Cloud AutoML Vision

Pricing

AutoML Vision pricing depends on what feature you are using: image classification, object detection, or AutoML Vision Edge.

Image classification

AutoML Vision Image Classification enables you to train custom machine learning models to classify images into a custom set of categories.

Prices for usage of AutoML Vision Image Classification are based on resource usage, for both training and online prediction.

Free Trial

The pricing effective on **November 21, 2019**, at 12 AM Pacific time, is:

You can try AutoML Vision Image Classification for free by using 40 free node hours each for training and online prediction, and 1 free node hour for batch prediction, per billing account. Your free node hours are issued right before you create your first model. For batch prediction, the free node hour is issued at the time of the first batch prediction is initiated. You have up to one year to use them.

Prices are listed in US Dollars (USD). If you pay in a currency other than USD, the prices listed in your currency on Cloud Platform SKUs (https://cloud.google.com/skus/) apply.

Image classification training costs

The cost for AutoML Vision Image Classification model training is $3.15 per node hour.

For each unit of time, we use 8 nodes in parallel, where each node is equivalent to a n1-standard-8 machine with an attached NVIDIA® Tesla® V100 GPU. See Table below*.

The time required to train your model depends on the size and complexity of your training data. Many customer find that 8 node hours (approximately 1 "wall clock" hour) is sufficient to build an experimental model. Additional training time increases accuracy to a production level. The

early stopping feature ensures that training stops when further accuracy improvement is not possible.

You pay only for the compute hours used; if training fails for any reason other than a user-initiated cancellation, you will not be billed for the time. You will be charged for training time if you cancel the operation.

Training example

Example 1 - Cloud model with resumable training

You trained a Cloud Image Classification model for experimental use with **40 node hours**, and two days later spent **16 node hours** of resumable training to get it ready for production use.

You will receive a bill that shows:

- ($3.15 per node hour) * (40 node hours) = **$126.00 for the initial training**
- ($3.15 per node hour) * (16 node hours) = **$50.40 for the resumable training**

Example 2 - Cloud model with early stopping

You trained a Cloud Image Classification model that required **32 node hours** to train, while you set a budget of **40 node hours** with early stopping enabled. Only **4 hours** elapsed in this example, but training is done on **8 nodes** in parallel. The training time accumulated was **32.12 node hours**, and hence the charge was:

- ($3.15 per node hour) * (32.12 node hours) = **$101.18 (USD) for training**

Image classification deployment and prediction costs

Models must be deployed before they can provide online predictions.

Important: You pay per node deployed, even if no prediction is made. There is no additional charge for each prediction served. **You must undeploy your model to stop incurring further charges** .

Note that GPUs and/or CPUs remain allocated for your model so that your predictions are not delayed by startup latency.

The cost for deployment and prediction is $1.25 per node hour. One node is usually sufficient for most experimental traffic. You can adjust the number of nodes when you deploy your model. When you select the number of nodes for deployment in the Integrated UI (https://cloud.google.com/vision/automl/docs/deploy#automl_vision_classification_deploy_model_node_count-web-integrated), you receive an estimate of the prediction queries per second your model will support.

For **batch prediction**, the pricing is $2.02 per node hour used, with the first node hour free per account (one time).

Based on the equivalent machine configuration for this node, an estimated charge would be approximately $40 for a batch of 1 million images. **It may be significantly higher when complex models or images take more compute time to generate predictions.**

Please note that the charge is only for the node hours consumed and *not* for wall-clock time. Cancelling the batch prediction request after the computation has started will not currently result in a charge for the node hours consumed. You may not get partial prediction results since the entire pipeline will be aborted without required post-processing, and resources will be released. Contact Google Cloud Platform Support (https://cloud.google.com/support) the following day if the operation has not returned results within your expected time frame.

Deployment and prediction examples

Example 1 - Online prediction on old prediction service

This applies only to models trained prior to the Beta Refresh release on October 9, 2019. These models are shown as deployed with '0 nodes' in the Integrated UI. They will serve online predictions under the old pricing structure until you redeploy. Online predictions from these models will stop on February 21, 2020, unless you redeploy (https://cloud.google.com/vision/automl/docs/deploy).

You sent **1 million images** to the Cloud Image Classification model for prediction in your monthly billing cycle. The first 1000 images are free. For the remaining 999,000 images, you will be billed:

- ($3 / 1000 images) * (999,000 images) = **$2997.00 for online prediction services**

You will receive the same bill if you deploy an Edge model to the cloud for the same duration.

Example 2 - Cloud model with automatic deployment

Charges for model deployment will stop only when you undeploy the trained model. Let's assume at training time you opted in to auto-deploy, so after training finished your model was automatically **deployed on 1 node**. Then you forgot about your automatic model deployment! In this case, you will receive a bill in one month for:

- ($1.25 per node hour) * (1 node) * (24 hours per day) * (30 days) = **$900 (USD) for deployment & online prediction**

Example 3 - Deployment, online prediction, and undeployment

On the other hand, you may choose not to auto-deploy and deploy (https://cloud.google.com/vision/automl/docs/deploy) later when needed. Your deployed model was used for online predictions, and you followed up by undeploying (https://cloud.google.com/vision/automl/docs/undeploy) it promptly. From the time deployed model was ready for online prediction until the API to undeploy was called, the wall clock time needed was **0.242 hours**. Hence you will received a bill for:

- $1.25 per node hour) * (1 node) * (0.242 hours) = **$0.30 (USD)**

Example 4 - Batch prediction

You submitted **100,000 images** in a single job for batch prediction in your monthly billing cycle. Let us assume that the batch pipeline got the predictions done by utilizing **3 nodes** in parallel for **0.75 hours**, resulting in **2.25 node hours compute time** being billed. You may notice that 1 wall clock hour elapsed before results were returned, rather than .75 hours. This happens because there are preprocessing and postprocessing stages before and after the batch predictions, and waiting time between stages.

You will be billed:

- ($2.02 per node hour) * (2.25 node hours) = **$4.55 for batch prediction services**

The pricing effective on **November 21, 2019**, at 12 AM Pacific time, is:

Image Classification	Free	Paid

Image Classification	Free	Paid
Training	First 40 node hours are free (one time)	USD$3.15 per node hour
Deployment and Online (Individual) Prediction	First 40 node hours are free (one time)	USD$1.25 per node hour
Batch prediction	First node hour is free (one time)	USD$2.02 per node hour

If you pay in a currency other than USD, the prices listed in your currency on Cloud Platform SKUs (https://cloud.google.com/skus/) apply.

Object detection

AutoML Vision Object Detection enables you to train custom object detection models to localize a custom set of objects in your images.

Prices for AutoML Vision Object Detection are based on resource usage, for both training and classification online prediction.

Free Trial

You can try AutoML Vision Object Detection for free by using 40 free node hours each for training and online prediction, and 1 free node hour for batch prediction, per billing account. Your free node hours are issued right before you create your first model. For batch prediction, the free node hour is issued at the time of the first batch prediction is initiated. You have up to one year to use them.

Prices are listed in US Dollars (USD). If you pay in a currency other than USD, the prices listed in your currency on Cloud Platform SKUs (https://cloud.google.com/skus/) apply.

Object detection training costs

The cost for training a AutoML Vision Object Detection model is $3.15 per node hour.

For each unit of time, we use 9 nodes in parallel, where each node is equivalent to a n1-standard-8 machine with an attached NVIDIA® Tesla® V100 GPU (https://www.nvidia.com/en-us/data-center/tesla-v100/). ***See Table below****.

The time required to train your model depends on the size and complexity of your training data. Many customers find that **40 node hours** (approximately 5 "wall clock" hours) are sufficient to build a model with 5,000 labeled images or less.

You pay only for the compute hours used; if training fails for any reason other than a user-initiated cancellation, you will not be billed for the time. You *will* be charged for training time if you cancel the operation.

Training example

You trained a Cloud Object Detection model that required **38.207 node hours** to train, while you set a budget of 40 node hours with early stopping enabled. *Even though the wall clock time elapsed during training may be 5 hours*, the training job will be using 9 nodes in parallel. This explains why node hours charged is significantly more at 38.207. You will receive a bill for:

- ($3.15 per node hour) * (38.207 node hours) = **$120.35 for training**

Object detection deployment and prediction costs

Models must be deployed before they can provide online predictions.

Important: You pay per node deployed, even if no prediction is made. There is no additional charge for each prediction served. **You must undeploy your model to stop incurring further charges**.

Note that GPUs remain allocated for your model so that your predictions are not delayed by startup latency.

The cost for deployment and prediction is $1.82 per node hour. For each unit of time, we use 1 node equivalent to a n1-standard-4 machine with an NVIDIA® P100 GPU. ***See Table below*****.

Many customers find that with one node hour they can serve maximum 1.5 QPS. You can adjust the number of nodes when you deploy your model.

Deployment and prediction examples

When possible, you should remove model deployments (https://cloud.google.com/vision/automl/object-detection/docs/undeploy#automl_vision_object_detection_undeploy_model-web) if they are not needed. You can deploy (https://cloud.google.com/vision/automl/object-detection/docs/deploy) models later when they are needed for prediction again.

Example 1 - Deployment & Online prediction

You deployed your Cloud Object Detection model on **10 nodes**, and sent 1 million images for prediction over a period of **20.25 hours**. After using the prediction service you then undeploy this Cloud-hosted model. Since you have undeployed the model, your billing will be limited to 20.25 hours for each of the 10 nodes, accumulating a total of **202.5 node hours**. Even though you sent 1 million images for prediction, there is no charge per image. So you will receive a bill for:

- ($1.82 per node hour) * (202.5 node hours) = **$368.55 for deployment & prediction**

Example 2 - Deployment & Online prediction

Charges for Object Detection model deployment can stop ***only when you undeploy the trained model***. Let's assume at training time you opted in to auto-deploy, so after training finished your model was **automatically deployed on 1 node**. Then you forgot about your automatic model deployment! In this case, you will receive a bill in **one month** for:

- ($1.82 per node hour) * (1 node) * (24 hours per day) * (30 days) = **$1310.40 for deployment & prediction**

Example 3 - Batch prediction

You submitted 100,000 images in a single job for batch prediction in your monthly billing cycle. Let us assume that the batch pipeline got the predictions done by utilizing **3 nodes** in parallel for **5.45 hours** on average, resulting in **16.35 node hours** compute time being billed. You may notice that 6 wall clock hours elapsed before results were returned. This is so because batch predictions execute between preprocessing and postprocessing stages. Furthermore, there is waiting time between stages.

You will be billed:

- ($2.02 per node hour) * (16.35 node hours) = **$33.03 for batch prediction**

Object detection	Free	Paid
Training	First 40 node hours are free (one time)	USD$3.15 per node hour
Deployment and Online (Individual) Prediction	First 40 node hours are free (one time)	USD$1.82 per node hour
Batch prediction	First node hour is free (one time)	USD$2.02 per node hour

If you pay in a currency other than USD, the prices listed in your currency on Cloud Platform SKUs (https://cloud.google.com/skus/) apply.

AutoML Vision Edge

Cloud pricing will be applicable to an Edge model when it is used in the Cloud for online prediction. The deployment charge per node hour will apply. This is true for *both* Image Classification and Object Detection.

The Edge models are trained on TPUs.

- **Image Classification**: The cost for training an AutoML Vision Edge model for image classification is $4.95 per hour.
- **Object Detection**: The cost for training an AutoML Vision Edge model for object detection is $18 per hour.

For each unit of time you use 1 node, where the node is equivalent to a Cloud TPU v2 machine (https://cloud.google.com/tpu/).

Free Trial

You can try Edge for free by using 15 free node hours for training per billing account. Your free node hours are issued right before you create your first model, and you have up to one year to use them.

Many customers find that 3 node hours is sufficient to build a model with 5k labeled images or less.

You pay only for the compute hours used; if training fails for any reason other than a user-initiated cancellation, you will not be billed for the time. You will be charged for training time if you cancel the operation. Trained models can be exported and downloaded for free.

AutoML Vision Edge	Free	Paid
Image Classification Training	15 node hours of free training per account (one time)*	Subsequent training node hours are USD$4.95 per hour
Object Detection Training	15 node hours of free training per account (one time)*	Subsequent training node hours are USD$18.00 per hour
Exporting models to edge devices	Free	Free

* Effective May 7, 2019

If you pay in a currency other than USD, the prices listed in your currency on Cloud Platform SKUs (https://cloud.google.com/skus/) apply.

Edge Image Classification training example

You trained an Image Classification Edge model for development that required **1.506 node hours** with early stopping enabled. You will receive a bill for:

- ($4.95 per node hour) * (1.506 node hours) = **$7.45 for training**

Edge Object Detection training example

You trained an Object Detection Edge model for development that required **1.506 node hours** with early stopping enabled. You will receive a bill for:

- ($18.00 per node hour) * (1.506 node hours) = **$27.11 for training**

Google Cloud Platform costs

Since you store images to be analyzed in Google Cloud Storage, and may use other Google Cloud Platform resources in tandem with the AutoML Vision, such as Google AI Platform, containers and database instances, then you will also be billed for the use of those services. The price for human labeling available through the AI Platform Data Labeling Service may be viewed on their pricing page (https://cloud.google.com/data-labeling/pricing). See the Google Cloud Platform Pricing Calculator (https://cloud.google.com/products/calculator/) to determine other costs based on current rates.

To view your current billing status in the Cloud Console, including usage and your current bill, see the Billing page (https://console.cloud.google.com/billing). For more details about managing your account, see the Cloud Billing Documentation (https://cloud.google.com/billing/docs/) or Billing and Payments Support (https://cloud.google.com/support/billing/).

Review quotas on Google Cloud Console

There are two main ways to view your current quota limits in the Google Cloud Console (https://console.cloud.google.com/):

- Using the Quotas (https://console.cloud.google.com/quotas?project=_) page, which gives you a list of all your project's quota usage and limits.
- Using the console (https://console.cloud.google.com/apis/dashboard), which gives you quota information for a particular API, including resource usage over time.

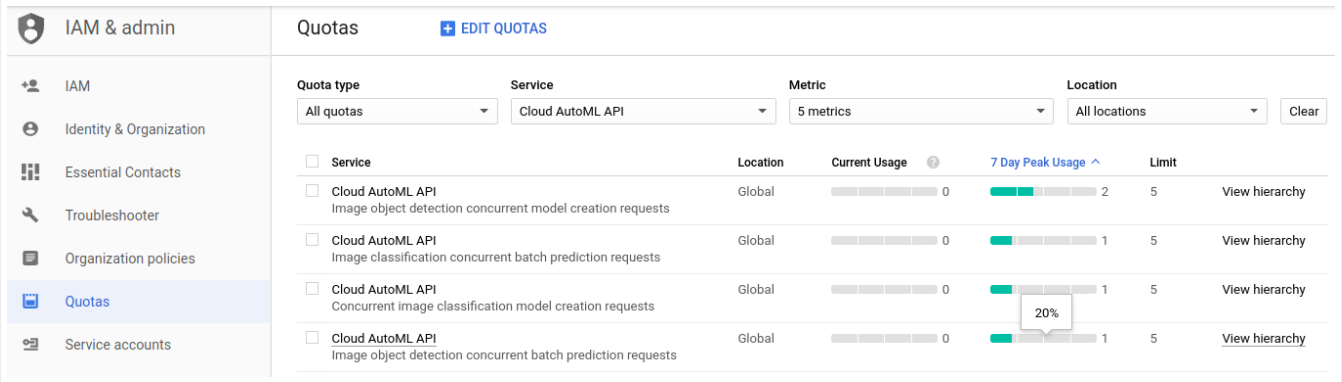
Locate specific operation quotas in the Quotas (https://console.cloud.google.com/quotas?project=_) page by first selecting `Cloud AutoML API` from the **Service** menu. With **Service:** `Cloud AutoML API` selected you can then select the appropriate **Metric**.

Examples:

Operation description	Method name
Image classification: Simultaneous model training	`projects.locations.models.create` (https://cloud.google.com/automl/docs/reference/r

Image classification: Online prediction	projects.locations.models.predict (https://cloud.google.com/automl/docs/reference/
Object detection: Simultaneous model training	projects.locations.models.create (https://cloud.google.com/automl/docs/reference/r
Object detection: Simultaneous offline batch prediction	projects.locations.models.batchPredict (https://cloud.google.com/automl/docs/refe

Quotas page:



Last updated January 14, 2020.