Guest Editorial Multimedia Computing With Interpretable Machine Learning

I. Introduction

MULTIMEDIA is increasingly becoming the "biggest big data," among the most important and valuable source for insight and information. Many powerful machine learning algorithms, especially deep learning models such as convolutional neural networks (CNNs), have recently achieved outstanding predictive performance in a wide range of multimedia applications, including visual object classification, scene understanding, speech recognition, and activity prediction. Nevertheless, most deep learning algorithms are generally conceived as black-box methods, and it is difficult to intuitively and quantitatively understand the results of their prediction and inference. Since this lack of interpretability is a major bottleneck in designing more successful predictive models and exploring wider-range useful applications, there has been an explosion of interest in interpreting the representations learned by these models, with profound implications for research into interpretable machine learning in the multimedia community.

According to Wikipedia and DARPA, explainable or interpretable artificial intelligence (XAI) refers to a suite of artificial intelligence (AI) methods and techniques – especially machine learning techniques – that produce more explainable models or results that enable human users to understand, appropriately trust, and effectively manage, while maintaining a high level of learning performance (prediction accuracy). These methods include, but are not limited to, visualizing feature maps, establishing architectures with mathematically, cognitively or biologically plausible theories, and explaining the results with mid-level attributes. Instead of general interpretability in AI, interpretable machine learning is defined by W. J. Murdoch *et al.* as the extraction of relevant knowledge from a machine learning model concerning relationships either contained in data or learned by the model. Tremendous data exists in multimedia, such as network videos, medical images, voices, texts, to name a few, providing vast applications for machine learning. Therefore, it is possible to boost the performances of interpretable machine learning with the help of rich and explainable information in multimedia, such as interpreting the model with attributes from texts and exploring the relationship among the multi-modal data. In turn, interpretable machine learning could also promote to multimedia applications, such as assisting in medical diagnosis.

The goal of this special issue is to broadly engage the machine learning and multimedia communities on the emerging yet challenging interpretable machine learning topic – tying together many threads which are deeply related but often considered in isolation. A total of 102 papers were submitted to this special issue, including two invited papers. After a rigorous review process, seventeen high-quality papers have been selected for publication, with an acceptance rate of around 17%. They roughly outline the research challenges and issues in the field, and present the state-of-the-art theory, methods, and technologies on multimedia computing with interpretable machine learning.

II. Resesarch Topics in This Special Issue

The accepted papers can be divided into five categories. In the following, we will present a brief description of these papers for the convenience of the readers.

A. Interpretable Multimedia Intelligence

AI has undergone a "new" wave of development since 1950 s, which should give credits to the success of deep learning. In the position paper, "Multimedia Intelligence: When Multimedia Meets Artificial Intelligence," W. Zhu, X. Wang, and W. Gao raised one question: What happens when multimedia meets AI? To answer this question, they introduced the concept of *multimedia intelligence*. They explored two aspects in which multimedia and AI interactively enhance each other: i) multimedia drives AI to experience a paradigm shift towards more explainability, and ii) AI in turn injects new ways of thinking for multimedia research. Then they discussed what and how efforts had been done in the literature and shared their insights on research directions deserving further study to produce a profound impact on multimedia intelligence.

B. Quantifying and Visualizing the Interpretability

How to quantify and visualize the interpretability is one key issue to develop interpretable learning algorithms for multimedia applications. In the first work, "Feature-flow Interpretation of Deep Convolutional Neural Networks," X. Cui, D. Wang, and Z. J. Wang proposed a novel Feature-fLOW INterpretation (FLOWIN) model to interpret a Deep CNN by its feature flow. The FLOWIN is able to provide an instance-specific interpretation, by presenting its feature flow units and their quantitative interpretations for its network decision. From the class-level view, networks can be further analyzed by studying feature flows

within and between classes. In the experiments, the FLOWIN was evaluated on different datasets and networks in quantitative and qualitative ways to show its interpretability.

By highlighting important features that contribute to model prediction, visual saliency is a natural form to interpret the working mechanism of Deep Neural Networks (DNNs). In the second paper entitled "Learning Reliable Visual Saliency for Model Explanations," Y. Wang, *et al.* found that the existing saliency estimation methods were not reliable enough to provide meaningful interpretation, mainly due to the attribution vanishing and adversarial noise. In order to learn reliable visual saliency, they proposed a simple method requiring the output of the model to be close to the original output while learning an explanatory saliency mask. Experimental results show that their approach helps to improve reliability by suppressing false saliency responses.

C. Interpretable Multimedia Processing Algorithms

Deep learning is revolutionizing the typical paradigm of many image and video processing tasks such as compression, reconstruction, colorization, and retargeting. However, the "deep" image and video processing systems should also have the ability to explain their rationale, characterize their strengths and weaknesses, and convey an understanding of how they will behave in the future.

In the first paper entitled "iWave: CNN-Based Wavelet-Like Transform for Image Compression," H. Ma, *et al.* proposed iWave as a framework for deriving wavelet-like transform more suitable for natural image compression. iWave adopts an update-first lifting scheme, where the prediction filter is a trained CNN and can be embedded into an end-to-end autoencoder. This trained wavelet-like transform still possesses the lifting structure, which ensures perfect reconstruction, supports multi-resolution analysis, and is more interpretable than the deep networks trained as "black boxes." They performed experiments to verify the generality as well as the specialty of iWave in comparison with JPEG-2000.

It is a research hotspot to restore decoded videos with existing bitstreams by applying a DNN to improve compression efficiency at the decoder end. To make use of the multi-scale similarity at the decoder end to enhance compression efficiency, in the second paper entitled "The Interpretable Fast Multi-Scale Deep Decoder for the Standard HEVC Bitstreams," W. Xiao, *et al.* considered the use of underused inter multiscale information and proposed the Fast Multi-Scale Deep Decoder (Fast MSDD) for the state-of-the-art video coding standard HEVC. Fast MSDD is interpretable, and can achieve a higher coding efficiency without modifying any encoding algorithm. Moreover, it guarantees the model's inference speed while fully using the multiscale similarity among video frames. Extensive experimental results verified its effectiveness, interpretability, and computational efficiency.

In the third paper, "Automated Colorization of a Grayscale Image with Seed Points Propagation," S. Wan, *et al.* proposed a fully automatic image colorization method for grayscale images using neural network and optimization. For a determined training set including the gray images and its corresponding color images, their method segments grayscale images into superpixels and then extracts features of particular points of interest in each superpixel. The obtained features and their RGB values are given as input for the training colorization neural network of each pixel. Then their method propagates the resulting color points to neighboring pixels and employs a guided image filter to refine the colorized image. Experiments on a wide variety of images showed that it could achieve superior performance over the state-of-the-art algorithms.

Content-aware image retargeting is used to manipulate media content to make it adapt to different aspect ratios of the device screens intelligently. In the fourth paper entitled "Cycle-IR: Deep Cyclic Image Retargeting," W. Tan, *et al.* proposed a deep cyclic image retargeting approach called Cycle-IR. Their idea is built on the reverse mapping from the retargeted images to the given images. If the retargeted image has serious distortion or excessive loss of important visual information, the reverse mapping is unlikely to restore the input image well. They constrained this forward-reverse consistency by introducing a cyclic perception coherence loss, and proposed an image retargeting network (IRNet) to produce visually pleasing target images directly. In the IRNet, a spatial and channel attention layer can discriminate visually important regions of input images effectively, especially in cluttered images. Extensive experiments on the standard RetargetMe dataset showed the superiority of Cycle-IR.

In the paper "PixelRL: Fully Convolutional Network with Reinforcement Learning for Image Processing," R. Furuta, N. Inoue, and T. Yamasaki proposed to tackle a new problem setting for image processing: reinforcement learning with pixel-wise rewards (pixelRL). In the pixelRL, each pixel has an agent, and the agent changes the pixel value by taking an action. An effective learning method was proposed for the pixelRL by considering not only the future states of the own pixel but also those of the neighboring pixels. When applied to some image processing tasks, it is possible to visualize what kind of operation is employed for each pixel at each iteration, which would help us understand why and how such an operation is chosen and easily identify which operations can be changed or modified if necessary. The experimental results demonstrated that pixelRL could achieve comparable or better performance, compared with state-of-the-art methods.

D. Interpretable Learning Models and Algorithms for Multimedia Analysis and Retrieval Tasks

When classifying or recognizing items of interest (e.g., visual objects, pedestrians, events) in multimedia data, the model or algorithm should help human users to understand and consequently trust its results. The strategy for achieving that goal is to develop new or modified learning techniques that will produce more explainable models on multimedia data, while maintaining a high level of learning performance.

In the first paper, "Bidirectional Attention-Recognition Model for Fine-grained Object Classification," C. Liu, *et al.* proposed a bidirectional attention-recognition model (BARM) to actualize the bidirectional reinforcement for fine-grained object

classification. The proposed BARM consists of one attention agent for discriminative part regions proposing and one recognition agent for feature extraction and recognition. Meanwhile, a feedback flow is creatively established to optimize the attention agent directly by a recognition agent. Therefore, in BARM the attention agent and the recognition agent can reinforce each other in a bidirectional way and the overall framework can be trained end-to-end without neither object nor parts annotations. On several benchmarks, BARM outperformed some state-of-the-art methods in classification accuracy, and exhibited clear interpretability and kept consistent with the human perception.

Zero-Shot Learning (ZSL) has received extensive attention in recent years. In the second paper, "Hierarchical Prototype Learning for Zero-Shot Recognition," X. Zhang, *et al.* proposed a hierarchical prototype learning formulation to provide a systematical solution (named HPL) for zero-shot recognition. Specifically, HPL is able to obtain discriminability on both seen and unseen class domains by learning visual prototypes respectively under the transductive setting. To narrow the gap between two domains, they further learned the interpretable super-prototypes in both visual and semantic spaces by maximizing their structural consistency. This not only facilitates the representativeness of visual prototypes, but also alleviates the loss of information of semantic prototypes. Extensive experiments demonstrated that HPL obtained more favorable efficiency and effectiveness, over some alternatives.

In the third paper, "Similarity-Aware and Variational Deep Adversarial Learning for Robust Facial Age Estimation," H. Liu, *et al.* proposed a similarity-aware deep adversarial learning (SADAL) approach for facial age estimation. During the learning process, they jointly optimized both procedures of generating hard negatives and learning discriminative age ranker via a sequence of adversarial-game iterations. To circumvent the model overfitting problem, particularly on unseen age classes with many individuals, they proposed an interpretable variational deep adversarial learning (VDAL) paradigm, which explicitly factorizes the raw input facial image into the intra-class variance and the intra-class invariance, in parallel with preserving the age-difference information in the variational factorized feature representations. Experimental results demonstrated that their approach achieved to generate meaningful samples as the complements to existing training samples, leading to the promising performance in contrast to most state-of-the-art age estimation methods.

The newly emerging task of person search with natural language query aims at retrieving the target pedestrian by a text description of the pedestrian. In the fourth paper, "Adversarial Attribute-Text Embedding for Person Search with Natural Language Query," Z. J. Zha, *et al.* proposed a novel Adversarial Attribute-Text Embedding (AATE) network for person search with text query. A cross-modal adversarial learning module was developed to learn discriminative and modality invariant visual-textual features. Moreover, an attribute graph convolutional network was utilized to learn the visual attributes of pedestrians, which possess better descriptiveness, interpretability, and robustness compared to pedestrian appearance. Extensive experimental results on two challenging benchmarks demonstrated the effectiveness of the proposed approach.

In the fifth paper entitled "Spatio-Temporal VLAD Encoding of Visual Events using Temporal Ordering of the Mid-Level Deep Semantics," M. Soltanian, S. Amini, and S. Ghaemmaghami showed that by considering the temporal inter-frame dependencies and tracking the chronological order of video sub-events, the accuracy of event recognition could be further improved. First, the frame-level descriptors are extracted using CNNs pre-trained on the ImageNet and fine-tuned on a portion of the training video frames. Then, a spatiotemporal encoding is applied to the derived descriptors in the form of a convex optimization problem. The experimental results showed its superiority in terms of recognition accuracy over both frame-level video encoding approaches and spatio-temporal video representations.

To facilitate the deployment of DNNs on resource-constrained devices, in the seventh paper entitled "Iterative Deep Neural Network Quantization with Lipschitz Constraint," Y. Xu, *et al.* proposed an interpretable framework for network quantization with arbitrary bit-widths. They presented two Lipschitz-constraint-based quantization strategies, namely width-level network quantization (WLQ) and multi-level network quantization (MLQ), for high-bit and extremely low-bit (ternary) quantization. In the WLQ, Lipschitz-based partition is developed to divide parameters in each layer into two groups, respectively for quantization and for re-training. The WLQ is further extended to MLQ by introducing the layer partition to suppress the quantization loss for extremely low bit-widths. Experimental results showed that it could improve the performance of tasks like classification, object detection, and semantic segmentation, with a guarantee of convergence.

E. New Applications With Interpretable Machine Learning

Interpretable machine learning will promote to multimedia applications, and even give birth to attractive applications. Three examples have been presented in this special issue.

In the first paper, "PointHop: An Explainable Machine Learning Method for Point Cloud Classification," M. Zhang, *et al.* proposed an explainable machine learning method for point cloud classification. Their method, called PointHop, consists of two stages. In the attribute building stage, they developed a robust descriptor to characterized the relationship between a point and its one-hop neighbor in a PointHop unit. When multiple PointHop units were put in cascade, the attributes of a point would grow by taking its relationship with one-hop neighbor points into account iteratively. To control the rapid dimension growth of the attribute vector associated with a point, they used the Saab transform to reduce the attribute dimension in each PointHop unit. In the classification stage, they fed the feature vector obtained from multiple PointHop units to a classifier, and explored ensemble methods to further improve the performance. Experimental results showed that the PointHop offered classification performance comparable to state-of-the-art methods, with much lower training complexity.

In the second paper, "VINet: A Visually Interpretable Image Diagnosis Network," D. Gu, *et al.* proposed a visually interpretable network (VINet) which could generate diagnostic visual interpretations while making accurate diagnoses. VINet is an end-to-end model consisting of an importance estimation network and a classification network. The former produces a diagnostic visual interpretation for each case, and the classifier diagnoses the case. In the classifier, by exploring the information in the diagnostic visual interpretation, the irrelevant information in the feature maps is eliminated by the proposed feature destruction process. This allows the classification network to concentrate on the important features. Through a joint optimization of higher classification accuracy and eliminating as many irrelevant features as possible, a precise, fine-grained diagnostic visual interpretation, along with an accurate diagnosis, can be produced by the proposed network simultaneously. Extensive experiments demonstrated that the VINet could produce state-of-the-art diagnostic visual interpretations over all baseline methods.

The attractiveness of a property is one of the most interesting, yet challenging, categories to model. Image characteristics can be used to examine the influence of visual factors on the price or timeframe of the listing. In the third paper entitled "What Image Features Boost Housing Market Predictions," Z. Kostic and A. Jevremovic proposed a set of techniques for the extraction of visual features for efficient numerical inclusion in modern-day predictive algorithms. After comparing different techniques as applied to a set of property-related images (indoor, outdoor, and satellite), they concluded the following: (i) the entropy is the most efficient single-digit visual measure for housing price prediction; (ii) image segmentation is the most important visual feature for the prediction of housing lifespan; and (iii) deep image features can be used to quantify interior characteristics and contribute to captivation modeling. The set of 40 image features selected in the paper carried a significant amount of predictive power and outperformed some of the strongest metadata predictors.

III. Concluding Remarks

From these accepted papers, we now leverage it to frame what we feel are the most important challenges and research directions moving forward in the field.

(1) Well-formed definition and technological roadmaps. Currently, there is no clear consensus in the community around what it means to be interpretable, and how to select, upgrade, or even modify methods for producing interpretations of machine learning models. In the absence of a well-posed definition and some well-formed technological roadmaps, a broad range of models and methods have been labeled as interpretation. Strictly speaking, for example, some papers in this special issue cannot be regarded as interpretable machine learning methods. Nevertheless, we believe that these works can give constructive insight to us for exploring the clearer definition and reasonable technological roadmaps of interpretable machine learning in the multimedia domain.

(2) Quantitative evaluation benchmarks. How to evaluate the interpretation methods quantitatively still remains an open issue. Consequently, the evaluation criterion varies considerably across different works, making it challenging both for researchers in the field to measure their progress and for potential users to select appropriate models. Technologically, the evaluation issue will involve the testing methodology and benchmark datasets. Just as the success of deep learning can be mostly attributed to the ImageNet benchmark, the field of multimedia can also promote the evaluation task of interpretable machine learning very well. However, due to the diversified scenarios and subjective nature of explanations, it is very challenging about how to have the ground truth for benchmark evaluation on the quality of generated explanations.

(3) Explainable crossmedia reasoning. Exploring more explainable reasoning procedures across different entries with various media tyeps will be another important research direction deserving further investigations. The first thing we can do is to equip the model with more and better reasoning-augmented layers or modules, such as utilizing the graph neural networks to enable the model to conduct human-like relation reasoning. Take a further step, given that current neural networks and the reasoning modules are optimized separately, the incorporation of neural network and reasoning through a joint-optimizing framework will play an important role in achieving the goal of explainable reasoning in multimedia.

(4) Attractive multimedia applications. Finally, just like the great success of AI is inseparable from its close integration with real-world industrial applications, the interpretable machine learning models and algorithms should validate themselves in real-world applications, even foster newly, attractive multimedia applications. Some works in this special issue have demonstrated this possibility, but not enough. The future development of interpretable machine learning depends on whether these models and methods play an important or even leading role in new applications.

Acknowledgment

At the end of this editorial, we would like to thank those individuals who have helped make this special issue possible, especially the former Editor-in-Chief, Prof. Wenwu Zhu, and the administrative assistant, Ms. Mikaela Langdon. Most importantly, thank you to all of the authors who submitted manuscripts for consideration, and to the many dedicated reviewers who helped us arrive at our final choices.

Y. Tian, *Guest Editor*
Department of Computer Science and Technology
School of EE&CS, Peking University
Beijing 100871, China,
The Artificial Intelligence Research Center
PengCheng Laboratory,
Shenzhen 518066, China

C. Snoek, *Guest Editor*
Informatics Institute
University of Amsterdam
Amsterdam 94323, The Netherlands

J. WANG, *Guest Editor*
Microsoft Research Asia
Beijing 100080, China

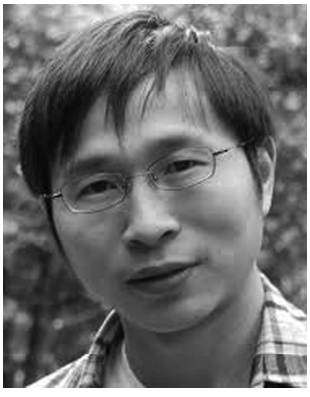
Z. LIU, *Guest Editor*
AT&T Labs Research
Middletown, NJ 07748 USA

R. LIENHART, *Guest Editor*
Computer Science Department
University of Augsburg
Augsburg 86159, Germany

S. BOLL, *GUEST EDITOR*
Department of Computing Science
University of Oldenburg
Oldenburg 26121, Germany

Yonghong Tian (Senior Member, IEEE) is currently a Boya Distinguished Professor with the Department of Computer Science and Technology, School of EE&CS, Peking University, China, and is also the deputy director of Artificial Intelligence Research Center, PengCheng Laboratory, Shenzhen, China. His research interests include computer vision, multimedia big data, and brain-inspired computation. He is the author or coauthor of over 200 technical articles in refereed journals and conferences. Prof. Tian was/is an Associate Editor of IEEE TCSVT (2018.1-), IEEE TMM (2014.8-2018.8), *IEEE Multimedia Mag.* (2018.1-), and IEEE ACCESS (2017.1-2019.12). He co-initiated IEEE Intl Conf. on Multimedia Big Data (BigMM) and served as the TPC Co-chair of BigMM 2015, and aslo served as the Technical Program Co-chair of IEEE ICME 2015, IEEE ISM 2015 and IEEE MIPR 2018/2019, and General Co-chair of IEEE MIPR 2020 and IEEE ICME 2021. He is the steering member of IEEE ICME (2018-) and IEEE BigMM (2015-), and is a TPC Member of more than ten conferences such as CVPR, ICCV, ACM KDD, AAAI, ACM MM and ECCV. He was the recipient of the Chinese National Science Foundation for Distinguished Young Scholars in 2018, two National Science and Technology Awards and three ministerial-level awards in China, and obtained the 2015 EURASIP Best Paper Award for Journal on Image and Video Processing, and the best paper award of IEEE BigMM 2018. He is a senior member of IEEE, CIE and CCF, a member of ACM.

Cees G. M. Snoek is a Full Professor in computer science with the University of Amsterdam, where he heads the Intelligent Sensory Information Systems Lab. He is also a director of three public-private AI research labs: QUVA Lab with Qual-comm, Atlas Lab with TomTom and AIM Lab with the Inception Institute of Artificial Intelligence. He received the M.Sc. degree in business information systems (2000) and the Ph.D. degree in computer science (2005) both from the University of Amsterdam, The Netherlands. He has published over 200 refereed journal and conference papers on video and image recognition in multimedia and computer vision venues. Several of his Ph.D. students and post-docs have won awards, including the IEEE Transactions on Multimedia Prize Paper Award (2012) the SIGMM Best Ph.D. Thesis Award (2013), the Best Paper Award of ACM Multimedia (2014), and the Best Paper Award of ACM Multimedia Retrieval (2016). He was general chair of ACM Multimedia 2016 in Amsterdam. He was previously on the editorial boards of ACM Transactions on Multimedia, IEEE Transaction on Multimedia and IEEE Multimedia.



Jingdong Wang is a Senior Principal Research Manager with the Visual Computing Group, Microsoft Research, Beijing, China. He received the B.Eng. and M.Eng. degrees from the Department of Automation, Tsinghua University, Beijing, China, in 2001 and 2004, respectively, and the PhD degree from the Department of Computer Science and Engineering, the Hong Kong University of Science and Technology, Hong Kong, in 2007. His areas of interest include deep learning, large-scale indexing, human understanding, and person re-identification. He is an Associate Editor of IEEE TPAMI, IEEE TMM and IEEE TCSVT, and is an area chair (or SPC) of some prestigious conferences, such as CVPR, ICCV, ECCV, ACM MM, IJCAI, and AAAI. He is a Fellow of IAPR and an ACM Distinguished Member.

Zhu Liu is a Principal Inventive Scientist at AT&T Labs Research. He received the B.S. and M.S. degrees in Electronic Engineering from Tsinghua University, Beijing, China, in 1994 and 1996, respectively, and the Ph.D. degree in Electrical Engineering from NYU Tandon School of Engineering in 2001. He is an Adjunct Associate Professor at Columbia University. His research interests include video and multimedia content analysis, machine learning, big data, and natural language understanding. In 2017, he received the AT&T Science & Technology Medal for his contributions and leadership in video analytics. He holds 133 granted U.S. patents and has published more than 70 conference and journal papers and book chapters. He is serving as an associate editor for IEEE Transactions on Multimedia and IEEE Signal Processing Letters. He is also on the organizing committee and technical committee for many IEEE International Conferences. Dr. Liu is a senior member of IEEE.

Rainer Lienhart is a Full Professor in the computer science department of the University of Augsburg and chair of the Multimedia Computing and Computer Vision Lab (MMC Lab). His group is focusing on all aspects of (1) large-scale image, video, humans pose, sensor and data mining algorithms, (2) object, human pose and action detection/recognition as well as (3) image, video, human pose and action retrieval. Since April 2010, he is also the executive director of the Institute for Computer Science at the University of Augsburg. The scientific work of Rainer Lienhart covers more than 80 refereed publications and more than 20+ patents. He was a general co-chair of ACM Multimedia 2017, ACM Multimedia 2007 and SPIE Storage and Retrieval of Media Databases 2004 & 2005. He serves in the editorial boards of 3 international journals. For more than a decade he is a committee member of ACM Multimedia, IEEE ICME, SPIE Storage and Retrieval of Media Databases, and many more conferences. From July 2009 till June 2017 he was the vice chair of ACM SIGMM.

Susanne Boll is a Full Professor in the Department of Computing Science at leads the Media Informatics and Multimedia Systems group at the University of Oldenburg, in Germany. She serves on the executive board of the OFFIS Institute for Information Technology, in Oldenburg, Germany. Her research interests lie in the field of semantic retrieval of digital media, interactive and intelligent user interfaces. She is a member of the editorial board of the IEEE Multimedia Magazine and Springer Multimedia Tools and Applications (MTAP). She has published over 250 articles in competitive conferences, workshops, journals and books; served as a member on the technical program committees of more than 150 renowned conferences and workshops; and successfully acquired, led, and managed 50 national and international projects. She has taken over leading roles in the organization and support for many events as was a Program Co-Chair of ACM Multimedia in 2017 and General Co-Chair of ACM Multimedia in 2019.