ImageNet Classification with Deep Convolutional Neural Networks

Summary written by Sahana Srihari
March 09, 2020

Summary: The paper presents a Deep CNN architecture which details winning top error rates in the LSVRC-2010 competition on the ImageNet dataset. They address the different layers and their implementation in detail for a 1000 class classification. Highly optimised GPU implementation of CNNs with new features reduces training time and boost performance. Overfitting rose as a problem which was tackled using dropout regularisation. The final network contained 5CNN layers and 3 FC layers and concurrent experimental proof was obtained.

Related work: The author's work is centered around the use of CNNs and its modification for efficient classification. The model they use for training deviates from the standard sigmoid function and instead adopts a non-saturating nonlinear function ReLU adapted from [5]. For an efficient GPU implementation for dealing with a large datasets, their modifications follow columnar CNNs as [2] and normalisation of ReLU response as [4]. When faced with overfitting, compared to usual data augmentation [6][1][2] they use dropout [3]

Approach: The images considered were downsampled to 256x256 and demeaned. The main structure consists 8 learned layers(5 CNNs and 3 FC). The outputs are modelled with ReLU, which trains faster than the traditional functions due to non-saturation which aids with large Neural Networks(NN).This is applied across all the layers. Performance of CNNs falls short with large numbers of images of high resolution if there is no optimization. The author's have uniquely tackled the issue is by dividing the workload between 2 GPU's and further limiting the interactions to certain layers. Adopting such a scheme helped reduce the top1 and top5 error rates by 1.7% and 1.2%. The 2nd, 4th and 5th convolutional layer is connected to the previous layer on the same GPU and the 3rd layer kernal is fully connected in the 2nd layer. Since ReLU's have unbounded activations, Local Response Normalisation is needed to normalise for local contract enhancement, following local inhibition of neurons. The neurons with larger excitement or response is emphasised for the succeeding layers whereas other neurons in the local neighbourhood (across the channel)are suppressed, they've implemented this in the 1st and 2nd convolutional layer. In a general framework of CNNs pooling layers are constructed to summarize the feature maps such that the stride and size of the patch considered do not overlap the pooling layers. But their approach implement overlapping pooling layers with a stride s = 2 and patch size z = 3 and in doing so it reduces the possibility of overfitting and also reduced the top error rates. Max-pooling is their preferred method which occurs after the normalization layers and the last convolution layer. Dealing with such a large dataset, methods to avoid overfitting was through 2 data augmentation techniques (i) image translation and horizontal relfection through random subsampling of image patches (224 x 224) (ii) Altering RGB channel intensities in the training images using PCA and using Dropout which involves setting some outputs of the neuron to 0 with a probability of 0.5. The final stage consists of a 1000-way softmax for clasification. *Datasets, Experiments and Results:* Top-1 and top-5 test error rates(ER) of 37.5% and 17.0% and averaging the predictions of 5 CNNs gives an ER of 16.4%.One CNN, with an extra sixth convolutional gives an ER of 16.6% Averaging the predictions of two CNNs pre-trained on Fall 2011 release gives an ER of 15.3% The network learnt frequency and orientation-selective kernels with kernels on GPU 1 being color-agnostic and on GPU 2 color-specific.

Strengths: It was one of the largest CNN with record top error rates.It exceeded in performance by a large margin by reducing error rates by 9.6% and 11.2% for top-1 and top-5 error rates. It tackles many issues though improved learning functions, overfitting methods and normalisation which makes this model robust. The efficient cross-parallelised GPU implementation profoundly impacted the training time on large datasets for neural networks efficiently. At the time of manual feature extraction AlexNet radically showed the positives of learned features and thereby surpassing traditional object recognition procedures. One of the advantages is also the rapid downsampling and intermediate representations.

Weaknesses: The network is incapable of handling varying dimension images and therefore needs downsampling. The learning rates in the model is set to be equal for all layers which needed to be manually tweaked during training.The model uses CNNs,it comes with problems associated with them such as class imbalance and reliance on human intervention. It is heavily reliant on the depth and removing any layer becomes detrimental. The paper does not explicitly show the performance for images that are heavily occluded, missing, variations in illumination etc.,

Reflections: The paper has not explicitly mentioned zero-padding by a factor of 3, without which an image (224x224) convolved with a filter of size 11, gives the output of size

54.25 and not 55. It is intriguing to see the stark difference in the way object recognition was handled using Deep CNNs instead of shallow networks. The architecture made ripples in the community and lead further grounds for improvements for reducing the gap to the human visual pathway.Further improvements could be achieved through variations of GPU implementation as they come up with better GPUs. Overall the framework is easy to comprehend and the authors have relayed the information succinctly.

References

[1] D. Ciregan, U. Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3642–3649. IEEE, 2012.

[2] D. C. Cireşan, U. Meier, J. Masci, L. M. Gambardella, and J. Schmidhuber. High-performance neural networks for visual object classification. *arXiv preprint arXiv:1102.0183*, 2011.

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

[4] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun. What is the best multi-stage architecture for object recognition? In *2009 IEEE 12th international conference on computer vision*, pages 2146–2153. IEEE, 2009.

[5] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[6] P. Y. Simard, D. Steinkraus, J. C. Platt, et al. Best practices for convolutional neural networks applied to visual document analysis. In *Icdar*, volume 3, 2003.