Cloud AutoML Vision

Preparing your training data

Preparing your images

- Images must use a supported file type (see list below).
- AutoML Vision models are optimized for photographs of objects in the real world.
- The training data should be as close as possible to the data on which predictions are to be made. For example, if your use case involves blurry and low-resolution images (such as from a security camera), your training data should be composed of blurry, low-resolution images. In general, you should also consider providing multiple angles, resolutions, and backgrounds for your training images.
- AutoML Vision models can't generally predict labels that humans can't assign. So, if a human can't be trained to assign labels by looking at the image for 1-2 seconds, the model likely can't be trained to do it either.
- We recommend about 1000 training images per label. The minimum per label is 10, or 50 for advanced models. In general it takes more examples per label to train models with multiple labels per image, and resulting scores are harder to interpret.
- The following image formats are supported when training your model. Maximum file size is 30MB.
    - JPEG
    - PNG
    - GIF
    - BMP
    - ICO

  The following image formats are supported when requesting a prediction from (querying) your model. Maximum file size is 1.5MB.
    - JPEG
    - PNG
    - GIF

★ **Note:** The AutoML API currently only supports sending base64-encoded image content to the `predict` method. For an example, see Make a prediction (https://cloud.google.com/vision/automl/docs/predict).

- The model works best when there are at most 100x more images for the most common label than for the least common label. We recommend removing very low frequency labels.
- Consider including a `None_of_the_above` label and images that don't match any of your defined labels. For example, for a flower dataset, include images of flowers outside of your labeled varieties, and label them as `None_of_the_above`. This can improve the accuracy of your model. Note that, while any label name will work, `None_of_the_above` is treated specially by the system and will always appear last in the label list in the UI.

Training vs. evaluation datasets

The data in a dataset is divided into three datasets when training a model: a training dataset, a validation dataset, and a test dataset.

A training dataset is used to build a model. The model tries multiple algorithms and parameters while searching for patterns in the training data. As the model identifies patterns, it uses the validation dataset to test the algorithms and patterns. The best performing algorithms and patterns are chosen from those identified during the training stage.

After the best performing algorithms and patterns have been identified, they are tested for error rate, quality, and accuracy using the test dataset.

Both a validation and a test dataset are used in order to avoid bias in the model. During the validation stage, optimal model parameters are used, which can result in biased metrics. Using the test dataset to assess the quality of the model after the validation stage provides an unbiased assessment of the quality of the model.

By default, AutoML Vision splits your dataset randomly into 3 separate sets:

- 80% of images are used for training.
- 10% of images are used for hyper-parameter tuning and/or to decide when to stop training.
- 10% of images are used for evaluating the model. These images are not used in training.

The maximum size of a test dataset is 50,000 images, even if 10% of the total dataset exceeds that maximum.

If you'd like to specify which dataset each image in your CSV file should belong to, you can use a .csv file as described in the next section

Create a CSV file with image URIs and labels

Once your files have been uploaded to Google Cloud Storage, you can create a CSV file that lists all of your training data and the category labels for that data. The .csv file can have any filename, must be in the same bucket as your image files, must be UTF-8 encoded, and must end with a `.csv` extension. The file has one row for each image in the set you are uploading, with these columns in each row:

1. **Which set to assign the content in this row to.** This field is optional and can be one of these values:
   - `TRAIN` - Use the image to train the model.
   - `VALIDATION` - Use the image to validate the results that the model returns during training.
   - `TEST` - Use the image to verify the model's results after the model has been trained.

★ **Note:** If you specify `TRAIN` and `VALIDATION` sets but do not include a `TEST` set, the **Evaluate** tab will not be available in the AutoML Vision UI.

   If you do not specify a set for the image in a row, then AutoML Vision automatically places it in one of the three sets to ensure that there is enough training, validation, and testing content. The AutoML Vision uses the 80% of your content documents for training, 10% for validating, and 10% for testing. The maximum size of a test dataset is 50,000 images, even if 10% of the total dataset exceeds that maximum.
2. **The content to be categorized.** This field contains Google Cloud Storage URI for the image. Google Cloud Storage URIs are case-sensitive.
3. **A comma-separated list of labels that identify how the image is categorized.** Labels must start with a letter and only contain letters, numbers, and underscores. You can

include up to 20 labels for each image. You can also leave labels blank for manual labeling through the UI or through the human labeling service.

For example:

- Labeled: `gs://my-storage-bucket-vcm/flowers/images/img100.jpg,daisy`
- Not labeled: `gs://my-storage-bucket-vcm/flowers/images/img403.jpg`
- Multi-label: `gs://my-storage-bucket-vcm/flowers/images/img384.jpg,dandelion,tulip,rose`
- Assigned to a set: `TEST,gs://my-storage-bucket-vcm/flowers/images/img805.jpg,daisy`

- labels shown: `daisy`, `dandelion`, `tulip`, `rose`
- possible sets: `TRAIN`, `VALIDATION`, `TEST` (shown)

Save the contents as a CSV file in your Google Cloud Storage bucket.

Common errors with CSV

- Using unicode characters in labels. E.g. Japanese characters are not supported.
- Using spaces and non-alphanumeric characters in labels.
- Empty lines.
- Empty columns (lines with two successive commas).
- Incorrect capitalization of Cloud Storage image paths.
- Incorrect access control configured for your image files. Your service account should have read or greater access, or files must be publicly-readable.
- References to non-image files (such as PDF or PSD files). Likewise, files that are not image files but that have been renamed with an image extension will cause an error.
- URI of image points to a different bucket than the current project. Only images in the project bucket can be accessed.
- Non-CSV-formatted files.

Last updated December 5, 2019.