ALEGION

Alegion | Understanding Scoring and Quality Metrics for Machine Learning Models

Understanding Scoring and Quality Metrics for Machine Learning Models

Introduction

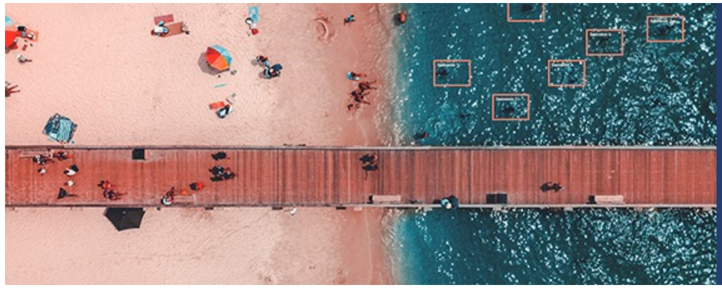
No matter what your project or use case is, the quality of your training data is critical to building an effective model. Quantifiable quality (measuring the quality of your data) allows you to understand and improve the performance of your model through the quality of your training data. Data quality is measured using some fundamental metrics. These metrics assess the difference between a known correct answer (called an authoritative answer, or sometimes a ground truth or gold answer) and an answer being tested (provided by a human worker or generated by a model). Using these metrics across a sample or set of data, and comparing provided answers to known correct answers, will provide an objective measurement for the quality of the dataset.

This article looks closely at individual quality metrics for object localization in computer vision, explaining how they are calculated, and showing how these metrics provide the basis for evaluating the quality of your dataset as a whole.

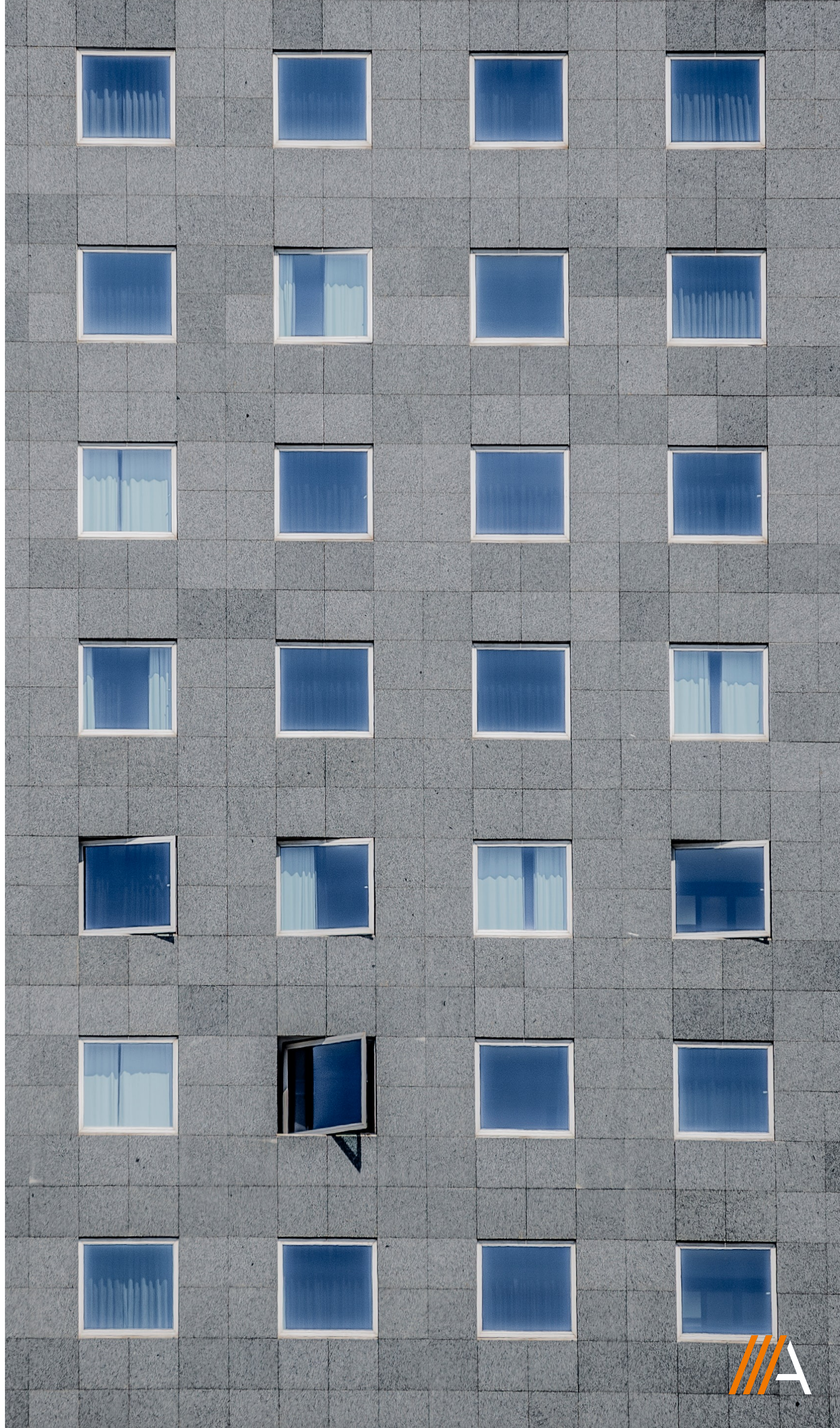
The Difference Between Individual and Batch Metrics

Individual metrics are used to score each instance of an annotation separately. Think of these metrics as the building blocks of quality measurement for your annotated data set. Typically, they are numerical measurements that, when compared to a chosen threshold, provide an assessment for the annotation as correct (True Positive or True Negative) or incorrect (False Positive or False Negative).

Batch metrics are the metrics used to evaluate a labeled dataset as a whole, and they rely on counts of the individual assessments (True Positives, False Positives, etc.). The most common ones are accuracy, recall, precision, and F1 Score (for a detailed run-down of how we calculate these batch metrics, check out our piece):



[How to Get High Quality Training Data from Your Data Labeling Platform]



Ratio vs. Difference Metrics

Ratio metrics measure difference using proportions, while distance metrics measure difference between ground truth answers and other provided answers using an absolute difference on a specific scale (such as pixels in the frame).

Both ratio and distance metrics are important depending on what you are measuring, but using a ratio metric provide a higher degree of accuracy across a wider variety of cases because it is scale-invariant and domain agnostic.

Intersection over union (IOU) is a ratio metric. It is the standard metric used in the machine learning discipline for bounding boxes and other shapes because it applies well to both large and small shapes. It correlates easily to a distance metric like pixel-tolerance. The scale invariance of IOU means that it requires tighter pixel tolerances for small shapes and allows larger pixel tolerances for large shapes.

DISTANCE

- Range is 0.0 to infinity
- Best match is smallest distance, 0.0
- Scaling with image/video resolution:
  - Pixel tolerance is constant at different resolutions
  - The same points (content) in images at different resolutions have different distance error measurements

VS

RATIO

- Range is 0.0 to 1.0
- Best match is smallest distance, 1.0
- Scaling with image/video resolution:
  - Pixel tolerance varies for the same ratio as resolutions changes
  - The same points (content) in images at different resolutions have the same ratio error measurements

Three Important Individual Metrics for Computer Vision

Intersection Over Union (IOU)

IOU is used to evaluate the accuracy of bounding boxes, polygons, ellipses, and 3D bounding boxes.

Intersection = the area covered by both boxes

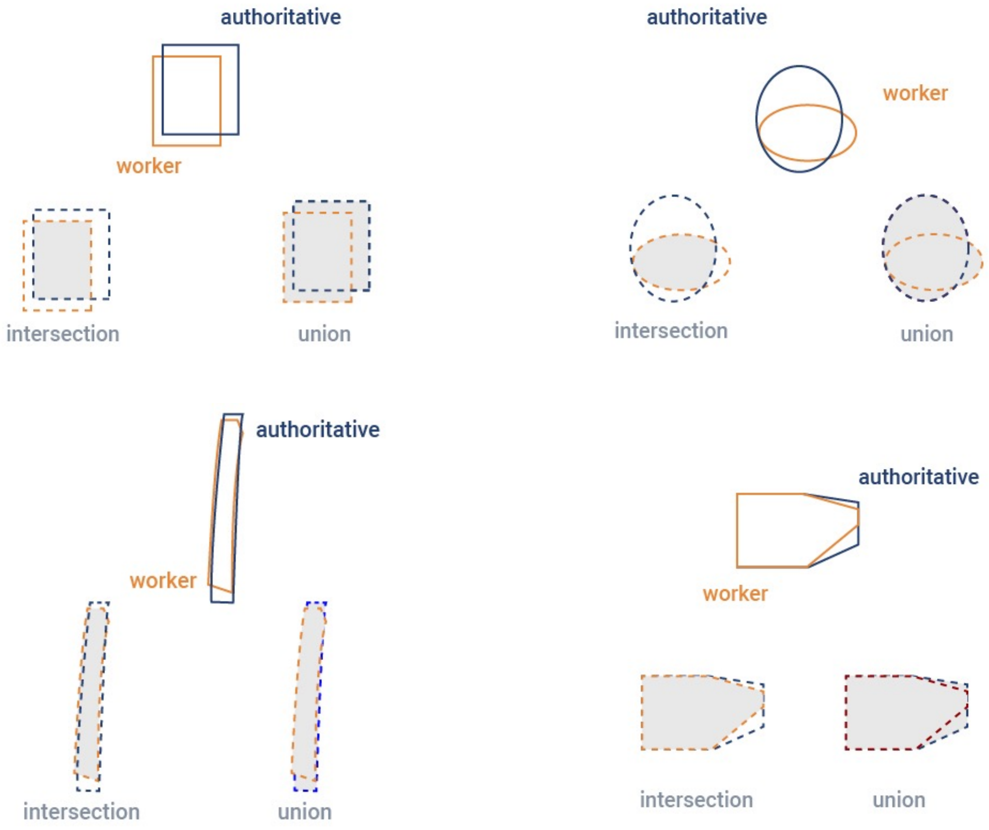
Union = the total area covered by at least one of the boxes

IOU = the intersection area divided by the union area

An IOU of 1.0 is perfect and an IOU of 0.0 indicates no overlap at all. Depending on the use case and shape being annotated, an acceptable IOU threshold can vary significantly. Without specific information about the expected uses cases, a good rule of thumb is to use an IOU threshold of 0.8.

Some conditions favor lower thresholds for shape IOUs, such as ambiguity in an object's boundaries (for example the exact boundaries of a cloud or dirt or the exact boundary between a hand and an arm). If shapes are small (less than 50x50 pixels), then lower IOUs tolerances may also be needed.

Application: You want to annotate food in the grocery store as a cCustomer selects it off the shelf and places it into their cart. Your ground truth data has established bounding boxes around each grocery item in the frame. IOU allows you to figure out how accurately the food items were localized in worker generated annotations for the same frame.



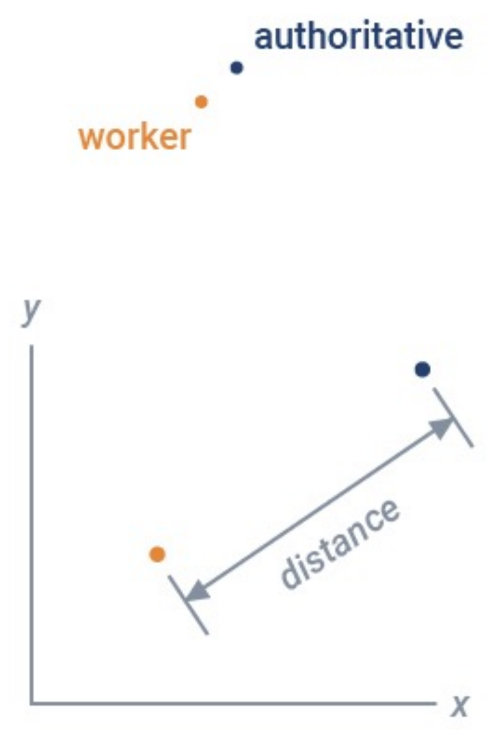
Three Important Individual Metrics for Computer Vision

Keypoint Distance

Keypoint distance is a metric that measures the Euclidean distance in pixel-coordinate-space between keypoints. For a perfect match, keypoint distance = 0.0. The maximum keypoint distance for a given image (diagonal of the image) varies for images at different resolutions. Missing keypoints are assigned a distance of [infinity] away from any valid keypoint. A good default threshold for pixel distance is 6-8 pixels, without specific information about a use case.

Ambiguity in the “correct” position of a keypoint (for example the exact position of a human joint or the center of an irregularly shaped object) favors higher thresholds for keypoint distance. Higher resolution images also favor higher keypoint distance thresholds because the number of pixels between any two visual points is higher in a higher resolution image.

Application: You’re interested in tracking the movement and position of a goalie as she takes a goal kick. Your ground truth data uses keypoints to identify where her shoulder joints, elbows, and kneecaps are in each frame. The keypoint distance metric allows you to evaluate the difference between where ground truth said these joints are located in the frame and where the worker generated annotations say the joints are located.



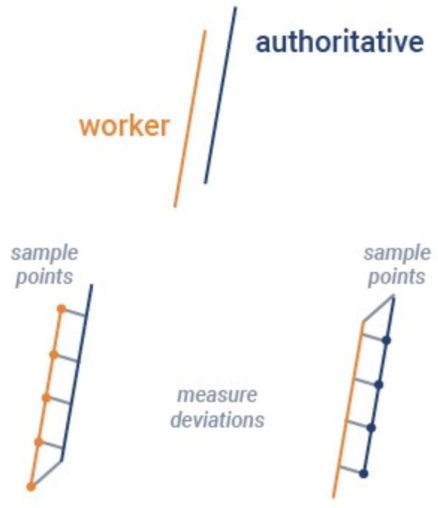
Keypoint Distance

Three Important Individual Metrics for Computer Vision

Root Mean Squared Error (RMSE)

Alegion has developed a unique approach to measuring the accuracy of line localization in images by applying a common statistical metric, root mean squared error (RMSE). RMSE is the standard deviation (square-root of the variance) when an estimator is unbiased (expected value is the true mean value). RMSE is often used to measure the accuracy of a model's fit to data, where the error is the difference in y-values from each data point to the predicted line for the same value of x. It is an excellent approach for measuring differences in both line length and displacement of points on one line compared to another line, and vice versa.

This approach is used for both lines and polylines (multiple connected points along a path that is not restricted to just one straight line). In Cartesian space (a coordinate system that specifies each point uniquely by a pair of numerical coordinates, typically X-Y), for lines and polylines, the error is the Euclidean distance (that is, the length of a line segment) from each sampled point on one line to the other line and vice versa. RMSE is an aggregate measure using the squares of these distances. For a perfect match, polyline RMSE = 0.0. The maximum polyline RMSE for a given image varies with the scale of the image. Missing polylines are assigned an RMSE of [infinity] away from any valid line.



Root Mean Square Error (RMSE)

An RMSE between 8 and 12 is considered good in most cases. Ambiguity in the correct position of a line (for example the midline of an object with unclear boundaries), favors higher tolerances for RMSE. In other cases, a combination metric using maximum and average deviations (using the same sampling/distance procedure) may also be used to determine whether a line or polyline meets the threshold for correctness.

Application: You're interested in training data for autonomous vehicles and you need to track the position of lane lines or lane dividers over a series of frames. Using RSME allows you to calculate how far off the worker annotation of the lane line is from your ground truth data.

Conclusion

Developing the capacity to annotate massive volumes of data while maintaining quality is a function of the model development lifecycle that enterprises often underestimate. It is resource intensive and requires specialized expertise. Your computer vision team needs partners and platforms it can trust to deliver the data quality you need.

At Alegion, we've established quality management best practices based on our experience labeling tens of millions of images, video, text, and audio records. Our breakthrough video annotation tools are optimized for:

- Precise object localization
- Object classification
- Attribute assignment
- Complex relationships to other objects (including composition, association)
- Frame-by-frame scene classification
- Instance recognition

To find out more about how we measure quality and how we can partner with you to create a reliable data pipeline for your model, check out our recent deep dive on the topic:
How to Get High Quality Training Data from Your Data Labeling Platform

Want to talk about your data quality needs with one of our specialists? Reach out to ***solutions@alegion.com*** *today.*