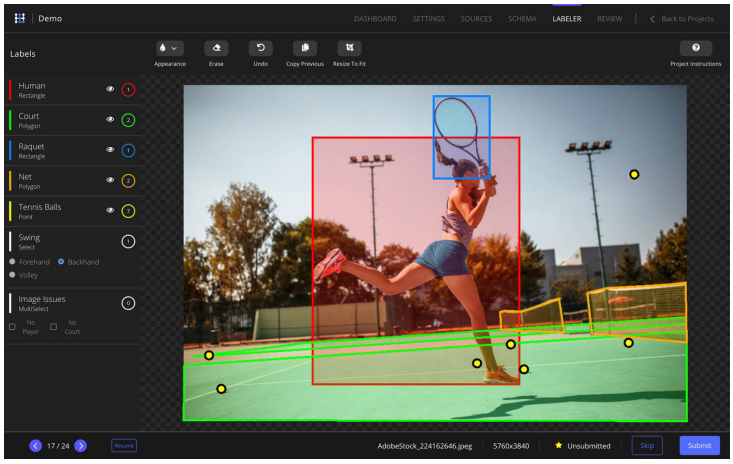
DATA LABELING

PAIN & LABEL:

Why and How We Built our Own ML Data Labeling Tool and Released it Free for Everyone



BY LOGAN SPEARS // INNOVATION CHIEF, SIXGILL, LLC

The Machine Learning (ML) revolution is here. It seems like every company and technical team wants to join this new wave of innovation. But what's the first step?

At Sixgill, after setting out to infuse ML capabilities throughout our data automation product suite, we hit an obstacle that surprised us. It wasn't figuring out ML itself. Nor was it defining best practices for deep neural net architectures, activation functions, or data augmentation techniques. That type of information is readily available. It wasn't even putting developed models into production; we were able to quickly deploy new models on multiple clouds, trained and served from Python, Javascript, and Go.

The bottleneck in our process was creating high-quality datasets that we could use to train our ML models for novel use cases. *Sound familiar?*

After exploring many different solutions and experiencing mostly frustration, we ultimately realized the best way to overcome this obstacle was to build our own data-labeling tool. While we built it to solve our own pain, we deeply understood the pain our peer developers, data scientists and engineers were also feeling. So, we launched our data labeling application publicly to share our faster, easier, better way to create visual ML training datasets. It's free to download and use with no user limits on label quantity.

We call our tool HyperLabel. We built it to be fast, easy to use, flexible for diverse use cases, and powerful enough to scale to high volume workflows.

We call our tool HyperLabel. We built it to be fast, easy to use, flexible for diverse use cases, and powerful enough to scale to high volume workflows.

And we made it a desktop application that uses ML itself to accelerate data labeling and end-to-end encryption, while protecting data privacy and IP for users.

But the decision to build HyperLabel wasn't easy. Here's some insight into our journey to launch, and the key questions we asked ourselves along the way.

SHOULD WE USE OPEN SOURCE OR NOT?

Like most startups, we first tried to solve our problem using open source solutions. A few seemed fine at first, but quickly proved insufficient once our needs became more complex and demanding.

For example, let's say you're trying to detect people in an image. To gain higher detection accuracy, you'll probably retrain Coco SSD mobile net, YOLO, or other object detection models. This type of training requires manually drawing bounding boxes around people in countless images. Unless you want to devote expensive engineering time to this mundane task, you're likely going to outsource it.

That's when things start to turn ugly, as these open source solutions are simply hard to use.

Engineers can figure them out pretty well, but people with less of a technical background need training. Lots of it. Most non-engineer users stop at "first install python". We had to write training materials for these users, and even then, the results were not ideal.

The second issue we ran into with open source solutions proved to be even more dire. Labelers often ask questions such as, "this image is mostly black and hard to see. How should I label it?" or "the image is rotated. How should I draw the boxes?"

After hearing these kinds of questions over and over, the light bulb finally clicked on and we realized that we needed to label these issues themselves and treat them as a multi-class classification problem. But when we went back to our labeling software, we realized that it didn't allow users to associate the data in a single workflow; instead, it required multiple tools to accomplish the task.

SHOULD WE GO COMMERCIAL?

After striking out with open source solutions, we tried a few commercial ones, including two of the leaders in the space. Both offered a highly configurable schema and customizable interface, which helped with the above-mentioned multi-label workflow problem. But there were significant downsides.

With both products, it was shocking how quickly we were forced to talk to the sales people about forking over some cash in order to keep going. One of them cuts off its free tier at 2,500 "labeled assets per year," which we quickly hit. With the other, it took us several calls to even get pricing, which seemed like a big waste of time.

And neither product was cheap. One charged $1,000 a month just to use its labeling software. The other was even more expensive. To us, the prices seemed astronomical for a SaaS offering.

Another huge downside was that we had to trust them with our datasets and labels, both of which are valuable IP for us. That was a deal breaker by itself.

In our experience, many companies—especially enterprise-scale organizations—are categorically not permitted to trust third parties with their data unless they perform extensive due diligence. This extra requirement would have hindered negotiations with our own clients and potentially cost us business. While the products we tried do offer on-premise installations, the cost is prohibitive.

MAKING THE DECISION TO BUILD OUR OWN APPLICATION

After facing these challenges and finding inadequate solutions, we decided that the best way to get what we really needed was to build it ourselves. With our own ML data labeling application (HyperLabel), we wanted to include all of the things we didn't get elsewhere, including these important qualities:

- **Make it easy and intuitive** for non-engineers to use, and get from project setup to label export in just a few steps (5 in our case).
- **Provide flexible schema selection** to enable quick iteration across various data labeling workflows. Included custom schemas are: rectangles, polygons, point, feature points, free text, select, and multi-select.
- **Let users to control their own data.** Users can import files from local drives or cloud storage – no need to use any external service.
- **Give it scalability** that allows users to manage labeling projects of almost any size or complexity.
- **Provide easy export** in formats such as JSON, COCO, Pascal VOC and YOLO.
- **Make it free.** By removing the cost barrier for the Developer version, we're making quality ML training datasets an easy reality for anyone.

As we iterated our own product, we also discovered a need to synchronize datasets and labels across machines, with end-to-end encryption. Other features that we feel are highly desirable and are currently working on include deep learning-based object tracking for speeding up video labeling, pretrained object detectors for identifying objects with additional taxonomy, and the ability to use GANs to guess object or scene outlines for semantic segmentation.

We're passionate about realizing the great potential of deep learning ourselves, as well as removing the pain from labeling for data experts and developers around the world.

We can't wait to experience and share the innovation yet to come as HyperLabel helps others take the fastest path to Machine Learning.

Follow HyperLabel on Twitter @HyperLabel

ABOUT THE AUTHOR

Logan Spears is Innovation Chief at Santa Monica, CA-based Sixgill, LLC, a provider of universal data automation and authenticity products and services for governing IoE. Since joining Sixgill in 2016, he has worked on and led nearly every part of the Sixgill stack, including mobile, server and ML.

ABOUT SIXGILL

Sixgill provides a full suite of universal data automation and authenticity products and services that enable organizations to govern IoE assets.

With the Sixgill® product suite, organizations easily acquire, analyze and act on IoE data, at any velocity or scale. Meeting the increasing necessity for end-to-end sensor data management, process automation and analytics for sensor-informed operations, Sixgill offers Sense™ for sensor data enrichment and automation, Sense Vision™ for ML-based camera data intelligence, and Integrity™ for blockchain-based authenticity.

HyperLabel™, by Sixgill, is a complete application for creating, automating, updating, and managing annotated datasets for Machine Learning. To learn more, visit Sixgill.com.