2015 IEEE International Conference on Multimedia Big Data

Multimedia Analysis with Deep Learning

Qingbo Wu[1] Hua Zhang[2] Si Liu[1] Xiaochun Cao[1]

Email: wuqingbo@iie.ac.cn *Email: huazhang@tju.edu.cn* *Email: liusi@iie.ac.cn* *Email: caoxiaochun@iie.ac.cn*

[1] *State Key Laboratory Of Information Security, Chinese Academy of Science*
[2] *School of Computer Science and Technology, Tianjin University*

***Abstract*—Recently, deep learning method has been attracting more and more researchers due to its great success in various computer vision tasks. Particularly, some researchers focus on the study of multimedia analysis by deep learning method, and the research tasks mainly include the following six aspects: classification, retrieval, segmentation, tracking, detection and recommendation. As far as we know, there is not any literature conducting on survey of these studies, and it is of great significance for the community to review this subject. In this paper, we discuss the application of deep learning method in the six multimedia analysis tasks, and also point out the future directions of deep learning in multimedia analysis.**

***Keywords*-deep learning; multimedia analysis; classification; retrieval; segmentation; tracking; detection; recommendation**

I. Introduction

The method of deep learning achieves astonishing success[1][2][3] in the field of machine learning in recent years, and gains great attention by the community of industry and academy.

Krizhevsky et al.[1] train a large deep convolutional neural network to classify the 1.2 million high-resolution images into the 1000 different classes in the ImageNet LSVRC-2010 contest. On the test data, they achieve top-1 and top-5 error rates of 37.5% and 17.0% which are considerably better than the previous state-of-the-art. K. Jarrett et al.[2] show that non-linearities (including rectification and local contrast normalization) is the most important ingredient for good accuracy on object recognition benchmarks. They also show that two stages of feature extraction yield better accuracy than one stage. And Y. Bengio[3] focuses on the role of unsupervised pre-training of representations, and the usage in the transfer learning scenario, where he cares about predictions on examples that are not from the same distribution as the training distribution.

Particularly, some researchers study the deep learning method in several tasks related to multimedia. According to our statistic, these tasks mainly consist of classification, recommendation, tracking, retrieval, detection and segmentation, which will be discussed in detail later. The performance of deep learning in these tasks consistently outperform those of the traditional methods. However, as far as we know, there isn't a review about these new works, and it is useful to make further researches on them.

The paper is organized as follows. Section II discuss the details of the deep learning method in six tasks of multimedia, and Section III is our conclusion and the future directions of deep learning in multimedia analysis.

II. Six Tasks in Multimedia with Deep Learning

A. Image/Video Classification

Classification, or namely categorization, aims to recognize whether the image/video contains the known objects, which is a hot-topic in the multimedia community. Considering the importance of this topic in multimedia analysis, a great number of researchers have put attentions on this problem.

As a novel manner of extracting the features and training the models of classification, deep learning method has been widely used in multimedia community for content analysis. Z. Peng et al. [4] propose a novel framework named deep boosting on the image classification task by extracting the discriminative features based on the multi-layers DBN. Similarly, S. Zhong et al. [5] propose to train the image classification model based on the bilinear deep belief network (BDBN). Different from [4], BDBN provides human-like judgment by referencing the architecture of the human visual system and the procedure of intelligent perception. Further, for video classification, Z.Wu et al. [6] focus on the challenging task of classifying videos based on the content of videos. Via rigorously imposing regulations in a deep neural network(DNN), inter-feature and inter-class relationships can be learned and utilized.

Beyond classifying the images or videos based on their contents, there are also existing a lot work on the specific task, e.g. emotion classification, action/activity classification and face classification. On the task of emotion classification, W. Zheng et al. [7] develop a model based on the deep belief network (DBN). This method firstly extracts the differential entropy features from multichannel electroencephalogram(EEG). On the task of activity classification in the video, Z. Liang et al. [8] develop an expressive configurable human activity model by using the deep learning technology and reconfigurable part-based models. This model regards

978-1-4799-8688-0/15 $31.00 © 2015 IEEE
DOI 10.1109/BigMM.2015.27

each human activity as an ensemble of cubic-like video segments, and learns to discover the temporal structures for a category of activities. On the task of face classification, K. Wang et al. [9] focus on the face verification, and propose a model, which learns an explicit mapping from the original space to an optimal subspace by the deep independent subspace analysis network(DISAN). Further, the proposed method is a deep and local learning architecture compared with the similar kernel methods. Moreover, T. Xiao et al. [10] develop a training algorithm that grows a network incrementally as well as hierarchically. Classes are grouped according to similarities, and self-organized into different levels. The newly added capacities are divided into component models that predict coarse-grained super-classes and those return final prediction within a superclass.

Different from the above-mentioned classification work, there exists some classification models based on incomplete training data. Inspired by human visual cortex and intelligent perception, X.Cai et al. [11] propose a deep architectures to simulate the laminar structure of human cerebral cortex, and the information delivered between multiple layers reproduces the neural loop in humans visual areas. They put forward a novel deep learning technique called semiconducting bilinear deep belief networks (SBDBN) based on the typical deep model called deep belief networks (DBN).

In the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 and 2013, the contribution of deep learning is demonstrated by the successes of the application of deep Convolutional Neural Networks (CNN) to image classification. However, the large space of parameters limit the application scenario such as mobile phone. Y. Bian et al. [12] propose a novel method which focuses on transferring the large model to the new dataset by fine-tuning of the big structure, normalized Google distance (NGD) and WordNet lexical semantic similarity(WNLSS).

As one of the important and useful tasks on multimedia analyzing, classification has gained a promoted performance with the deep learning. However, there are still existing a lot of problem to solve. For example, a novel topic is how to use the weakly labeled data or unlabeled data to help for training the model, instead of the supervised label.

B. Image Retrieval

The image retrieval has been widely studied for years in multimedia. In this subfield, deep learning has also been discussed by many researchers, such as J. Wang et al.[13], Y. Bai et al.[14], X. Ou et al.[15], P. Wu et al.[16], Q. Ma et al.[17] and J. Wan et al.[18].

A regression based cross modal deep learning model was proposed by J. Wang et al.[13]. This model can only tackle the queries that have highly semantical or identical queries in the training set. The image features and the query features are separately fed into the regression based on cross modal deep learning model, and then the deep network outputs the relevance scores directly, providing an end-to-end scoring fashion.

For the same purpose of learning high-level image representation, a bag-of-words based deep neural network was given by Y. Bai et al.[14]. The DNN model is trained on the large scale click-through data. In this data-set, query textual keyword is corresponding to clicked images. The cosine similarity of query's bag-of-words representation and image's bag-of-words representation predicted by DNN is used to measure the relevance between query and image.

Different from the two models above, an Inductive Transfer Deep Hashing (ITDH) is developed by X. Ou et al.[15], who concentrate on the semantic hashing for image retrieval issue. According to this method, this work uses a transfer deep learning algorithm to learn the robust image representation, and the neighborhood structure preserved method to map the image into discriminative hash codes in hamming space.

Deep learning is also considered by some researchers to be put together with online learning techniques or 3D shape. The online learning techniques are explored by P. Wu et al.[16] to learn the multi-modal similarity functions from the streams of triplet constraints. This work intends to learn the similarity function in the learning phase, so that the image ranking task in the retrieval phase will be facilitated. During the learning phase, they assume that triplet training data instances arrive sequentially. A fresh framework building a bridge between 3D shape and deep learning was put forward by Q. Ma et al.[17]. Their main work is about extracting a middle-level position-independent feature from any low-level 3D descriptors, and generating high-level features for 3D shape retrieval via deep learning. The benefit of these middle-level features lies in that they encode the geometric information as well as their spatial relationship.

Besides, J. Wan et al[18] discuss an open problem: Although deep learning is a hope for bridging the semantic gap in CBIR, how much improvements in CBIR tasks can be achieved by exploring the state-of-the-art deep learning techniques for learning feature representations and similarity measures?

There exist many content-based image retrieval systems based on the deep learning method and each system gives the impressive retrieval results. But, some disadvantages of the deep learning method limit the performance and the speed of these systems. For example, the manners of encoding images influence the retrieval speed and precision.

C. Video Segmentation

Video segmentation is a key technology in the new generation of video coding, video browsing, internet multimedia interacting, etc. In the video segmentation, deep learning method is deeply researched by K. Teng et al.[19] and P. Xu et al.[20].

An encoding method is developed by K. Teng et al.[19] with the help of highly nonlinear mapping in a deep neural network. This method add some occluded noise into the learning process for the purpose of enhancing the robustness of dealing with background noise and partial occlusions. During the training phase, they perform background subtraction and multi-object tracking for surveillance videos to extract active objects such as a person and a car. In the retrieval phase, user can select an interesting object from the query image through an interactive operation. Afterwards the selected object is encoded with the learned deep neural network.

Different from [19], a novel and practical method based on deep auto-encoder networks is proposed by P. Xu et al.[20] with a technically designed cost function. Dynamic background images are extracted through a deep auto-encoder network called Background Extraction Network, from video frames containing motion objects. They view dynamic background and foregrounds as clean data and noise data respectively and use a deep auto-encoder network to get background images.

Based on the video segmentation, we could exact the main information from the images. However, the complex video content and the dynamal background still limit the performance of video segmentation. So, an open problem in video segmentation is how to use the deep learning to suppress the influence of background and model the foreground.

D. Motion Tracking

Motion tracking, aiming at detecting and tracking moving objects through a sequence of images, has gained great success in monitoring activity in public places, and become a key element for further analysis of video image.

Hand gesture tracking is a hot topic on motion tracking. A method that uses deep learning to train a set of gestures (81 gestures) is presented by J. S. Riera et al.[21], and is used as a rough estimate of the hand pose and orientation. The method will serve as a registration of non-rigid model algorithm that will find the parameters of hand, even when temporal assumption of smooth movements of hands is violated. By pre-learning a complete set of gestures, the proposed algorithm obtains more flexibility (recover from mismatch of tracking or important drifts on pose parameters) and less tedious initialization of hand gesture parameters.

Motion tracking can help to locate the object of interest, and some promising results have been achieved based on the deep learning. But, the existing methods are mostly assuming that the objects of interest are in the same camera. So there is an important open problem of how to track the object of interest in different camera.

E. Saliency Detection

Saliency detection has been a very active research area in recent years. Most traditional methods suffering from the problem that existing visual features are not discriminative or not robust enough to predict salient locations.

To solve this problem, S. Wen et al.[22] propose a two-layer Deep Boltzmann Machine (DBM) to learn enhanced features from existing contrast-based low-level features, which are more discriminative and reliable. A saliency computation model is then trained to build a mapping from those enhanced features to eye fixation data.

Saliency detection is an important preprocessing step and has been explored for many years. But, there are also open problems: how to combine the image saliency with the eye attention is an open problem, and how to model the image saliency and the object location.

F. Music Recommendation

In many existing content-based music recommendation systems, the traditional features, originally not created for music recommendation, cannot capture all relevant information in the audio and thus put a cap on recommendation performance.

Using a novel model based on deep belief network and probabilistic graphical model, X. Wang et al.[23] unify the two stages into an automated process that simultaneously learns features from audio content and makes personalized recommendations. They develop a novel content based recommendation model based on probabilistic graphical model and the deep belief network (DBN) proposed by the deep learning community. It unifies feature learning and recommendation.

Recommendation is very useful in real scenario, which plays an important role in multimedia community. Although the results of music recommendation have achieved a better results based on the deep learning, some open problems still need to be solved. One of the problems is how to explore the topic of some related music and how to model the correlations among different music.

III. Conclusion

In this paper, a preliminary survey of the deep learning methods in multimedia analysis has been done in detail. Many new research results have been obtained in the six tasks of multimedia. However, there are three aspects to be improved in the future. Firstly, the big data produced by the real multimedia application would be discussed in the future research of deep learning method. Secondly, the structure of deep neutral network should be finely designed to improve the performance for specific task. Finally, a better learning algorithms would be studied to speed up the learning of deep neural networks.

References

[1] A. Krizhevsky, I. Sutskever and G. Hinton, *ImageNet Classification with Deep Convolutional Neural Networks*, In NIPS, 2012.

[2] K.Jarrett, K. Kavukcuoglu, M.A. Ranzato and Y.LeCun, *What is the best multi-stage architecture for object recognition?*, ICCV, 2009

[3] Y. Bengio, *Deep Learning of Representations for Unsupervised and Transfer Learning*, Journal of Machine Learning Research-Proceedings Track, 2012

[4] Z. Peng, L. Lin, R. Zhang and J. Xu, *Deep boosting: layered feature mining for general image classification*, ICME, 2014

[5] S. Zhong, Y. Liu and Y.Liu, *Bilinear Deep Learning for Image Classification*, Proc. ACM MM,2011

[6] Z. Wu, Y.-G. Jiang, J.Pu and X. Xue, *Exploring inter-feature and inter-class relationships with deep neural networks for video classification*, ACM MM, 2014

[7] W. Zheng, J. Zhu, Y. Peng and B.-L. Lu, *EEG-based emotion classification using deep belief networks*, ICME, 2014

[8] K. Wang, X. Wang, L.Lin, M. Wang and W. Zuo, *3D human activity recognition with reconfigurable convolutional neural networks*, ACM MM, 2014

[9] X. Cai, C. Wang, B. Xiao, X. Chen and J. Zhou, *Deep Nonlinear Metric Learning with Independent Subspace Analysis for Face Verification*, ACM MM, 2012

[10] T. Xiao, J. Zhang, K. Yang, Y. Peng, and Z. Zhang, *Error-driven incremental learning in deep convolutional neural network for large-scale iamge classification*, ACM MM, 2014

[11] S. Zhong, Y. Liu, F.Chung and G. Wu, *Semiconducting bilinear deep learning for incomplete image recognition*, ICMR, 2012

[12] Y. Bian, Y. Dong, H. Bai, B. Liu, K. Wang and Y. Liu, *Reducing Structure of Deep Convolutional Neural Networks for HUAWEI Accurate and Fast Mobile Video Annotation Challenge*, ICEMW, 2014

[13] J. Wang, C. Kang, Y. He, S. Xiang and C. Pan, *Cross Modal Deep Model and Gaussian Process Based Model for MSR-Bing Challenge*, ACM MM, 2014

[14] Y. Bai, W. Yu, T. Xiao, C. Xu, K. Yang, W. Ma and T. Zhao, *Bag-of-Words Based Deep Neural Network for Image Retrieval*, ACM MM, 2014

[15] X. Ou, L. Yan, H. Ling, C. Liu and M. Liu, *Inductive Transfer Deep Hashing for Image Retrieval*, ACM MM, 2014

[16] P. Wu, S. C.H. Hoi, H. Xia, P. Zhao, D. Wang and C. Miao, *Online multimodal deep similarity learning with application to image retrieval*, ACM MM, 2013

[17] S. Bu, Z. Liu, J. Han, J. Wu and R. Ji, *Learning High-Level Feature by Deep Belief Networks for 3-D Model Retrieval and Recognition*, IEEE Transactions on Multimedia, 2013

[18] J. Wan, D. Wang, S. C.H. Hio, P. Wu, J. Zhu, Y. Zhang, and J. Li, *Deep Learning for Content-Based Image Retrieval: a comprehensive study*, ACM MM, 2014

[19] K. Teng, J. Wang, M. Xu and H. Lu, *Mask Assisted Object Coding with Deep Learning for Object Retrieval in Surveillance Videos*, ACM MM, 2013

[20] P. Xu, M. Ye, X. Li, Q. Liu, Y. Yang and J. Ding, *Dynamic Background Learning through Deep Auto-encoder Networks*, ACM MM, 2014

[21] J. S. Riera, Y. Hsiao, T. Lim, K. Hua and W. Cheng, *A Robust Tracking Algorithm for 3D Hand Gesture with Rapid Hand Motion through Deep Learning*, ICMEW, 2014

[22] S. Wen, J. Han, D. Zhang and L. Guo, *Saliency Detection based on Feature Learning using Deep Boltzmann Machines*, ICME, 2014

[23] X. Wang and Y. Wang, *Improving content-based and hybrid music recommendation using deep learning*, ACM MM, 2014