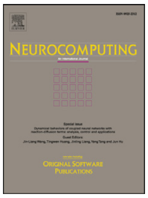
Neurocomputing 000 (2017) 1–4

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Editorial

Machine learning and signal processing for big multimedia analysis

1. Introduction

The emergence of Big Data has brought about a paradigm shift to many fields of data analytics. Multimedia is typical big data: not just big in volume, but also unstructured, noisy, redundant, and heterogeneous. Problems we did not see before are becoming critical now for big multimedia analysis, e.g., the scalability and high computational cost of sophisticated algorithms, the incompleteness and shortage of well-annotated raw data, the heterogeneity in integrating data from different sources, the difficulty in discovering valuable knowledge from noisy and redundant data, etc.

This special issue aims to demonstrate how machine learning algorithms and signal processing techniques have contributed, and are contributing to the research and applications of big multimedia analysis. Many machine learning algorithms and signal processing techniques have already been successfully applied to address the corresponding problems. For example, online learning and parallel programming have been successfully applied to improve the scalability in large-scale multimedia data analysis; semi-supervised learning and weakly-supervised learning have significantly improved the performance when limited annotated data or only weakly labeled data is available; correlation analysis methods, transfer learning and multi-task learning models have shown the potential in integrating and utilizing severely heterogeneous data; sparse machine learning methods and clustering models have been exploited in denoising and selecting exemplary samples from the raw data; deep learning has already been widely adopted in visual understanding and feature extraction. In total, we received 127 submissions from all around the world. The submissions cover a wide variety of areas including large scale face recognition, large scale visual surveillance, web-scale image retrieval/classification, massive object recognition etc. After two rounds of vigorous review by at least two expert reviewers for each paper, we finally selected 22 high-quality articles to be included in this highly-popular special issue.

2. Overview of articles

We give an overview of all the included papers summarizing the contributions and novel aspects of each paper. All the articles are divided into 3 subsections roughly according to their areas, namely, Big Learning Algorithms for Multimedia, Novel Deep Learning Models and other related topics.

2.1. Big learning algorithms for multimedia

The article entitled “Deep binary codes for large scale image retrieval” [1] proposes a novel and effective method to create compact binary codes based on deep convolutional features for image retrieval. Deep binary codes are generated by comparing the response from each feature map and the average response across all the feature maps on the deep convolutional layer. Additionally, a spatial cross-summing strategy is proposed to directly generate bit-scalable binary codes. As the deep binary codes on different deep layers can be obtained by passing the image through the CNN and each of them makes a different contribution to the search accuracy, they then present a dynamic, on-the-fly late fusion approach where the top N high quality search scores from deep binary codes are automatically determined online and fused to further enhance the retrieval precision.

The article entitled “A multi-scale kernel learning method and its application in image classification” [2] proposes kernel centered polarization to construct an optimization problem which was used to learn the multi scale kernel function and select the optimal parameters. A thorough analysis and proofs are provided. Experimental results show that the proposed kernel learning method and algorithm are reasonable and effective and have very good generalization performance.

The article entitled “Large-scale image retrieval with sparse embedded hashing” [3] proposes a novel sparsity-based hashing framework termed Sparse Embedded Hashing (SEH), exploring the technique of sparse coding. Specifically, SEH firstly generates sparse representations in a data-driven way, and then learns a projection matrix, taking sparse representing, affinity preserving and linear embedding into account. In order to make the learned compact features locality sensitive, SEH employs the matrix factorization technique to approximate the Euclidean structures of the original data. The usage of the matrix factorization enables the decomposed matrix to be constructed from either visual or textual features depending on which kind of Euclidean structure is preserved. Due to this flexibility, the SEH framework could handle both single-modal retrieval and cross-modal retrieval simultaneously.

The article entitled “Synergistic integration of graph-cut and cloud model strategies for image segmentation” [4] proposes a new graph cut image partitioning method that calculates image data using cloud model for constructing the objective functions (GC-CM). In the objective function, it contains a boundary preserving smooth term and a data item which evaluates the deviation of each pixel

http://dx.doi.org/10.1016/j.neucom.2017.01.091

that belongs to different regions. The core method models the foreground object and background of the images as cloud models by the back cloud generator. The data item is calculated with the X-condition cloud generator. They use the membership degree between each pixel to calculate the similarity of the neighbor pixel established as the smooth term. The energy minimization is completed with the minimum cut theory and the graph cut iterations.

The article entitled “SPA: spatially pooled attributes for image retrieval” [5] attempts to encode weak spatial information into attribute embedding for effective image retrieval. Specifically, they partition the image into regular grids and extract Classemes attribute vector from each patch, which results in a large pool of Classemes descriptors followed by VLAD aggregation for generating holistic representation. In order to produce a compact and discriminative code, they employ a piecewise Fisher Discriminant Analysis (FDA) for dimension reduction and concatenate all the compressed Classemes into a single vector coined Spatially Pooled Attributes.

2.2. Novel deep learning models

The article entitled “Joint entity and relation extraction based on a hybrid neural network” [6] proposes a hybrid neural network model to extract entities and their relationships without any handcrafted features. The hybrid neural network contains a novel bidirectional encoder-decoder LSTM module for entity extraction (BiLSTM-ED) and a CNN module for relation classification. The contextual information of entities obtained in BiLSTM-ED further passes through to CNN module to improve the relation classification. They conduct experiments on the public dataset ACE05 to verify the effectiveness of the method.

The article entitled “Learning 3D faces from 2D images via stacked contractive autoencoder” [7] proposes a deep learning framework for 3D face reconstruction. The framework is designed to compute subspace feature of arbitrary face image, then map the feature to its counterpart in another subspace learned with 3D faces, and reconstruct the 3D face using the counterpart feature. During the course of training, they learn 2D and 3D subspaces through Stacked Contractive Autoencoders (SCAE), use a two-layer one-layer fully connected neural network to learn the mapping, and use the pre-trained parameters of the SCAEs and the two-layer one-layer network to initialize a deep feedforward neural network whose input are face images and output are 3D faces. The network is optimized by gradient descent algorithm with back-propagation. Extensive experimental results on various data sets indicate the effectiveness of the proposed SCAE-based 3D face reconstruction method.

The article entitled “An unsupervised deep domain adaptation approach for robust speech recognition” [8] proposes an unsupervised deep domain adaptation (DDA) approach to acoustic modeling in order to eliminate the training-testing mismatch that is common in real-world use of speech recognition. Under a multi-task learning framework, the approach jointly learns two discriminative classifiers using one deep neural network (DNN). As the main task, a label predictor predicts phoneme labels and is used during training and at test time. As the second task, a domain classifier discriminates between the source and the target domains during training. The network is optimized by minimizing the loss of the label classifier and to maximize the loss of the domain classifier at the same time.

The article entitled “Hierarchical deep semantic representation for visual categorization” [9] proposes Hierarchical Deep Semantic Representation (H-DSR), a hierarchical framework which combines semantic context modeling with visual features. First, the input image is sampled with spatially fixed grids. Deep features are then extracted for each sample in particular location. Second, using pre-learned classifiers, a detection response map is constructed for each patch. Semantic representation is then extracted from the map, which has a sense of latent semantic context. They combine the semantic and visual representations for joint representation. Third, a hierarchical deep semantic representation is built with recurrent reconstructions using three layers. The concatenated visual and semantic representations are used as the inputs of subsequent layers.

The article entitled “A model for fine-grained vehicle classification based on deep learning” [10] proposes a novel model to combine two parts of vehicle detection model and vehicle fine-grained detection and classification model. Faster R-CNN method is adopted in vehicle detection model to extract single vehicle images from an image with clutter background which may contain serval vehicles. This step provides data for the next classification model. In vehicle fine-grained classification model, an image contains only one vehicle is fed into a CNN model to produce a feature, then a joint Bayesian network is used to implement the fine-grained classification process. Experiments show that vehicle’s make and model can be recognized from transportation images effectively by using this method.

The article entitled “DeepSim: deep similarity for image quality assessment” [11] proposes a novel deep learning model to measure the local similarities between the features of the test image and those of the reference image; afterward, the local quality indices are gradually pooled together to estimate the overall quality score. In addition, various factors that may affect the IQA performance are investigated. Thorough experiments conducted on standard databases show that: (1) DeepSim can accurately predict human perceived image quality and outperforms previous state-of-the-art; (2) mid-level representations are most effective for quality prediction; and (3) preprocessing, the restricted linear units and max-pooling operations are beneficial for the IQA performance.

The article entitled “Online object tracking based on CNN with spatial-temporal saliency guided sampling” [12] incorporates spatial-temporal saliency detection to guide a more accurate target localization for qualified sampling within an inter-frame motion flow map. With an optional strategy for the output combination of intra-frame appearance correlations and inter-frame motion saliency based on a compositional energy optimization, the proposed tracking has shown a superior performance in comparison to the other state-of-the-art trackers on both challenging non-rigid and generic tracking benchmark datasets.

The article entitled “Ultrasonic signal classification and imaging system for composite materials via deep convolutional neural networks” [13] proposes a deep learning based framework to classify ultrasonic signals from carbon fiber reinforced polymer (CFRP) specimens with void and delamination. In our proposed algorithm, deep Convolutional Neural Networks (CNNs) are used to learn a compact and effective representation for each signal from wavelet coefficients. To yield superior results, they propose to use a linear SVM top layer in the training process of signal classification task. The experimental results demonstrated the excellent performance of the proposed algorithm against the classical classifier with manually generated attributes. In addition, a post processing scheme is developed to interpret the classifier outputs with a C-scan imaging process and visualize the locations of defects using a 3D model representation.

2.3. Other related topics

The article entitled “Combining paper cooperative network and topic model for expert topic analysis and extraction” [14] proposes a novel method for expert topic analysis and extraction by combining paper cooperation network and topic model. In the method, they extract each paper’s author information and construct an expert cooperation network. At the same time, by means of LDA