Guest Editorial: Deep Learning for Multimedia Computing

CONVENTIONAL multimedia computing is often built on top of handcrafted features, which are often much restrictive in capturing complex multimedia content such as images, audios, text and user-generated data with domain-specific knowledge. Recent progress on deep learning opens an exciting new era, placing multimedia computing on a more rigorous foundation with automatically learned representations to model the multimodal data and the cross-media interactions. This represents unprecedented opportunity for new researches on deep network architecture, learning and inference algorithms, and computational systems and infrastructures for multimedia applications. Existing studies have revealed promising results that have greatly advanced the state-of-the-art performance in a series of multimedia research areas, from the multimedia content analysis, to modeling the interactions between multimodal data, to multimedia content recommendation systems, to name a few here.

This Special Issue aims at providing a forum to present recent advancements in deep learning research that directly concerns the multimedia community. Specifically, deep learning has successfully designed algorithms that can build deep nonlinear representations to mimic how the brain perceives and understands multimodal information, ranging from low-level signals like images and audios, to high-level semantic data like natural language. For multimedia research, it is especially important to develop deep networks to capture the dependencies between different genres of data, building joint deep representation for diverse modalities.

A total of 43 papers were submitted to this Special Issue, including 1 invited survey paper. After three rounds of rigorous reviews, 20 papers have been selected for publication. We categorize the accepted papers into four groups. The first group consists of five papers studying the deep learning and inference algorithms applied to data of multi-modalities. The second group comprises of three papers which index and retrieve images on various levels of image structures. The third group of four papers use the deep network to analyze the user-generated data in social media and search engines. Finally, the last group of six papers study a variety of deep learning applications into multimedia problems, which have potential of impacts on our everyday lives.

At the core of the research on "deep learning for multimedia computing" is how to bridge the cross-modality gap between a large varieties of multimedia data. Cho, Courville, and Bengio review this problem in an invited paper "Describing multimedia content using attention-based encoder-decoder networks." This paper proposes a new family of deep networks with rich input and output structures that are related with each other. A novel attention mechanism is presented, which shows effective performance on machine translation, video clip description and speech recognition.

Many multimedia problems contain multimodal data, making it nontrivial to develop proper deep learning algorithms of multi-modalities. To address this challenge, Wang *et al.* in their paper "Large-margin multi-modal deep learning for RGB-D object recognition" present a general Convolutional Neural Network (CNN) based multi-modal learning framework for RGB-D object recognition problem. This framework not only discovers the most discriminative features for each modality, but also is able to harness the complementary relationship between modalities. In addition, Tang *et al.* in "RGB-D object recognition via incorporating latent data structure and prior knowledge" incorporate the multimodal data structure to construct the CNN for object recognition problem by transferring the prior knowledge from rich feature hierarchies trained on ImageNet. Both methods demonstrate competitive object recognition performance on RGB-D datasets. In "A continuous learning framework for activity recognition using deep hybrid feature models," Hasan and Roy-Chowdbury propose a continuous activity learning framework by intricately tying together deep hybrid feature models and active learning. It takes the advantage of both the local hand-engineered features and the deep model in an efficient way. It also uses active learning to train the activity classifier with a reduced amount of labeled instances, and retrains the model by selecting the best subset of accumulated training examples.

In addition to the above three papers aiming at specific problems, the other two papers in the first group study the multimodal deep learning problems for general applications. Huang *et al.* consider a general problem where the multiple labels present dependency and some modalities are likely to be missing. In their paper "Unconstrained multimodal multi-label learning," a multi-label conditional restricted Boltzmann machine is proposed to handle modality completion, fusion and multi-label prediction in a unified framework. In "Heterogeneous feature selection with multi-modal deep neural networks and sparse group LASSO," Zhao *et al.* present a discriminative feature selection framework through solving a sparse group LASSO problem in a unified multimodal deep neural network. The results show that the approach is effective in selecting the relevant feature groups and achieves competitive classification performance.

Digital Object Identifier 10.1109/TMM.2015.2485538

The second group of papers present the deep learning approaches to model the images on various levels. In “Multi-task CNN model for attribute prediction,” Abdulnabi *et al.* propose a joint multi-task CNN model to learn attribute-specific feature representation for images. Natural grouping of attributes is applied in a way that attributes in the same group share more knowledge and the attributes from different groups compete with each other. In “Learning representative deep features for image set analysis” and “Cross indexing with grouplets,” two groups of researchers independently present an idea of treating images by groups in learning the deep feature representations. Wu *et al.* propose to learn features from sets of labeled raw images to reduce the overfitting risk, while Zhang *et al.* present an image indexing system by viewing the image database as a set of grouplets. Both approaches show improvement in efficiency of modeling large-scale image database.

In the third group, the authors put their attentions to user-generated multimedia data. In “Understanding blooming human groups in social networks,” Hong *et al.* propose an approach to understanding the new concepts of human group with few positive samples. Two different CNNs based on face and upper body are constructed separately, while the surrounding texts are represented by semantic vectors as image labels. Li *et al.* presents a new distance metric learning algorithm under the deep learning framework in “Weakly supervised deep metric learning for community-contributed image retrieval.” It utilizes a progressive learning approach to jointly exploit the heterogeneous data structures from visual contents as well as user-provided tags of social images. Finally, in “Learning cross space mapping via DNN using large scale click-through logs,” Yu *et al.* extends the ability of Deep Neural Networks (DNNs) to image retrieval task by leveraging image-query similarity calculation. A Cross Space Mapping DNN model is developed by mapping image and queries to a common vector space where image-query similarity is naturally defined as their inner product. A large scale experiment with 23 million clicked image-query pairs between 1 million images and 11.7 million queries shows superior results. Finally, Pang *et al.* present using Deep Bolzmann Machine to understand the perceived emotions inherent in the social media data, in absence of labeling effort in their paper “Deep multimodal learning for affective analysis and retrieval.”

The last group of papers showcase several applications of deep learning approaches. In “Rating image aesthetics using deep learning,” Lu *et al.* authors propose a double-column CNN to support heterogeneous inputs to reveal the style and semantic attribtes of images to boost the aesthetics categorization performance. Tian *et al.* in “Query-dependent aesthetic model with deep learning for photo quality assessment” present a novel query-dependent aesthetic deep learning model instead. Both methods draw definitive conclusion that deep learning features outperform the handcrafted features in image aesthetics and quality assessment. Ding *et al.* propose a comprehensive deep learning framework to jointly model face representation with multimodal information in “Robust face recognition via multimodal deep face representation.” A set of elaborately designed CNNs and a three-level auto-encoder are structured to extract facial features. Kereliuk *et al.* apply deep learning framework to study the problem of music adversaries in “Deep learning and music adversaries.”

They find the CNNs are more robust compared with the systems based on a majority vote over individually classified author frames. Finally, Wang *et al.*, in "DeepBag: Recognizing handbag models," show the application of a new feature selective joint classification-regression CNN model into fashion industry for handbag recognition. The research leads to a better handbag classifier even with existence of large inter-class similarity. Shen *et al.*, in their paper “Predicting eye fixations on webpage with an ensemble of early features and high-level representations from deep network,” apply deep neural networks to predict fixations on webpage. Mukherjee and Robertson present a CNN-based model for human head pose estimation in low-resolution multi-modal RGB-D data in “Deep head pose: Gaze-direction estimation in multimodal video.”

We would like to extend our appreciation to those individual reviewers for their timely and professional comments. We also thank Ms. A. Fisher who assisted us in the past few months through the long publication process for this special issue. Many thanks shall go to Prof. C. W. Chen, the Editor-in-Chief of the IEEE TRANSACTIONS ON MULTIMEDIA. Without his strong support of this issue, nothing would have happened.

Most importantly, thanks to all the authors who submitted their manuscripts for consideration. Their support makes the special issue successful.

Guo-Jun Qi, *Guest Editor*
Department of Computer Science
University of Central Florida
Orlando, FL 32816 USA

Hugo Larochelle, *Guest Editor*
Department of Computer Science
Université de Sherbrooke
Sherbrooke, QCJ1K 2R1 Canada

Benoit Huet, *Guest Editor*
Multimedia Communications Department
EURECOM
Biot, 06410 France

Jiebo Luo, *Guest Editor*
Department of Computer Science
University of Rochester
Rochester, NY 14627 USA

Kai Yu, *Guest Editor*
Multimedia Department
Baidu Inc.
Beijing, China