**Title:**

**Automatic Labeling of Data – Why and How?**

**Abstract:**

While data is the "fuel" of Machine Learning models, labeled data could be a vital component for success. We present a method for automatically labeling data, which could be very useful for model training as well as resource conservation.

**Description (talk outline):**

In the world of artificial intelligence, labeled data is virtually the most basic commodity – images categorized by classes, objects in an image marked according to their location, categorization of sentences and documents, etc. There is a great need for labeled data but labeling itself is a very challenging task – manual labeling is routine and tedious, and obtaining quality labeling of large quantities of data quickly is difficult, and therefore expensive.

We focus on the domain of computer vision, and specifically on classification and detection. In order to train a model to classify an image according to objects that appear within, there is a need for image data labeled in such a way that each image is accompanied by a label identifying the object’s class. Training a model to recognize the position of the object in the image as well, requires a different data labeling – label not only the object’s class, but accurately mark its position. For example, bounding the object within a rectangle. Marking such data manually is significantly more challenging. What's more, there can be several objects in the image, and we would like to identify the location of each one. Although methods exist that try to reduce the dependence on labeled data, such as semi- or self-supervised learning, often the results of such models are less accurate.

To overcome this lack of data without sacrificing the model’s quality, we propose taking another approach – labeling data automatically. The automatic labeling is based on the combination of a classical algorithm with a deep learning algorithm. This work emphasizes two main principles (along with open-source code):

1. The idea behind automatic data labeling – Why it's important and how to do it qualitatively.
2. The combination of a classical algorithm with a neural network-based algorithm – Why it is necessary and how to perform it.

For the purpose of illustration, we focus on identifying the locations of people’s eyes within an image, including the presence of multiple people in the image. Motivation for this may be training a real-time camera detector to focus on the center of people’s faces (i.e. the area of the eyes) when photographed in order to obtain optimal photo quality.

In practice, no data providing such a detailed quality labeling exists, and manually generating such a labeling for tens of thousands of images is an almost impossible task that would be very costly in resources. Existing solutions, and combinations thereof, can be utilized to automatically generate such a labeling on a large scale. Today, there exist more than a few methods for extracting landmarks, from which the exact location of a specific objects can be deduced. Due to inherent issues encountered in complex situations (e.g. identifying faces wearing masks, semi-occluded objects), their results are unsatisfactory. Therefore, it is possible to combine a classic algorithm built on identifying object boundaries, together with a deep learning-based algorithm built on finding an image’s feature map, to qualitatively identify landmarks. The combination of the two also gives a measure for when the algorithms fail, so we know in such cases not to use the labeling. Overall, auto-labeling allowed us to collect many labeled objects to use for our overall model.

Once landmarks and labeling have been extracted and their quality has been verified, they can then be used for downstream tasks. Going back to our camera detector example, we can extract facial landmarks which are converted to bounding boxes of the eyes. We then train our camera detector to identify eyes using this labeled data.

We supply a simple framework of an easily run Python file (open source) that retrieve the requested labels. From this, we can learn how to qualitatively auto-label data, and how combining a classic algorithm with a neural network can lead to improved results and reliable labels, which can later be used. If one requires large quantities of labeled data and does not have the resources to collect it manually (either independently or by purchase) – they needn’t settle for solutions that have less need for labeled data and thus harm quality, but can attempt auto-labeling data themselves.