tf.fake_quant_with_min_max_args

```
fake_quant_with_min_max_args(
    inputs,
    min=-6,
    max=6,
    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, type float to 'outputs' tensor of same type.

Attributes `[min; max]` define the clamping range for the `inputs` data. `inputs` values are quantized into the quantization range (`[0; 2^num_bits - 1]` when `narrow_range` is false and `[1; 2^num_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.

Quantization is called fake since the output is still in floating point.

Args:

- `inputs` : A **Tensor** of type `float32` .
- `min` : An optional `float` . Defaults to `-6` .
- `max` : An optional `float` . Defaults to `6` .
- `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` .

Last updated November 2, 2017.

Stay Connected

Blog

GitHub

Twitter

Support

Issue Tracker

Release Notes

Stack Overflow

English

Terms | **Privacy**