arXiv:1512.03131v1 [cs.CV] 10 Dec 2015

Deep Learning Algorithms with Applications to Video Analytics for A Smart City: A Survey

Li Wang, *Member, IEEE,* and Dennis Sng

***Abstract*—Deep learning has recently achieved very promising results in a wide range of areas such as computer vision, speech recognition and natural language processing. It aims to learn hierarchical representations of data by using deep architecture models. In a smart city, a lot of data (e.g. videos captured from many distributed sensors) need to be automatically processed and analyzed. In this paper, we review the deep learning algorithms applied to video analytics of smart city in terms of different research topics: object detection, object tracking, face recognition, image classification and scene labeling.**

***Index Terms*—Deep learning, smart city, video analytics.**

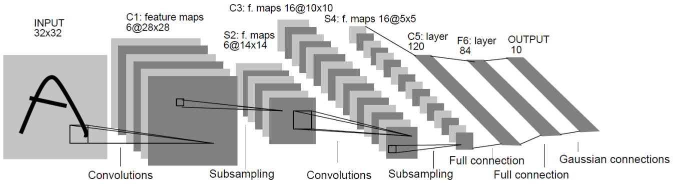

Fig. 1. Illustration of the CNN proposed by LeCun et al. [9] for digits recognition.

I. Introduction

A smart city aims to improve quality and performance of urban services by using digital technologies or information and communication technologies. Data analytics plays an important role in smart cities. Many sensors are installed in a smart city to capture a huge volume of data such as surveillance videos, environment and transportation data. To capture useful information from such big data, machine learning algorithms are often used and have achieved very promising results in a wide range of applications, e.g. video analytics. Therefore, leveraging on machine learning can facilitate smart city development.

Machine learning aims to develop the computer algorithms which can learn experience from example inputs and make data-driven predictions on unknown test data. Such algorithms can be divided into two categories: supervised learning (e.g. [1] [2]) and unsupervised learning (e.g. [3] [4]). Given labeled input and output pairs, supervised learning aims to find a mapping rule for predicting outputs of unknown inputs. In contrast, unsupervised learning focuses on exploring intrinsic characteristics of inputs. Supervised learning and unsupervised learning are complementary to each other. Since supervised learning leverages labels of inputs which are meaningful to human, it is easy to apply this kind of learning algorithms to pattern classification and data regression problems. However, supervised learning relies on labeled data which could cost a lot of manual works. Moreover, there are uncertainties and ambiguities in labels. In other words, the label for an object is not unique. To mitigate these problems, unsupervised learning can be used to handle intra-class variation as it does not require labels of data. In the past decades, machine learning methods have been applied to a wide range of applications such as bioinformatics, computer vision, medical diagnosis, natural language processing, robotics, sentiment analysis, speech recognition and stock market analysis, etc. In this paper, we focus on reviewing the recent achievements of deep learning which is a subfield of machine learning.

In contrast with shallow learning algorithms, deep learning aims to extract hierarchical representations from large-scale data (e.g. images and videos) by using deep architecture models with multiple layers of non-linear transformations. With such learned feature representations, it becomes easier to achieve better performance than using raw pixel values or hand-crafted features. The principle behind this success is that deep learning is able to disentangle different levels of abstractions embedded in observed data by elaborately designing the layer depth and width, and properly selecting features that are beneficial for learning tasks.

In fact, the history of deep learning starts at least from 1980, when Neocognitron [5] is proposed by Fukushima. In 1989, LeCun et al. [6] propose to apply backpropagation onto a deep neural network for handwritten ZIP code recognition. However, the training time on the network was too long for practical use. Also, deep neural networks have been studied in speech recognition for many years, but can hardly surpass the shallow generative models. It is due to the fact that deep learning architectures require large training data which was scarce in those early days. Hinton et al. [7] review these difficulties and claim their confidence of solving these issues for applying deep learning to speech recognition, since Hinton [8] has achieved breakthroughs on training multi-layer neural networks by pre-training one layer at a time as an unsupervised restricted Boltzmann machine and then using supervised backpropagation for fine-tuning. Since the breakthrough of deep learning, it has been applied to many other research areas besides speech recognition.

Deep learning architectures have different variants such as Deep Belief Networks (DBN) [10], Convolutional Neural Networks (CNN) [11], Deep Boltzmann Machines (DBM) [12] and Stacked Denoising Auto-Encoders (SDAE) [13], etc. The most attractive model is Convolutional Neural Networks which

L. Wang and D. Sng are with the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University, Singapore 637553 (e-mail: wa0002li@e.ntu.edu.sg; dennis.sng@ntu.edu.sg).

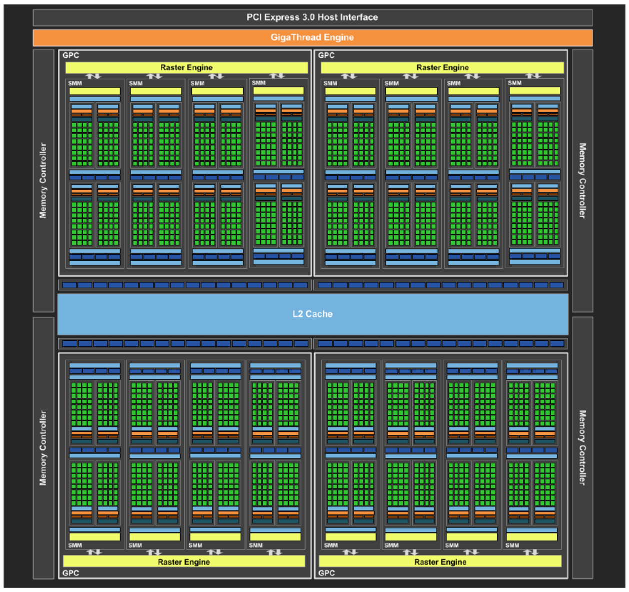

Fig. 2. Illustration of the parallel structure of the Nvidia GeForce GTX 980.

have achieved very promising results in both computer vision and speech recognition. An illustration of the CNN proposed by LeCun et al. [9] for digits recognition is presented in Figure 1. As shown in the figure, a CNN usually consists of convolutional layers, pooling layers and fully connected layers. With loss layers on the top of the CNN, the whole network can be trained end-to-end by using the backpropagation algorithm. Compared to other deep feed-forward neural networks, a CNN is easier to train as it has fewer parameters to estimate. As a result, CNN has a wide range of applications such as image classification [14], face recognition [15], object tracking [16], pedestrian detection [17], attribute prediction [18], scene labeling [19], person re-identification [20], RGB-D object recognition [21], image labeling [22], scene image classification [23], speech recognition [24] and natural language processing [25], etc.

Deep learning has received much attention from not only academic but also industry. For example, Geoff Hinton and Li Deng started their collaborations from 2009 in the focus of applying deep learning to large-scale speech recognition, in which the performance is significantly improved against the traditional generative models by using big training data and the correspondingly designed deep neural networks. Another exciting example is that Andrew Ng and Jeff Dean from the Google Brain team successfully extract object-level semantics (e.g. cats) from unlabeled YouTube videos by using a neural network with the self-taught capacity. In future, deep learning will have more and more real applications in industry.

Besides methodology breakthroughs and available big training data, the recent success for deep learning is also due to advances in hardware. Specifically, an electronic circuit called Graphics Processor Unit (GPU) is designed for accelerating the algorithms that need to process a large number of blocks of data, and is characteristic of its highly parallel structure.

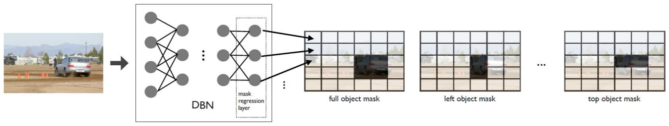

Fig. 3. Illustration of object detection as DNN-based regression proposed by Szegedy et al. [28].

For example, Nvidia's latest GPU, the GTX 980, is based on their 10th generation GPU architecture, called Maxwell, which delivers double the performance per watt compared to the previous generation. The GM 204 chip is composed of an array of 4 Graphics Processing Clusters (GPCs), 16 Streaming Multiprocessors (SMs), and 4 memory controllers. The GeForce GTX 980 uses the full complement of these architectural components. Its parallel structure is illustrated in Figure 2. It is necessary to mention that the Nvidia company has made a lot of contributions for popularizing GPUs, e.g. Nvidia marketed the GeForce 256 as the world's first GPU. Recently, GPUs are very popular in machine learning. For a general instance, Raina et al. [26] propose to accelerate the sparse coding algorithm [27] by using GPUs and finally speed up the previous method using a dual-core CPU up to 15 times. In particular, Raina et al. [26] also report that the GPU speedup on learning Deep Belief Networks (DBNs) achieves up to 70 times against a dual-core CPU implementation. For learning a four-layer DBN with 100 million parameters, using GPU rather than CPU is able to reduce the required time from several weeks to around a single day.

Smart city is so-called "smart" as it has the capability of computing and analyzing urban data collected from e.g. monitoring systems, government agencies, commercial companies and social networking websites. Since deep learning is suitable for handling large-scale data, it can be used to process and analyze millions of video data captured from the distributed sensors in a smart city. Regarding such data, there are many active research topics such as object detection, object tracking, face recognition, image classification and scene labeling. In the following sections, we review the state-of-the-art deep learning algorithms in these application areas.

II. Object Detection

Object detection aims to precisely locate interested objects in video frames. Many interesting works have been proposed for object detection by using deep learning algorithms. We review some representative works as follows.

Szegedy et al. [28] modify deep convolutional neural networks [11] by replacing the last layer with a regression layer to produce a binary mask of the object bounding box as illustrated in Figure 3. Moreover, a multi-scale strategy (see Figure 4) is proposed for the DNN mask generation to refine detection precision. As a result, the average precision 0.305 over 20 classes on Pascal Visual Object Challenge (VOC) 2007 can be achieved by applying the proposed network a few dozen times per input image.

Different from the Szegedy's method [28], Girshick et al. [29] propose a bottom-up region proposal based deep model

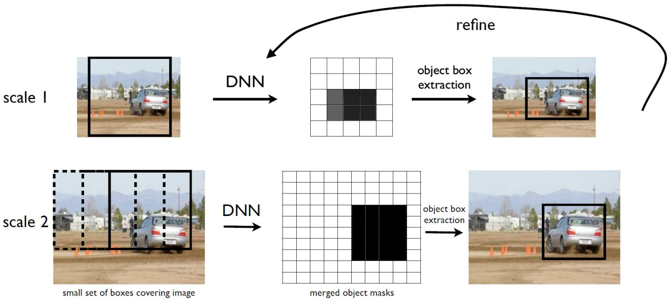

Fig. 4. Illustration of multi-scale strategy for refining detection precision proposed by Szegedy et al. [28].

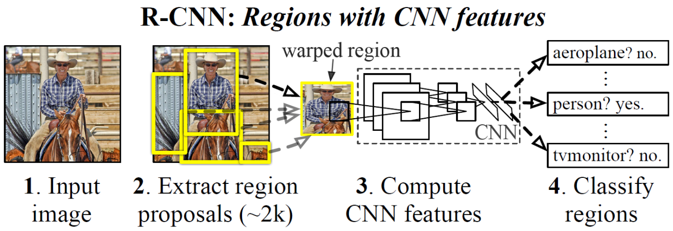

Fig. 5. Illustration of R-CNN: Regions with CNN features proposed by Girshick et al. [29].

for object detection. Figure 5 presents an overview of the method. First, around 2000 region proposals are generated within the input image. For each proposal, a large convolutional neural network is used to extract features. Finally, each region is classified by using class-specific linear SVMs. It is reported that using this method can significantly improve detection performance on VOC 2012 by more than 30% in terms of mean average precision (mAP).

Similarly, Erhan et al. [30] propose a saliency-inspired deep neural networks to detect any object of interest. However, a small number of bounding boxes are generated as object candidates by using a deep neural network in a class agnostic manner. In this work, object detection is defined as a regression problem to the coordinates of bounding boxes. For the training part, an assignment problem between predictions and groundtruth boxes is solved to update box coordinates, their confidences and the learned features by using backpropagation. In summary, a deep neural network is customized towards object localization problem.

Object detection has a wide range of smart city applications, such as pedestrian detection, on-road vehicle detection, unattended object detection. Deep learning algorithms are able to handle large variations of different objects. As a result, the smart city systems using deep models are more robust to large-scale real data.

III. Object Tracking

Object tracking is intended to locate a target object in a video sequence given its location in the first frame. Recently, some deep learning based tracking algorithms have achieved very promising results. We review some representative works as follows.

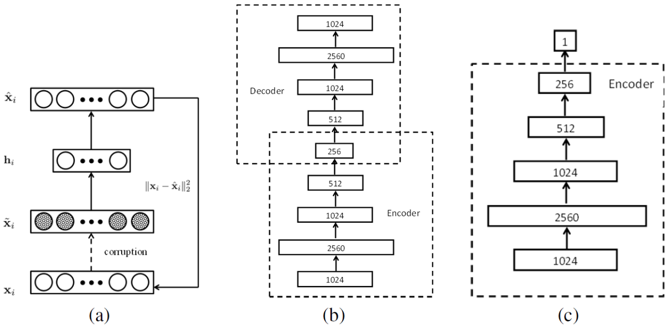

Fig. 6. Illustration of the network architecture proposed by Wang and Yeung [29]: (a) denoising autoencoder; (b) stacked denoising autoencoder; (c) network for online tracking.

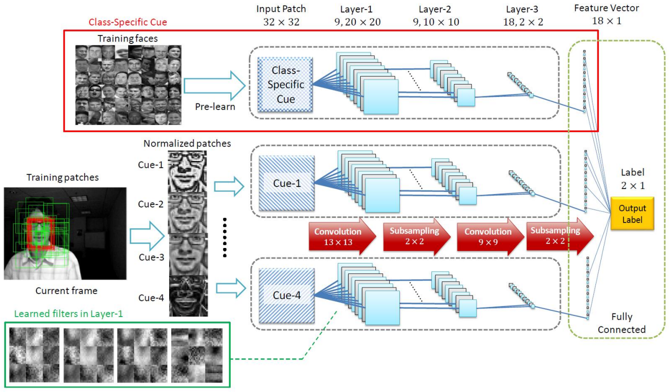

Fig. 7. Overview of the network architecture proposed by Li et al. [16].

Wang and Yeung [31] propose to learn deep and compact features for visual tracking by using stacked denoising autoencoder [32]. The network architecture is illustrated in Figure 6. It is reported that using the learned deep features can outperform other 7 state-of-the-art trackers on 10 sequences in terms of two tracking measurements: average center error (7.3 pixels) and average success rate (85.5%). Additionally, the proposed tracker with deep features achieved an average frame rate of 15fps which is suitable for real-time applications.

Li et al. [16] propose to learn discriminative feature representations for visual tracking by using convolutional neural networks. An overview of the method is illustrated in Figure 7. It can be observed that a pool of multiple convolutional neural networks are utilized to maintain different kernels regarding all possible low-level cues, which aim to discriminate object patches from their surrounding background. Given a frame, the most prospective convolutional neural network in the pool is used to predict the new location of the target object. Meanwhile, the selected network is retrained using a warm-start backpropagation scheme. Also, we can observe from Figure 7 that a class-specific convolutional neural network is involved in the whole architecture, which is helpful for tracking a certain class of objects (e.g. a person's face).

Wang et al. [33] propose to learn hierarchical features robust to both complicated motion transformations and target appearance changes. The stacked architecture and the adap-

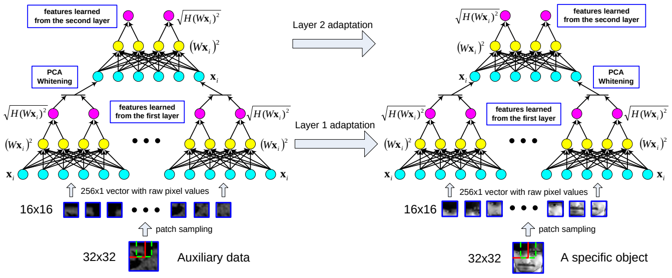

Fig. 8. Illustration of the stacked architecture and the adaptation module proposed by Wang et al. [33].

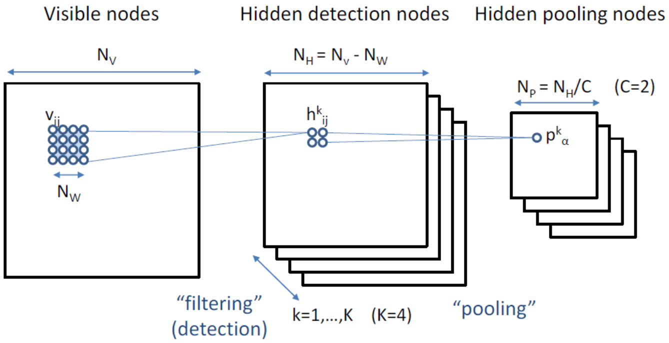

Fig. 9. Illustration of the convolutional restricted Boltzmann machine used in Huang et al. [35].

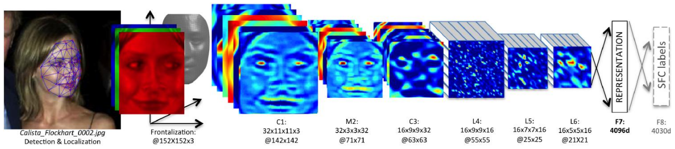

Fig. 10. Overview of the nine-layer deep neural network used in Taigman et al. [15].

tation module are illustrated in Figure 8. First, the generic features robust to complicated motion transformations are offline learned from auxiliary video data by using a two-layer neural networks under the temporal constraints [34]. Then, the pre-learned features are online adapted according to the specific target object sequence. As a result, the adapted features are able to capture appearance changes of target objects. For example, the proposed tracker can handle not only non-rigid deformation of a basketball player's body but also the specific appearance changes. It is reported that using the feature learning algorithm can significantly improve tracking performance, especially on the sequences with complex motion transformations.

Object tracking can be applied to surveillance systems of smart cities. It is important to automatically track suspected people or target vehicles for safety monitoring and urban flow management. Deep learning can leverage smart city big data to train deep models which are more robust to visual variations of target objects than traditional models. Therefore, smart city tracking systems can be enhanced by using deep learning algorithms for handling large amount of video data.

IV. Face Recognition

Face recognition consists of two main tasks: face verification and face identification. The former aims to determine whether the given two faces belong to the same person. The latter is intended to find the identities of the given faces from the known face set. Recently, many deep learning based algorithms have achieved very promising results in these two face recognition tasks. We review some of them as follows.

Huang et al. [35] propose to learn hierarchical features for face verification by using convolutional deep belief networks. The main contributions of this work are as follows: i) a local convolutional restricted Boltzmann machine is developed to adapt to the global structure in an object class (e.g. face); ii) deep learning is applied to local binary pattern representation [36] rather than raw pixel values to capture more complex characteristics of hand-crafted features; iii) learning the network architecture parameters is evaluated to be necessary for enhancing the multi-layer networks. The convolutional restricted Boltzmann machine used in the proposed method is illustrated in Figure 9. It is reported that using the learned representations can achieve comparable performance with state-of-the-art methods using hand-crafted features. Actually, the subsequent works have shown that deep features outperform hand-crafted features significantly.

Taigman et al. [15] propose a 3D face model based face alignment algorithm and a face representation learned from a nine-layer deep neural network. An overview of the architecture of the deep network is illustrated in Figure 10. The first three convolutional layers are used to extract low-level features (e.g. edges and textures). The next three layers are locally connected to learn a different set of filters for each location of a face image since different regions have different local statistics. The top two layers are fully connected to capture correlations between features captured in different parts of a face image. At last, the output of the last layer is fed to a K-way softmax which predicts class labels. The objective of training is to maximize the probability of correct class by minimizing the cross-entropy loss for each training sample. It is shown that using the learned representations can achieve the near-human performance on the Labeled Faces in the Wild benchmark (LFW).

Sun et al. [37] propose to learn so-called Deep hidden IDentity features (DeepID) for face verification. Figure 11 illustrates the feature extraction process. First, the local low-level features of an input face patch are extracted and fed into a ConvNet [9]. Then, the feature dimension gradually decreases to 160 through several feed-forward layers, during which more global and high-level features are learned. Last, the identity class (among $10,000$ classes) of the face patch is predicted directly by using the 160-dimensional DeepID. Rather than training a binary classifier for each face class, Sun et al. simultaneously classify all ConvNets regarding $10,000$ face identities. The advantages of this manipulation are as follows: i) effective features are extracted for face recognition by using the super learning capacity of neural networks; ii) the hidden features among all identities are shared by adding

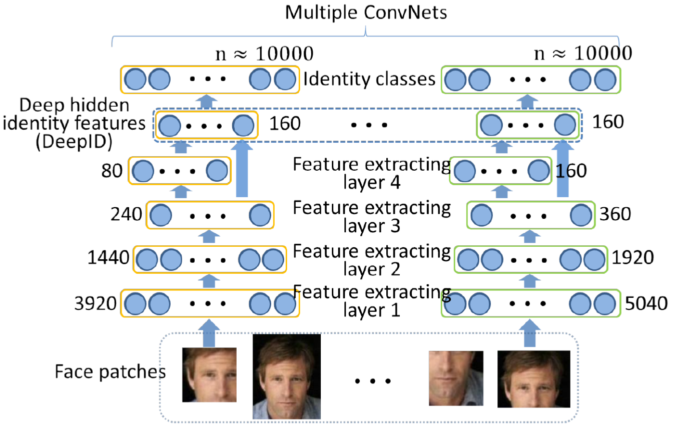
Multiple ConvNets
n ≈ 10000
n ≈ 10000
Identity classes
Deep hidden identity features (DeepID)
160
160
80
Feature extracting layer 4
160
240
Feature extracting layer 3
360
1440
Feature extracting layer 2
1920
3920
Feature extracting layer 1
5040
Face patches

Fig. 11. Illustration of the feature extraction process proposed in Sun et al. [37].

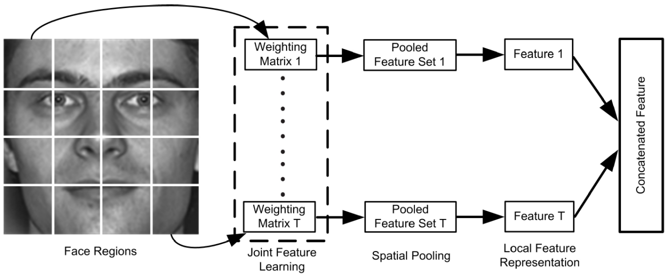

Fig. 12. Illustration of the basic idea of the approach [38].

a strong regularization to ConvNets. It is reported that using the learned DeepID can achieve the near-human performance on the LFW dataset although only weakly aligned faces are used.

Lu et al. [38] develop a joint feature learning approach to automatically learn hierarchical representation from raw pixels for face recognition. Figure 12 illustrates the basic idea of the proposed method. First, each face image is divided into several non-overlapping regions and feature weighting matrices are jointly learned. Then, the learned features in each region are pooled and represented as local histogram feature descriptors. Lastly, these local features are combined and concatenated into a longer feature vector for face representation. Moreover, the joint learning model is stacked into a deep architecture exploiting hierarchical information. As a result, the effectiveness of the proposed approach is demonstrated on five widely used face datasets.

Face recognition has been widely used in security systems and human-machine interaction systems. It is still a challenge for computer to automatically identify or verify a person due to large variations, e.g. illumination, pose and expression. Deep learning can utilize big data for training deep architecture models so as to obtain more powerful features for representing faces. In future, face recognition systems in smart cities will largely rely on hierarchical features learned from deep models.

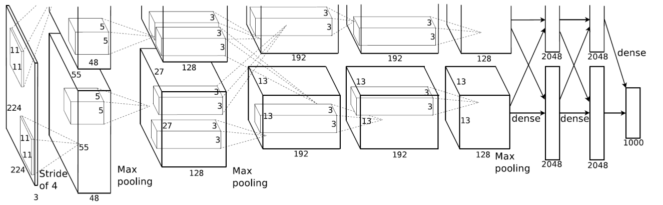

Fig. 13. Illustration of the architecture of the deep CNN proposed by Krizhevsky et al. [11]. It consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully connected layers with a 1000-way softmax.

V. Image Classification

Image classification has been an active research topic in the past few decades. Many methods have been proposed to achieve very promising results by using Bag-of-Words (BoW) representation [39], spatial pyramid matching [40], topic models [41], part-based models [42] and sparse coding [43], etc. However, these methods utilize raw pixel values or hand-crafted features which cannot capture data-driven representations for specific input data. Recently, deep learning has achieved very promising results in image classification. We review some representative works as follows.

Krizhevsky et al. [11] develop a deep convolutional neural network which has 60 million parameters and $650,000$ neurons. The deep model includes five convolutional layers followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax. The architecture of the deep convolutional neural network proposed by Krizhevsky et al. [11] is illustrated in Figure 13. The deep model achieved very impressive results on the ImageNet LSVRC-2010 contest by reducing the error rates down to 8% against the then-state-of-the-art. The dataset includes 1.2 million high-resolution images in 1000 different classes. This competition has become one of the largest and most challenging computer vision challenges and is held annually. The challenge has attracted not only academic research groups but also industry companies. For example, Google has won the classification challenge in the ImageNet 2014 with a error rate 6.66%. Nowadays, high performance computing plays a very important role in deep learning. Recently, the Chinese search engine company Baidu has obtained a 5.98% error rate on the ImageNet classification dataset by using a supercomputer called Minwa, which consists of 36 server nodes, each with 4 Nvidia Tesla K40m GPUs. Baidu also claims that Minwa can handle higher-resolution images and the larger training dataset (2 billion images) which is generated by distorting, flipping and changing colors of the original 1.2 million images. As a result, the system is able to work on real-world photos.

Lu et al. [44] present a multi-manifold deep metric learning (MMDML) method for image set classification. First, each image set is modeled as a manifold which is passed into multiple layers of deep neural networks and mapped into another feature space. Specifically, the deep network is class-specific so that different classes have different parameters in their networks. Then, the maximal manifold margin criterion

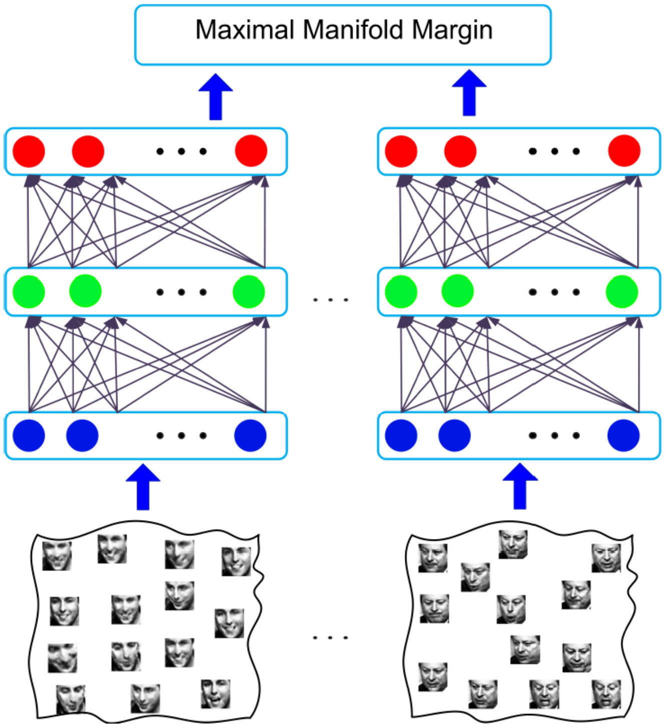

Fig. 14. Illustration of the basic idea of the method [44].

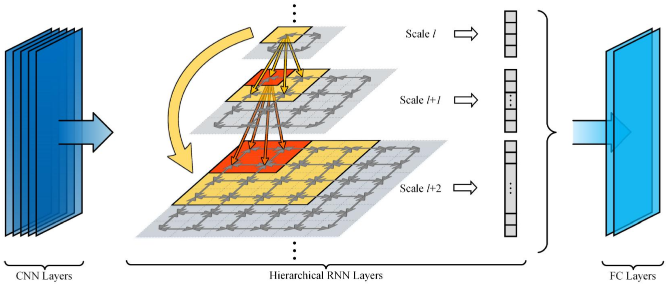

Fig. 15. Illustration of the overall framework of C-HRNNs [14].

is used to learn the parameters of these manifold. In the testing stage, these class-specific deep networks are applied to compute the similarity between the testing image set and all training classes. Finally, the smallest distance is used for classification. As a result, the proposed method can achieve the state-of-the-art performance on five widely used datasets by exploiting both discriminative and class-specific information.

Zuo et al. [14] develop an end-to-end Convolutional Hierarchical Recurrent Neural Networks (C-HRNNs) to explore contextual dependencies for image classification. Figure 15 illustrates the overall framework of C-HRNNs. First, mid-level representations for image regions are extracted by using five layers of Convolutional Neural Networks (CNNs). Then, the output of the fifth CNN layer is pooled into multiple scales. For each scale, spatial dependencies are captured by direct or indirect connections between each region and its surrounding neighbors. For different scales, scale dependencies are encoded by transferring information from the higher level scales to corresponding areas at the lower level scales. Finally, different scale HRNN outputs are collected and put into two fully connected layers. C-HRNNs not only make use of the representation power of CNNs, but also efficiently encodes spatial and scale dependencies among different image regions. As a result, the proposed model achieves state-of-the-art performance on four challenging image classification benchmarks.

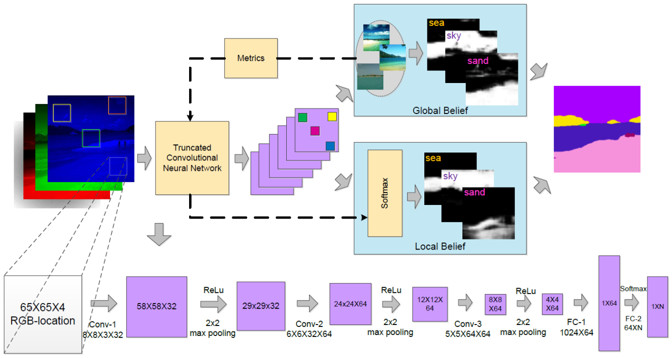

Fig. 16. Illustration of the framework of the approach [19].

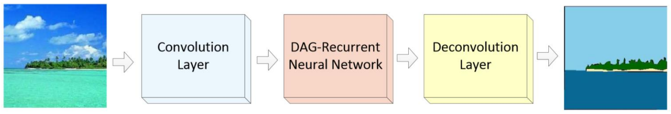

Fig. 17. Illustration of the architecture of the full labeling network in the approach [45].

VI. Scene Labeling

Scene labeling aims to assign one of a set of semantic labels to each pixel of a scene image. It is very challenging due to the fact that some classes may be indistinguishable in a close-up view. Generally, “thing” pixels (cars, person, etc) in real world images can be quite different due to their scale, illumination and pose variation. Recently, deep learning based methods have achieved very promising results for scene labeling. We review some of them as follows.

Shuai et al. [19] propose to adopt Convolutional Neural Networks (CNNs) as a parametric model to learn discriminative features and classifier for scene labeling. Figure 16 illustrates the framework of the proposed method. First, global scene semantics are used to remove the ambiguity of local context by transferring class dependencies and priors from similar exemplars. Then, the global potential is decoupled to the aggregation of global beliefs over pixels. The labeling result can be obtained by integrating the local and global beliefs. Finally, a large margin based metric learning is introduced to make the estimation of global belief more accurate. As a result, the proposed model is able to achieve state-of-the-art results on the SiftFlow benchmark and very competitive results on the Stanford Background dataset.

Shuai et al. [45] present a directed acyclic graph RNNs (DAG-RNNs) for scene labeling to model long-range semantic dependencies among image units. Figure 17 illustrates the

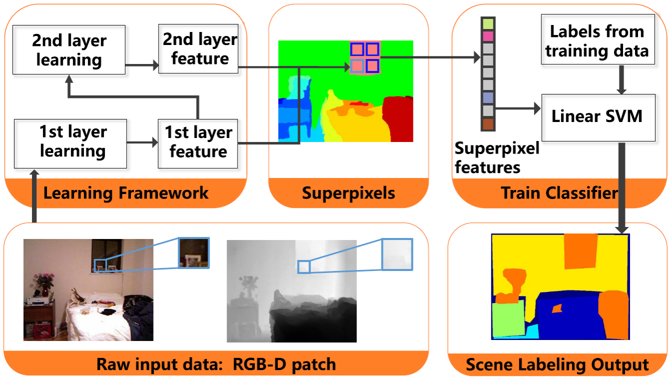

Fig. 18. Illustration of the framework for RGB-D indoor scene labeling in the approach [46].

architecture of the full labeling network in the proposed method. First, undirected cyclic graphs (UCG) are adopted to model interactions among image units. Due to the loopy property of UCGs, RNNs are not directly applicable to UCG-structured images. Thus, the UCG is decomposed to several directed acyclic graphs (DAGs). Then, each hidden layer is generated independently through applying DAG-RNNs to the corresponding DAG-structured image, and they are integrated to produce the context-aware feature maps. As a result, the local representations are able to embed the abstract gist of the image, so their discriminative power are enhanced remarkably. It is reported that the DAG-RNNs achieve new state-of-the-art results on the challenging SiftFlow, CamVid and Barcelona benchmarks.

Wang et al. [46] develop an unsupervised joint feature learning and encoding (JFLE) framework for RGB-D scene labeling. Figure 18 illustrates the framework for RGB-D indoor scene labeling in the proposed method. First, feature learning and encoding are jointly performed in a two-layer stacked structure, called joint feature learning and encoding framework (JFLE). To further extend the JFLE framework to a more general framework called joint deep feature learning and encoding (JDFLE), a deep model is used with stacked nonlinear layers to model the input data. The input to the learning structure (either JFLE or JDFLE) is a set of patches densely sampled from RGB-D images, and the learning output is the set of corresponding path features, which are then combined to generate superpixel features. Finally, linear SVMs are trained to map superpixel features to scene labels. It is reported that the proposed feature learning framework can achieve competitive performance on the benchmark NYU depth dataset.

Scene labeling can be used to understand urban images captured from surveillance cameras. It is a very important component of smart city, in which city elements such as “road” and “building” are required to be recognized. Deep models can leverage on big data from smart city to learn hierarchical features for labeling each pixel of scene images.

VII. Conclusions

In this paper, we have reviewed the recent progress of deep learning in object detection, object tracking, face recognition, image classification and scene labeling. The deep models have significantly improved the performance in these areas, often approaching human capabilities. The reasons for this success are two-folded. First, big training data are becoming increasingly available (e.g. data streams from a multitude of smart city sensors) for building up large deep neural networks. Second, new advanced hardware (e.g. GPU) has largely reduced the training time for deep networks. We believe that deep learning will have a more prospective future in a wide range of applications.

Acknowledgment

This research was carried out at the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University, Singapore. The ROSE Lab is supported by the National Research Foundation, Prime Ministers Office, Singapore, under its IDM Futures Funding Initiative and administered by the Interactive and Digital Media Programme Office.

References

[1] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[2] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *Cognitive modeling*, vol. 5, 1988.

[3] T. Kohonen, “Self-organized formation of topologically correct feature maps,” *Biological cybernetics*, vol. 43, no. 1, pp. 59–69, 1982.

[4] H. Hotelling, “Analysis of a complex of statistical variables into principal components.” *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933.

[5] K. Fukushima, “Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position,” *Biological cybernetics*, vol. 36, no. 4, pp. 193–202, 1980.

[6] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel, “Backpropagation applied to handwritten zip code recognition,” *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.

[7] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[8] G. E. Hinton, “Learning multiple layers of representation,” *Trends in cognitive sciences*, vol. 11, no. 10, pp. 428–434, 2007.

[9] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[10] G. E. Hinton, “Deep belief networks,” *Scholarpedia*, vol. 4, no. 5, p. 5947, 2009.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Annual Conference on Neural Information Processing Systems*, 2012, pp. 1106–1114.

[12] R. Salakhutdinov and G. E. Hinton, “Deep boltzmann machines,” in *International Conference on Artificial Intelligence and Statistics*, 2009, pp. 448–455.

[13] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, “Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion,” *The Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, 2010.

[14] Z. Zuo, B. Shuai, G. Wang, X. Liu, X. Wang, and B. Wang, “Learning contextual dependencies with convolutional hierarchical recurrent neural networks,” *CoRR*, vol. abs/1509.03877, 2015.

[15] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “Deepface: Closing the gap to human-level performance in face verification,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.

[16] H. Li, Y. Li, and F. Porikli, "Deeptrack: Learning discriminative feature representations by convolutional neural networks for visual tracking," in *British Machine Vision Conference*, 2014.

[17] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun, "Pedestrian detection with unsupervised multi-stage feature learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3626–3633.

[18] A. H. Abdulnabi, G. Wang, J. Lu, and K. Jia, "Multi-task CNN model for attribute prediction," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1949–1959, 2015.

[19] B. Shuai, G. Wang, Z. Zuo, B. Wang, and L. Zhao, "Integrating parametric and non-parametric models for scene labeling," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4249–4258.

[20] R. R. Varior, G. Wang, and J. Lu, "Learning invariant color features for person re-identification," *CoRR*, vol. abs/1410.1035, 2014.

[21] A. Wang, J. Lu, J. Cai, T. Cham, and G. Wang, "Large-margin multimodal deep learning for RGB-D object recognition," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1887–1898, 2015.

[22] B. Shuai, Z. Zuo, and G. Wang, "Quaddirectional 2d-recurrent neural networks for image labeling," *IEEE Signal Process. Lett.*, vol. 22, no. 11, pp. 1990–1994, 2015.

[23] Z. Zuo, G. Wang, B. Shuai, L. Zhao, and Q. Yang, "Exemplar based deep discriminative and shareable feature learning for scene image classification," *Pattern Recognition*, vol. 48, no. 10, pp. 3004–3015, 2015.

[24] L. Deng, O. Abdel-Hamid, and D. Yu, "A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 6669–6673.

[25] R. Collobert and J. Weston, "A unified architecture for natural language processing: deep neural networks with multitask learning," in *International Conference on Machine Learning*, 2008, pp. 160–167.

[26] R. Raina, A. Madhavan, and A. Y. Ng, "Large-scale deep unsupervised learning using graphics processors," in *International Conference on Machine Learning*, 2009, pp. 873–880.

[27] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Annual Conference on Neural Information Processing Systems*, 2006, pp. 801–808.

[28] C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," in *Annual Conference on Neural Information Processing Systems*, 2013, pp. 2553–2561.

[29] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.

[30] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, "Scalable object detection using deep neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2155–2162.

[31] N. Wang and D. Yeung, "Learning a deep compact image representation for visual tracking," in *Annual Conference on Neural Information Processing Systems*, 2013, pp. 809–817.

[32] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, 2010.

[33] L. Wang, T. Liu, G. Wang, K. L. Chan, and Q. Yang, "Video tracking using learned hierarchical features," *IEEE Transactions on Image Processing*, vol. 24, no. 4, pp. 1424–1435, 2015.

[34] W. Y. Zou, A. Y. Ng, S. Zhu, and K. Yu, "Deep learning of invariant features via simulated fixations in video," in *Annual Conference on Neural Information Processing Systems*, 2012, pp. 3212–3220.

[35] G. B. Huang, H. Lee, and E. G. Learned-Miller, "Learning hierarchical representations for face verification with convolutional deep belief networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2518–2525.

[36] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[37] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10, 000 classes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1891–1898.

[38] J. Lu, V. E. Liong, G. Wang, and P. Moulin, "Joint feature learning for face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 7, pp. 1371–1383, 2015.

[39] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Workshop on statistical learning in computer vision, ECCV*, vol. 1, no. 1-22. Prague, 2004, pp. 1–2.

[40] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006, pp. 2169–2178.

[41] F. Li and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 524–531.

[42] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2003, pp. 264–271.

[43] J. Yang, K. Yu, Y. Gong, and T. S. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1794–1801.

[44] J. Lu, G. Wang, W. Deng, P. Moulin, and J. Zhou, "Multi-manifold deep metric learning for image set classification," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 2015, pp. 1137–1145.

[45] B. Shuai, Z. Zuo, G. Wang, and B. Wang, "Dag-recurrent neural networks for scene labeling," *CoRR*, vol. abs/1509.00552, 2015.

[46] A. Wang, J. Lu, J. Cai, G. Wang, and T. Cham, "Unsupervised joint feature learning and encoding for RGB-D scene labeling," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4459–4473, 2015.