tf.fake_quant_with_min_max_vars_per_channel

```
fake_quant_with_min_max_vars_per_channel(
    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 and one of the shapes: `[d]`,

`[b, d]` `[b, h, w, d]` via per-channel floats `min` and `max` of shape `[d]` 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; 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.

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`.

Last updated November 2, 2017.