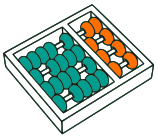

The Revolution will not be Supervised by Alyosha Efros

Lesson No. 14

Rosa Yuliana Gabriela Paccotacya Yanque - RA:263068

This work presents a summary of the presentation [1] given by Alyosha Efros.

1 Introduction

Even though there are algorithms or procedures that have a good performance on certain tasks, the results are rather poor when applied to situations that are less specific. For example, algorithms for image classification have better results on the ImageNet dataset compared to images from the real world.

A direct semantic supervision might be harmful and be equal to memorization. Classifiers accomplish to do the task, but they do it typically in the most simple way. There is a dataset bias that will not go away due to the data being finite. In this setting, we must make better use of the data we have.

2 Why not use labels

Using semantics might not be the best way to represent the world. A word can have different meanings that have nothing in common. Thus, the sensory world does not always have a direct correspondence to words, creating a language bottleneck.

Categorization is used for knowledge transfer and communication. However, the classical view of categories is not really shared by humans. Instead of categories, hierarchies could be used, but they also have some problems like intransitivity.

Categories not always naturally arise. The concept space might vary smoothly. The context plays an important role in defining concepts.

If we need to communicate, categories need to be used. That means decisions are made ahead of time. Only when we know the task, categorization should be used and communication should be dropped.

3 Embeddings

Association should be used instead of categorization. If connections are made between things, they could be represented by the features they share.

From the point of view of the learning spectrum (Fig. 1), there are two extremes. We can try to generalize from very few data (extrapolation problem). On the other hand, if we have a large quantity of data, we want to find the closest example to the one at hand (interpolation problem). For this end, the nearest neighbor could be found. However, in the real world, the changing environment makes that method impractical.

Categorization compress the information too much. Humans can remember many details, like what differentiates the state of an object from another state, or exemplars of the same kind of thing. Our memory is driven by an embedding that is much more detailed than categories. These details are non-linguistic and non-semantic. They are in a lower level.

Therefore, machine learning methods should work without semantics and get to that embedding, where things are represented by what is important to us.

4 Self-supervision

Self-supervision comes from this vision. Data is used as its own supervision, but what it is supervised is some structure of the data (meta-supervision).

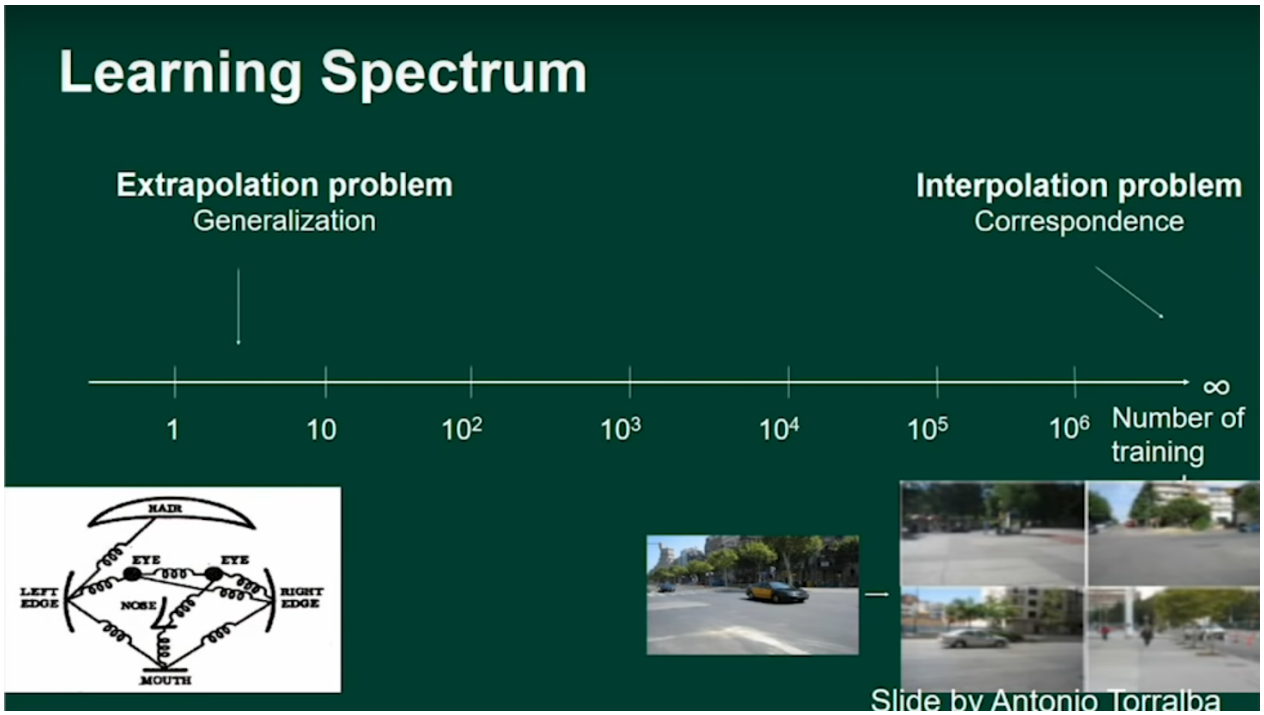

Figure 1. Extremes in the learning spectrum.

An example of this self-supervision can be taken from a patch embedding created by a CNN. Given input data and no labels, it can found the nearest neighbors of, for example, a cat. The concept of "cat" was not taught directly, but the system figured it out from first principals. Context was used as a way to supervise grouping cats together in the embedding space. So, in this way learning can be faster because we classify a whole group in "one-pass".

The first time that the word self-supervision was mentioned was in 1994 by Virginia de Sa [2] that derived the label from co-ocurring input to another modality e.g the image of a cow and mooing. In multi-sensory learning, different sensory inputs can be used. That is the case when talking about learning audio-visual correspondences. One could use random pairs, but they do not require motion analysis, leading to a simplistic learning. However, in time-shifted pairs, patterns in video and audio are shifted apart. The system detects the shift, taking a more complicated and meaningful way to solve the problem of misaligned audio in videos. Another example is the task of on/off-screen audio-visual source separation.

The visual world is smooth and continuous. To learn the correspondence between different features, we can try to detect their association in time using the temporal continuity as a signal.

References

[1] *Week 9 (b): Cs294-158 deep unsupervised learning (4/10/19) - youtube*, https://www.youtube.com/watch?v=PX11C5Vfo9U&feature=youtu.be&fbclid=IwAR116sq-8yaQdGiuS31ooIogh-2EoZwvp3W-mEXPYWM8FeiYArV51Qziqyw (Accessed on 07/08/2020).

[2] V. R. de Sa, "Learning classification with unlabeled data," in *Advances in neural information processing systems*, 1994, pp. 112–119.