tf.contrib.tpu.shard

```
shard(
    computation,
    inputs=None,
    num_shards=1,
    input_shard_axes=None,
    outputs_from_all_shards=True,
    output_shard_axes=None,
    infeed_queue=None,
    global_tpu_id=None,
    name=None
)
```

Defined in `tensorflow/contrib/tpu/python/tpu/tpu.py`.

Shards `computation` for parallel execution.

`inputs` must be a list of Tensors or None (equivalent to an empty list), each of which has a corresponding split axis (from `input_shard_axes`). Each input is split into `num_shards` pieces along the corresponding axis, and computation is applied to each shard in parallel.

Tensors are broadcast to all shards if they are lexically captured by `computation`. e.g.,

x = tf.constant(7) def computation(): return x + 3 ... = shard(computation, ...)

TODO(phawkins): consider adding support for broadcasting Tensors passed as inputs.

If `outputs_from_all_shards` is true, the outputs from all shards of `computation` are concatenated back together along their `output_shards_axes`. Otherwise, each output is taken from an arbitrary shard.

Inputs and outputs of the computation must be at least rank-1 Tensors.

Args:

- `computation`: a Python function that builds a computation to apply to each shard of the input.
- `inputs`: a list of input tensors or None (equivalent to an empty list). Each input tensor has a corresponding shard axes, given by `input_shard_axes`, which must have size divisible by `num_shards`.
- `num_shards`: the number of shards.
- `input_shard_axes`: a list of dimensions along which to shard `inputs`, or `None`. `None` means "shard all inputs along dimension 0". If not `None`, there must be one dimension per input.
- `outputs_from_all_shards`: boolean or list of boolean. For each output, if `True`, outputs from all shards are concatenated along the corresponding `output_shard_axes` entry. Otherwise, each output is taken from an arbitrary shard. If the argument is a boolean, the argument's value is used for each output.
- `output_shard_axes`: a list of dimensions along which to concatenate the outputs of `computation`, or `None`. `None` means "concatenate all outputs along dimension 0". If not `None`, there must be one dimension per output. Ignored if `outputs_from_all_shards` is False.
- `infeed_queue`: if not None, the InfeedQueue to use to augment the inputs of `computation`.
- `global_tpu_id`: if not None, a Numpy 2D array indicating the global id of each TPU device in the system. The outer dimension of the array is host task id, and the inner dimension is device ordinal, so e.g., global_tpu_id[x][y] indicates

the global id of device /task:x/device:TPU_NODE:y.

- `name`: name of the operator.

Returns:

A list of output tensors.

Raises:

- `ValueError`: if num_shards <= 0
- `ValueError`: if len(input_shard_axes) != len(inputs)
- `ValueError`: if len(output_shard_axes) != len(outputs from `computation`)

Last updated November 2, 2017.