Preliminary Evaluation of CNN Classification by Objective Testers

Li Huang

Leiden University, Niels Bohrweg 1
2300 CA Leiden, The Netherlands
lihuang864@gmail.com

Abstract In general the most recent methods such as ResNet and the residual based methods are considered state-of-the-art and are frequently used for benchmarking and initial features. One of the dangers of using standardized datasets is that it is possible to accidentally mix the test set with the training set, which will usually give higher accuracy. This is particularly difficult to check for in neural networks because the model is simply a set of millions of weights. The real goal here is to do an acid test between VGG and ResNet which is getting results from real users with imagery that has not been seen by the classifier before. When accidental mistakes are not possible in the training procedure, which one achieves the highest accuracy?

1 Introduction

In this recent generation of deep neural networks, one of the noteworthy aspects is that typically the models, the source code, and the design of the architecture are open and free. In many ways this is a golden age for the free and open distribution of scientific achievements and contributions. At the same time it is possible for scientists to make mistakes, perhaps even be biased whether it is conscious or subconscious. One of the easiest ways to get the highest accuracies on ImageNet test is simply to have overlap between the training set and the test set.

Another way which can be done accidentally is to use another test set from another group in another country which also used the same Flickr images. Here there is the question of: if you have one data set made in London and another in New York, how similar is the imagery? In fact, could a number of images be exactly the same or be near duplicates with very slight differences such as different compression levels or different resolutions. As far as I know these things are not looked into carefully. This means that some of the lot of datasets could be getting images from big databases such as Flickr; or it could also be the case that the people who submit images to Flickr may also submit to other sources. So even if the scientist goes out of his way to collect images from a Twitter website feed, he might accidentally also get the same images that were available from the Flickr website.

What this means is that there is some uncertainty as to which deep neural networks actually perform the best. So, with students we designed a small scale test (the results can be seen as a rough indication but should not be seen as rigorous or definitive) to look at how well several of the extremely well-known convolutional neural networks do in a challenging and also realistic test which is the "top-1" or R@1 accuracy.

2 Related Work

There are numerous important areas related to machine learning and computer vision. Some of the popular ones include human-computer interaction [1][2], content-based retrieval [3][4][5][6][7][8][9]10], object detection and activity recognition [11][12][13][14] and many others.

In 2012 there was the famous work AlexNet [15] which introduced a 7-layer convolutional network which won an early ImageNet competition. this network used a number of new ideas which included different activation functions different kinds of layers to add some invariance to translations and in general attempted to reduce the problem of gradient explosion.

These early neural networks required weeks to months to train. They also achieved rather spectacular improvements over the competition. The next level up was VGG [16] by the Oxford group which gave significant improvements to the AlexNet which included changes to allow even deeper networks. This brings us to one of the most impactful neural networks called ResNet [17] or *residual network* which was based upon skip connections and extended the depth of the network from 20 layers to 150 layers. It also won the ImageNet competition and achieved the best accuracy of its time within the parameters of the competition.

Research quickly moved onwards and the state-of-the-art advanced in multiple directions at once. For example, many researchers have continued to improve ResNet (see [18]) and also combining competitors such as Inception with ResNet [19] or proposing new architectures [20].

Deep networks have been found to be useful in diverse areas. These areas cover fields such as big data [8], salient point detection [11], activity recognition [12], generic object recognition [13], edge detection [21], content-based retrieval [3]. For a good general overview, see [22][23].

3 Fair Testing

Here I want to briefly point out that one has to be careful in comparing different accuracy numbers. One has to perform all steps in exactly the same way in order to have a fair comparison. For example, can one use additional training data in training a new network?

Should the training data be fixed to a certain set of images? Should there be a limit to the processing time?. For example, is it fair to compare a neural network there was trained on only the official training set for the competition or can one also use another 15 million images from another dataset?

This is why in most of the performance tables these days there is a notation on whether the method used additional data and additional data means data outside of the original test. A good summary of modern benchmarking is [24].

4 Evaluation

So now we get to the performance and how we are going to measure it. First we will only use well known, freely downloadable, published models from VGG and ResNet. This means we are not training the models ourselves - we are not making mistakes in training the models. We are using the author's models as intended and as published by the authors also for reproducibility.

Next, we asked 50 students in the computer science department to select images based upon Google image searches - 60 images per student. This means that the students will also be directing the search, not the authors of the neural networks. There will be in fact no communication or influence from the authors to the users. The image selected by the students are checked for being near copies of images from the ImageNet dataset. If they are near copies, then the students must find replacements. Here is the reality check, the real evaluation with unbiased images from objective uses.

So, we have a test set of 3,000 images which are divided uniformly in the three categories of

- Hummingbird
- Labrador
- Coffee maker

5 Results

From the Imagenet challenge, we would expect ResNet to beat VGG. The results were surprising.

When each student is treated as a separate test for 60 images, then in 82% of the tests, VGG had a higher average accuracy than ResNet at R@1 and in 18% of the tests, ResNet had a higher average accuracy at R@1.

Table 1. VGG vs ResNet for each student

	VGG won	**ResNet won**
50 students	41	9
Percentage wins	82	18

Next, when we look at the total percentage of correctly classified images, we get

Table 2. Overall accuracy for VGG and ResNet

VGG	**ResNet**
0.804	0.796

6 Discussion & Conclusions

In both Table 1 and 2, the results are surprising. Instead of ResNet winning the majority of the time, we see that VGG has won 82%.

However, when we look at the actual accuracy in Table 2, then we see that they are very close for the 3,000 image test set.

In our tests we found the following:

When one uses the pretrained VGG and ResNet models and uses them for images which have not been seen in the training process, then both VGG and ResNet have a similar R@1 in our experiments. VGG actually does slightly better but the difference is certainly within the realm of plausible noise.

At the same time, we understand the limits and weaknesses of our test. First, this test only used 3 categories, not 1000. It is possible that these 3 categories happen to be ones where the two classifiers are neck and neck. Second, we did not use a very large test set. Our test set was only 3,000 images, not a hundred thousand or a million. For these reasons, we recommend taking the results of these experiments as only “preliminary”. Our results only lightly suggest that more objective benchmarking should be done. This was an eye-opening result for all of the students because beforehand they all would have thought ResNet would easily win the tests.

In future research we will look deeper at more categories and also popular transformations such as rotation.

References

[1] Dix, A., 2017. Human–computer interaction, foundations and new paradigms. Journal of Visual Languages & Computing, 42.

[2] Sebe, N., Lew, M.S. and Huang, T.S., 2004. The state-of-the-art in human-computer interaction. In International Workshop on Computer Vision in Human-Computer Interaction (pp. 1-6). Springer, Berlin, Heidelberg.

[3] Zhou, W., Li, H. and Tian, Q., 2017. Recent advance in content-based image retrieval: A literature survey. arXiv preprint arXiv:1706.06064.

[4] Azad, H.K. and Deepak, A., 2019. Query expansion techniques for information retrieval: a survey. Information Processing & Management, 56(5).

[5] Huiskes, M.J. and Lew, M.S., 2008. Performance evaluation of relevance feedback methods. In Proceedings of the 2008 international conference on Content-based image and video retrieval.

[6] Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., Martinez-Gonzalez, P. and Garcia-Rodriguez, J., 2018. A survey on deep learning techniques for image and video semantic segmentation. Applied Soft Computing, 70.

[7] Sebe, N. and Lew, M.S., 2001. Salient Points for Content-Based Retrieval. In BMVC.

[8] Zhang, Q., Yang, L.T., Chen, Z. and Li, P., 2018. A survey on deep learning for big data. Information Fusion, 42.

[9] Sebe, N., Lew, M.S. and Smeulders, A.W., 2003. Editorial introduction: video retrieval and summarization. Computer Vision and Image Understanding.

[10] Han, J., Zhang, D., Cheng, G., Liu, N. and Xu, D., 2018. Advanced deep-learning techniques for salient and category-specific object detection: a survey. IEEE Signal Processing Magazine, 35(1).

[11] Han, J., Zhang, D., Cheng, G., Liu, N. and Xu, D., 2018. Advanced deep-learning techniques for salient and category-specific object detection: a survey. IEEE Signal Processing Magazine, 35(1).

[12] Wang, J., Chen, Y., Hao, S., Peng, X. and Hu, L., 2019. Deep learning for sensor-based activity recognition: A survey. Pattern Recognition Letters, 119.

[13] Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X. and Pietikäinen, M., 2018. Deep learning for generic object detection: A survey. arXiv preprint arXiv:1809.02165.

[14] Gout, A., Lifchitz, Y., Cottencin, T., Groshens, Q., Fix, J. and Pradalier, C., 2017. Evaluation of off-the-shelf cnns for the representation of natural scenes with large seasonal variations.

[15] Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems.

[16] Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[17] He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition.

[18] Wu, Z., Shen, C. and Van Den Hengel, A., 2019. Wider or deeper: Revisiting the resnet model for visual recognition. Pattern Recognition, 90.

[19] Szegedy, C., Ioffe, S., Vanhoucke, V. and Alemi, A.A., 2017, February. Inception-v4, inception-resnet and the impact of residual connections on learning. In Thirty-First AAAI Conference on Artificial Intelligence.

[20] Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition.

[21] Liu, Y. and Lew, M.S., 2016. Learning relaxed deep supervision for better edge detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

[22] Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M.P., Shyu, M.L., Chen, S.C. and Iyengar, S.S., 2018. A survey on deep learning: Algorithms, techniques, and applications. ACM Computing Surveys (CSUR), 51(5).

[23] Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y. and Alsaadi, F.E., 2017. A survey of deep neural network architectures and their applications. Neurocomputing.

[24] Shi, S., Wang, Q., Xu, P. and Chu, X., 2016. Benchmarking state-of-the-art deep learning software tools. In 2016 7th International Conference on Cloud Computing and Big Data (CCBD).