Handling Data Sources for Effective Machine Learning

Machine learning models are highly reliant on training data. In this talk, I would like to share my personal experience working with various data sources and the lessons I've learned along the way. There are two crucial aspects of data: quality and quantity. When it comes to data obtained through human raters, there is an additional practical factor to consider: economic feasibility. In the talk, I will discuss how to strike a balance among these three factors

Types of Data Sources

There are two major types of data depending on their source that require different processing approaches before they can be used for machine learning. First, there is user-generated data. This can be obtained by carefully logging user activity and behaviour. For instance, in online advertising, user actions like clicks, conversions can be collected, logged and finalised as labelled data.
Second, there is data where no labelling originally exists, however there is still a need to train an ML model. One way to tackle this issue is employing human annotators to label data manually. For example MNIST dataset, one of the earliest computer vision datasets compiled and labelled by humans.

Another strategy of handling non-labelled data is useful when there are no resources for staffing enough annotators to label the training data. In such a case, we can opt for the proxy-label method. For example, users can report content that they find inappropriate and then it can be labelled as such, although sometimes such data can be noisy and might require some attention of the annotators' team.

Let us focus on the ways to handle the most challenging type, the non-labelled data.

Gathering Datasets with Human Annotators

Building datasets employing human annotators is a commonly used technique in machine learning. Its two main limitations are costs and time consumption.

Costs can turn into a major challenge if a dataset is large and processing it means hiring many people. Also, the nature of data may require a high level of expertise and, consequently, employing expensive specialists. For instance, annotating medical images or legal documents can only be done by highly trained staff.

Time can be an issue for several reasons. First, it is obvious that large datasets take a long while (or an impractical number of people) to process. Sometimes it can also be difficult to recruit annotators willing to engage in long-time projects. Second, annotators should be adequately trained to deliver the desired level of quality. Such training can be time consuming, especially when high expertise levels are required.

Here are the lessons I learned while working on an image quality classifier for IG&FB shop. Our solution was built to automatically detect blurry, cropped, and images with an unprofessional background. In order to accomplish this, we needed to gather a training sample. Here is what made the process more effective

- **Annotate data in batches**
  As always, we had a limited budget for human raters, so splitting the annotation process into batches was extremely useful. Initially, our raters weren't producing high-quality results, and when we discovered this, we had only spent about 5% of the budget. To address the issue we collaborated with the team of annotators to refine the guidelines and process in order to reduce the number of errors (more on this in the next section). Detecting the data quality issue early on allowed us to allocate the remaining 95% of the budget to a well-trained team of human annotators.
  Another advantage of sending data in batches is that it enabled us to implement active learning. If we had sent the entire dataset to the annotators all at once, we wouldn't have been able to utilise this strategy

- **Sample batches with active learning**
  Random Sampling is generally a good initial approach, but with this particular project there were two problems:
    - The negative class was overrepresented, because good images were a majority in the database
    - Also, with this approach we could not easily add “difficult” images (with borderline predictions of about 0.5) to a training sample

  The solution was to sample with Active Learning. We followed three steps:
    - Randomly sampled some images
    - Got predictions for every image using the current model
    - Sorted images by predictions in descending order, sent images with the highest probabilities OR borderline images to human annotators

  Switching from random sampling to active learning strategy enabled us to increase percentage of the positive class:
    - 2x for images with an unprofessionally made background
    - 4x for partially displayed images
    - 33% for blurry images

- **Track annotators’ quality**
  It’s quite common when working with human annotators to collect several responses

per item and assign the final label using the supermajority rule.
For example, let's assume we collect three responses per image and if two answers indicate that the image is blurry we assign it as the final label. Intuitively it seems like this procedure should significantly increase accuracy of the data.
However, if we make simple calculations we will see that the increase in probability of the final label to be correct is small and it's more important to work on the accuracy of the individuals.
Let's assume each annotator has a minimum probability to give a correct answer for a given question; let us put this probability as *p*. Now let us calculate what is the minimum probability for at least two raters to give a correct answer; we shall put this probability as *q*. Then, *q* is a sum of probabilities of two events: all three raters giving a correct answer (p^3) and any two raters giving a correct answer 3*(1-p)*p^2. In the table below we can see how *q* changes depending on *p*.

Annotator's minimum probability of providing a correct answer (*p*)	Final label's minimum probability of being correct (*q*)
0.6	0.648
0.7	0.784
0.8	0.896

As we can see, the final probability (*q*) does not change much if we apply the supermajority rule with a minimum of three responses. That is why it is very important to track the quality of each annotator.
In our case the following process helped

- Introducing a "golden set", created by well trained annotators. We used this set to calculate accuracy for each annotator assigned to our project
- Having a biweekly AMA where annotators could ask the questions about controversial images
- Introducing final exam and minimum performance threshold. An annotator could only start rating if they pass the final exam and got the score higher than threshold

Reducing human involvement

Sometimes, a human-based approach may fail or become too impractical to employ. In many cases, there might be an elegant solution to bail you out. Let us have a look at some of them:

- **Proxy Values**
  Sometimes we could be creative and use a proxy value for a label instead of building a dataset employing human annotators. For instance, we used well known brands as a source of high quality images, it was especially beneficial for "unprofessional background dtection"

- **Data augmentation**
  In this case, data from an already processed dataset is reproduced in an altered form and then fed back into the model. Working on my image quality project, I took the existing good photos and used basic graphic tools to make the good photos bad. In particular, I blurred the good images and added these "bad" images to the training set. I did the same with randomly cropping the images as well, so it displayed only part of the product. These simple transformations brought me +8.08% to ROC AUC.

As you see, human input in machine learning is essential. However managing it may be a matter of survival for many tech projects. The rule of thumb is opting for less but more qualified staff, setting reference datasets and extracting as much as possible from actions of users, employing them basically as free annotators.