ImageNet Classification with Deep Convolutional Neural Networks

Authors: Alex Krizhevsky, Ilya Sutskever, G.E.Hinton
Year: 2012
Conference: Neural Information Processing Systems Conference.
Report:

1. **Abstract:**

"We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax. To make training faster, we used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully-connected layers we employed a recently-developed regularization method called "dropout" that proved to be very effective. We also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry."

2. **Introduction:**
   - The training datasets will be of the size in thousands to train and test, in general. The datasets like MNIST had already reached the human performance with so far implementing algorithms.
   - But, in the real world, the trained images will not be any more sufficient as there would be innumerable possibilities of structures and placements of the objects.
   - So, the large datasets should be required to increase the interaction of computer vision with the real world is needed, which they had done in this article.
   - They have created a dataset called as ImageNet, trained with CNN and got better error results with 37.5% and 17% which are better than the former works.

3. **Content and Approach:**
   - MNIST, one of the fundamental datasets had already approached the human performance with the current best results.
   - Now, the more requirement of training data is required for the better realworld understanding of the model.
   - So, a new dataset called ImageNet is introduced to train the model. To the date, it had been the largest dataset with 14 million images and 22000 categories labelled manually with hand.

- This size of data would be hard enough to train with the contemporary hardware and the algorithms. So, they require some phenomenal characteristics in their process to train the model.
- The architecture of the neural networks, they designed, had eight learned layers in which five are CNN and the remaining three are fully connected neural networks.
- They have used neural networks with ReLU instead of tanh function, as the former is a way better and faster than the latter. So, that makes a good progress with this size of data in the neural networks.
- The authors also noted that the faster learning also had influence on the performance of the large data.
- Regarding Hardware, a single GPU is not sufficient for the process. So, two GPUs are used, in which they communicate only in the corresponding layers of neural network and that even saves the time when compared to single GPU.
- Local Response Normalization is a process which helps in generalizing the data and then ReLU non linearity is applied on the neural network.
- Pooling gives the output of the neighbouring groups of neurons of the same Kernel map.So, this is also adapted through the process.
- The architecture contains eight layers in which the first layer filters the image with 96 kernels, converts each into 4 pixels.
- The input of the second layer is the output of the first layer. It takes the output and filters with 256 kernels. The remaining third, fourth and fifth also filters the image with corresponding kernels.
- Overall the network has 4096 neurons and all the layers are connected with different CNN networks.
- One of the problem faced through the process is overfitting. It is very common while training the data. There are two major procedures to overcome the problem. They are:
    - **Data Augmentation:** The data is enlarged artificially either by the orientation of the image or altering the RGB values of the training data. This would decrease the overfitting because although the image changes in orientation and contrast, the features would be almost same.
    - **Dropout**: Initial probability of weights are given 0.5 through out the network. While iterating the weights update and the features with lower weights will be dropped out. This process is very good for large training data.
- The batch size of 128 is iterated over 90 cycles with the initial weights of 0.01 through the network.
- It took 5 to 6 days on two NVIDIA GTX 580 3GB GPUs through the training set of 1.2 million images.

4. **Results:**
    - This process had given the better error results with 37.5% and 17% for top-1 and top-5 respectively. Whereas the Sparse coding and SIFT+FV models had given 47.1% and 45.2% of top-1 error rates and 28.2% and 25.7% of top-5 error rates.
    - The removal of a layer in CNN has also caused a drastic change in error rate which caused a rise in 2% error rate.
    - The better results also given for oddly placed objects in the picture. For example a mite in the top left corner of the image is also resulted correctly.

5. **Discussion and Thoughts:**
    - The authors have discussed the various features affecting the training and testing data through the process.
    - However, they have not reached the range of human performance through the process, there is more to come after the article.
    - Moreover, they suggested that the increase in memory handling through hardware will increase the performance.
    - It can be envisioned to use the method to different data types, despite the fact that the images will be having more memory than the other data types.

6. **Conclusion:**
    - The article declared that the large deep convolutional networks will be greatly accessible for the large sets of training data.
    - It also given better results compared to the former critical models, despite of being simple.
    - The advancement would definitely lead to the increase in performance of interaction between the computer vision models and the real world.
    - The future work which they had represented that the deep convolutional networks on the video sequences which are not static would also be a good progress.