Image and Vision Computing 55 (2016) 3–5

Contents lists available at ScienceDirect

Image and Vision Computing

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

The changing fortunes of pattern recognition and computer vision☆

Rama Chellappa

Department of Electrical and Computer Engineering and the Center for Automation Research, UMIACS, University of Maryland, College Park, MD 201742, USA

ARTICLE INFO

Article history:
Received 31 March 2016
Accepted 13 April 2016
Available online 22 April 2016

Keywords:
Convolutional Neural Networks
Deep Learning
Image Recognition
Biometrics

ABSTRACT

As someone who had been attending conferences on pattern recognition and computer vision since 1978, I have watched with interest the ups and downs of pattern recognition and computer vision areas and how they have been presented at various conferences such as PRIP, CVPR, ECCV, ICCV, ACCV, ICPR, IJCAI, AAAI, NIPS, ICASSP and ICIP. Given the recent successes of deep learning networks, it appears that the scale is tilting towards pattern recognition as is commonly understood. A good number of papers in recent vision conferences seem to push data through one or other deep learning networks and report improvements over state of the art. While one cannot argue against the remarkable (magical?) performance improvements obtained by deep learning network-based approaches, I worry that five years from now, most students in computer vision will only be aware of software that implements some deep learning network. After all, 2-D based detection and recognition problems for which the deep learning networks have shown their mettle are only a subset of the computer vision field. While enjoying the ride, I would like to caution that understanding of scene and image formation, invariants, interaction between light and matter, human vision, 3D recovery, and emerging concepts like common sense reasoning are too important to ignore for the long-term viability of the computer vision field. It will be a dream come true if we manage to integrate these computer vision concepts into deep learning networks so that more robust performance can be obtained with less data and cheaper computers.

1. A brief history of developments in image recognition using computer vision and neural network-based methods

Since the early sixties, when Robert's edge operator was introduced, computer vision researchers have been working on designing various object recognition systems. The goal has been to design an end-to-end automated system that will simply accept 2D, 3D or video inputs and spit out the class labels or identities of objects. Beginning with template matching approaches in the sixties and seventies, methods based on global and local shape descriptors were developed. In the seventies, methods based on representations such as Fourier descriptors, moments, Markov models, and statistical pattern recognizers were developed. Even in the early years, the need for making the global recognition approaches be invariant to various transformations such as scale, rotation, etc. were recognized. Unlike these global descriptors, local descriptors based on primitives such as line segments, arcs etc. were used in either structural or syntactic pattern recognition engines. For example, generative grammars of various types were designed to parse the given object contour into one of many classes. More information on these developments can be found in [1–2].

In the eighties, statistical pattern recognition methods were seen as not being able to handle occlusions or geometric representations. Graph matching or relaxation approaches became popular for addressing problems such as partial object matching. In the mid-eighties, 3D range data of objects became available leading to surface-based descriptors, jump edges (edges separating the object and background) and crease edges (edges between surfaces). These representations naturally led to graph-based or structural matching algorithms. Another approach based on interpretation trees yielded a class of algorithms for object recognition. The theory of invariants became popular with the goal of recognizing objects over large viewpoints. More information on these developments can be found in [3–8].

While these approaches were being developed, methods based on artificial neural networks (ANNs) made a comeback in the mid-eighties. The emergence of ANNs was largely motivated by the excitement generated by Hopfield network's ability to address the traveling salesman problem and the rediscovery of back-propagation algorithm for training the ANNs. ANNs were not welcomed by most computer vision researchers. Computer vision researchers were brought up with the notion that representations derived from geometric, photometric as well as human vision points of view were critical for the success of object recognition systems [32]. The approach of simply feeding images into a 3-layer ANN and getting the labels out using training data was not appealing to most computer vision researchers. For one thing, it was not clear how general invariances could be integrated into ANNs, despite early attempts of Fukushima in designing the Neocognitron. Also, computer vision researchers were more interested in 3D object recognition problems and were not into optical character recognition

☆ This paper has been recommended for acceptance by Sinisa Todorovic.

http://dx.doi.org/10.1016/j.imavis.2016.04.005
0262-8856/

(OCR) where the ANNs were being applied. More information on these developments can be found in [9–13].

While the ANNs were becoming popular, a class of networks known as Convolutional Neural Networks (CNNs) was developed by LeCun and associates. The CNNs showed much promise in the domain of OCR. The CNNs represented the idea that one can learn the representations from data using learning algorithms. The tension between learning representations directly from data vs handcrafting the representations and applying appropriate preprocessing steps has been ever present. The emergence of representations such as the scale-invariant feature transform (SIFT), which showed an order of magnitude improvement when compared to interest points developed more than three decades ago, the histogram of gradients (HoG) operator and the local binary pattern [14] are good examples of hand-crafted features. In contrast, CNNs left the feature extraction work to a learning algorithm. Irrespective of whether hand-crafted or data-driven features extracted, there was a common agreement on the effectiveness of support vector machines as classifiers. More information on these developments can be found in [15–17].

The undaunted stalwarts of ANNs continued their quest for improving the performance by increasing the number of layers. Since the effectiveness of backpropagation algorithm was diminishing as the number of layers increased, unsupervised methods based on Boltzmann machines [18] and autoencoders [19] were suggested for obtaining good initial values of network parameters which were then fed into deep ANNs. As these developments were being made, the "Eureka" moment came about when deep CNNs were first deployed for the ImageNet challenge a mere four years back. The performance improvements obtained by DCNNs [20] for the ImageNet challenge were quite good. The power of depth, the availability of GPUs and large annotated data, replacement of traditional sigmoidal nonlinearities by Rectified Linear Units (ReLU) and drop-out strategies were embodied in the network now known as AlexNet [20]. The life of computer vision researchers has not been the same since!

The success of AlexNet motivated researchers from companies and numerous universities [21–23] to design various versions of DCNNs by changing the number of layers, the amount of training data being used and modifications to the nonlinearities, etc. While some may be dismayed by the reemergence of the so called "black box" approach to computer vision problems such as object detection/recognition and face verification/recognition, the simple fact of life is that it is hard to argue against performance. What is comforting though, the original issues in object recognition problems, such as robustness to pose, illumination variations, degradations due to low-resolution, blur and occlusion still remain. The current capability of DCCNs can be seen as close to that of a Model-T when it was introduced many decades ago.

2. Open issues

Given sufficient amount of annotated data and GPUs, DCNNs have been shown to yield impressive performance improvements. Still many issues remain to be addressed to make the DCNN-based recognition systems robust and practical. These are briefly discussed below.

2.1. Reliance on large training data sets

As discussed before, one of the top-performing networks in the MegaFace challenge needs 500 million faces of about 10 million subjects. Such large annotated training sets may not be always available (e.g. expression recognition, age estimation). So networks that can perform well with reasonable-sized training data are needed.

2.2. Invariance

While limited invariance to translation is possible with existing DCNNs, networks that can incorporate more general 3D invariances are needed.

2.3. Training time

The training time even when GPUs are used can be several tens to hundreds of hours, depending on the number of layers used and the training data size. More efficient implementations of learning algorithms are desired, preferably implemented using CPUs.

2.4. Number of parameters

The number of parameters can be several tens of millions. Novel strategies that reduce the number of parameters need to be developed.

2.5. Handling degradations in training data

DCNNs robust to low-resolution, blur, illumination and pose variations, occlusion, erroneous annotation, etc. are needed to handle degradations in data.

2.6. Domain adaptation of DCNNs

While having large volumes of data may help with processing test data from a different distribution than that of the training data, systematic methods for adapting the deep features to test data are needed.

2.7. Theoretical considerations

While DCNNs have been around for a few years, detailed theoretical understanding is just starting to develop [24–27]. Methods are needed for deciding the number of layers, nonlinearities, and the neighborhood size over which max pooling operations are performed.

2.8. Incorporating domain knowledge

The current practice is to rely on fine tuning. For example, for the age estimation problem, one can start with one of the standard networks such as the AlexNet and fine tune it using aging data. While this may be reasonable for somewhat related problems (face recognition and facial expression recognition), such fine tuning strategies may not always be effective. Methods that incorporate context may make the DCNNs more applicable to a wider variety of problems.

2.9. Memory

Although recurrent CNNs are on the rise, they still consume a lot of time and memory for training and deployment. Efficient DCNN algorithms are needed to handle videos and other data streams as blocks.

References

[1] K.S. Fu, Syntactic Pattern Recognition and Applications, Prentice-Hall, 1981.
[2] K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, 1990.
[3] O. Faugeras, Three-Dimensional Computer Vision: a Geometric View Point, The MIT Press, 1993.
[4] W. Grimson, Object Recognition by Computer: the Role of Geometric Constraints, MIT Press, Cambridge, MA, 1990.
[5] B.K.P. Horn, Robot Vision, The MIT Press, Cambridge, MA, 1986.
[6] T. Kanade, Three Dimensional Vision, Kluwer Academic Publishers, Boston, MA, 1987.
[7] A. Rosenfeld, A. Kak, Digital Picture Processing, vol 1 and 2, Academic Press, 1982.
[8] R. Szeliski, Computer Vision: Algorithms and Applications, Springer, 2010.
[9] G. Carpenter, S. Grossberg, Pattern recognition by self-organizing neural networks, A Bradford Book, 1991.
[10] K. Fukushima, Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, Biol. Cybern. 36 (4) (1980) 93–202.
[11] D.E. Rumelhart, J.L. McClelland, the PDP Research Group, Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol 1, MIT Press, Cambridge, 1986.
[12] P. Werbos, The Roots of Backpropagation: from Ordered Derivatives to Neural Networks and Political Forecasting, Wiley-Interscience, New York, NY, 1994.

[13] Y. Zhou, R. Chellappa, Artificial Neural Networks for Computer Vision, Springer-Verlag, 1991.
[14] T. Ahonen, A. Hadid, M. Pietikainen, Face description with local binary patterns: application to face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 28 (12) (2006) 2037–2041.
[15] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.
[16] Y. LeCun, F.-J. Huang, L. Bottou, Learning methods for generic object recognition with invariance to pose and lighting, IEEE Conf. Comput. Vis. Pattern Recognit. (2004) 97–104.
[17] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer, 1995.
[18] G.E. Hinton, T.J. Sejnowski, Learning and Relearning in Boltzmann Machines in parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1, Foundations. MIT Press, Cambridge, MA, 1986.
[19] Y. Bengio, Learning deep architectures for AI, Found. Trends Mach. Learn. 2 (1) (Jan. 2009) 1–127.
[20] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: F. Pereira, C. Burges, L. Bottou, K. Weinberger (Eds.),Advances in Neural Information Processing Systems, 25 (2012), pp. 1097–1105.
[21] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: a unified embedding for face recognition and clustering, IEEE Conf. Comput. Vis. Pattern Recognit. (June 2015) 815–823.
[22] Y. Sun, X. Wang, X. Tang, Deeply learned face representations are sparse, selective, and robust, IEEE Conf. Comput. Vis. Pattern Recognit. (June 2015) 2892–2900.
[23] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Deepface: closing the gap to human-level performance in face verification, IEEE Conf. Comput. Vis. Pattern Recognit. (June 2014) 1701–1708.
[24] J. Bruna, S. Mallat, Invariant scattering convolution networks, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1872–1886.
[25] R. Giryes, G. Sapiro, A.M. Bronstein, On the stability of deep networksarXiv preprint arXiv:1412.58962014.
[26] B.D. Haeffele, R. Vidal, Global optimality in tensor factorization, deep learning, and beyondarXiv preprint arXiv:1506.075402015.
[27] S. Mallat, Understanding deep convolutional networksarXiv preprint arXiv:1601.049202016 (arXiv:1411.7923, 2014).