1 notes

SPP is trying to avoid data augmentation . But the latest one introducing it model more sophisticated system. Terms:dropout,maxout network Reducing filter size help to improve system.Increasing depth will help.

2 Multi-stage Hubel-Wiesel architecture

In 1962, Hubel DH and Wiesel TN[1],[2] was studied visual cortex of anaesthetized cats with spots of white light of various shapes. They classified the cells in the visual system in to simple,complex and hypercomplex . Simple cells are influenced by the arrangement of excitatory and inhibitory regions of the receptive field and position of stimulus is important . This cells receives input from cells of the lateral geniculate nucleus(LGN), which is connected to retina. But the complex cells will responds to a properly oriented stimulus regardless of the cell position in the receptive field. Complex cells are activated by edge,dark bar,slit and mixed stimuli . Hypercomplex cells are activated by edge,single-stopped(corner),double-stopped (tongue),slit(double-stopped)and dark bar(double-stopped).

Hubel DH and Wiesel TN [2] presented a functional hierarchical structure of the visual cortex . According to their model, visual perception cells are in the order , $simple \Longrightarrow complex \Longrightarrow lower\ order\ hypercomplex \Longrightarrow higher\ order\ hypercomplex$. Activation of lower stage is influenced by the position of the input patterns and higher stages are position-invariant. There are several contradictory to this structure, but no one completely deny this hierarchical model.

Inspired by this work , Fukushima, K [3] proposed a neural network model for pattern recognition called neocognitron. In neocognitron , cells are arranged in a number of cascded structure. Each structure U include a simple cell layer U_s and a complex cell layer U_c. This network is not affected by change in position or small distortion in the shape of patterns. It is also capable to do self organization based on an unsupervised competitive learning algorithm[4] in the first two layers and classification based on supervised learning in the output layer.

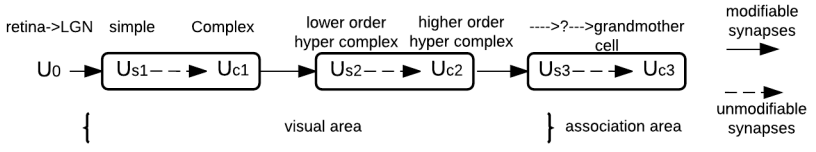

Figure 1: Neocognitron[3]

This design was improved by Yann LeCun [5],[6], [7],[8], [9] using backpropagation algorithm[10] to train the entire system.

3 Network In Network

Inspired by the work of Ian J. Goodfellow et al.[11] on maxout networks, Min Lin et al.[12] introduced a micro network in each convolution layer so that it will compute more abstract features.This network gave a state-of-the-art performance in ILSVRC 2013 competition with an error rate of 12.95%.They used NVIDIA TITAN GPU to train the newtork.

4 Visualizing and Understanding Convolutional Networks

Matthew D. Zeiler and Rob Fergus[13] presented a method to visualize the function of intermediate feature layers of CNNs and used as a diagnostic tool to improve the model proposed by Krizhevsky et al.[14]. This method helped them to understand activation in the feature maps with respect to the input patterns. It shows Krizhevsky et al.'s architecture does not have enough mid frequency coverage in the first layer filters and aliasing artifacts caused by large stride in the first layer convolutions. Authors solved this problems by decreasing filter size to 7×7 and reducing stride to 2.This implementation won the ILSVRC 2013 competition with an error rate of 11.74%

5 Spatial pyramid pooling in Deep Convolutional Networks

Instead of using fixed input size in CNNs, Kaiming He et al.[15] suggested the use of a spooling strategy called spatial pyramid pooling(SPP)[16][17] to avoid cropping or warping of images. It introduced a new layer on top of the convolution layer and perform aggregation based on Bag-of-Words (BoW) model [18]. But the classical back propagation training methods expect layers to have a fixed size. To overcome this problem author implemented two fixed size networks with shared parameters and switch the network on alternate epochs. This network is trained using a single GeForce GTX Titan GPU with a starting learning rate of 0.01 and achieved a less error rate of 8.06% on ILSVRC 2014 data set.

This implementation improves the performance of baseline architectures including ZF-5[13] , Convnet [14] and Overfeat-5/7 [19]. Their study shows, accuracy of CNNs will improve on multi-size training, multi-level pooling, and full-image representations.

6 Going deeper with convolutions

Christian Szegedy and et al.[20] proposed a network named GoogLeNet with receptive field(input layer) of size 244×244 with the number of layers around 100. Network is trained using asynchronous stochastic gradient descent with 0.9 momentum and fixed learning rate schedule based on no of epochs . Learning procedure took advantage of model and data-parallelism in a CPU-based cluster environment. This network gave an error rate of 6.67% on ILSVRC 2014 data set. Their result shows that use of existing dense blocks to build the sparse structure can improve the performance of convolutional networks.

7 VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION

Karen Simonyan and Andrew Zisserman [21] evaluated the effect of network depth in image classification using very small convolution filters. Their deep network architecture comprise of fixed size input layers , a stack of convolution layers , three Fully-Connected (FC) layers and 5 max-pooling layers for spatial spooling over a 2×2 pixel window with stride 2. Hidden layers are modeled using Rectified Linear Units(ReLU)[22]. On the hardware side, it uses a multi-GPU system with NVIDIA Titan Black GPUs. Network is trained using multinomial logistic regression based on backpropagation with momentum of 0.9 and batch size 256.

In this work authors formed a conclusion that greater depth with small convolution filters and initialization of certain layers will cause the learning process to converge in less number of epochs. This model of the convolution network does not differ from the classical architecture proposed by LeCun et al.[23]. But the authors reported a significant improvement in the performance using an increased depth. This implementation results in a significant improvement in accuracy with an error rate of 6.8% in ILSVRC 2014 of ImageNet.

8 Deep Image: Scaling up Image Recognition

The latest attempt in image classification with an error 5.98% in ImageNet data set is reported by Ren Wu et al.[24] of Baidu research.They developed an end to end deep learning system named Deep Image. It uses a highly optimized parallel algorithm to implement large deep neural network with augmented input data. The network is trained using stochastic gradient decent algorithms (SGD)[ref] on a custom built high performance system comprised of 36 server nodes, each with 2 six-core Intel Xeon E5-2620 processors and 4 NVIDIA Tesla K40m GPUs . System uses an InfiniBand network for interconnections. Parallelism strategies used in their network are model-data parallelism and data parallelism. These methods have been proposed by Alex Krizhevsky [25] and Omry Yadan et al.[26] for training convolutional neural networks with SGD on

a multiple GPU systems. But it is not easy extend the same strategies to multiple GPU clusters because of the communication overhead. So the Baidu Team focused on minimizing network data transfers and overlapping the computation. They use butterfly synchronization and lazy update strategies to achieve data parallelism in the gradient computation. Their results show model-data parallelism is better when number of GPUs is less than 16. Implementation of Data parallelism in a large number of GPU cluster is better because of the constant communication requirements.

The authors have explored different data augmentation techniques to increase the number of labeled images in the training set. This includes color casting, Vignetting, Lens distortion, Rotation, Flipping, and Cropping. Instead of using the same resolution on all images, they have trained separate models at different scales, combined results by averaging soft max class posteriors. Data set used in this experiment was subset of ImageNet dataset, used in the competition ImageNet Large-Scale Visual Recognition Challenge (ILSVRC)[27]. This data set includes 1.2 million images which contain 1,000 categories.

Major contribution of this work is the demonstration of tremendous computational power to achieve high accuracy in image classification. It also shows augmented multi-scale images can be combined to achieve less error rate in convolutional network in the context of the image classification.

9 REFERENCES

References

[1] Wiesel TN Hubel DH, "Receptive fields, binocular interaction and functional architecture in the cats visual cortex," *The Journal of Physiology*, vol. 160, no. 1, pp. 106–154, 1962.

[2] D H Hubel and T N Wiesel, "Receptive Fields and Functional Architecture in Two Nonstriate Visual Areas (18 and 19) of the Cat.," *Journal of neurophysiology*, vol. 28, pp. 229–289, 1965.

[3] K Fukushima, "Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position.," *Biological cybernetics*, vol. 36, pp. 193–202, 1980.

[4] Kunihiko Fukushima and Sei Miyake, "Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position," *Pattern Recognition*, vol. 15, no. 6, pp. 455–469, 1982.

[5] Y. LeCun, "Learning processes in an asymmetric threshold network," in *Disordered systems and biological organization*, E. Bienenstock, F. Fogelman-Soulié, and G. Weisbuch, Eds., Les Houches, France, 1986, pp. 233–240, Springer-Verlag.

[6] Y. LeCun, L. D. Jackel, B. Boser, J. S. Denker, H. P. Graf, I. Guyon, D. Henderson, R. E. Howard, and W. Hubbard, "Handwritten digit recognition: Applications of neural net chips and automatic learning," in *Neurocomputing, Algorithms, Architectures and Applications*, F. Fogelman, J. Herault, and Y. Burnod, Eds., Les Arcs, France, 1989, Springer.

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

[8] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Handwritten digit recognition with a backpropagation network," in *Advances in Neural Information Processing Systems (NIPS 1989)*, David Touretzky, Ed., Denver, CO, 1990, vol. 2, Morgan Kaufman.

[9] Y. LeCun, O. Matan, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel, and H. S. Baird, "Handwritten zip code recognition with multilayer networks," in *Proc. of the International Conference on Pattern Recognition*, IAPR, Ed., Atlantic City, 1990, vol. II, pp. 35–40, IEEE, invited paper.

[10] A E BRYSON, W F DENHAM, and S E DREYFUS, "OPTIMAL PROGRAMMING PROBLEMS WITH INEQUALITY CONSTRAINTS," *AIAA Journal*, vol. 1, no. 11, pp. 2544–2550, Nov. 1963.

[11] Ian J. Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio, "Maxout Networks," *arXiv preprint*, pp. 1319–1327, 2013.

[12] Min Lin, Qiang Chen, and Shuicheng Yan, "Network In Network," *arXiv preprint*, p. 10, 2013.

[13] Matthew D Zeiler and Rob Fergus, "Visualizing and Understanding Convolutional Networks," *arXiv preprint arXiv:1311.2901*, pp. 818–833, 2013.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances In Neural Information Processing Systems*, pp. 1–9, 2012.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," *arXiv preprint arXiv ...*, vol. cs.CV, pp. 1–14, 2014.

[16] Kristen Grauman and Trevor Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," *Proceedings of the IEEE International Conference on Computer Vision*, vol. II, no. October, pp. 1458–1465, 2005.

[17] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, 2006, vol. 2, pp. 2169–2178.

[18] J Sivic and A Zisserman, "A text retrieval approach to object matching in videos," *Proc. CVPR*, , no. Iccv, pp. 2–9, 2003.

[19] Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus, and Yann LeCun, "OverFeat : Integrated Recognition , Localization and Detection using Convolutional Networks," *arXiv preprint arXiv:1312.6229*, pp. 1–15, 2013.

[20] Christian Szegedy, Scott Reed, Pierre Sermanet, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," pp. 1–12.

[21] Karen Simonyan and Andrew Zisserman, "V d c n l -s i r," pp. 1–13, 2015.

[22] Vinod Nair and Geoffrey E Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines," *Proc. 27th Int. Conf. Mach. Learn.*, , no. 3, pp. 807–814, 2010.

[23] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, pp. 2278–2323, 1998.

[24] Ren Wu, Shengen Yan, Yi Shan, Qingqing Dang, and Gang Sun, "Deep Image: Scaling up Image Recognition," *arXiv Prepr. arXiv1501.02876*, 2015.

[25] Alex Krizhevsky, "One weird trick for parallelizing convolutional neural networks," pp. 1–7, 2014.

[26] O Yadan and Keith Adams, "Multi-GPU Training of ConvNets," *arXiv*, pp. 10–13, 2013.

[27] Alex Berg and J Deng, "Imagenet large scale visual recognition challenge 2010," *Challenge*, 2010.