Murat Ambarkutuk {murata@vt.edu}, 04/14/2016

Visual Recognition with Humans in the Loop

The authors claim that object recognition algorithms are not developed enough to provide practical applications, despite the amount of effort that has been given to object recognition task(s). Thus, the authors approach the problem from another angle and try to solve two basic questions to reason about what lies behind the lower performance of the objection recognition algorithms. The authors propose a "human-in-the-loop" method for object classification where the distinction line between two phenomena "what is easy for human to recognize" and "what is easy for a computer to recognize" is drawn. The authors aim at leveraging long-term memory of a computer system, while using humans as the supervisors as the system learns how to distinguish classes finely-related to each other (clearly needs expertise).

The paper formulates the problem by employing the famous 20 questions games where the whole system is required to distinguish the object class by asking one questions for each step, and less than 20 questions in total. The method is essentially an application of the Bayesian Framework where whenever a new answer provided by the user it is regarded as a new observation and it is incorporated into prior distribution with Bayes' formula. Along with user responses, the computer vision output is incorporated in similar manner. One of the significant property of the contribution is that any existent multi-class object recognition algorithm can be plugged into the system. The ultimate goal of the system is to reduce the number questions needed for any classes to be distinguished.

Results and Conclusions

The paper present a parameteric study on two different datasets, Birds-200 and Animals with attributes. Deterministically speaking, $\approx 7.643 (= \log_2(200))$ bits are enough to encode all the classes and distinguish one from another for 200 cluster scenario. If a deterministic approach would be employed, the approach should come up with about 8 question to identify the class of interest then others, which sets the baseline for the classification task.

Figure-5 shows the correlation between the question needed versus classification accuracy. Due to limitations of the expertise in birds of Mechanical Turks, the classification percentage levels off 5% level, where the classes are finely-related to each other, while deterministically acquired responses make accuracy level skyrocketed with 8 questions. The proposed technique, on the other hand, shows a linear correlation between two variables until 30 questions to be asked. This validates that where some form of expertise is needed to achieve classification task, the proposed method can provide a fair amount of service with enough responses from the humans in the loop. Moreover, the authors claim that the proposed system can achieve around 66% classicafication accuracy, the contribution of the computer vision part of the system discarded.

The figure-6 illustrates the classification accuracy versus how many binary questions needed to be asked if the output of the vision algorithm is incorporated. The graph clearly shows that in order to reach about the same level of accuracy, the system requires more questions if the vision part would be discarded. This implies that the vision algorithm reduces the manual work, which was an intended goal. This phenomenon validates the contribution of the vision part of the system to the

overall performance. Furthermore, the same figure also shows that the overall accuracy increases by 3% if the vision part is incorporated.