

## tf.fake\_quant\_with\_min\_max\_vars

```
fake_quant_with_min_max_vars(  
    inputs,  
    min,  
    max,  
    num_bits=8,  
    narrow_range=False,  
    name=None  
)
```

Defined in `tensorflow/python/ops/gen_array_ops.py`.

See the guide: [Tensor Transformations > Fake quantization](#)

Fake-quantize the 'inputs' tensor of type float via global float scalars `min`

and `max` to 'outputs' tensor of same shape as `inputs`.

`[min; max]` define the clamping range for the `inputs` data. `inputs` values are quantized into the quantization range `[0; 2num_bits - 1]` when `narrow_range` is false and `[1; 2num_bits - 1]` when it is true) and then de-quantized and output as floats in `[min; max]` interval. `num_bits` is the bitwidth of the quantization, between 2 and 8, inclusive.

This operation has a gradient and thus allows for training `min` and `max` values.

### Args:

- `inputs`: A `Tensor` of type `float32`.
- `min`: A `Tensor` of type `float32`.
- `max`: A `Tensor` of type `float32`.
- `num_bits`: An optional `int`. Defaults to `8`.
- `narrow_range`: An optional `bool`. Defaults to `False`.
- `name`: A name for the operation (optional).

### Returns:

A `Tensor` of type `float32`.

---

Except as otherwise noted, the content of this page is licensed under the [Creative Commons Attribution 3.0 License](#), and code samples are licensed under the [Apache 2.0 License](#). For details, see our [Site Policies](#). Java is a registered trademark of Oracle and/or its affiliates.

Last updated November 2, 2017.

### Stay Connected

[Blog](#)

[GitHub](#)

[Twitter](#)

**Support**

[Issue Tracker](#)

[Release Notes](#)

[Stack Overflow](#)

**English**

[Terms](#) | [Privacy](#)