

[Cloud AutoML Vision Object Detection](#)

Deploying your model

Initial model deployment

After you have created (trained) a model, you must deploy the model before you can make online (or synchronous) calls to it.

You can now also update model deployment if you need additional online prediction capacity.

Deploying an Object Detection model incurs charges. For more information, see the [pricing page](https://cloud.google.com/vision/automl/pricing) (<https://cloud.google.com/vision/automl/pricing>).

Node count will be subject to quota listed at the quota page; **by default, a user can get up to 10 nodes.**

- 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.

WEB UI

REST & CMD LINE

MORE ▾

1. Open the [Cloud AutoML Vision Object Detection UI](#) (<https://console.cloud.google.com/vision>) and select 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.
2. Select the row for the model you want to use to label your images.
3. Select the **Test & Use** tab just below the title bar.
4. Select **Deploy model** from the banner beneath your model name to open the model deployment option window.

The screenshot shows the 'TEST & USE' tab of the AutoML console. At the top, there are navigation tabs: IMPORT, IMAGES, TRAIN, EVALUATE, and TEST & USE. Below the tabs, a dropdown menu shows the selected model: 'salad_dataset_20190331034505'. A message states: 'Deploy your model to send prediction requests to it. [Pricing guide](#)' with a 'DEPLOY MODEL' button. The main content area is titled 'Deploy 'salad_dataset_20190331034505?'. It explains that deployment hosts the model in the cloud and sends REST prediction requests. It asks the user to select the number of compute nodes, noting that more nodes support more queries per second. A dropdown menu is set to '1 node'. Below this, it states: 'Your model will be able to support 0.68 prediction queries per second.' At the bottom right, there are 'CANCEL' and 'DEPLOY' buttons.

In this window you can select the number of nodes to deploy on and view the available prediction queries per second (QPS).

5. Select **Deploy** to begin model deployment.

This screenshot shows the same console interface as the previous one, but the 'DEPLOY' button has been clicked. The 'DEPLOY MODEL' button is now disabled and greyed out. A large grey banner at the bottom of the main content area contains a circular loading icon and the text 'Deploying model...'. The rest of the interface, including the model name and navigation tabs, remains the same.

6. You will receive an email when model deployment has completed.

AutoML Vision finished deploying model "salad_dataset_20190331034505"

Inbox x



AutoML Vision <noreply-automl-vision@google.com> •

8:00 AM (24 minutes ago)



to me ▾

Hello AutoML Vision Customer,

AutoML Vision finished deploying model "salad_dataset_20190331034505".

Additional Details:

Resource Name:

projects/547439493576/locations/us-central1/models/IOD4758387257851772928

Operation State: Succeeded

To continue your progress, go back to your model using

<https://console.cloud.google.com/vision/datasets/IOD8792275517836361728;modelId=IOD4758387257851772928/predict?project=547439493576>

Sincerely,

The Google Cloud AI Team

Update a model's node number

Once you have a trained deployed model you can update the number of nodes the model is deployed on to respond to your specific amount of traffic. For example, if you experience a higher amount of queries per second (QPS) than expected.

You can change this node number *without* first having to undeploy the model. Updating deployment will change the node number without interrupting your served prediction traffic.

Node count will be subject to quota listed at the quota page; **by default, a user can get up to 10 nodes.**

WEB UI

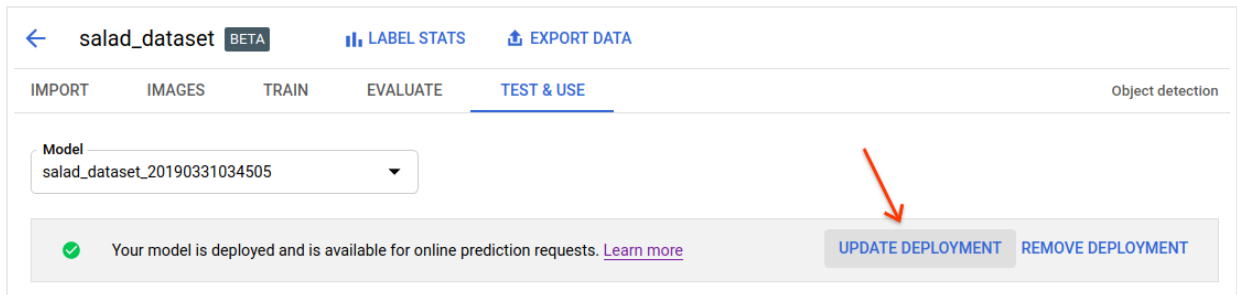
REST & CMD LINE

MORE ▾

1. In the Cloud AutoML Vision Object Detection UI (<https://console.cloud.google.com/vision>) and select 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.

2. Select your trained model that has been deployed.
3. Select the **Test & Use** tab just below the title bar.
4. A message is displayed in a box at the top of the page that says "Your model is deployed and is available for online prediction requests". Select the **Update deployment** option to the side of this text.



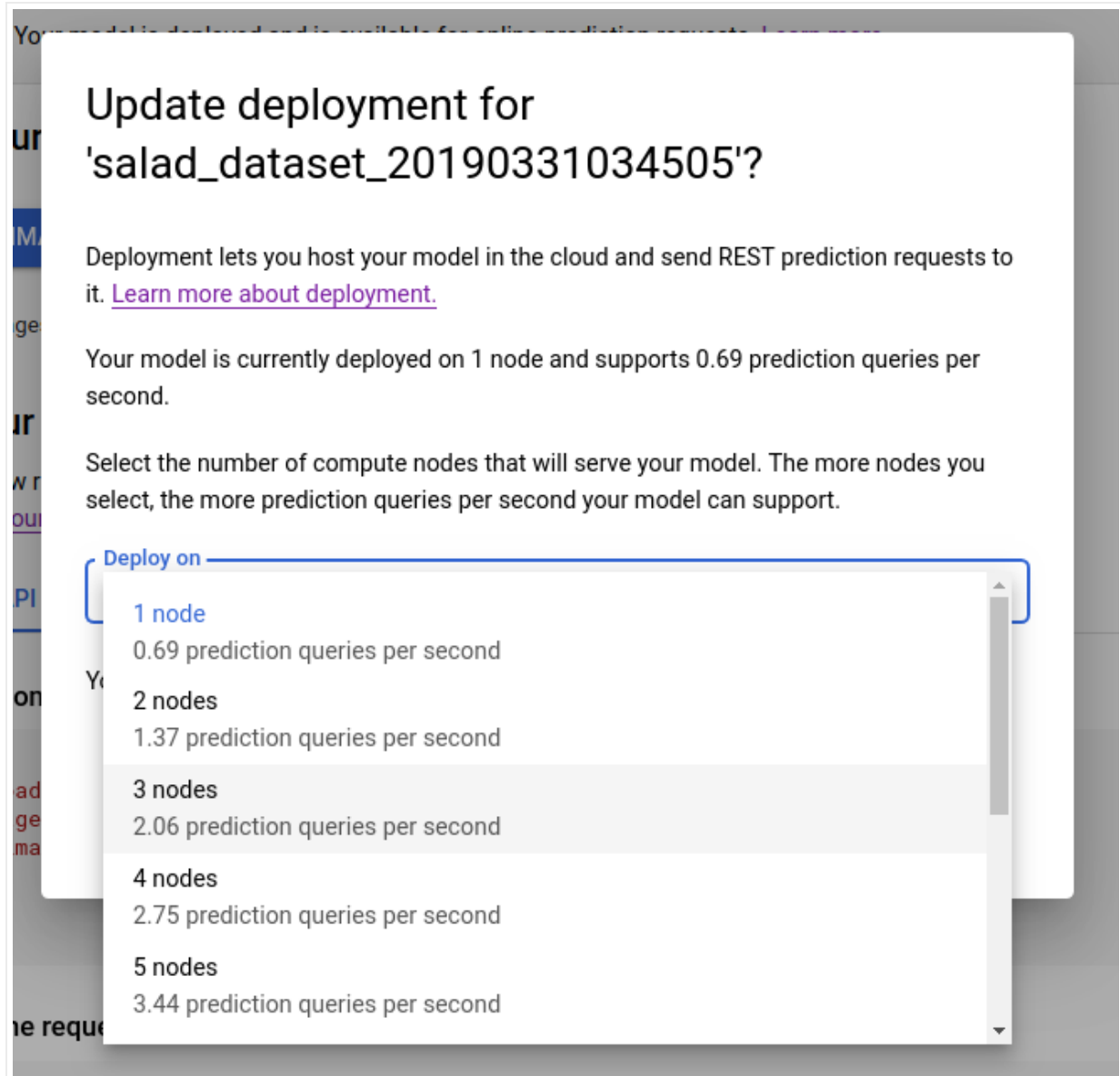
salad_dataset BETA LABEL STATS EXPORT DATA

IMPORT IMAGES TRAIN EVALUATE TEST & USE Object detection

Model
salad_dataset_20190331034505

✔ Your model is deployed and is available for online prediction requests. [Learn more](#) UPDATE DEPLOYMENT REMOVE DEPLOYMENT

5. In the **Update deployment** window that opens select the new node number to deploy your model on from the list. Node numbers display their estimated prediction queries per second (QPS).



Update deployment for 'salad_dataset_20190331034505'?

Deployment lets you host your model in the cloud and send REST prediction requests to it. [Learn more about deployment.](#)

Your model is currently deployed on 1 node and supports 0.69 prediction queries per second.

Select the number of compute nodes that will serve your model. The more nodes you select, the more prediction queries per second your model can support.

Deploy on

- 1 node
0.69 prediction queries per second
- 2 nodes
1.37 prediction queries per second
- 3 nodes
2.06 prediction queries per second
- 4 nodes
2.75 prediction queries per second
- 5 nodes
3.44 prediction queries per second

6. After selecting a new node number from the list select **Update deployment** to update the node number the model is deployed on.

Update deployment for 'salad_dataset_20190331034505'?

Deployment lets you host your model in the cloud and send REST prediction requests to it. [Learn more about deployment.](#)

Your model is currently deployed on 1 node and supports 0.69 prediction queries per second.

Select the number of compute nodes that will serve your model. The more nodes you select, the more prediction queries per second your model can support.

Deploy on
3 nodes

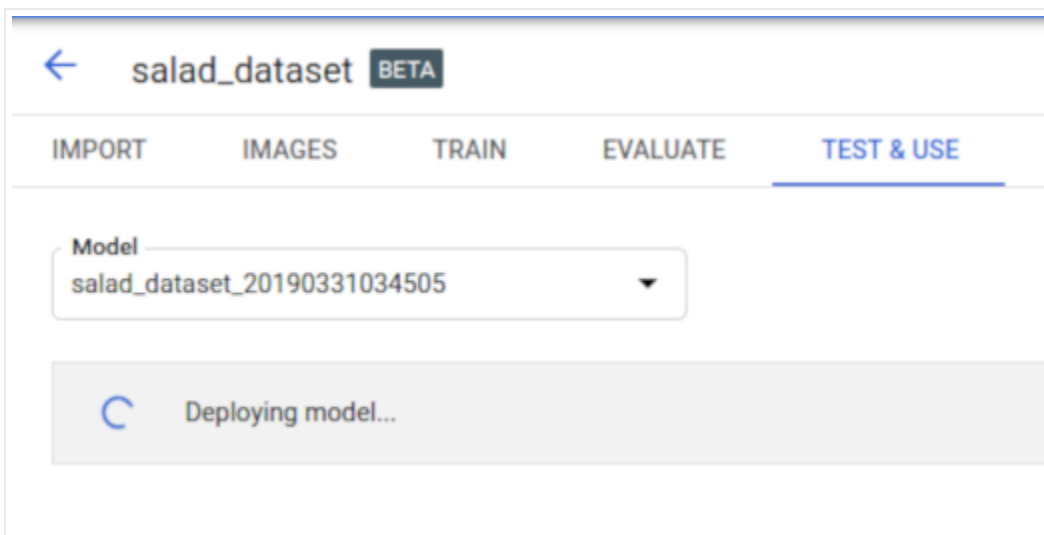
Your model will be able to support 2.06 prediction queries per second.

CANCEL

UPDATE DEPLOYMENT

7. You will be returned to the **Test & Use** window where you see the text box now displaying "Deploying model...".

★ **Note:** This is similar to the initial deployment of your model.



8. After your model has successfully deployed on the new node number you will receive an email at the address associated with your project.

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 November 20, 2019.