Charles Ollion
Aug 2, 2018 · 7 min read

What's easy/hard in AI & Computer Vision these days?

When talking about AI (particularly about Computer Vision), I spend half of my time saying how much the field has progressed in the last few years, and the other half *debunking* and *diminishing* what's possible today.

Xkcd nailed it a few years ago. It changed. A bit.

Recently, I came accross an article by Pete Warden showing a plant disease classifier. It seems very accurate at detecting different types of diseases the human eye would have trouble to; however it has spectacular failures when used on random pictures (non plant), that a human would never make.

It seems that capabilities of Computer Vision systems are usually very different from our human intelligence, and this is what I decided to illustrate with a little quiz.

Quiz solutions

Disclaimer: If you disagree with the answer I provide to this question, I'm happy to discuss as there is still plenty to learn in that field and I don't think I have all the answers as of now!

Diabetic Retinopathy: Classification should be fairly **Easy** as the problem is well constrained in terms of input and output (Google claim really good performance in their blogpost). Difficulties arise when putting such a system into production, UX & the way you handle the interaction with the doctor is key, as there could be huge imbalance between the different classes.

Webcam Gesture Recognition: The problem is rather well defined, but the variability makes it quite hard: webcam videos have people with varying distance, gesture duration, etc… Furthermore, natural difficulties arise from the analysis of video which bring more engineering problems for training. I'd say this problem is **Quite Hard** but solvable.

Handbag detection on Instagram: The problem seems easy and solved, but the input domain is open/unconstrained (instagram) and the class definition is wide (handbag could mean practically anything, there are no clear visual patterns associated with handbags). This make this problem unexpectedly **Very Hard**, see by yourself…

Pedestrian Detection from Camera: The problem is rather **Easy**: The input domain is quite constrained (fixed camera), and the class (pedestrian) is quite standard. There will be problems related to occlusions, but globally the problem is easilly solvable (you could even do it without Deep Learning). However, modify slightly the problem scope and it can become much harder: if the Camera is moving (in a robot, in a car…) ; or has several points of view, angles, scale — the problem becomes more and more open and difficult.

Robotic Object Grasping: This problem is **Very Hard**. It goes beyond a standard classification or regression problem as the output is a robotic policy, usually trained using reinforcement learning, which is much less mature than supervised learning. Moreover, objects vary in size, shape, and the way you have to grasp them may require semantic understanding. So even though this problem is solved by a 2-year old child in a much less restrictive setting (the camera here is fixed and the background is always the same), it's still a long way before we solve this problem!

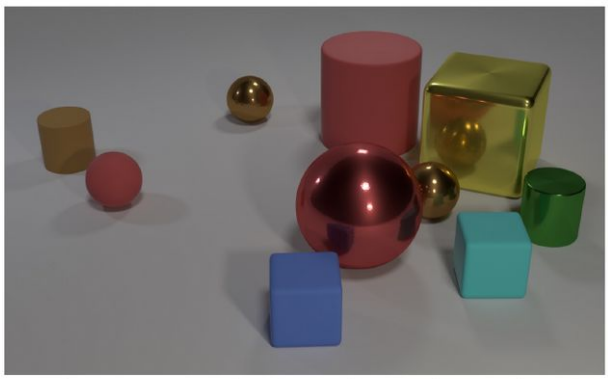
Charles Ollion
317 Followers
CTO Heuritech #machinelearning #deeplearning

Expectations in Computer Vision & AI

> The notion of "difficulty" is very different for a Computer Vision system or us humans. This is one of the main points which leads to wrong expectations in the field of AI.

Engineers and researchers have to be realistic and educational about the performance of systems in open domains.

We showed that there are problems in understanding the progress of AI systems today. Take **Autonomous Driving** for instance: there is a vast difference between being able to drive in well constrained definitions (i.e. motorways) versus driving in open domain, under any condition (i.e. within cities, small roads, …). Most of the industry today evaluate autonomous driving progress based of the number of miles driven without alerting the driver. This creates an incentive on putting the cars in easy conditions, while we should instead have metrics which focus on broadening the scope in which they can operate correctly. More generally,
Very narrow problems may be solved provided you have enough labeled data, well defined and constrained classes. But incorporating commonsense knowledge of the world to computer vision systems is still a complete challenge.



Q: Are there an equal number of large things and metal spheres? **Q:** What size is the cylinder that is left of the brown metal thing that is left of the big sphere? **Q:** There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere? **Q:** How many objects are either small cylinders or metal things?

ClevR, a dataset for evaluation Visual Question Answering and Visual Reasoning

Plenty of researchers acknoledge this and several interesting research fields are booming these days, such as Visual Reasoning & Grounding, Laws of Physics Discovery, Representation learning through unsupervised/self-supervised learning, or even reinforcement learning tasks starting from raw pixels (if interested, see references below).

Finally, this was about Computer Vision as this is where I have most experience, but I believe the same reasonning applies to any Machine Learning, especially Deep/Machine learning based NLP.

Charles Ollion is CoFounder & Head of AI at Heuritech. Also teaches Deep Learning at Master Datascience (Ecole Polytechnique/Paris Saclay) and EPITA. *Thanks to Hedi Ben Younes and Alexandre Ramé for helpful comments!*