Active Learning and Why All Data Is Not Created Equal

How we can build better machine learning models by using less — but more carefully curated — data

Drew Gray Follow
Feb 7 · 12 min read ★

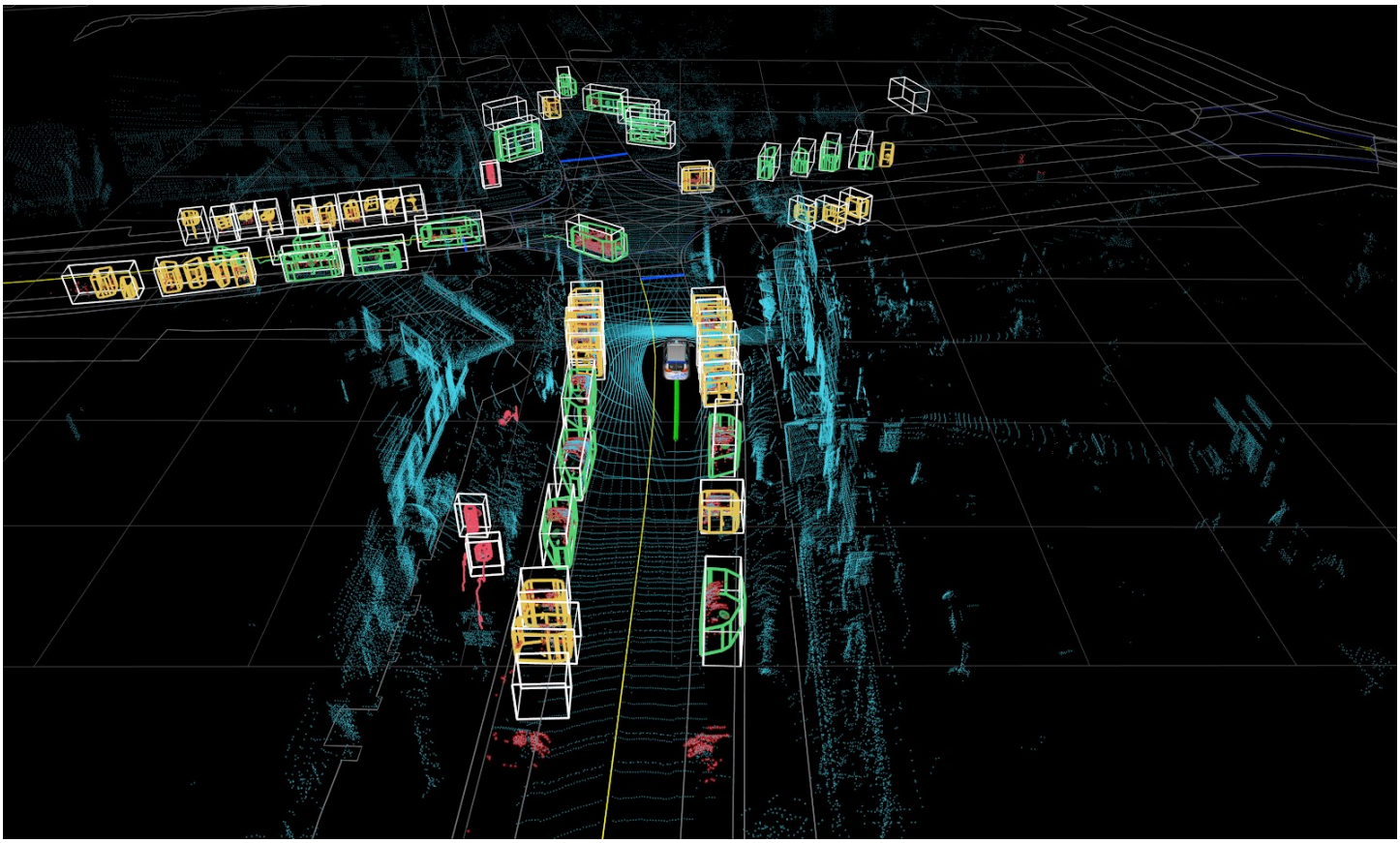

Figure 1: A typical fully labeled scene from a self-driving car. The 3D bounding boxes with class type are the labels

and CEO of Alectio.

The Big Data Labeling Crisis

The field of computer vision reached a tipping point when the size and quality of available datasets finally met the needs of theoretical machine learning algorithms. The release of ImageNet, a fully-labeled dataset of 14-million-images, played a critical role. When ImageNet was realized, it was close to impossible for most companies to generate such a large, clean, and (most importantly) labeled dataset for computer vision. The reason for that was not that collecting the actual data was challenging (in fact, data collection and storage had already gotten much easier and cheaper), it was because obtaining and validating such a large volume of labels was slow, tedious, and expensive.

Data scientists know all too well that data preparation takes up most of their time, yet many people — including seasoned engineers — do not fully grasp the challenges related to data labeling. Data labeling can be viewed as the step that consists of encoding and integrating human knowledge into the algorithms. Getting it right is critical.

Some industries are lucky in that aspect because the data they are working with is “naturally” labeled. For example, in e-commerce, the label itself comes directly from the customer (e.g. Did this customer buy this specific product? What rating did the reviewer give to a specific product?). However, if you are working on a machine

Our Experiment

Finding What Matters

It's no secret to machine learning engineers that more data leads to better models. In fact, increasing the size of their training set is what most researchers do whenever they are disappointed in the performance of their current model. This effect can easily be shown by building a plot called a learning curve. A learning curve is a plot of a model's performance versus the quantity of data trained on. The metric we use to measure model performance for object detection is the industry standard of mean average precision (MAP). People usually build learning curves by incrementally increasing the size of their training set and then retraining their model. The goal is to identify how data quantity impacts model performance. In most cases, the larger samples include all records included in earlier iterations, and hence the represented samples are not mutually exclusive.

The major limitation to a typical learning curve, however, is it doesn't capture the many different ways to sample the additional data to grow a training set. In order to understand more in-depth how the size of our training set impacted our model, we ran a slightly different experiment.

- **We started by creating a small, experimental dataset of 10,000 total frames (much smaller than our overall dataset) to study how our model was performing as a baseline.** Because of the overall size of this dataset, we were interested in finding

Figure 2: Results of our Active Learning feasibility study with Alectio. Representing the relationship between the model's performance and the size of the training set with a regular learning curve has its limits since there are many different ways to select the training samples. Here, each violin plot represents the distribution (or the range) of the MAP values obtained across 30 different training processes run with the exact same amount of data (but a different training sample). We can see, for example, that training with 20 percent of the data might lead to a MAP as high as 69 percent, or as low as 45 percent, which supports the fact not all data is created equal.

Enter Active Learning

Now that we know that curation can help us tremendously, let's go back to our original question. If data isn't equally valuable, how do we find the "good stuff?" One way of doing this is through a process known as Active Learning. While this technique is not new, it is underutilized and misunderstood by machine learning engineers across the industry.

Active Learning consists of adding data volume progressively, re-training the model with a larger batch, and using the trained model to identify which data is the most valuable to add next. The underlying idea is to examine the model's learnings at each iteration. One way to do this is to use the model in its current state to infer on the remaining 'unpicked' data and select the records that were predicted with the highest level of uncertainty. A fairly popular way to do this is to select the records inferred with the lowest confidence — a strategy known as "least-confidence." We usually stop just as we reach the desired performance (or when no more improvement is observed as per the learning curve). Intuitively, it is like teaching a child something new by using only a small set of examples, before testing them to identify their weaknesses in order to know what examples to show next.

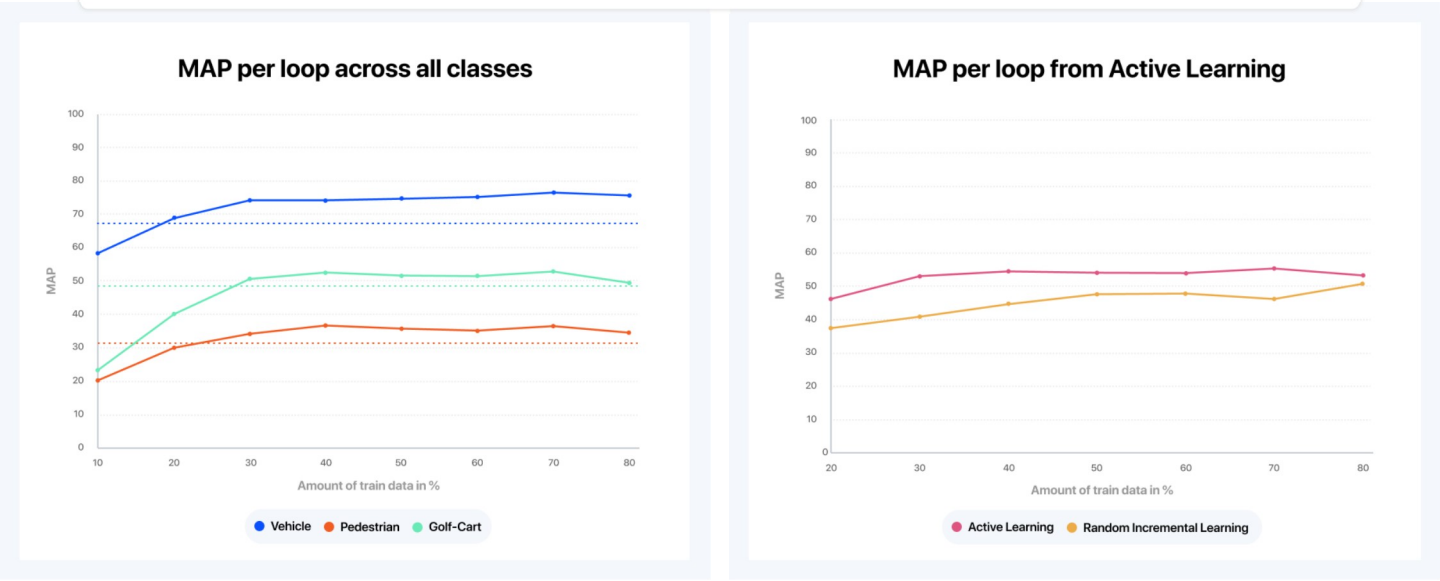

Figure 4 (left): Learning curve split by class using Alectio's Active Learning strategy. Figure 5 (right): A comparison of a random learning curve (created by randomly selecting the next batch of data, as is normal practice) versus the smart Active Learning strategy.

Looking closely at Figure 4 above, our querying strategy seems to be doing a really good job. We are able to reach the nominal accuracy with only 30 to 40 percent of the data. Figure 5 gives us an alternative view on the same data. It shows the derivative of the learning curve to compare learning speeds between classes. A few interesting observations can be made from the graph:

- The golf cart class, while not the most accurately predicted one, is the one that the model learns the fastest; this is followed by the pedestrian class. This is due to the fact that the vehicle class is already reasonably well understood after the initial random loop (as vehicles are the most ubiquitous in the dataset)
- When going from 30 to 40 percent of the data, the pedestrian class becomes the class the model learns the fastest
- Eventually, all of the classes end up at the same learning rate
- Finally, above a certain point, it looks like adding more data is hurting the performance of the model for pedestrians and golf carts (given the limited size of this experimental dataset, it is likely overfitting on the training data)

Rapid Progress for Our Machine Learning Models

We have seen in this study that the application of Active Learning matches well the intention of a human who would try to dynamically select useful data without the need to label the data. But, this project has actually allowed us to do much more than just explore opportunities to label our data more efficiently. Thanks to the insights provided with Alectio, we are able to tune our data collection process in order to get more of the data that truly matters to our model. We're also able to better understand how our model works, and eventually to iterate on it in a more educated manner.

Until recently, we thought of Active Learning primarily as a way to reduce labeling costs. Now we know there is much more to the story — Active Learning can help us improve performance, increase the speed of development, and reduce R&D costs. We can't wait to see the impact this research will have on our self-driving technology.

Thanks to Justin Erlich.

Machine Learning Robotics Autonomous Cars Startup Autonomous Vehicles