EVOTEGRA
CONCEPTS & INTEGRATION

Semi-supervised data labeling for Deep Learning

Human compared to machine labeling based on the Tsinghua-Tencent 100k dataset

Abstract: The costs related to training data currently prevent a wide adoption of AI. Semi-supervised data labeling has the potential for significant cost reduction. After a short introduction over the current state of AI and an analysis how machines learn compared to humans, we compare the performance of human and machines in labeling visual data on cost, quality and quantity level. The evaluation is based on the public Tsinghua-Tencent 100k dataset and analyzes the typical errors of both machines and humans.

1 The data aspects for Deep Learning

1.1 The current state of AI

Today's AI technology is fit for production. Neural networks can be reliably trained and executed, C++ and network optimization enable reliable operation and process integration. A wide range of available hardware, from embedded systems to high-end data center solutions, enables the use of AI solutions in a large number of use-cases and scenarios.

1.2 Data is key

While deep learning scales almost infinitely with additional data, it typically performs worse than classic computer vision methods with insufficient amount of data. To solve a general recognition problem with deep learning, we typically recommend 1000-10,000 training samples *per class*.

So while all other requirements are met, the required amount of data remains the main obstacle to the introduction of AI.

We currently see 3 different options that attempt to solve the data problem:
1.) Sparse Modeling
2.) Transfer learning
3.) Deep learning with highly automated data acquisition

1.3 Better results with less data?

Both sparse modeling and transfer learning try to significantly reduce the amount of data required to train an AI. The basic idea is to mimic human learning. But to better understand the potential of these methods it is important to understand the key differences in the learning process of both humans and machines.

1.4 How humans learn

The human brain has around 86 Billion neurons aligned in a 3 dimensional fashion. While it takes humans years to develop a basic understanding of their environment, humans acquire knowledge in a continuous learning process throughout their lifetime. This prior knowledge enables humans as adults to learn new content from a low number of new samples.

1.5 How machines learn

Today's weak AI typically has just 5-150 million neurons aligned in a 2 dimensional unidirectional structure. With less than 0.17% the capacity of the human brain neuronal networks have no prior knowledge and can learn only from the training data. The learning process is limited to a maximum of a few days during the training stage.
Without prior knowledge lean data methods are mathematically limited to the information content or entropy of the training data. The lower the entropy or the amount of data the higher is the risk for the AI to learn pseudo features. Pseudo features describe the different classes within the training data, but not in practice.

1.6 Human Labeling

Our estimation is that a single person can label up to 500 images per day. The issues affecting the quality are the monotone workload as well as the subjectivity of the individuals. Human perception is generally affected by factors like assumptions, fatigue and mood. Additional factors are both the limited attention span and the limited capability to deal with complexity. Therefore we expect the quality of human labeling to degrade with the higher number of human annotators and classes. Important to understand is the implied cost of human labeling, even if we just consider quantity as a measure. To label a minimum production grade dataset based on the 157 classes in Tsinghua-Tencent 100k, human annotators would require more than one man-year effort to label 157,000 signs.

1.7 Semi-supervised machine labeling

Semi-Supervised labeling is used to build high quality dataset in the order of millions of unique training samples and hundreds of individual classes. With a batch size of typically 100,000 unlabeled images it is an iterative process supervised by humans. Starting point is typically a public network, dataset or manually labeled data. But before data can be automatically labeled, the data pipeline needs to be setup. Depending on the project phase and requirements highly automated labeling is using different techniques based on (un)supervised machine learning and even classical algorithms. The effort for each process iteration is initially higher and decreases over time. But on average a single person can supervise the labeling of 10,000 to 20,000 images per day.

1.8 Data value

The value of training data for AI must be evaluated on quantity, quality, information content and distribution level. While low quantity can lead to poor generalization and detection performance also incorrect or missing labels negatively affect the learning process and hence prevent to leverage the full potential of the data.
The performance of individual classes typically follows the data distribution. So while the uniform distribution is the ideal distribution for deep learning data in reality the data is typically distributed in a Gaussian form. While imbalance can be addressed to some extend, very unbalanced datasets negatively impact the accuracy of underrepresented classes.

2 Tsinghua-Tencent 100k dataset

According to the authors of the Tsinghua-Tencent 100k, the dataset provides 100,000 high quality images at a resolution of 4 megapixels from diverse scenes in China that contain 30,000 traffic-sign instances. These images cover large variations in illuminance and weather conditions. Each traffic-sign in the benchmark is annotated with a class label, its bounding box and pixel mask.

The dataset covers 127 base classes and is divided into 3 categories:

Warning signs | Prohibition signs | Information signs

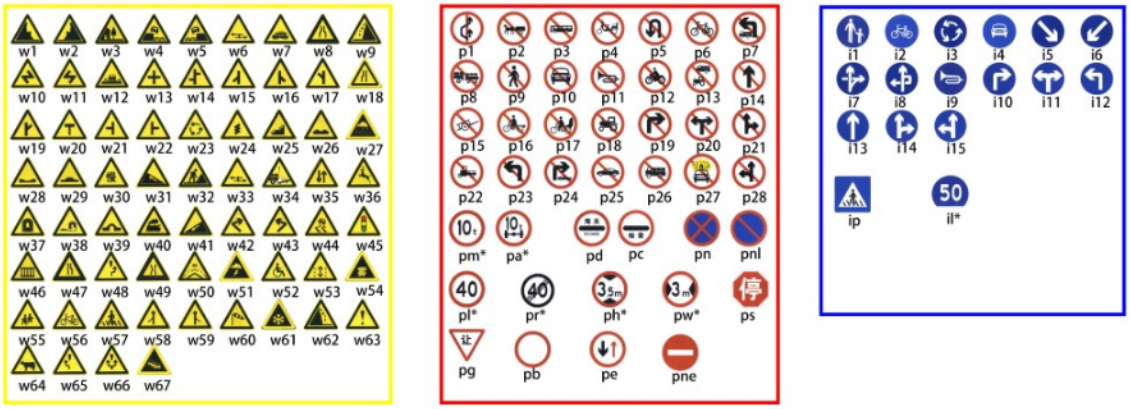


Figure 1 **Overview over the dataset**

As classes marked with a * are additionally separated by the individual face value of the sign the total number of classes in the dataset is 157. For information, warning or prohibition sign that do not fit to any specific class, each category has an additional unspecific class "unknown".
Looking into the dataset we found 22,235 labeled instances of specific signs and 2067 label instances in all 3 categories of unknown signs.
The dataset is split into 9176 positive images containing signs and 82,094 negative images that are not supposed to contain any sign.

2.1 Information content

Quantity, resolution, diversity and the overall quality of the images in the dataset are in fact very good. Therefore the potential information content of the dataset is considered to be good for training high quality deep learning networks.

2.2 Distribution

The distribution of the training data is as expected very uneven. The average number of instances per class is 143 with a standard deviation of 367.
With 72 classes having less than 10 training samples almost half of classes are not suited to train a deep learning algorithm. Additional 43 classes have less than a hundred samples, which is still a critical low number. Just 5 classes fulfill our recommended minimum quantity of 1000 training samples.

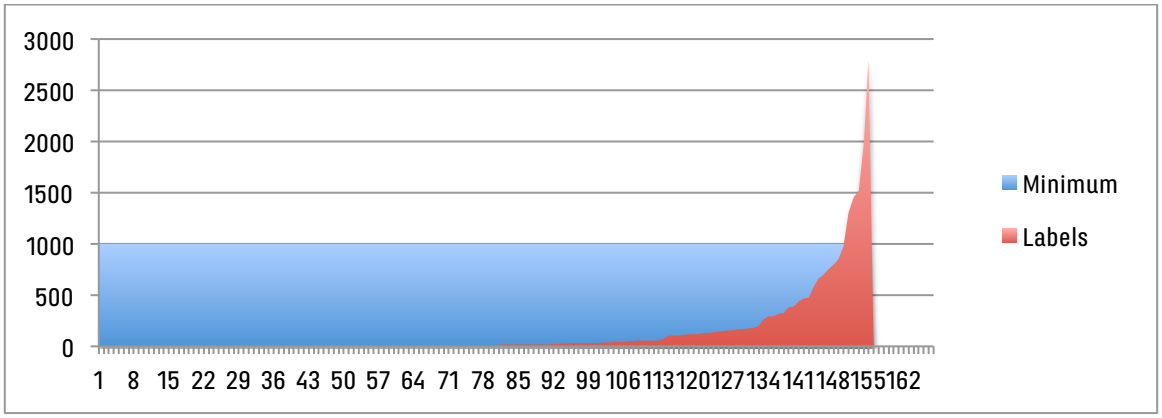


Figure 2 **Data distribution over classes sorted by quantity**

If we apply the recommended quantities on the total 157 classes of the Tsinghua-Tencent dataset this would mean a total of 157,000 to 1,570,000 million instances. While the authors of the Tsinghua-Tencent 100k consider the dataset to be a "large benchmark" it represents just 1.5% to 15% of the recommended amount of data for a production grade dataset. Most classes are in fact below a critical threshold that would allow an objective evaluation of different Deep Learning algorithms.

2.3 Semi-supervised labeling of the positive dataset

The batch size for a single iteration in semi-supervised labeling typically has the same size as the Tsinghua-Tencent 100k dataset as a whole. Therefore in order to demonstrate the potential of semi-supervised labeling on large-scale datasets (>1 Million) we choose a pre-trained network with a quality you achieve typically after a couple iterations.

The total effort to label the 9176 positive images semi-supervised in 157 classes was 2 days.

2.3.1 Quantity

While human annotators labeled 22,235 specific instances of traffic signs, semi-supervised labeling found 24,863 instances, which is a surplus of 2,628 instances or 12%. The trend was valid for any group of signs.

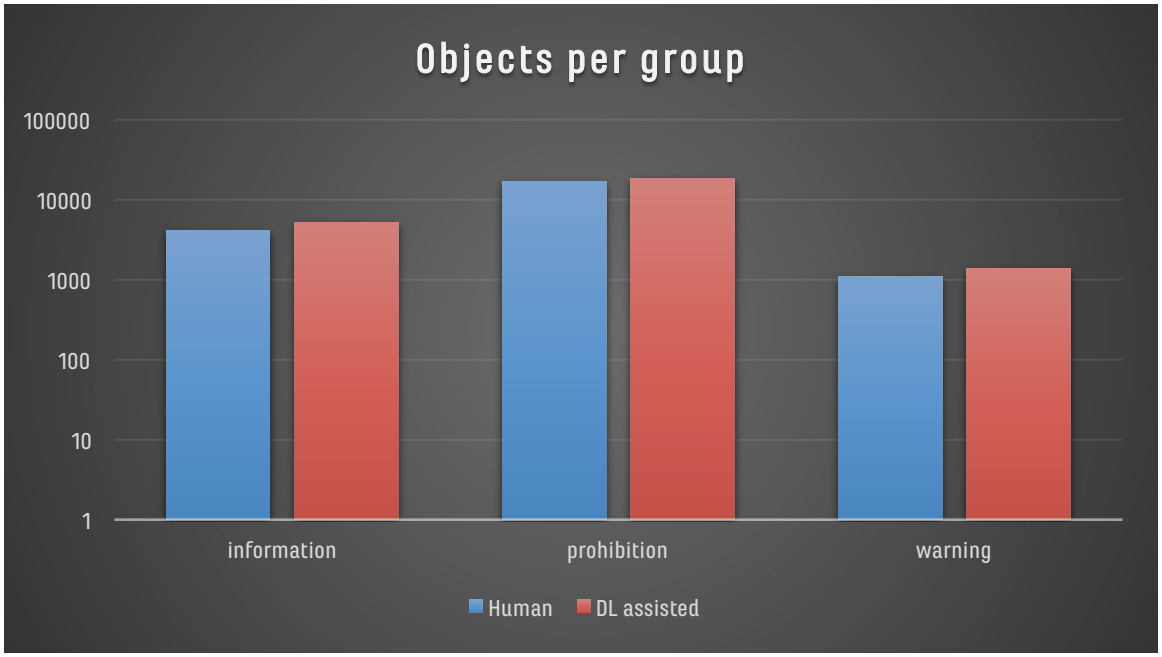


Figure 3 **Labeled objects by group on log scale**

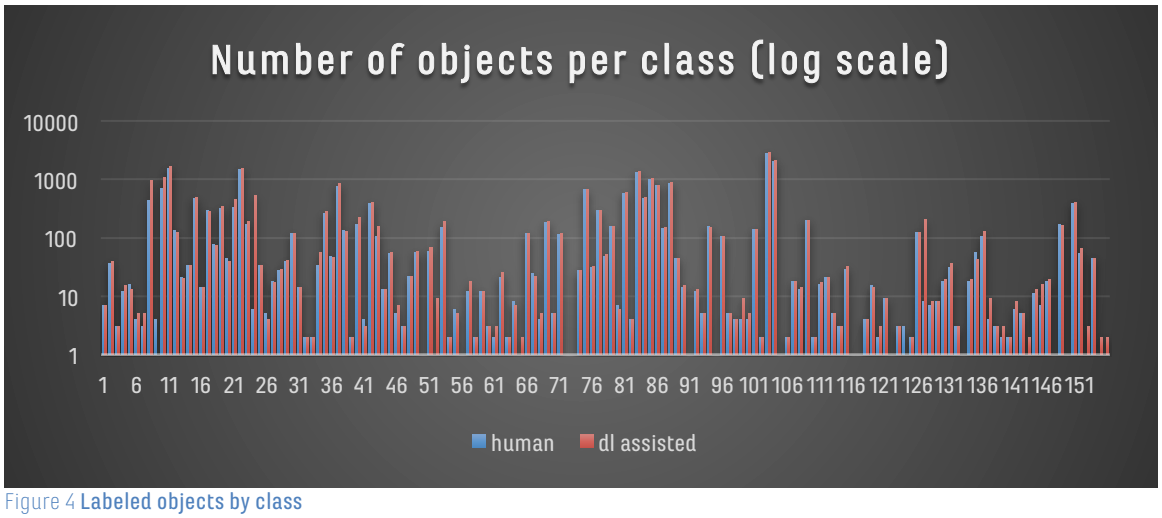


Figure 4 **Labeled objects by class**

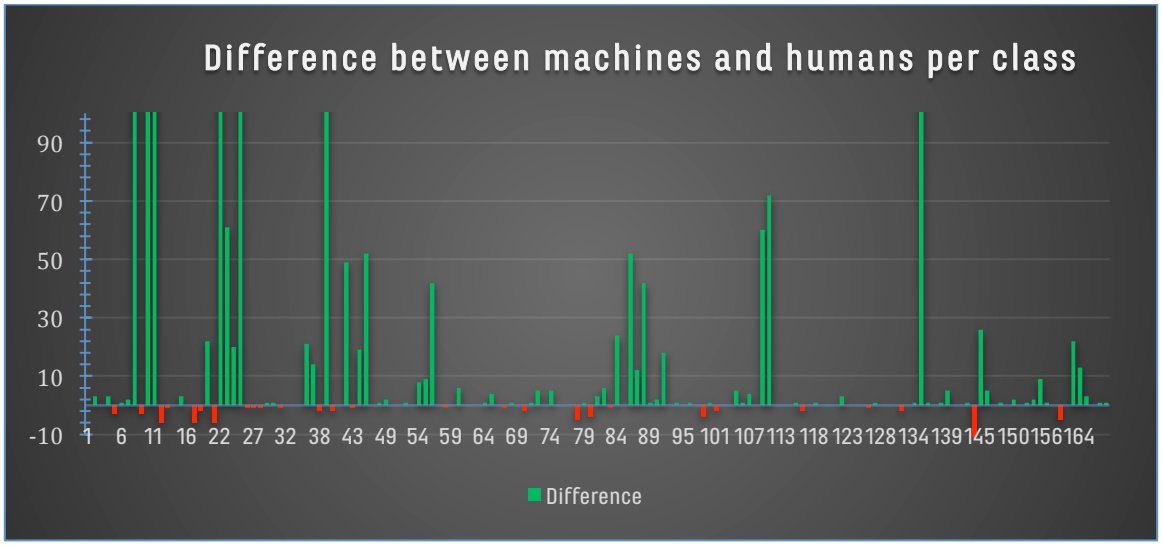


Figure 5 Difference between machines and humans per class capped at +100

2.3.2 Quality

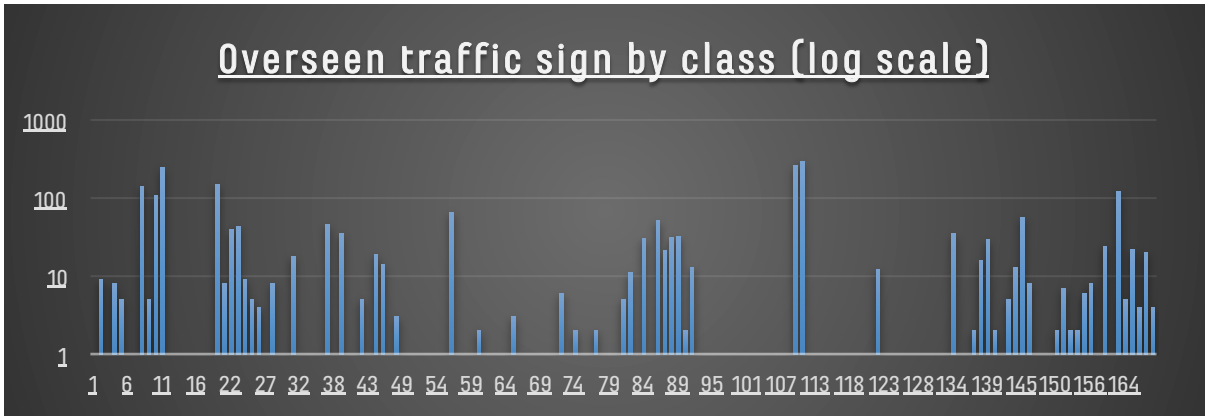
To evaluate the quality we used a Deep Learning classifier to automatically verify the human labels. The deep learning classifier found in total 1699 wrong labels which results in a human accuracy of 93%. The predominant error was to annotate known classes as their corresponding "unknown" counterpart. The accuracy of the Deep Learning classifier was 98.1%.

Figure 6 Examples of wrong label

2.4 The negative dataset

According to its definition the negative dataset should not contain any instances of traffic signs. Analyzing 82,094 negative images took 2 days. As a result we found 2137 traffic signs in the negative

data. This is a human false negative rate of 2.5% or a recall of 97.5%. The AI achieves a false negative rate of 1% or a recall of 99%.



3 Analysis

3.1 Human errors

3.1.1 Class complexity and subjective judgment

Keeping the full overview over 157 classes is a difficult task for humans. Especially rare known signs are prone to be labeled as their "unknown" counterpart. Another reason is subjective judgment related to minor variations of known signs e.g. small arrows below the symbol of known signs.

Figure 7 **Examples of multiple known signs labeled as „unknown"**

3.1.2 Assumption driven perception

While humans have a total field of vision of about 175 degrees the area that allow humans to see sharp is only 0.016 degrees. Only due to rapid eye movement this area is perceived to be larger. Therefore to solve the labeling task more efficiently humans focus the attention to areas where traffic signs are typically expected. However objects that appear in unusual locations are prone to be overseen.

Figure 8 **Humans oversee signs in the side street**

Figure 9 **Semi supervised labeling finds the signs in the side street**

3.1.3 Small objects in complex scenes

Small objects in complex or noisy scenes have an increased risk to be overseen by humans.

Figure 10 **Example of a small warning sign in a complex scene**

3.1.4 Low contrast

In low contrast scenes machines typically perform better than humans.

Figure 11 **Example of low contrast image**

3.2 Machine errors

3.2.1 Very small objects

Humans sometimes label very small objects down to 10x10 pixels despite the fact the face value cannot be objectively determined anymore with high confidence. Sets of images are typically organized in sequences and therefore human annotators sometimes remember the face value of the sign from a prior image. Another effect is that the human brain amends information that is missing from raw perception. This effect can be observed when sometimes we seem to recognize objects better in smaller images than in bigger ones. Yet it deems questionable if such subjective information has positive effect on the learning process. Therefore we generally recommend to not label very small objects.

Figure 12 **Examples of very small sign annotations**

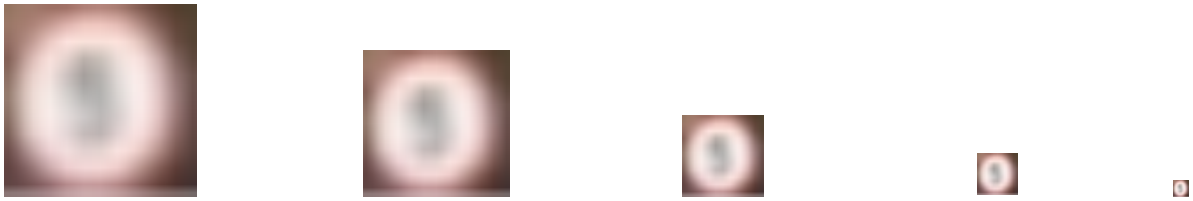


Figure 13 **Same image at different sizes**

3.2.2 Unknown reason

In rare cases Deep Learning fails to detect objects for no apparent reason. This typically happens for classes with a lower representation in the dataset.

Figure 14 **Human sees both signs**

Figure 15 **AI does not see right sign**

3.3 Remaining data potential

While the dataset already comprises 157 classes, many types of Chinese traffic signs are still missing in the dataset. This includes especially highly variant gantry signs which are significantly more difficult to detect reliably than the typical low variant signs in the dataset. Professional networks for Chinese traffic signs detect close to 400 different classes based on several million unique training samples.

Figure 16 **Examples of potential new classes in the data**

4 Summary

With a higher than 70% lower error rate semi-supervised data labeling yields a reduction of up to 90% in effort, costs and duration compared to human labeling. Next to costs this enables datasets in the order of millions of unique samples and hundreds of classes that enable deep learning networks of unparalleled accuracy.
Yet the biggest advantage of semi-supervised over human labeling is data change management. The initial class model is typically based on human experience also known as assumptions. So the probability that a class model requires adaptation during a project is high. With its huge advantage in quantity semi-supervised labeling allows adaptations of class models during a project. So next to improved quality and quantity semi-supervised labeling helps to significantly reduce the project cost and risks.

	Human	Machine	Difference
Classification Accuracy	93%	98,1%	+5,1%
Detection Recall	97.5%	99%	+1.5%
Labeling effort positive dataset	Est. 18 days	2 days	+900%
Labeling effort negative dataset	Est. 40 days	2 days	+4000%

Based on our experience, we assumed that human annotators can process 500 positive images and 2000 negative images per day.

5 Evotegra GmbH

As a full stack solution provider for customer specific Deep Learning solutions, Evotegra provides the full range from consulting and data services to process integration. With the ability to provide low cost solutions at scale we target small and mid tier businesses that do not build up their own AI competence.

Special thanks to Bruno Marotta and Nikolas Markou / Elekti as well as the whole team of Evotegra.

EvoTegra GmbH
Kreuzwingert 11
D – 55296
Gau-Bischofsheim

Tel.: +49 171 / 402 4242
E-Mail: info@evotegra.de
Homepage: www.evotegra.de