ImageNet Classification with Deep Convolutional Neural Networks

Summary written by Nicholas Hinke
March 01, 2022

Summary: This paper presents the use of one of the deepest *convolutional neural networks* (CNNs) ever created for image classification. Later nicknamed "AlexNet", the network as proposed by the authors was comprised of eight layers–five convolutional layers to act as high level feature extractors, and three fully-connected layers including a softmax layer to act as a more traditional classifier–which totaled over 650 thousand neurons with 60 million learned parameters [6]. The construction of this network was only possible due to other recent work on methods to more efficiently train CNNs [1, 4, 7] as well as methods to prevent model overfitting–especially for a network of this size and depth [1, 2, 3, 4, 9]. After demonstrating superior performance in the authors' experiments when compared to previous competition winners, the network was entered in the 2012 ImageNet Large-Scale Visual Recognition Challenge (ILSVRC), ultimately winning by a margin of more than 10% over the second place finisher [6].

Approach: As alluded to above, the authors primarily tested their eight layer network on the 2010 ILSVRC dataset, which contains approximately 1.2 million images of 1000 different object classes. As consequence, given their choice to attempt to utilize such a deep network with so many trainable parameters, the authors were forced to use a variety of techniques to both improve the efficiency of training and to combat potentially severe overfitting [6].

Most notably, the authors chose to use the *Rectified Linear Unit* (ReLU) activation function rather than the more classical choice of hyperbolic tangent, as recent experiments demonstrated a potential of up to 6x speedup during training [6, 7]. Additionally, the authors chose to train the network in parallel on two 3GB GPUs, rather than using just a single GPU. This not only aided in the speed of training, but more importantly allowed them to utilize substantially more parameters–by almost a factor of two–in the convolutional layers of the network [6].

In addition to those listed above, the authors made several more architectural decisions that increased the performance of the network. Namely, they added local-response normalization and overlapping max pooling layers after each of the first two convolutional layers. These design choices were shown to further prevent ReLU saturation and help to combat overfitting [6, 7]. It should also be noted that another max pooling layer was added after the final convolutional layer [6].

Beyond architecture design, the authors also implemented several interesting strategies during network training in the hopes of mitigating overfitting as much as possible. First, they used the technique of data augmentation to generate many more training images on the CPU during training–thus removing the need to store them locally–in order to offset the effects of the huge number of parameters in the network [1, 2, 9]. Moreover, the authors included the then-recent technique of dropout regularization in order to force the network to learn more robust features and prevent neuron co-adaptation [3, 6].

Upon completion of the network, the authors tested it on the available 2010 ILSVRC dataset and found that it outperformed the next-best published results by more than 8% [8]. After performing some of these experiments, a version of the network then went on to win the 2012 competition with a classification error rate of only 15.3%–a substantial margin over second place at 26.2% [6].

Strengths: Most importantly, AlexNet achieved the best ever published results on a variety of subsets of the ImageNet dataset. It goes without saying that this marked a historic moment in the evolution of deep neural networks. Moreover, the authors were able to successfully demonstrate the importance of *depth* in networks of this nature, as they showed that removing any of the layers would be a detriment to the network's performance [6]. Finally, the authors greatly influenced future advances in the field, as they made all of their code and trained parameters publicly available [5].

Weaknesses: Although considered by the authors, they failed to implement any unsupervised pre-training of the model due to the added computational cost, despite the potential unrealized gains in the network's performance [6]. Moreover, the architecture of the network required a consistently-sized square input image, which forced the rectangular images in the dataset to be cropped accordingly. This cropping (and downsampling!) may have caused several misclassifications due to the loss of information, especially for objects that were highly shifted in the image.

Reflections: Since the importance of network depth was made clear throughout the paper, it may prove extremely valuable to consider adding more convolutional layers. Additionally, considering the convolutional layers just as high level feature extractors, it would be beneficial to study the performance of other classifiers (*e.g.* an SVM) in comparison to the three fully-connected layers with a softmax as utilized by the authors. Finally, it would be quite interesting to

test the network's efficiency and performance on higher resolution images using today's modern hardware and GPUs for training.

References

[1] D. C. Ciresan, U. Meier, J. Masci, L. M. Gambardella, and J. Schmidhuber. High-performance neural networks for visual object classification. *ArXiv*, abs/1102.0183, 2011.

[2] D. C. Ciresan, U. Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3642–3649, 2012.

[3] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *ArXiv*, abs/1207.0580, 2012.

[4] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun. What is the best multi-stage architecture for object recognition? *2009 IEEE 12th International Conference on Computer Vision*, pages 2146–2153, 2009.

[5] A. Krizhevsky. One weird trick for parallelizing convolutional neural networks. *CoRR*, abs/1404.5997, 2014.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[7] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010.

[8] J. Sánchez and F. Perronnin. High-dimensional signature compression for large-scale image classification. *CVPR 2011*, pages 1665–1672, 2011.

[9] P. Y. Simard, D. Steinkraus, and J. C. Platt. Best practices for convolutional neural networks applied to visual document analysis. *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.*, pages 958–963, 2003.