval, and multimodal intelligence. Dr. He has held Editorial Positions for IEEE TASLP, JSTSP, SPM, SPL and for the Transactions of the ACL (TACL). He has also served in the organizing committees/program committees of major speech and language processing conferences. He was a Member of the IEEE SLTC for the term of 2015 to 2017 and the Chair of the IEEE Seattle Section in 2016–2017.

Li Deng has been the Chief Artificial Intelligence Officer of Citadel since May 2017. Prior to Citadel, he was the Chief Scientist of AI, the founder of the Deep Learning Technology Center, and Partner Research Manager at Microsoft and Microsoft Research , Redmond, from 2000 to 2017. Prior to Microsoft, he was an Assistant Professor from 1989 to 1992, Tenured Associate from 1992 to 1996, and Full Professor from 1996 to 1999 at the University of Waterloo in Ontario, Canada. He also held faculty or research positions at Massachusetts Institute of Technology, Cambridge, U.K. from 1992 to1993 Advanced Telecommunications Research Institute, ATR, Kyoto, Japan, from 1997 to 1998, and HK University of Science and Technology, Hong Kong, in 1995. He is a Fellow of the Academy of Engineering of Canada, a Fellow of the Washington State Academy of Sciences, a Fellow of the IEEE, a Fellow of the Acoustical Society of America, and a Fellow of the International Speech Communication Association. He has also been an Affiliate Professor at the University of Washington, Seattle. He was an Elected Member of the Board of Governors of the IEEE Signal Processing Society, and was Editors-in-Chief of IEEE SIGNAL PROCESSING MAGAZINE and of IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING from 2008 to 2014, for which he received the IEEE SPS Meritorious Service Award. In recognition of the pioneering work on disrupting speech recognition industry using large-scale deep learning, he received the 2015 IEEE SPS Technical Achievement Award for "Outstanding Contributions to Automatic Speech Recognition and to Deep Learning". He also received dozens of best paper and patent awards for the contributions to artificial intelligence, machine learning, information retrieval, multimedia signal processing, speech processing and recognition, and human language technology. He is an author or co-author of six technical books on deep learning, speech processing, pattern recognition and machine learning, and, the latest, natural language processing (Springer, June 2018).

Richard Rose (Fellow, IEEE) received his PhD degree in electrical engineering from the Georgia Institute of Technology. He has been a Research Scientist at Google in New York City since October, 2014. While at Google he has contributed to efforts in speech recognition, language and video processing. More recently, he has been working on signal processing in YouTube videos. Before coming to Google, he served as a Professor of Electrical and Computer Engineering at McGill University, Montreal, since 2004, as a member of research staff at AT&T Labs/Bell Labs, and member of staff at MIT Lincoln Labs. Dr. Rose has been active in the IEEE Signal Processing Society. He has served twice as an Associate Editor of IEEE SPS TRANSACTIONS, twice as a Member of the Speech Technical Committee, as member of the SPS Board of Directors, and several times on organizing committees of IEEE workshops.

Minlie Huang received the Ph.D. degree, 2016. He is currently an Associate Professor, and Deputy director of the AI Lab. of the Department of Computer Science and Technology, Tsinghua University. Dr. Huang was selected by the "Beijing Century Young Elite Program" in 2013, and won the Hanvon Youngth Innovation Award in 2018. He won the IJCAI-ECAI 2018 distinguished paper, NLPCC 2015 best paper, and CCL 2018 best demo award. His work on Emotional Chatting Machine was reported by MIT Technology Review, the Guardian, NVIDIA, Cankao Xiaoxi, Xinhua News Agency, etc. He has published 60+ papers in premier conferences such as ACL, AAAI, IJCAI, EMNLP, KDD, and highly-impacted journals like IEEE TASLP, ACM TOIS, *Bioinformatics*, JAMIA, etc. He served as Area Chairs for ACL 2016, EMNLP 2014, EMNLP 2011, and IJCNLP 2017, and Senior PC of IJCAI 2017/IJCAI 2018(Distinguished SPC)/AAAI 2019, and Reviewers for ACL, IJCAI, AAAI, EACL, COLING, EMNLP, NAACL and journals such as TOIS, TKDE, TPAMI, etc..

Isabel Trancoso (Fellow. IEEE) received the Licenciado, Mestre, Doutor and Agregado degrees in electrical and computer engineering from IST, Lisbon, Portugal in 1979, 1984, 1987, and 2002, respectively. She is a Full Professor at IST, and the President of the Scientific Council of INESC-ID. Her research covers many different topics in spoken language processing. She was the Chair of the ECE Department of IST. She was elected Editor in Chief of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, and President of International Speech Communication Association (ISCA). She was a member of the IEEE Fellows Committee, and Vice-President of the ELRA Board. She chaired the IEEE James Flanagan Award Committee, ISCA Distinguished Lecturer Selection Committee, and Fellow Evaluation Committee of the Signal Processing Society of IEEE. She currently integrates the Editorial Board of the Proceedings of IEEE, and chairs the ISCA Fellow Selection Committee. She was elevated to IEEE Fellow in 2011, and to ISCA Fellow in 2014.



Chao Zhang received his B.E. and M.S. degrees in 2009 and 2012 respectively, both from the Department of Computer Science and Technology, Tsinghua University, and a Ph.D. degree from Cambridge University Engineering Department (CUED), in 2017. He is currently is a technique advisor of JD AI speech team and a Research Associate in speech and language processing at the University of Cambridge. He is an author of the Hidden Markov Model Toolkit (HTK) and developed HTK 3.5 and PyHTK with the native deep learning functions in C. He has published more than forty conference and journal papers, including the best student papers from NCMMSC 2011, ICASSP 2014, and ASRU 2019, along with a best paper candidate from ASRU 2015 and other awards and grants, *etc*. As a member of the CUED team, he attended multiple speech recognition evaluations including iARPA Babel 2013, DARPA BOLT 2014, and ASRU 2015 MGB.