Cloud AutoML Vision

Training Edge exportable models

You create a custom model by training it using a prepared dataset (https://cloud.google.com/vision/automl/docs/create-datasets). AutoML API uses the items from the dataset to train the model, test it, and evaluate (https://cloud.google.com/vision/automl/docs/evaluate) its performance. You review the results, adjust the training dataset as needed, and train a new model using the improved dataset.

Training a model can take several hours to complete. The AutoML API enables you to check the status (https://cloud.google.com/automl/docs/reference/rest/v1/projects.locations.operations/get) of training.

Since AutoML Vision creates a new model each time you start training, your project may include numerous models. You can get a list of the models in your project (https://cloud.google.com/vision/automl/docs/models#list-models) can delete models (https://cloud.google.com/vision/automl/docs/models#delete-model) you no longer need. Alternatively, you can use the Cloud AutoML Vision UI to list and delete models created via the AutoML API that you do not need anymore.

Note:

- Unless otherwise specified in applicable terms of service or documentation, custom models created in Cloud AutoML products cannot be exported.
- The maximum lifespan for a custom model is 18 months as of the GA release. You must create and train a new model to continue classifying content after that amount of time.
- Edge models are optimized for inference on an Edge device. Consequently, Edge model accuracy ***will differ*** from Cloud model accuracy.

Models are based on state-of-the-art research (https://ai.googleblog.com/2018/08/mnasnet-towards-automating-design-of.html) at Google. Your model will be available as a TF Lite package. For more information about how to integrate a TensorFlow Lite model using the TensorFlow Lite SDK reference the following links for iOS (https://www.tensorflow.org/lite/demo_ios) and Android (https://www.tensorflow.org/lite/demo_android) .

Training Edge models

When you have a dataset with a solid set of labeled training items, you are ready to create and train your custom Edge model.

TensorFlow serving and TF Lite models

At training time you can choose the type of Edge model you want, depending on your specific use case:

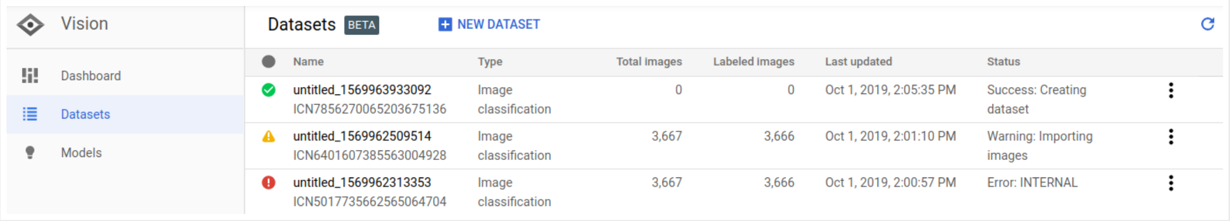
- low latency (`mobile-low-latency-1`)
- general purpose usage (`mobile-versatile-1`)
- higher prediction quality (`mobile-high-accuracy-1`)

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.

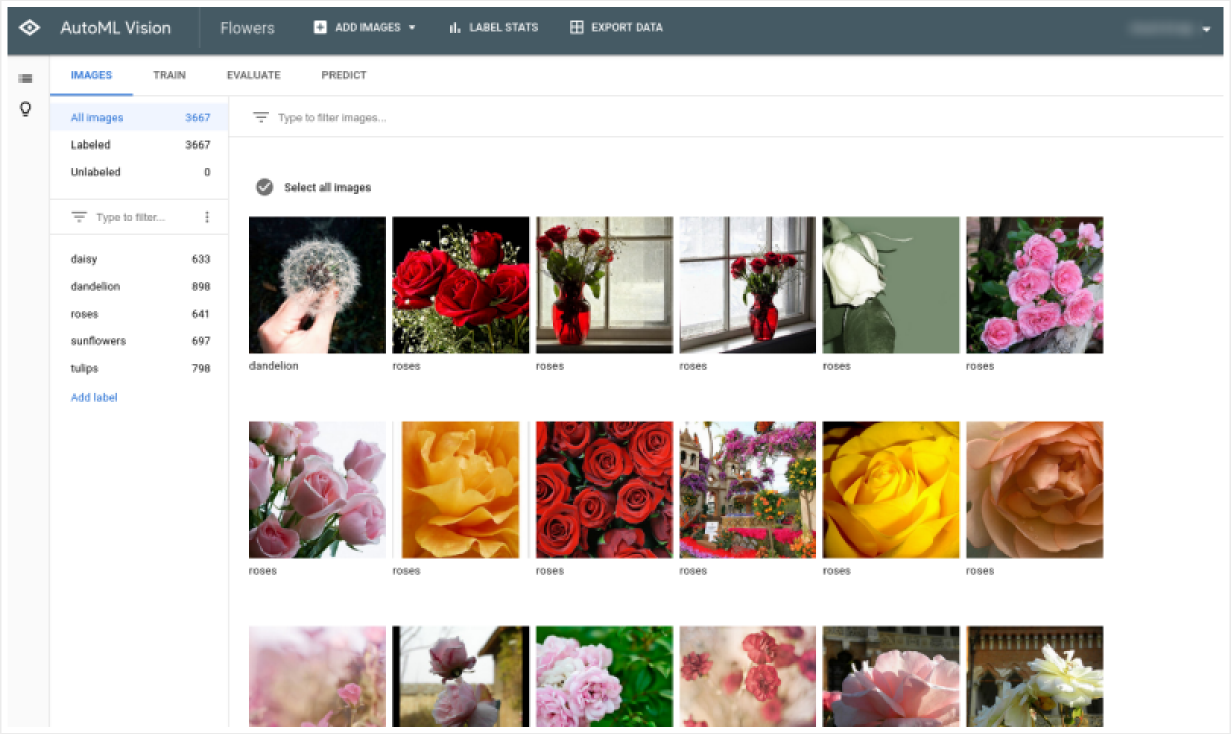
1. Open the AutoML Vision UI (https://console.cloud.google.com/vision).

   The **Datasets** page shows the available datasets for the current project.



	Name	Type	Total images	Labeled images	Last updated	Status
✓	untitled_1569963933092 ICN7856270065203675136	Image classification	0	0	Oct 1, 2019, 2:05:35 PM	Success: Creating dataset
⚠	untitled_1569962509514 ICN6401607385563004928	Image classification	3,667	3,666	Oct 1, 2019, 2:01:10 PM	Warning: Importing images
!	untitled_1569962313353 ICN5017735662565064704	Image classification	3,667	3,666	Oct 1, 2019, 2:00:57 PM	Error: INTERNAL

2. Select the dataset you want to use to train the custom model.
3. The display name of the selected dataset appears in the title bar, and the page lists the individual items in the dataset along with their labels.



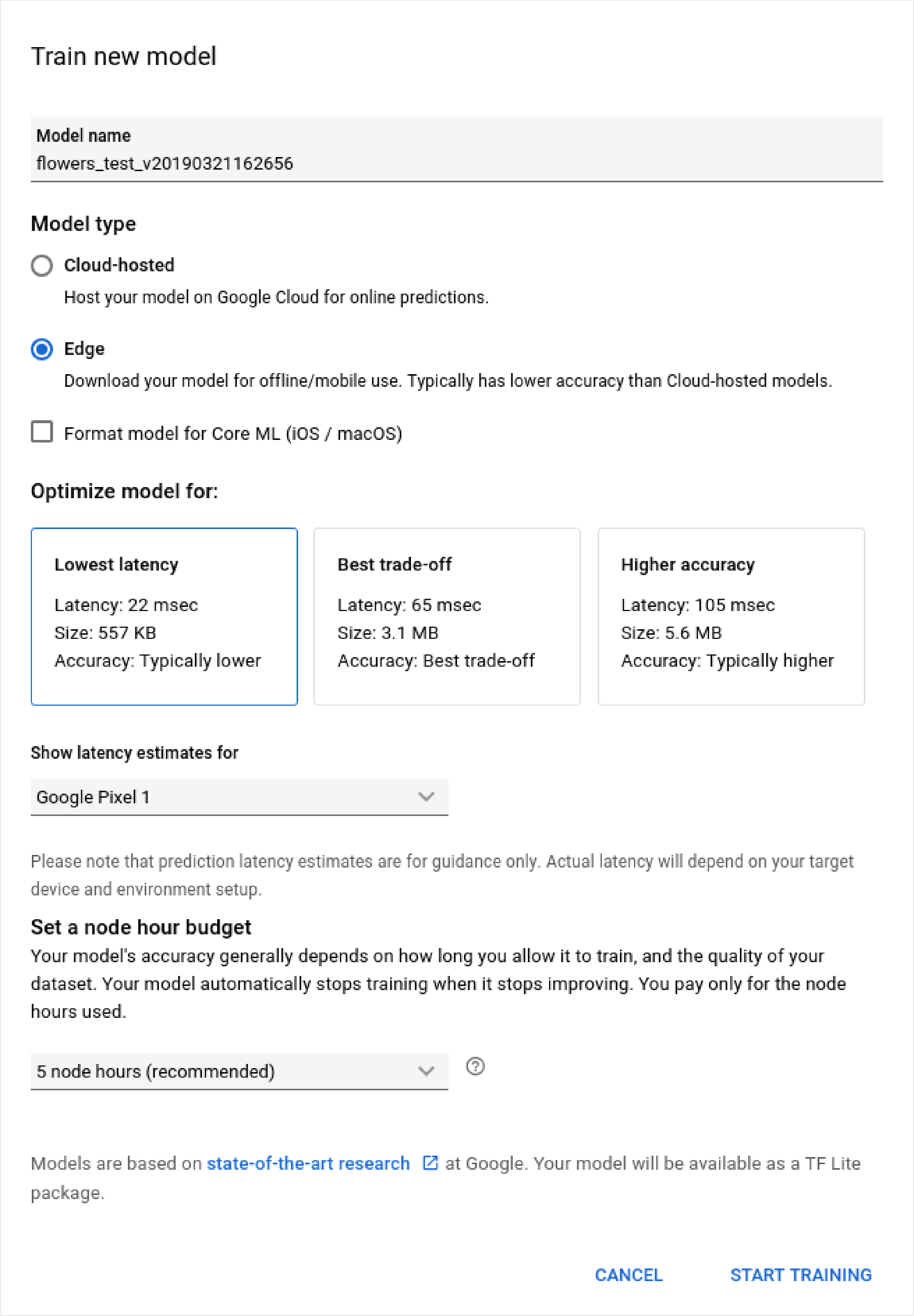
All images	3667
Labeled	3667
Unlabeled	0
daisy	633
dandelion	898
roses	641
sunflowers	697
tulips	798

4. When you are done reviewing the dataset, select the **Train** tab.

   The training page provides a basic analysis of your dataset and advises you about whether it is adequate for training. If AutoML Vision suggests changes, consider returning to the **Images** page and adding items or labels.

5. When the dataset is ready, choose "Edge" from the **Model type** options. After selecting to train an Edge model, choose from the three Edge options based on your model needs: ◉ **Higher accuracy**, **Best tradeoff**, or **Faster prediction**. You can use the **Estimate latency for** selector to get estimated latency, size and accuracy values for different devices. Latency values are estimated for an input image of size 224px by 224px.

   You can also specify your training budget in terms of compute hours in this window.



Train new model

Model name
flowers_test_v20190321162656

Model type

○ Cloud-hosted
Host your model on Google Cloud for online predictions.

● Edge
Download your model for offline/mobile use. Typically has lower accuracy than Cloud-hosted models.

□ Format model for Core ML (iOS / macOS)

Optimize model for:

Lowest latency	Best trade-off	Higher accuracy
Latency: 22 msec	Latency: 65 msec	Latency: 105 msec
Size: 557 KB	Size: 3.1 MB	Size: 5.6 MB
Accuracy: Typically lower	Accuracy: Best trade-off	Accuracy: Typically higher

Show latency estimates for
Google Pixel 1

Please note that prediction latency estimates are for guidance only. Actual latency will depend on your target device and environment setup.

Set a node hour budget
Your model's accuracy generally depends on how long you allow it to train, and the quality of your dataset. Your model automatically stops training when it stops improving. You pay only for the node hours used.

5 node hours (recommended)

Models are based on state-of-the-art research at Google. Your model will be available as a TF Lite package.

After specifying all your settings, select **Start Training**.

Training a model can take several hours to complete.

Core ML models

Similar to a regular Edge model, at training time you can choose the type of Core ML model you want, depending on your specific use case:

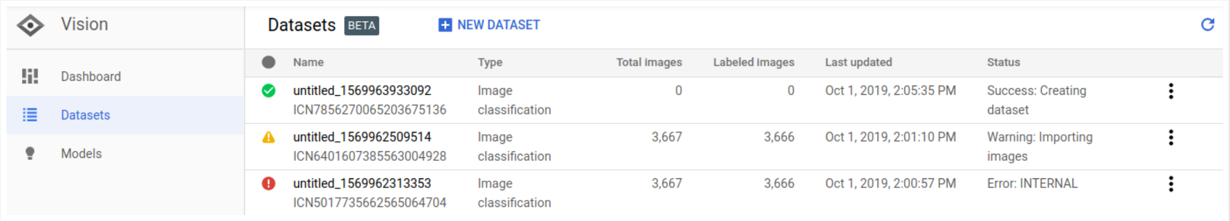
- low latency (`mobile-core-ml-low-latency-1`)
- general purpose usage (`mobile-core-ml-versatile-1`)
- higher prediction quality (`mobile-core-ml-high-accuracy-1`)

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).

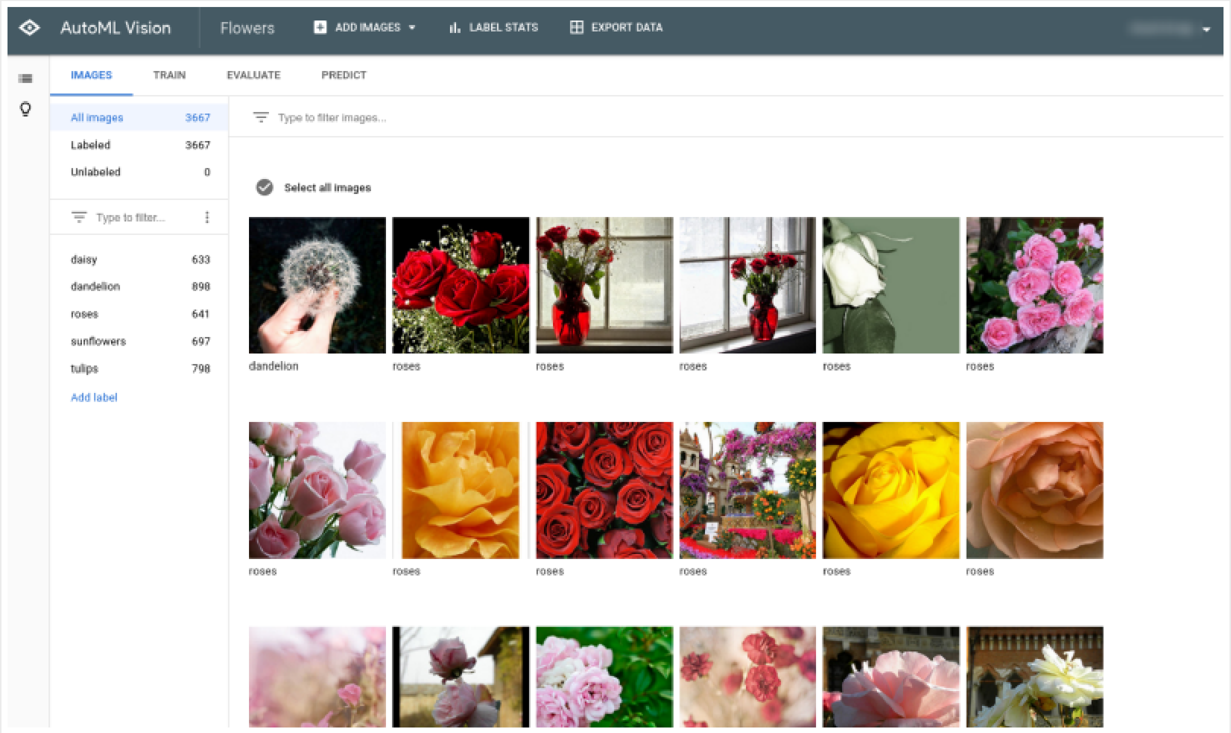
   The **Datasets** page shows the available datasets for the current project.



	Name	Type	Total images	Labeled images	Last updated	Status
✓	untitled_1569963933092 ICN7856270065203675136	Image classification	0	0	Oct 1, 2019, 2:05:35 PM	Success: Creating dataset
⚠	untitled_1569962509514 ICN6401607385563004928	Image classification	3,667	3,666	Oct 1, 2019, 2:01:10 PM	Warning: Importing images
!	untitled_1569962313353 ICN5017735662565064704	Image classification	3,667	3,666	Oct 1, 2019, 2:00:57 PM	Error: INTERNAL

2. Select the dataset you want to use to train the custom model.

   The display name of the selected dataset appears in the title bar, and the page lists the individual items in the dataset along with their labels.



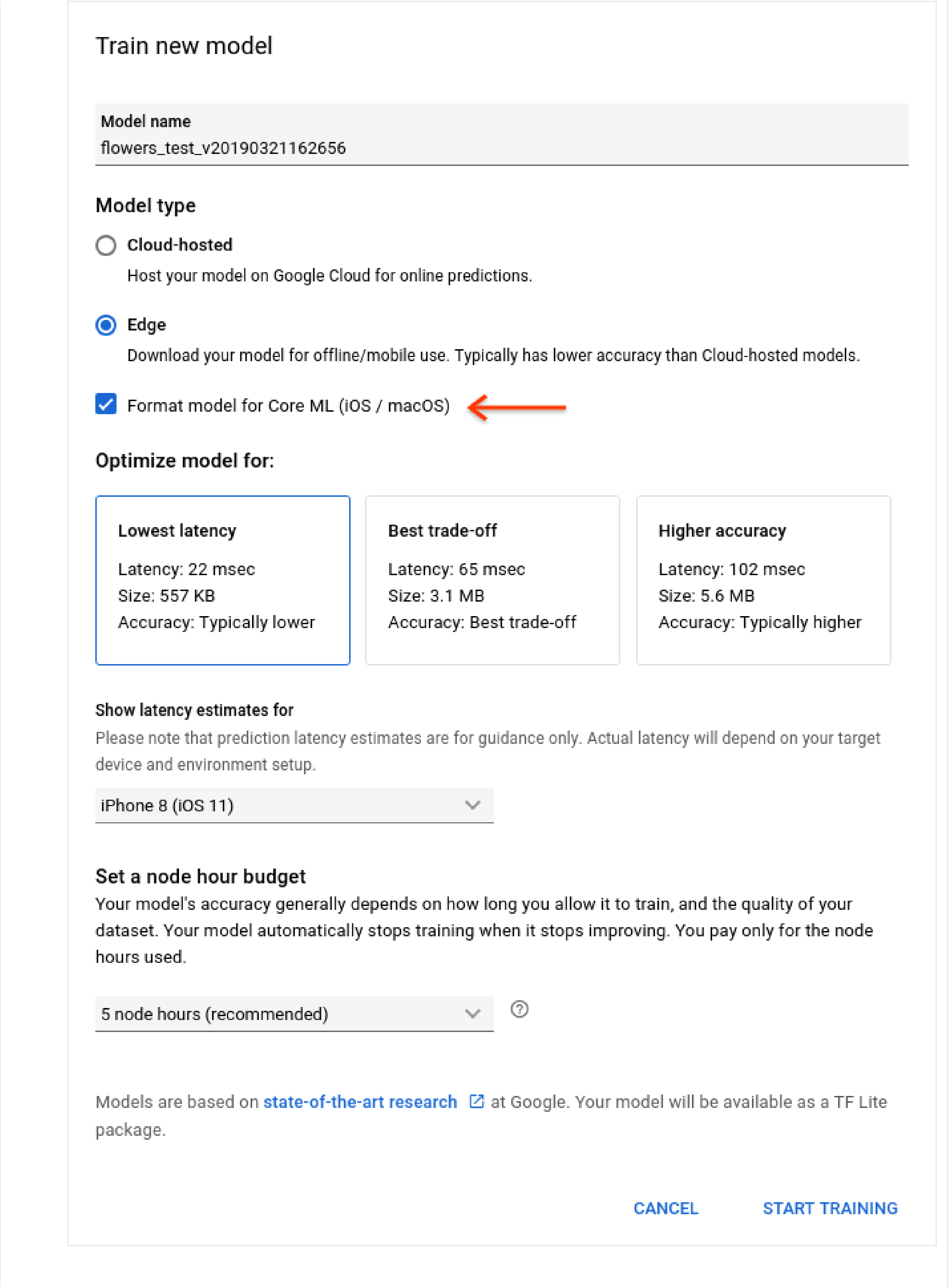
All images	3667
Labeled	3667
Unlabeled	0

daisy	633
dandelion	898
roses	641
sunflowers	697
tulips	798

3. When you are done reviewing the dataset, select the **Train** tab.

   The training page provides a basic analysis of your dataset and advises you about whether it is adequate for training. If AutoML Vision suggests changes, consider returning to the **Images** page and adding items or labels.

4. When the dataset is ready, select "Edge" from the **Model type** options. After selecting to train an Edge model, a checkbox will appear with the option to ☑ **"Format model for Core ML"**. Select the box.



Train new model

Model name
flowers_test_v20190321162656

Model type

○ Cloud-hosted
Host your model on Google Cloud for online predictions.

● Edge
Download your model for offline/mobile use. Typically has lower accuracy than Cloud-hosted models.

☑ Format model for Core ML (iOS / macOS)

Optimize model for:

Lowest latency	Best trade-off	Higher accuracy
Latency: 22 msec	Latency: 65 msec	Latency: 102 msec
Size: 557 KB	Size: 3.1 MB	Size: 5.6 MB
Accuracy: Typically lower	Accuracy: Best trade-off	Accuracy: Typically higher

Show latency estimates for
Please note that prediction latency estimates are for guidance only. Actual latency will depend on your target device and environment setup.

iPhone 8 (iOS 11)

Set a node hour budget
Your model's accuracy generally depends on how long you allow it to train, and the quality of your dataset. Your model automatically stops training when it stops improving. You pay only for the node hours used.

5 node hours (recommended)

Models are based on state-of-the-art research at Google. Your model will be available as a TF Lite package.

5. After selecting the box to train a Core ML model, choose from the three Edge options based on your model needs. You can use the **Estimate latency for** selector to get estimated latency, size and accuracy values for different devices. Latency values are estimated for an input image of size 224px by 224px.

   You can also specify your training budget in terms of compute hours in this window.

   After specifying all your settings, click **Start Training**.

Training a model can take several hours to complete.

List operations status

You can list your project's operations, and filter results.

REST & CMD LINE C# GO MORE

Before using any of the request data below, make the following replacements:

- ***project-id***: your GCP project ID.

HTTP method and URL:

```
GET https://automl.googleapis.com/v1/projects/project-id/locations/us-central1/ope
```

To send your request, choose one of these options:

CURL POWERSHELL

Note: Ensure you have set the GOOGLE_APPLICATION_CREDENTIALS (https://cloud.google.com/docs/authentication/production) environment variable to your service account private key file path.

Execute the following command:

```
curl -X GET \
-H "Authorization: Bearer "$(gcloud auth application-default print-access-token)
https://automl.googleapis.com/v1/projects/project-id/locations/us-central1/opera
```

The output you see will vary depending on the operations you have requested.

You can also filter the operations returned by using select query parameters (`operationId`, `done`, and `worksOn`). For example, to return a list of operations that have finished running modify the URL:

```
GET https://automl.googleapis.com/v1/projects/project-id/locations/us-central1/oper
```

Getting the status of an operation

REST & CMD LINE C# GO MORE

Before using any of the request data below, make the following replacements:

- ***project-id***: your GCP project ID.
- ***operation-id***: the ID of your operation. The ID is the last element of the name of your operation. For example:
  - operation name: `projects/project-id/locations/location-id/operations/IOD5281059901324392598`
  - operation id: `IOD5281059901324392598`

HTTP method and URL:

```
GET https://automl.googleapis.com/v1/projects/project-id/locations/us-central1/ope
```

To send your request, choose one of these options:

CURL POWERSHELL

Note: Ensure you have set the `GOOGLE_APPLICATION_CREDENTIALS` (https://cloud.google.com/docs/authentication/production) environment variable to your service account private key file path.

Execute the following command:

```
curl -X GET \
-H "Authorization: Bearer "$(gcloud auth application-default print-access-token)
https://automl.googleapis.com/v1/projects/project-id/locations/us-central1/opera
```

You should see output similar to the following for a completed **import operation**:

```
{
  "name": "projects/project-id/locations/us-central1/operations/operation-id",
  "metadata": {
    "@type": "type.googleapis.com/google.cloud.automl.v1.OperationMetadata",
    "createTime": "2018-10-29T15:56:29.176485Z",
    "updateTime": "2018-10-29T16:10:41.326614Z",
    "importDataDetails": {}
  },
  "done": true,
  "response": {
    "@type": "type.googleapis.com/google.protobuf.Empty"
  }
}
```

You should see output similar to the following for a completed **create model operation**:

```
{
  "name": "projects/project-id/locations/us-central1/operations/operation-id",
  "metadata": {
    "@type": "type.googleapis.com/google.cloud.automl.v1.OperationMetadata",
    "createTime": "2019-07-22T18:35:06.881193Z",
    "updateTime": "2019-07-22T19:58:44.972235Z",
    "createModelDetails": {}
  },
  "done": true,
  "response": {
    "@type": "type.googleapis.com/google.cloud.automl.v1.Model",
    "name": "projects/project-id/locations/us-central1/models/model-id"
  }
}
```

Cancelling an Operation

You can cancel an import or training task using the operation ID.

REST & CMD LINE

Before using any of the request data below, make the following replacements:

- ***project-id***: your GCP project ID.

- ***operation-id***: the ID of your operation. The ID is the last element of the name of your operation. For example:
  - operation name: `projects/`***`project-id`***`/locations/`***`location-id`***`/operations/IOD5281059901324392598`
  - operation id: `IOD5281059901324392598`

HTTP method and URL:

```
POST https://automl.googleapis.com/v1/projects/project-id/locations/us-central1/op
```

To send your request, choose one of these options:

CURL POWERSHELL

Note: Ensure you have set the [`GOOGLE_APPLICATION_CREDENTIALS`](https://cloud.google.com/docs/authentication/production) (https://cloud.google.com/docs/authentication/production) environment variable to your service account private key file path.

Execute the following command:

```
curl -X POST \
-H "Authorization: Bearer "$(gcloud auth application-default print-access-token)
-H "Content-Type: application/json; charset=utf-8" \
-d "" \
https://automl.googleapis.com/v1/projects/project-id/locations/us-central1/opera
```

You will see an empty JSON object returned from a successful request:

```
{}
```

Getting information about a model

When training is complete, you can get information about the newly created model.

The examples in this section return the basic metadata about a model. To get details about a model's accuracy and readiness, see the "Evaluating models" topic.

REST & CMD LINE C# GO MORE

Before using any of the request data below, make the following replacements:

- ***project-id***: your GCP project ID.
- ***model-id***: the ID of your model, from the response when you created the model. The ID is the last element of the name of your model. For example:
  - model name: `projects/project-id/locations/location-id/models/IOD4412217016962778756`
  - model id: `IOD4412217016962778756`

HTTP method and URL:

```
GET https://automl.googleapis.com/v1/projects/project-id/locations/us-central1/mod
```

To send your request, choose one of these options:

CURL POWERSHELL

Note: Ensure you have set the [`GOOGLE_APPLICATION_CREDENTIALS`](https://cloud.google.com/docs/authentication/production) (https://cloud.google.com/docs/authentication/production) environment variable to your service account private key file path.

Execute the following command:

```
curl -X GET \
-H "Authorization: Bearer "$(gcloud auth application-default print-access-token)
https://automl.googleapis.com/v1/projects/project-id/locations/us-central1/model
```

You should receive a JSON response similar to the following:

```
{
  "name": "projects/project-id/locations/us-central1/models/model-id",
  "displayName": "display-name",
  "datasetId": "dataset-id",
  "createTime": "2019-10-30T20:06:08.253243Z",
  "deploymentState": "UNDEPLOYED",
  "updateTime": "2019-10-30T20:54:50.472328Z",
  "imageClassificationModelMetadata": {
    "trainBudget": "1",
    "modelType": "mobile-low-latency-1",
    "nodeQps": 3.2
  }
}
```

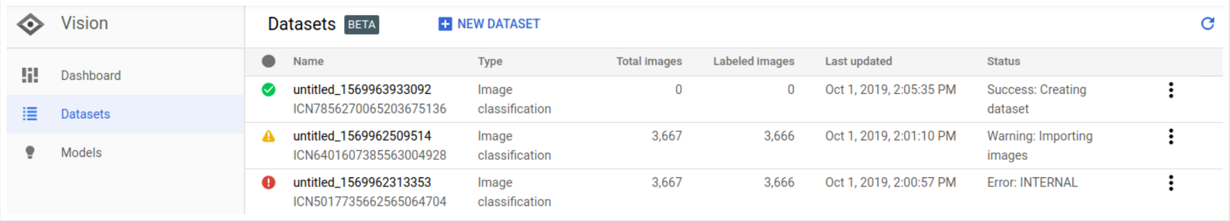
Resumable training

You can now pause and resume your custom model training for large datasets (with more than one thousand images).

WEB UI REST & CMD LINE

1. Open the AutoML Vision UI (https://console.cloud.google.com/vision).

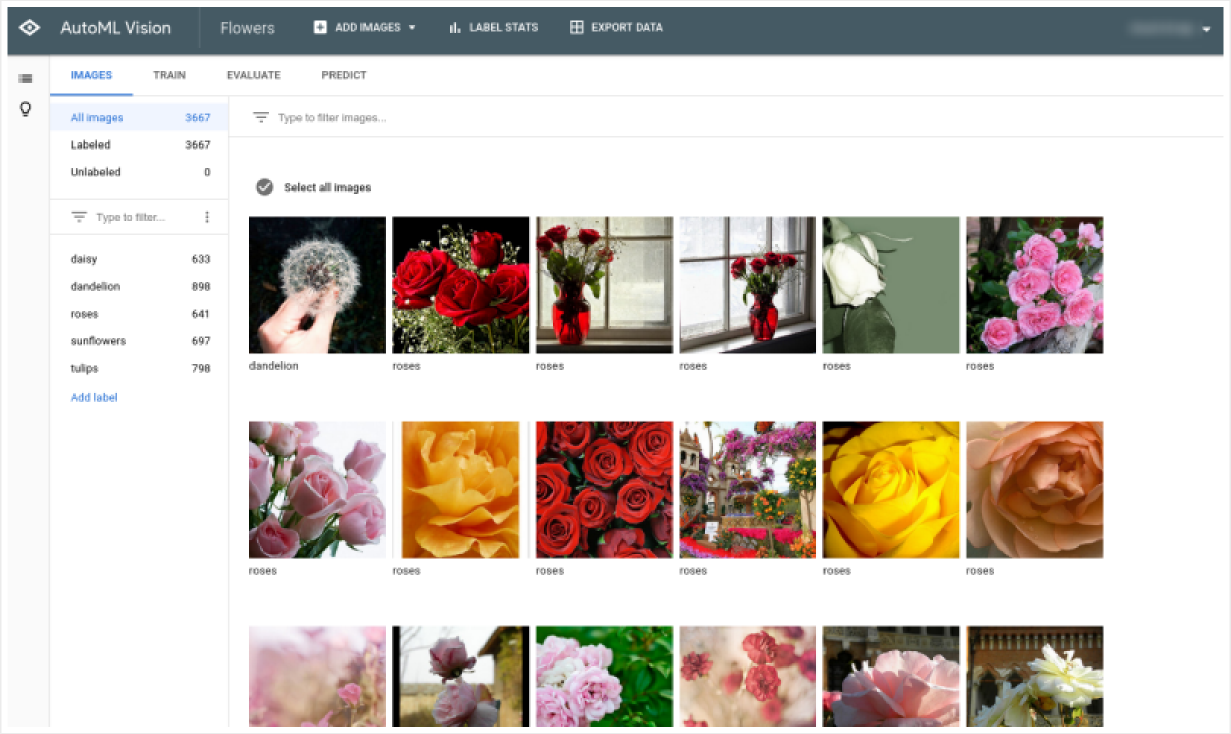
   The **Datasets** page shows the available datasets for the current project.

●	Name	Type	Total images	Labeled images	Last updated	Status
✓	untitled_1569963933092 ICN7856270065203675136	Image classification	0	0	Oct 1, 2019, 2:05:35 PM	Success: Creating dataset
⚠	untitled_1569962509514 ICN6401607385563004928	Image classification	3,667	3,666	Oct 1, 2019, 2:01:10 PM	Warning: Importing images
!	untitled_1569962313353 ICN5017735662565064704	Image classification	3,667	3,666	Oct 1, 2019, 2:00:57 PM	Error: INTERNAL

2. Select the dataset you want to use to train the custom model.

   The display name of the selected dataset appears in the title bar, and the page lists the individual items in the dataset along with their labels.



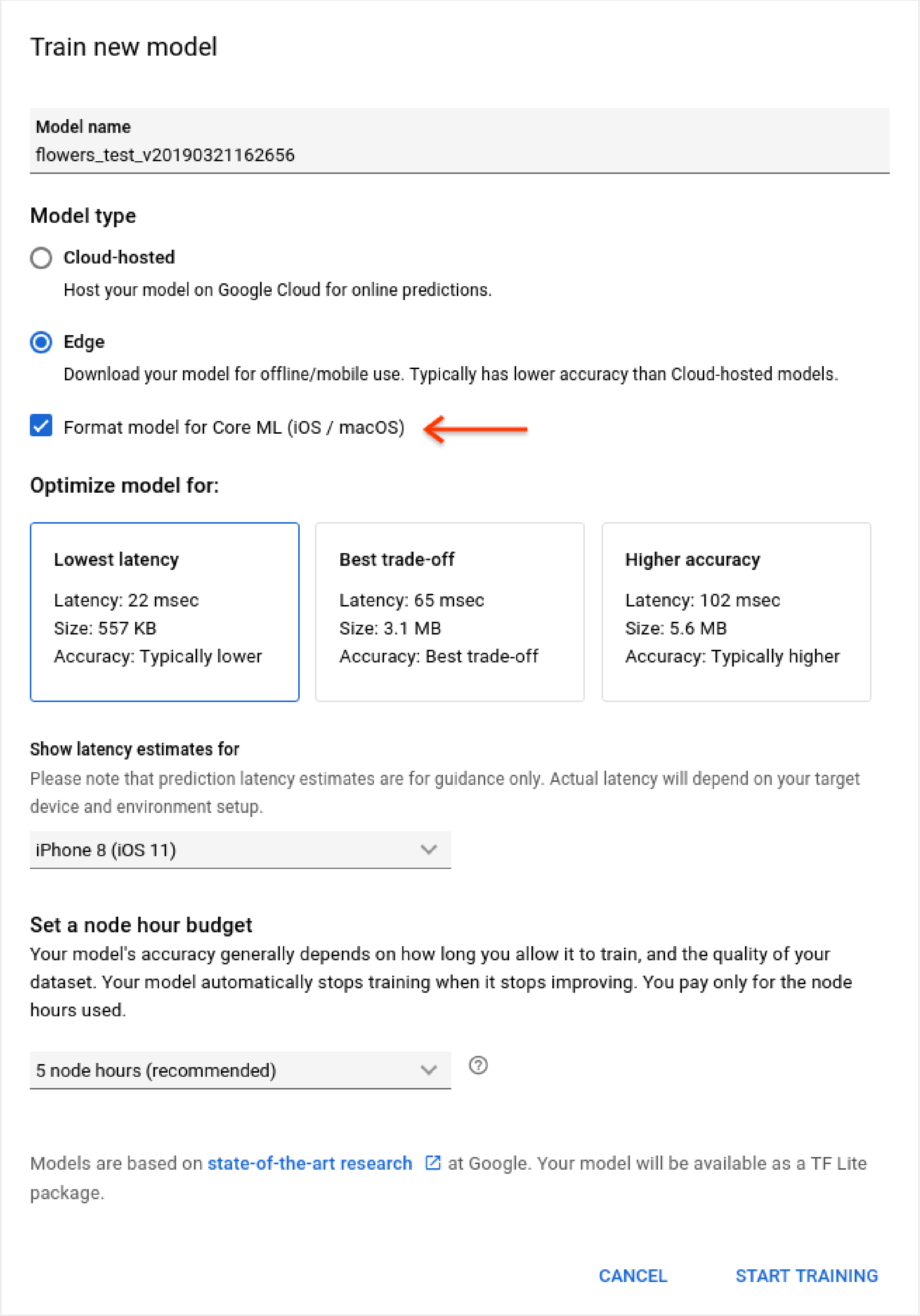
All images	3667
Labeled	3667
Unlabeled	0

daisy	633
dandelion	898
roses	641
sunflowers	697
tulips	798

3. When you are done reviewing the dataset, select the **Train** tab.

   The training page provides a basic analysis of your dataset and advises you about whether it is adequate for training. If AutoML Vision suggests changes, consider returning to the **Images** page and adding items or labels.

4. When the dataset is ready, select "Edge" from the **Model type** options. After selecting to train an Edge model, a checkbox will appear with the option to ☑ **"Format model for Core ML"**. Select the box.



Train new model

Model name
flowers_test_v20190321162656

Model type

○ Cloud-hosted
Host your model on Google Cloud for online predictions.

● Edge
Download your model for offline/mobile use. Typically has lower accuracy than Cloud-hosted models.

☑ Format model for Core ML (iOS / macOS)

Optimize model for:

Lowest latency
Latency: 22 msec
Size: 557 KB
Accuracy: Typically lower

Best trade-off
Latency: 65 msec
Size: 3.1 MB
Accuracy: Best trade-off

Higher accuracy
Latency: 102 msec
Size: 5.6 MB
Accuracy: Typically higher

Show latency estimates for
Please note that prediction latency estimates are for guidance only. Actual latency will depend on your target device and environment setup.

iPhone 8 (iOS 11)

Set a node hour budget
Your model's accuracy generally depends on how long you allow it to train, and the quality of your dataset. Your model automatically stops training when it stops improving. You pay only for the node hours used.

5 node hours (recommended)

Models are based on state-of-the-art research at Google. Your model will be available as a TF Lite package.

5. After selecting the box to train a Core ML model, choose from the three Edge options based on your model needs. You can use the **Estimate latency for** selector to get estimated latency, size and accuracy values for different devices. Latency values are estimated for an input image of size 224px by 224px.

   You can also specify your training budget in terms of compute hours in this window.

   After specifying all your settings, click **Start Training**.

Training a model can take several hours to complete.

Last updated January 22, 2020.