MJRET
OPEN ACCESS
Multidisciplinary Journal of Research in Engineering and Technology
www.mjret.in
eISSN:2348-6953

Deep Learning For Computer Vision

[1]Amit Narute, [2]Roshan Jaiswal

[1]*Student, Department of Institute of Computer Science, MET College Maharashtra, India*
[2]*Assistance Professor,* Department of Institute of Computer Science, MET College Maharashtra, India

Abstract: *Over the last years deep learning methods have been shown to outperform previous state-of-the-art machine learning techniques in several fields, with computer vision being one of the most prominent cases. This review paper provides a quick overview of a number of the foremost significant deep learning schemes utilized in computer vision problems, that is, Convolutional Neural Networks, Deep Boltzmann Machines and Deep Belief Networks, and Stacked Denoising Autoencoders. A brief account of their history, structure, advantages, and limitations is given, followed by an outline of their applications in various computer vision tasks, like object detection, face recognition, action and activity recognition, and human pose estimation. Finally, a quick overview is given of future directions in designing deep learning schemes for computer vision problems and therefore the challenges involved therein.*

Keywords: *Deep Learning, Computer Vision, Convolutional Neural Networks, Deep Belief Networks*

1. INTRODUCTION

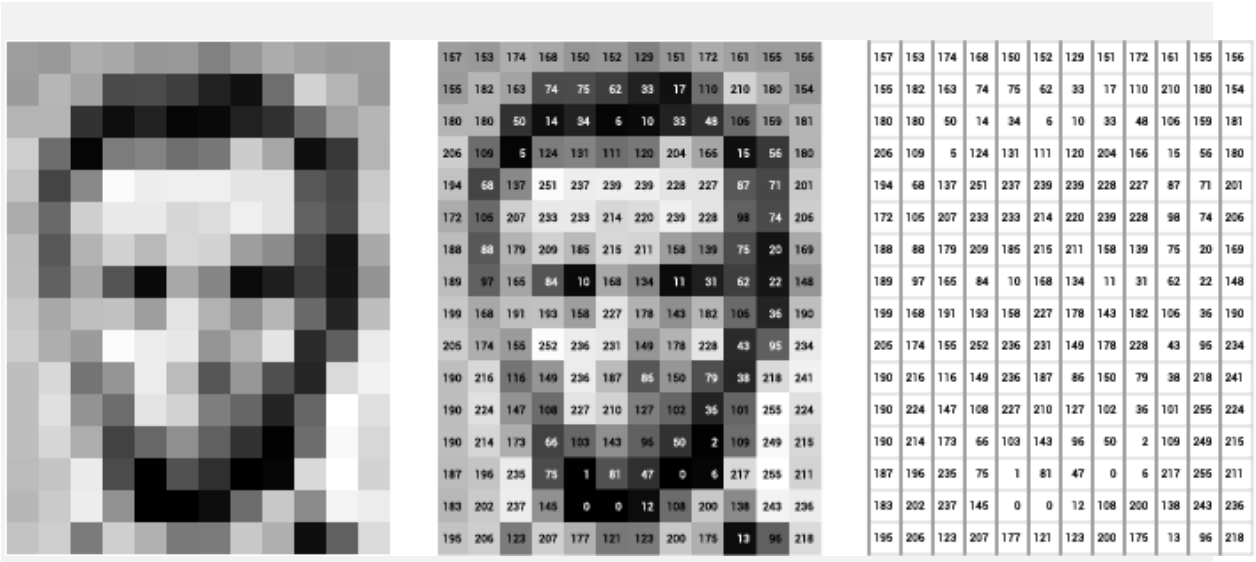
DL allows computational models of multiple processing layers to seek out and represent data with multiple levels of abstraction mimicking how the brain perceives and understands multimodal information, thus implicitly capturing intricate structures of large-scale data. Deep learning could also be an upscale family of methods, encompassing neural networks, hierarchical probabilistic models, and a selection of unsupervised and supervised feature learning algorithms. The recent surge of interest in deep learning methods is thanks to the very fact that they need been shown to outperform previous state-of-the-art techniques in several tasks, as well because the abundance of complex data from different sources (e.g., visual, audio, medical, social, and sensor). It is reasonable to say that the biggest difference with deep learning systems is that they no longer need to be programmed to specifically look for features. Rather than checking out specific features by way of a carefully programmed algorithm, the neural networks inside deep learning systems are trained. For example, if cars in an image keep being misclassified as motorcycles then you don't fine-tune parameters or re-write the algorithm. Instead, you continue training until the system gets it right. With the increased computational power offered by modern-day deep learning systems, there is steady and noticeable progress towards the point where a computer will be able to recognize and react to everything that it sees.

2. DEFINITION

Computer vision is an interdisciplinary field that deals with how computers are often made to understand high-level understanding from digital images or videos. From the attitude of engineering, it seeks to automate tasks that the human sensory system can do.[1][2][3] "Computer vision cares with the automated extraction, analysis and understanding of useful information from one image or a sequence of images. It involves the event of a theoretical and algorithmic basis to realize automatic visual understanding."[9] As a science, computer vision cares with the idea behind artificial systems that extract information from images. The image data can take many forms, like video sequences, views from multiple cameras, or multi-dimensional data from a medical scanner.[10] As a technological discipline, computer vision seeks to apply its theories and models for the development of computer vision systems.

How Does Computer Vision Work?

How exactly do our brains work, and how can we approximate that with our own algorithms? The reality is that there are very few working and comprehensive theories of brain computation; so despite the fact that Neural Nets are supposed to "mimic the way the brain works," nobody is kind of sure if that's actually true. The same paradox holds true for computer vision — since we're not selected how the brain and eyes process images, it's difficult to say how well the algorithms utilized in production approximate our own internal mental processes. On a particular level Computer vision is all about pattern recognition. So a method to coach a computer the way to understand visual data is to feed it images, many images thousands, millions if possible that are labeled, and then subject those to varied software techniques, or algorithms, that allow the pc to seek out patterns altogether the weather that relate to those labels. Below may be a simple illustration of the grayscale image buffer which stores our image of Lincoln . Each pixel's brightness is represented by a single 8-bit number, whose range is from 0 (black) to 255 (white):



Pixel data diagram. At left, our image of Lincoln; at center, the pixels labeled with numbers from 0–255, representing their brightness; and at right, these numbers by themselves.

In point of fact, pixel values are almost universally stored, at the hardware level, in a *one-dimensional array*. For example, the data from the image above is stored in a manner similar to this long list of unsigned chars:

```
{157, 153, 174, 168, 150, 152, 129, 151, 172, 161, 155, 156,
155, 182, 163, 74, 75, 62, 33, 17, 110, 210, 180, 154,
180, 180, 50, 14, 34, 6, 10, 33, 48, 106, 159, 181,
206, 109, 5, 124, 131, 111, 120, 204, 166, 15, 56, 180,
194, 68, 137, 251, 237, 239, 239, 228, 227, 87, 71, 201,
172, 105, 207, 233, 233, 214, 220, 239, 228, 98, 74, 206,
188, 88, 179, 209, 185, 215, 211, 158, 139, 75, 20, 169,
189, 97, 165, 84, 10, 168, 134, 11, 31, 62, 22, 148,
199, 168, 191, 193, 158, 227, 178, 143, 182, 106, 36, 190,
205, 174, 155, 252, 236, 231, 149, 178, 228, 43, 95, 234,
190, 216, 116, 149, 236, 187, 86, 150, 79, 38, 218, 241,
190, 224, 147, 108, 227, 210, 127, 102, 36, 101, 255, 224,
190, 214, 173, 66, 103, 143, 96, 50, 2, 109, 249, 215,
187, 196, 235, 75, 1, 81, 47, 0, 6, 217, 255, 211,
183, 202, 237, 145, 0, 0, 12, 108, 200, 138, 243, 236,
195, 206, 123, 207, 177, 121, 123, 200, 175, 13, 96, 218};
```

This way of storing image data may run counter to your expectations, since the data certainly *appears* to be two-dimensional when it is displayed. Yet, this is the case, since computer memory consists simply of an ever-increasing linear list of address spaces.

3. COMPUTER VISION TECHNIQUES

Image Classification

Classification is the process of predicting a specific class, or label, for something that is defined by a set of data points. Machine learning systems build predictive models that have enormous, yet often unseen benefits for people. For example, the reliable classification of spam email means that the average inbox is less burdened and more manageable. While the average end-user is likely unaware of the complexity of the problem and the vast amount of processing required to mitigate it, the benefits are clear.

Image classification is a subset of the classification problem, where an entire image is assigned a label. Perhaps a picture will be classified as a daytime or nighttime shot. Or, in a similar way, images of cars and motorcycles will be automatically placed into their own groups. There are countless categories, or classes, in which a specific image can be classified. Consider a manual process where images are compared and similar ones are grouped according to like-characteristics, but without necessarily knowing in advance what you are looking for. Obviously, this is an onerous task. To make it even more so, assume that the set of images numbers in the hundreds of thousands. It becomes readily apparent that an automatic system is needed in order to do this quickly and efficiently.

The deep learning architecture for image classification generally includes convolutional layers, making it a convolutional neural network (CNN). Several hyperparameters, such that the number of convolutional layers and the activation function for each layer, will have to be set. This is a non-trivial part of the process that it outside of the scope of this discussion. However, as a starting point, one can usually select these values based on existing research.

On such system is AlexNet, which is a CNN that gained attention when it won the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Another well-studied model is the Residual Neural Network (ResNet), which later won the same challenge, as well as the Microsoft Common Objects in Context (MS COCO) competition, in 2015.

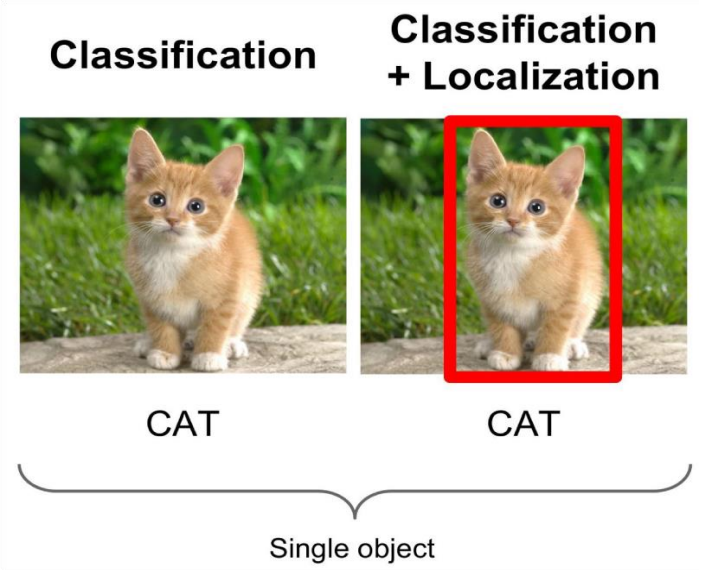
Image Classification with Localization

The second application of deep learning for computer vision is Image Classification with Localization. This problem is a specialization of image classification, with the additional requirement that the object within the picture is first located, and then a bounding box is drawn around it.

This is a more difficult problem than image classification, and it begins with determining whether there is only a single object depicted. If so, or if the number of objects is known, then the goal is to locate each object and identify the four corners of the corresponding bounding box. This process would be a necessary step in a system responsible for vehicle identification. Consider an automated system that browses pictures of cars, and it is guaranteed that there is a single vehicle contained within the scene. Once the vehicle has been located, properties such as the make, model, and color can be identified.

This task are often accomplished by employing a popular deep learning model, like AlexNet or ResNet, and modifying the fully connected layer to make the bounding box. As mentioned previously, there may be some fine-tuning to do in terms of setting hyperparameters or modifying the architecture for efficiency in a particular domain, but in practice, the basic architectures perform well. It will be necessary to have sufficient training data that includes examples with both the object description and the bounding box clearly defined, although sample datasets are available for this purpose.

The difficulty with this task comes about when there is an unknown number of objects in the picture. In the majority of images, especially those taken in public areas, there will be many possibilities such as different people, vehicles, trees, and animals. For this kind of environment, the problem becomes one of object detection.



Object Detection

Object Detection is image classification with localization, but in pictures which will contain multiple objects. This is a lively and important area of research because the pc vision systems which will be utilized in robotics and self-driving vehicles are going to be subjected to very complex images. Locating and identifying every object will undoubtedly be a critical part of their autonomy.

The architecture required for object detection differs in a crucial way. Namely, the size of the output vector is not fixed. If there is a single object in the picture, for example, then there will be four coordinates that define the bounding box. This static and predefined value works using the previously mentioned architectures. However, as the number of objects increases, the number of coordinates increases as well. Especially given that the number of objects is not known in advance, this requires adjustments in the makeup of the neural network. One such modified architecture is that the R-CNN: Regions with CNN features. This approach involves generating regions of interest that are scaled to a fixed size and then forwarding these regions into a model such as AlexNet. While this system produces good results, it is computationally expensive, and too slow for a real-time computer vision system. With the goal of speeding up R-CNN, there have been various adjustments made to the architecture. The first is Fast R-CNN, which contains optimizations and other innovations that improve both speed and detection accuracy. Taking it one step further, the next generation, Faster R-CNN model, includes an additional CNN named the Region Proposal Network (RPN). The RPN is trained to generate high-quality regions that are submitted to the Fast R-CNN model. The combination of these algorithms leads to an impressive increase in speed and is truly on the path towards real-time object detection in computer vision systems.

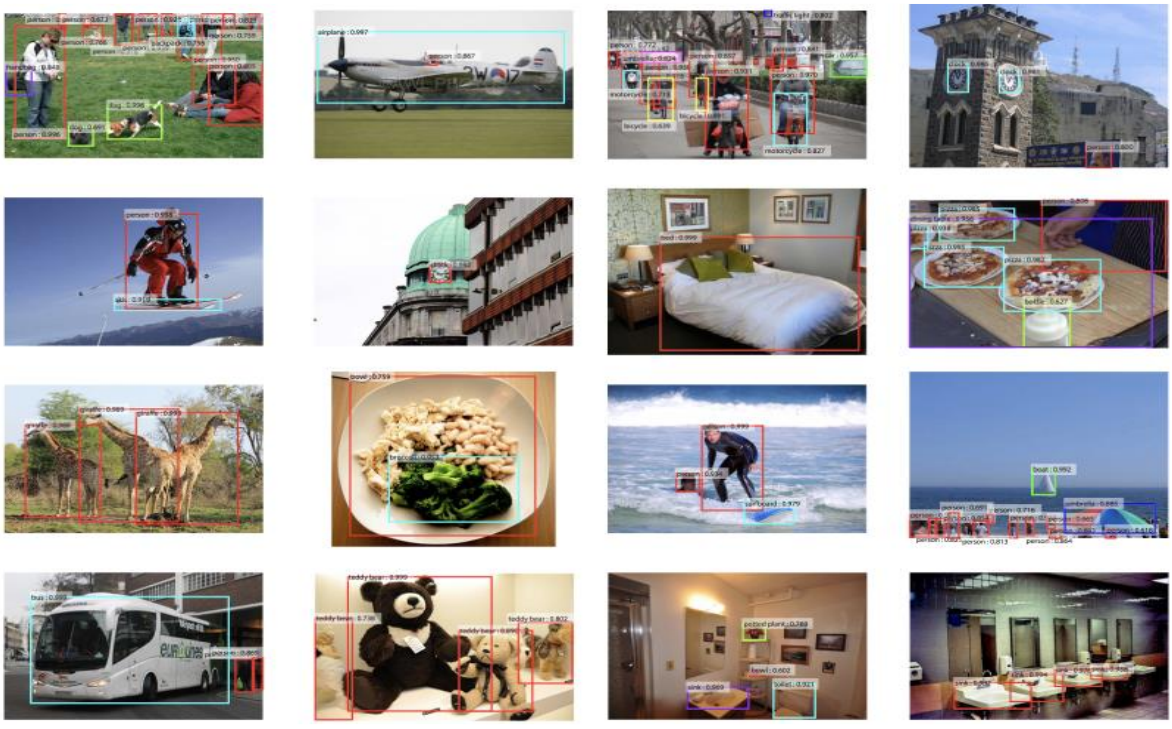
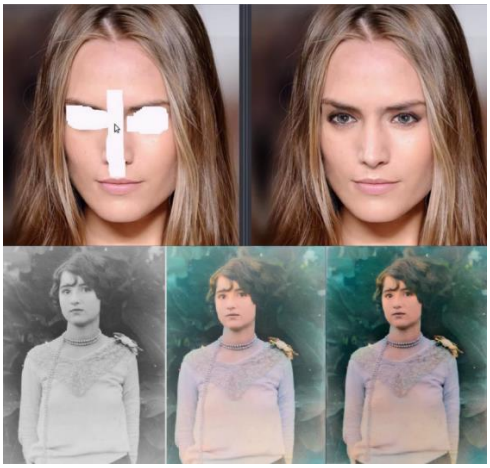


Image Reconstruction

Image Reconstruction is the task of recreating the missing or corrupt parts of an image. This is a difficult undertaking that can be thought of a transformation or filter, that may not have an objective evaluation. While it is indeed possible to ensure that the visible properties of an image can be closely matched, it is clearly unreasonable to demand that the computer re-create details for which there is no reference. As such, image reconstruction systems have limits that very much depend on how much of the original image is available to learn from. One model that performs image reconstruction is known as Pixel Recurrent Neural Networks. This is a system that makes use of a Recurrent Neural Network (RNN) to predict the missing pixels in an image along two spatial dimensions. Examples of applications for image reconstruction are restoration of photos, or black and white movies. In a self-driving vehicle, image reconstruction may be used to look beyond small obstructions, such as a signpost between the vehicle and a pedestrian that is being tracked.



(Image Reconstruction and Colorization. Source: NVIDIA and blog.floydhub.com)

Object Tracking

To this point, the tasks have been focused on operations that can work with a single, still image. A vital goal in computer vision, however, is to have the ability to recognize an event that is occurring over a period of time. With a single picture to visually describe the events at one instant in time, it requires a series of pictures to gain a greater understanding of the whole.

Object Tracking is one such example, where the goal is to keep track of a specific object in a sequence of images, or a video. The snapshot that begins the sequence contains the object with a bounding box, and the tracking algorithm outputs a bounding box for all of the subsequent frames. Ideally, the bounding box will perfectly encapsulate the same object for as long as it is visible. Moreover, if the object should become obscured and then re-appear, the tracking should be maintained. For the purpose of this discussion, we can assume that the input to an object tracking algorithm is the output from an object detection algorithm.

Object tracking is vital for virtually every computer vision system that contains multiple images. In self-driving cars, for example, pedestrians and other vehicles generally have to be avoided at a very high priority. Tracking objects as they move will not only help to avoid collisions through the use of split-second maneuvers, but also, the model can supply relevant information to other systems that will attempt to predict their next move.

The Open Source Computer Vision Library, OpenCV, contains an object tracking API. There are several algorithms available, each of which performs differently depending on the characteristics of the video, as well as the object itself. For example, some algorithms perform better when the object being tracked becomes momentarily obstructed. OpenCV contains both classic and state-of-the-art algorithms to handle many tasks in computer vision, and may be a useful resource for developing such systems.

4. CONCLUSION

Computer vision is a stimulating and important field that features a sort of applications across domains. Their effective use isn't simply relevant, but rather, required and important for further developing applications like autonomous robots and vehicles.

Traditional computer vision systems aren't only slow but rather inflexible. They require a great deal of input from the developer and do not easily adjust to new environments. Deep learning systems, on the opposite hand, handle computer vision tasks end-to-end and don't require external information or coaching to an equivalent degree.

Advancements in deep learning systems and computing power have helped to enhance the speed, accuracy, and overall reliability of computer vision systems. As deep learning models improve and computing power becomes more readily available, we'll still make steady progress towards autonomous systems which will truly interpret and react to what they perceive.

REFERENCES

[1]. *W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," Bulletin of Mathematical Biology, vol. 5, no. 4, pp. 115–133, 1943.View at: Publisher Site | Google Scholar*

[2]. *Y. LeCun, B. Boser, J. Denker et al., "Handwritten digit recognition with a back-propagation network," in Advances in Neural Information Processing Systems 2 (NIPS*89), D. Touretzky, Ed., Denver, CO, USA, 1990.View at: Google Scholar*

[3]. *S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.View at: Publisher Site | Google Scholar*

[4]. *G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," Neural Computation, vol. 18, no. 7, pp. 1527–1554, 2006.View at: Publisher Site | Google Scholar | MathSciNet*

[5]. *TensorFlow, Available online: https://www.tensorflow.org.*

[6]. *B. Frederic, P. Lamblin, R. Pascanu et al., "Theano: new features and speed improvements," in Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop, 2012, http://deeplearning.net/software/theano/.View at: Google Scholar*

[7]. *Mxnet, Available online: http://mxnet.io.*

[8]. *W. Ouyang, X. Zeng, X. Wang et al., "DeepID-Net: Object Detection with Deformable Part Based Convolutional Neural Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 7, pp. 1320–1334, 2017.View at: Publisher Site | Google Scholar*

[9]. *A. Diba, V. Sharma, A. Pazandeh, H. Pirsiavash, and L. V. Gool, "Weakly Supervised Cascaded Convolutional Networks," in Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5131–5139, Honolulu, HI, July 2017.View at: Publisher Site | Google Scholar*

[10]. *N. Doulamis and A. Voulodimos, "FAST-MDL: Fast Adaptive Supervised Training of multi-layered deep learning models for consistent object tracking and classification," in Proceedings of the 2016 IEEE International Conference on Imaging Systems and Techniques, IST 2016, pp. 318–323, October 2016.View at: Publisher Site | Google Scholar*

[11]. *N. Doulamis, "Adaptable deep learning structures for object labeling/tracking under dynamic visual environments," Multimedia Tools and Applications, pp. 1–39, 2017.View at: Publisher Site | Google Scholar*

[12]. *L. Lin, K. Wang, W. Zuo, M. Wang, J. Luo, and L. Zhang, "A deep structured model with radius-margin bound for 3D human activity recognition," International Journal of Computer Vision, vol. 118, no. 2, pp. 256–273, 2016.View at: Publisher Site | Google Scholar | MathSciNet*

[13]. *S. Cao and R. Nevatia, "Exploring deep learning based solutions in fine grained activity recognition in the wild," in Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), pp. 384–389, Cancun, December 2016.View at: Publisher Site | Google Scholar*

[14]. *A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, pp. 1653–1660, USA, June 2014.View at: Publisher Site | Google Scholar*

[15]. *X. Chen and A. L. Yuille, "Articulated pose estimation by a graphical model with image dependent pairwise relations," in Proceedings of the NIPS, 2014.View at: Google Scholar*

[16]. *H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in Proceedings of the 15th IEEE International Conference on Computer Vision, ICCV 2015, pp. 1520–1528, Santiago, Chile, December 2015.View at: Publisher Site | Google Scholar*

[17]. *J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '15), pp. 3431–3440, IEEE, Boston, Mass, USA, June 2015.View at: Publisher Site | Google Scholar*

[18]. *D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex," The Journal of Physiology, vol. 160, pp. 106–154, 1962.View at: Publisher Site | Google Scholar*

[19]. *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.View at: Publisher Site | Google Scholar*

[20]. *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. 2278–2323, 1998.View at: Publisher Site | Google Scholar*

[21]. *Y. LeCun, B. Boser, J. S. Denker et al., "Backpropagation applied to handwritten zip code recognition," Neural Computation, vol. 1, no. 4, pp. 541–551, 1989.View at: Publisher Site | Google Scholar*

[22]. *M. Tygert, J. Bruna, S. Chintala, Y. LeCun, S. Piantino, and A. Szlam, "A mathematical motivation for complex-valued convolutional networks," Neural Computation, vol. 28, no. 5, pp. 815–825, 2016.View at: Publisher Site | Google Scholar*

[23]. *M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free? - Weakly-supervised learning with convolutional neural networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, pp. 685–694, June 2015.View at: Publisher Site | Google Scholar*