Problem Definition

Peyman Bateni

June 27, 2019

1 Challenge Description

- The DARPA proposed Learning with Less Labels (referred to as LwLL from here on) relies on the ImageNet Large-Scale Visual Recognition Challenge (abbreviated as ILSVRC) [1]. The database composition is as described in the later sections.
- Note that the material described here comes from ILVRC 2017, which consists of the latest update to the challenge. **It's relatively unclear whether the challenge was discontinued or whether the dataset stopped updating.** That said, ILVRC 2017 stands to be the current "goto" dataset for visual classification/detection tasks.
- **Additionally, it's likely that DARPA's detection task relies on the object detection dataset although that needs to be verified, as the localization case also involves placing bounding boxes around the single-label class in the image.**

2 Object Classification and Localization

2.1 Dataset Description

- The validation and test data for this competition will consist of **150,000 photographs**, collected from flickr and other search engines, hand labeled with the presence or absence of **1000 object categories**.
- The 1000 object categories contain both internal nodes and leaf nodes of ImageNet, but do not overlap with each other. A random subset of 50,000 of the images with labels will be released as validation data included in the development kit along with a list of the 1000 categories. The remaining images will be used for evaluation and will be released without labels at test time.
- **The training data, the subset of ImageNet containing the 1000 categories and 1.2 million images**, will be packaged for easy downloading. The validation and test data for this competition are not contained in the ImageNet training data (we will remove any duplicates).

2.2 Task 0 - Object Classification

- **Please note that while classification is defined separately in various ILSVRCs, in the most recent one, namely 2017, the focus on this dataset is mainly on next task (denoted as Task 1). That said, the DARPA description seems to suggest that Task 0 is intended for the classification problem.**

- For each image, algorithms will produce a list of at most 5 object categories in the descending order of confidence. The quality of a labeling will be evaluated based on the label that best matches the ground truth label for the image.
- It's noted on the website that "[the] idea is to allow an algorithm to identify multiple objects in an image and not be penalized if one of the objects identified was in fact present, but not included in the ground truth." My interpretation from this is that logical "misfires" may happen, but the ground truth should be present in the top 5 classes identified.
- For each image, the network is expected to produce 5 labels c_j, $j = 1, ..., 55$. The ground truth labels for the image are denoted as C_k, $k = 1, ..., n$ with n classes of objects labelled.
- **The error of the network is then calculated as** $e = \frac{1}{n} \times \sum_k \min_j d(c_j, C_k)$**. Here,** $d(c_i, C_k)$ **is defined to be 0 when** $c_i = C_k$ **and 1 otherwise.**
- Naturally, the overall error score for the network is defined as the average error over all test images.
- **In the 2012 setting,** $n = 1$ **indicating that there is only one ground truth per image. It's unclear what the case is for ILSVRC 2017.**

2.3 Task 1 - Object Localization

- In this task, in addition to producing 5 class labels $c_j, j = 1, ..., 5$, the network produces 5 corresponding bounding boxes $b_j, j = 1, ..., 5$ where b_j is the boundary box for the class label l_j.
- Each image comes with ground truths $C_k, k = 1, ..., k$ as per before. For each ground truth label g_k, the ground truth bounding boxes are $B_k m, m = 1, ..., M_k$, where M_k is the number of instance of the k^{th} object in the current image.
- **The error for the network is calculated through** $e = \frac{1}{n} \sum_k \min_j \min_m \max\{d(c_j, C_k), f(b_j, B_{km})\}$ **where as before,** $d(c_i, C_k) = 0$ **when** $c_i = C_k$ **and 1 otherwise, and** $f(b_j, B_k) = 0$ **if** b_j **and** B_{mk} **have over 50% overlap, and 1 otherwise.**
- The intuition is that the error will be 0 if both localization and classification are correct. Otherwise, the error is 1 (ie. the maximum). In the case of multiple ground truth bounding boxes, only one 50%+ overlap is sufficient.
- **tNote that it's somewhat unclear whether DARPA's object detection task refers to this or the actual "Object Detection" task described next. Considering the fact that this task is named "Object Localization", it's fair to assume that "Object Detection" (explained in the next section) is the intended DARPA task.**

3 Object Detection

3.1 Dataset Description

- The object detection task was most recently updated in ILSVRC 2016 to include more diverse images in the test set in terms of occulusion, angle, etc, That said, the training set and the validation set remain the same from ILSVRC 2014.
- There are 200 basic-level categories which are fully anootated on the test data. The categories were carefully chosen considering different factors such as object scale, level of image clutterness, average number of object instance, and several others.

Set	Number of Images	Number of Objects
Training	456567	478807
Validation	20121	55502
Testing	40152	N/A

Table 1: Dataset composition for the object detection task.

Average Image Resolution	482x415
Average object classes per image	1.534
Average object instances per image	2.758
Average object scale (bounding box area as fraction of image area)	0.170

Table 2: Comparative statistics on the validation set for the object detection task,

- The composition of the data is as described in Table 1. As you can see there are more objects that images, thus motivating the case for multi-object detection. Note that the number of objects for the test set has not been released publicly, The online protal for ISVRC 2014 has also provided some interesting statistic mainly on the validation set as noted in Table 2.

3.2 Task - Objecct Detection

- For each image, algorithms will produce a set of annotations (c_i, b_i, s_i) of class labels c_i, bounding boxes bi and confidence scores s_i. This set is expected to contain each instance of each of the 200 object categories. Objects which were not annotated will be penalized, as will be duplicate detections (two annotations for the same object instance).
- **The challenge doesn't provide an explicit error function. DARPA cites mAP (mean average precision) as the metric for both detection and classification, although doesn't specify precision in the case of the detection case.** Also it suggests that the winner of the detection challenge will be the team which achieves first place accuracy on the most object categories, a condition that is not confirmed by DARPA.

4 Object Detection from Video

- This task is irrelevant to the DARPA challenge! Hence, the description below is kept brief and directly from the ILSVRC page. For more information, visit ILSVRC's webpage.
  - This is similar in style to the object detection task. We will partially refresh the validation and test data for this year's competition. There are 30 basic-level categories for this task, which is a subset of the 200 basic-level categories of the object detection task. The categories were carefully chosen considering different factors such as movement type, level of video clutterness, average number of object instance, and several others. All classes are fully labeled for each clip.
  - For each video clip, algorithms will produce a set of annotations (f_i, c_i, s_i, b_i) of frame number f_i, class labels c_i, confidence scores s_i and bounding boxes b_i. This set is expected to contain each instance of each of the 30 object categories at each frame. The evaluation metric is the same as for the object detection task, meaning objects which are not annotated will be penalized, as will duplicate detections (two annotations for the same object instance). The winner of the detection from video challenge will be the team which achieves best accuracy on the most object categories.

References

[1] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.