Department

Multimedia Deep Learning

Shu-Ching Chen
Florida International University

***Abstract*—By achieving breakthrough results on domains such as speech recognition, natural language processing, and computer vision, it is no surprise that deep neural networks are receiving a lot of attention these days. Specifically in the field of multimedia data analysis, there is a tremendous amount of multimedia big data that is being generated every day. Deep learning has the potential to overcome the issue of multimedia data having massive and heterogeneous characteristics that make it a challenge to store and analyze the data. This can be accomplished by allowing computers to easily and automatically extract features from unstructured data without the need to rely on human intervention. Although recent multimedia deep learning methods have achieved some remarkable results, deep learning challenges such as interpretability and generalization still make it difficult to be fit for critical decision-making tasks from fields such as medicine and defense.**

With the advent of social networks and new mobile technologies, a tremendous amount of multimedia data (also referred to as multimedia big data) is generated day to day. This explosion of multimedia data, including video, image, audio, and text, has provided numerous opportunities in various applications, such as healthcare, education, environment, and automotive industry. However, it is difficult to collect, store, process, and manage such massive and diverse data effectively. Due to these challenges and opportunities, multimedia big data analytics has become a demanding research topic and attracted significant attention in the multimedia community.[1]

In recent years, deep learning (DL) has been successfully explored in various multimedia applications such as natural language processing, visual data analytics, and speech recognition.[2] DL takes inspiration from the field of neuroscience, building neural networks structured in a way that resembles the brain. It is a subset of a broader field known as machine learning whose purpose is to automatically learn data representations. Considering multimedia data are characterized as large, unstructured, and heterogeneous, DL has the potential to overcome these issues by

Digital Object Identifier 10.1109/MMUL.2019.2897471
Date of current version 27 March 2018.

allowing computers to easily and automatically extract features from unstructured data without the need to rely on human intervention.[3] However, there are still several critical questions in multimedia DL that researchers are trying to answer: 1) how to handle multimedia big data efficiently; 2) how to utilize various data modalities using DL; and 3) how to gain insights and understand the decision-making process from deep neural networks.

The first challenge refers to the big amounts of data; despite the great success of DL, it is still highly dependent on large-scale labeled datasets. Providing such datasets is laborious and time-consuming. Transfer learning[4,5] and generative adversarial networks[6] have been proposed in the recent few years to alleviate this problem. In addition, unsupervised ML algorithms, such as deep reinforcement learning and variational autoencoders, have become more prominent these days.[2] Once the dataset is in hand, training deep neural networks may require days on powerful CPU and GPU clusters. To overcome this issue, many researchers have focused on parallel and scalable DL models.[2] Some researchers have also shifted the focus to build low-power DL models or to utilize DL accelerators using field-programmable gate array (FPGA).[7] Yet, big data and computational efficiency are major challenges in multimedia DL.

Moreover, traditional DL architectures have mostly worked on developing solutions with a focus on single modalities (e.g., text, image, or audio). Convolutional neural network (CNN) architectures have, for the most part, been used for visual feature extraction and word embedding models for textual analysis. In the study known as multimodal learning, DL frameworks must analyze all information extracted from different data modalities and find the underlying logical connections between them. By leveraging the knowledge acquired from different data sources, models can make decisions that are more precise. Thus, most proposed techniques rely on fusion techniques to integrate information from various input modalities. However, creating effective integrations to generate a common representation between different data types is a major challenge. The fusion approach must find the best alignment between long-range dependencies while exploiting the complementarity and redundancy of multiple modalities to create common representations between multimodal data.

Another main challenge in this area is DL interpretability. With enough training samples, deep neural networks will discover features that are too complicated for a human to understand. Therefore, the inner workings of a neural network are often referred to as a blackbox. Critical decision-making tasks require a clear insight into the steps that led the neural network to a certain solution. Hence, this reasoning capability is typically most vital in the fields that involve medicine and defense. For instance, medical image analysis comes in a combination of multiple image modalities such as magnetic resonance imaging (MRI) and computed tomography (CT) scan. Interpreting features generated from medical imaging is a major challenge, considering it requires the validation from highly trained human experts.[8,9]

In summary, despite the early achievements of utilizing DL in multimedia applications, there still exist several critical challenges that require more attention in future studies. Nonetheless, with the rapid improvements in computing power, as well as the advent of novel algorithms and techniques in this area, multimedia DL offers new opportunities that were not possible before.

I hope you enjoy reading the articles in this issue. You are welcome to submit your work to *IEEE MultiMedia*.

References

1. S. Pouyanfar, Y. Yang, S.-C. Chen, M.-L. Shyu, and S. S. Iyengar, "Multimedia big data analytics: A survey," *ACM Comput. Surveys*, vol. 51, no. 92, pp. 1–36, 2018.
2. I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, vol. 1. Cambridge, MA, USA: MIT Press, 2016.
3. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
4. Y. Mroueh, E. Marcheret, and V. Goel, "Deep multimodal learning for audio-visual speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 2130–2134.

5. J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3320–3328.
6. P. Jing, Y. Su, L. Nie, and H. Gu, "Predicting image memorability through adaptive transfer learning from external sources," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 1050–1062, May 2017.
7. I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
8. C. Wang, L. Gong, Q. Yu, X. Li, Y. Xie, and X. Zhou, "DLAU: A scalable deep learning accelerator unit on FPGA," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 36, no. 3, pp. 513–517, Mar. 2017.
9. D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 96–108, Nov. 2017.
10. L. Lu and Adam P. Harrison, "Deep medical image computing in preventive and precision medicine," *IEEE MultiMedia*, vol. 25, no. 3, pp. 109–113, Jul.–Sep. 2018.

Shu-Ching Chen is an Eminent Scholar Chaired Professor with Florida International University. Contact him at chens@cs.fiu.edu.