EECE 7370 Assignment 1

Review for the paper:

ImageNet Large Scale Visual Recognition Challenge

1. **Reviewer:** Sakshi Bhatia, Reviewed on: 22 September, 2024

2. **Paper Details:**
    a. Paper Title: ImageNet Large Scale Visual Recognition Challenge
    b. Authors: Olga Russakovsky, Jia Deng, Hao Su1, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, Li Fei-Fei
    c. Published Online: 11 April 2015
    d. Citation: Russakovsky, O., Deng, J., Su, H. *et al.* ImageNet Large Scale Visual Recognition Challenge. *Int J Comput Vis* **115**, 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y

3. **Paper Summary:**
    a. The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) is an object category classification and detection challenge that consists of a publicly available dataset, an annual competition, and a corresponding workshop.
    b. The challenge dataset consists of approximately 1.2M training images, 50 thousand validation images, 100 thousand test images, and 1000 object classes. The challenge has three tasks: image classification, single-object localization, and object detection. Different datasets have been carefully curated for each of these tasks.
    c. The paper talks about dataset collection for the challenge, manual large-scale annotation techniques used, challenges encountered in creating this dataset, and highlights the most successful algorithms in this challenge run over 5 years. Finally, they compare the accuracy of these algorithms to human-level accuracy.

4. **Main Contribution:**

a. The paper explains in detail the procedures used for creating the ILSVRC dataset. These procedures, and hence this paper, will be a significant resource for anyone attempting to create a benchmark dataset for Computer Vision tasks such as image classification and object detection with tight bounding boxes in the future.
b. Various means of data collection have been described, the major ones being the existing ImageNet dataset, and querying Flicker and other search engines for images according to categories. To collect lots of candidate images, expanded sets of queries in multiple languages were used.
c. Since the 1.4M images contained in this dataset are manually annotated using well-designed crowdsourcing techniques. These annotations have been labeled using a labeling algorithm with hand-designed queries that help reduce associated annotation efforts, and hence costs.
d. A summary of the most successful algorithms and the techniques used in those algorithms, on each of the tasks for all five years of the challenge has been discussed. Major breakthroughs in image classification and object recognition have resulted from the efforts in this dataset.
e. For the evaluation of algorithms submitted in this challenge, standardized procedures defined by Caltech 101 and PASCAL VOC datasets have been extended to adapt to the large scale of the dataset.
f. The authors then performed a detailed analysis of where the "optimistic" algorithms fail and how certain object properties impact the accuracy of recognition algorithms. For example, algorithms usually perform better on large and extra-large real-world objects than on smaller ones, irrespective of the object's scale in the dataset.
g. The paper then compared the accuracy of GoogLeNet, the most successful algorithm of ILSVRC2014 image detection with provided data track and object detection with external data track, to that of human annotators. Overall, a trained human annotator is found capable of outperforming GoogLeNet by approximately 1.7%. Further, specific features were identified where GoogLeNet did better vs where human annotators did better.

5. **Paper Strengths:**
    a. The paper discusses in depth the various challenges encountered when constructing a dataset with more than a million images and one thousand object categories, and evaluating algorithms on three different tasks on a large scale.
    b. Various novel techniques have resulted from this. Specifically:
        i. They developed multi-step annotation strategies for obtaining robust annotations manually on a large-scale dataset.
        ii. New evaluation criteria were developed for evaluating algorithms on a large dataset. New methods of calculating errors have been formulated for image classification and single-object localization tasks, and a method of calculating average precision has been adopted from PASCAL VOC for object detection.
    c. The paper identified patterns where algorithms usually do not do well in the challenge. Some of the object categories identified to be harder to recognize include small objects, untextured objects, rigid objects, images with filters, etc. This analysis would be a huge benefit for people looking to build more challenging datasets in the Computer Vision community.

6. **Paper Weaknesses:**
    a. The dataset for the image classification task should be improved. The way the evaluation is designed, there should not be more than 5 object classes in an image in the classification task, for the algorithm evaluation criteria to be more fair.
    b. Approximately 0.3% of ground truth annotations were found to be incorrect, as mentioned by the authors. Although it is a small percentage and justified by the authors in the paper, this can be quite a large number considering the scale of the dataset. Hence, there is a scope for improvement here.

7. **Experimental Procedures:**

a. While preparing the dataset for image classification, users are asked to classify whether each image contains the object category. Labeling is done until a pre-determined confidence score is reached. For manual annotation, a generic multi-class labeling algorithm has been adopted. A sequence of queries is designed such that 200 leaf node questions at the end of the queries correspond to 200 categories in the object detection task. Quality control is done through additional verification tasks.
b. On studying the effect of object properties on the performance of various successful algorithms in the challenge, it was observed that objects that occupy more space in an image tend to be easier to recognize by algorithms. An experiment was designed to ensure that this is not being influenced by the differences in the object scale. Each bin is normalized by object scale until the average scale of classes in each bin across one property is approximately same. It was observed that average precision of XS, S and M objects is statistically insignificant from average precision on L objects.
c. Similar experiments were designed to study the effects of deformability and object texture on performances of the optimistic models.
d. Another experiment was designed to study human accuracy on this challenge. Two human annotators with different levels of training were asked to perform the challenge tasks. The optimistic human error was defined and found to be 2.4%. The average pace of labeling turns out to be a bimodal distribution. Also, it is concluded that a significant training time is required for a human annotator to achieve good performance on this dataset.

8. **Future Work:**
    a. One recommendation is a modification in the challenge such that the categories that are identified as hard to classify and localize are included more. This might give us better future algorithms out of this challenge.
    b. For future work, the work could be extended to implement the suggestions of the authors in the paper. Future challenges could focus

on precision (of the predictions the algorithm made, how many were termed correct by humans), rather than accuracy and recall. This might reduce elaborate manual annotation efforts and the resulting costs.

9. **Final Recommendation:** Given the efforts put into this challenge, various analyses and comparisons done in the paper, and the conclusions coming out of these methods, this paper will be a significant contribution to the Computer Vision community and deserves to be accepted.