

[Cloud AutoML Vision](#)

Evaluating models

After training a model, AutoML Vision uses items from the TEST set (https://cloud.google.com/vision/automl/docs/prepare#training_vs_evaluation_datasets) to evaluate the quality and accuracy of the new model.

Evaluation overview

AutoML Vision provides an aggregate set of evaluation metrics indicating how well the model performs overall, as well as evaluation metrics for each category label, indicating how well the model performs for that label.

- **AuPRC** : Area under Precision/Recall curve ([https://en.wikipedia.org/wiki/Evaluation_measures_\(information_retrieval\)#Average_precision](https://en.wikipedia.org/wiki/Evaluation_measures_(information_retrieval)#Average_precision)), also referred to as "average precision." Generally between 0.5 and 1.0. Higher values indicate more accurate models.
- The **Confidence threshold curves** show how different confidence thresholds would affect precision, recall, true and false positive rates. Read about the relationship of precision and recall (https://en.wikipedia.org/wiki/Precision_and_recall).
- **Confusion matrix**: Only present for single-label-per-image models. Represents the percentage of times each label was predicted for each label in the training set during evaluation.

Confusion matrix

This table shows how often the model classified each label correctly (in blue), and which labels were most often confused for that label (in orange).

True label	Predicted label				
	roses	tulips	dandelion	daisy	sunflowers
roses	89.9%	10.1%	-	-	-
tulips	-	97.5%	1.3%	1.3%	-
dandelion	1.0%	-	98.1%	1.0%	-
daisy	-	-	1.7%	98.3%	-
sunflowers	-	-	-	-	100.0%

Ideally, label one would be assigned only to images classified as label one, etc, so a perfect matrix would look like:

```
100 0 0 0
0 100 0 0
0 0 100 0
0 0 0 100
```

In the example above, if an image was classified as one but the model predicted two, the first row would instead look like:

```
99 1 0 0
```

More information can be found by searching for 'confusion matrix machine learning' (<https://www.google.com/search?q=confusion+matrix+machine+learning>).

- ★ AutoML Vision creates the confusion matrix for up to 10 labels. If you have more than 10 labels, the matrix includes the 10 labels with the most confusion (incorrect predictions).

Use this data to evaluate your model's readiness. High confusion, low AUC scores, or low precision and recall scores can indicate that your model needs additional training data or has inconsistent labels. A very high AUC score and perfect precision and recall can indicate that the data is too easy and may not generalize well.

List model evaluations

Once you have trained a model, you can list evaluation metrics for that model.

WEB UI
REST & CMD LINE
MORE ▾

- Open the [AutoML Vision UI](https://console.cloud.google.com/vision) (https://console.cloud.google.com/vision) and click the **Models** tab (with lightbulb icon) in the left navigation bar to display the available models.
To view the models for a different project, select the project from the drop-down list in the upper right of the title bar.
- Click the row for the model you want to evaluate.
- If necessary, click the **Evaluate** tab just below the title bar.

If training has been completed for the model, AutoML Vision shows its evaluation metrics.

IMPORT
IMAGES
TRAIN
EVALUATE
TEST & USE
Single-Label Classification

Model
cloud_model

Confidence threshold 0.5

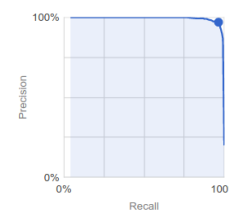
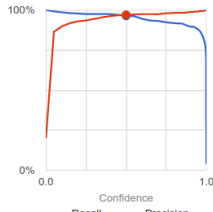
Filter labels

- All labels
- daisy
- dandelion
- roses
- sunflowers
- tulips

All labels

Total images	3,299
Test items	367
Precision	96.99%
Recall	96.46%

Use the slider to see which confidence threshold works best for your model on the precision-recall tradeoff curve.
[Learn more about these metrics and graphs.](#)

Confusion matrix

True Label \ Predicted Label	tulips	dandelion	sunflowers	daisy	roses
tulips	98%	-	-	-	3%
dandelion	-	99%	-	1%	-
sunflowers	-	-	96%	1%	3%
daisy	-	-	-	100%	-
roses	6%	-	-	2%	92%

Get model evaluation values

You can also get a specific model evaluation for a label (`displayName`) using an evaluation ID.

WEB UI

INTEGRATED UI

MORE ▾

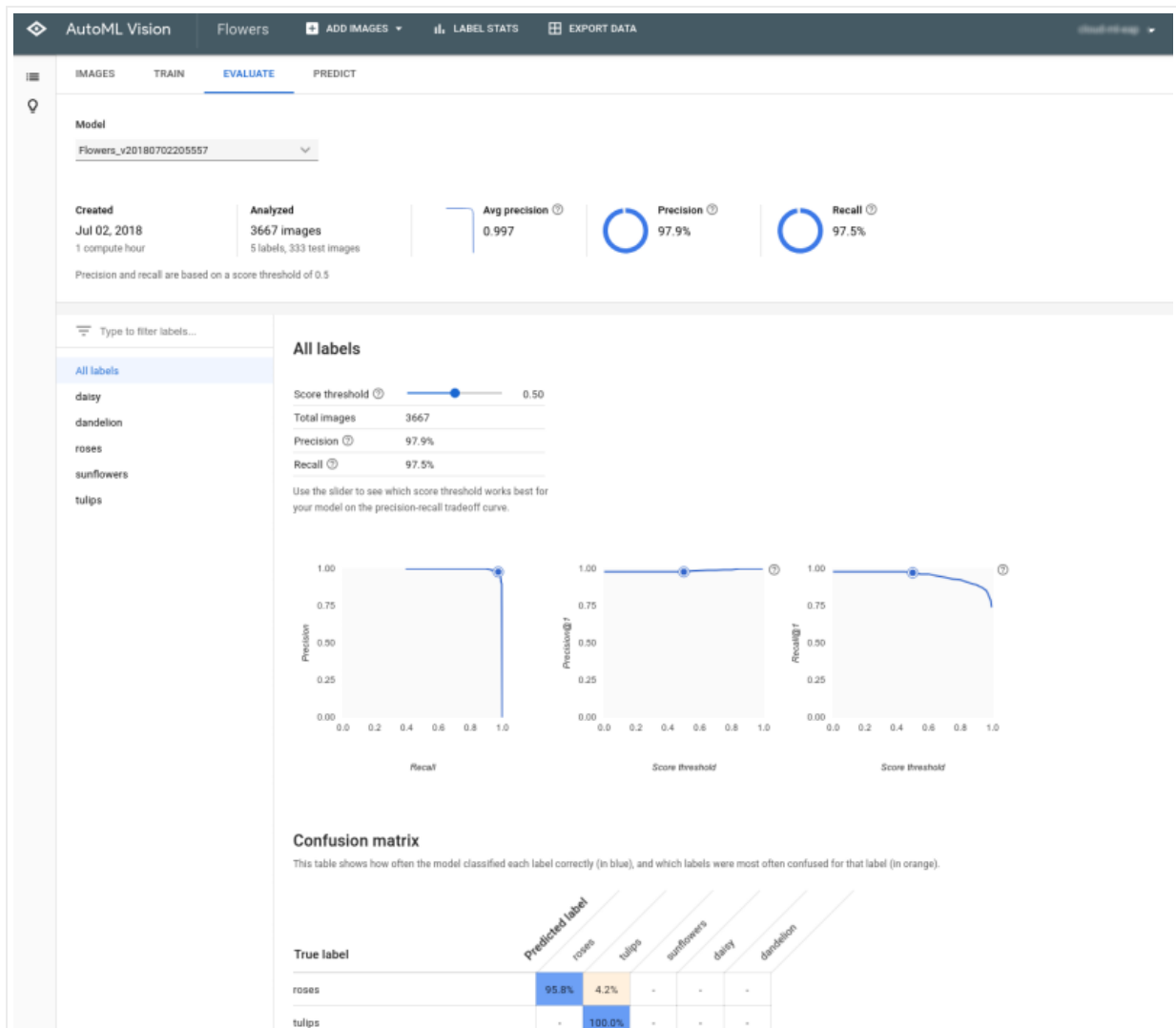
Note: Starting September 2019 we will start migrating AutoML Vision users to a new user interface that may affect the steps in this operation. This migration will occur in an on-going basis. See the "**Integrated UI**" tab for instructions using the updated interface.

1. Open the [AutoML Vision UI](https://console.cloud.google.com/vision) (<https://console.cloud.google.com/vision>) and click the lightbulb icon in the left navigation bar to display the available models.

To view the models for a different project, select the project from the drop-down list in the upper right of the title bar.

2. Click the row for the model you want to evaluate.
3. If necessary, click the **Evaluate** tab just below the title bar.

If training has been completed for the model, AutoML Vision shows its evaluation metrics.



4. To view the metrics for a specific label, select the label name from the list of labels in the lower part of the page.

IMPORT IMAGES TRAIN **EVALUATE** TEST & USE Single-Label Classification

Model: cloud_model Confidence threshold: 0.5

Filter labels: All labels, daisy, dandelion, **roses**, sunflowers, tulips

roses

Total images	3,299
Test items	0
Precision	93.55%
Recall	90.63%

Use the slider to see which confidence threshold works best for your model on the precision-recall tradeoff curve. [Learn more about these metrics and graphs.](#)

All test images are evaluated at the time of training. If you modify your dataset after training, these results will not be accurate.

True positives
Your model correctly predicted **roses** on these images

Score: 0.57167983 Score: 0.8095671 Score: 0.9303203 Score: 0.95381385 Score: 0.9647706 Score: 0.9732198

1 - 6 of many < >

False negatives

True Positives, False Negatives, and False Positives (UI only)

Note: This functionality is only available in the user interface (UI).

In the user interface you can observe specific examples of model performance, namely **true positive (TP)**, **false negative (FN)**, and **false positive (FP)** instances from your TRAINING and VALIDATION sets.

WEB UI

You can access the TP, FN, and FP view in the UI by selecting the **Evaluate** tab, and then selecting any specific label.

By viewing trends in these predictions, you can modify your training set to improve model performance.

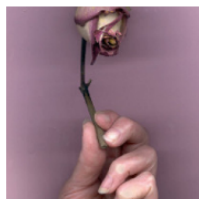
True positive images are sample images provided to the trained model that the model correctly annotated:

True positives

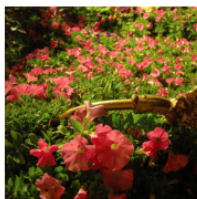
Your model correctly predicted roses on these images



Score: 0.57167983



Score: 0.8095671



Score: 0.9303203



Score: 0.95381385



Score: 0.9647706



Score: 0.9732198

1 - 6 of many < >

False negative images are similarly provided to the trained model, but the model failed to correctly annotate the image for the given label:

False negatives

Your model should have predicted roses on these images



Score(s): 0.21851009



Score(s): 0.006234663



Score(s): 0.14619568



Score(s): 0.4918342



Score(s): 0.16650382

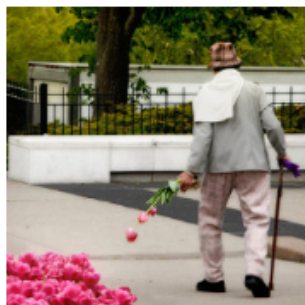


Score(s): 0.01657869

Lastly, **false positive** images are those provided to the trained model that were annotated with the given label, but *should not* have been annotated:

False positives

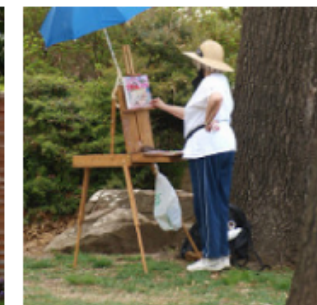
Your model incorrectly predicted roses on these images



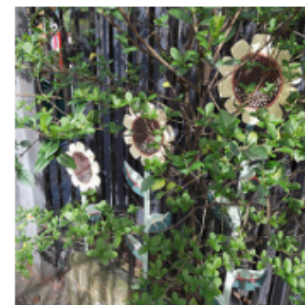
Score: 0.59571296



Score: 0.74452376



Score: 0.9224467




Score: 0.957296



The model is selecting interesting corner cases, which presents an opportunity to refine your definitions and labels to help the model understand your label interpretations. For example, a stricter definition would help the model understand if you consider an abstract painting of a rose a "rose" (or not).

With repeated label, train, and evaluate loops your model will surface other such ambiguities in your data.

You can also adjust the score threshold in this view in the user interface, and the TP, FN, and FP images displayed will reflect the threshold change:

Confidence threshold  0.84

roses


Total images	3,299
Test items	0
Precision 	96.55%
Recall 	87.5%


Use the slider to see which confidence threshold works best for your model on the precision-recall tradeoff curve. [Learn more about these metrics and graphs.](#)


All test images are evaluated at the time of training. If you modify your dataset after training, these results will not be accurate.

True positives


Your model correctly predicted roses on these images




Score: 0.9303203 



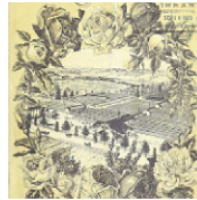
Score: 0.95381385




Score: 0.9647706



Score: 0.9732198



Score: 0.9757776



Score: 0.9807829

1 - 6 of many <

Iterate on your model

If you're not happy with the quality levels, you can go back to earlier steps to improve the quality:

- AutoML Vision allows you to sort the images by how “confused” the model is, by the true label and its predicted label. Look through these images and make sure they're labeled correctly.
- Consider adding more images to any labels with low quality.

- You may need to add different types of images (e.g. wider angle, higher or lower resolution, different points of view).
- Consider removing labels altogether if you don't have enough training images.
- Remember that machines can't read your label name; it's just a random string of letters to them. If you have one label that says "door" and another that says "door_with_knob" the machine has no way of figuring out the nuance other than the images you provide it.
- Augment your data with more examples of true positives and negatives. Especially important examples are the ones that are close to the decision boundary (i.e. likely to produce confusion, but still correctly labeled).
- Specify your own TRAIN, TEST, VALIDATION split. The tool randomly assigns images, but near-duplicates may end up in TRAIN and VALIDATION which could lead to overfitting and then poor performance on the TEST set.

Once you've made changes, train and evaluate a new model until you reach a high enough quality level.

Except as otherwise noted, the content of this page is licensed under the [Creative Commons Attribution 4.0 License](https://creativecommons.org/licenses/by/4.0/) (https://creativecommons.org/licenses/by/4.0/), and code samples are licensed under the [Apache 2.0 License](https://www.apache.org/licenses/LICENSE-2.0) (https://www.apache.org/licenses/LICENSE-2.0). For details, see our [Site Policies](https://developers.google.com/terms/site-policies) (https://developers.google.com/terms/site-policies). Java is a registered trademark of Oracle and/or its affiliates.

Last updated January 22, 2020.