EML Task 10.1

1. Task 10.1.1

1.1. compute_encodings

compute_encodings calculates all encoding parameters for all the quantization nodes. These encodings are used to simulate how the model will perform when it is quantized for deployment on hardware that supports lower precision arithmetic (such as INT8). It determines the optimal scale and offset for each quantizer, and ensures that the model maintains as much accuracy as possible after quantization.

1.2. forward_pass_callback

forward_pass_callback is a function provided by the user that runs forward passes on the model using given test data. The idea is to expose the model to typical input data, which helps in accurately calculating the encoding parameters for quantization. We want to simulate how the model will process real data, allowing the compute_encodings method to gather the necessary information to calculate optimal quantization parameters.

2. Task 10.1.3

2.1. Terminal output of simple_linear.py

```
Running simple linear quantization example
SimpleLinear(
  (m_layer): Linear(in_features=4, out_features=4, bias=False)
)
FP32 Result:
tensor([ 0.6733,  2.1158, -2.4068, -1.1887])
2024-05-18 18:49:31,862 - Quant - INFO - No config file provided, defaulting to config
file at /mnt/hd1/conda/aimet/lib/python3.8/site-packages/aimet_common/quantsim_config/
default_config.json
2024-05-18 18:49:31,875 - Quant - INFO - Unsupported op type Squeeze
2024-05-18 18:49:31,875 - Quant - INFO - Unsupported op type Mean
2024-05-18 18:49:31,876 - Quant - INFO - Selecting DefaultOpInstanceConfigGenerator to
compute the specialized config. hw_version:default
Sim:
------------------------
Quantized Model Report
------------------------
-------------------------------------------------------------
Layer: m_layer
  Input[0]: bw=4, encoding-present=True
    StaticGrid TensorQuantizer:
     quant-scheme:QuantScheme.post_training_tf, round_mode=RoundingMode.ROUND_NEAREST,
bitwidth=4, enabled=True
         min:-0.9839999914169312,  max=0.6559999942779541,  delta=0.10933333237965902,
offset=-9.0
  -------
  Param[weight]: bw=4, encoding-present=True
    StaticGrid TensorQuantizer:
     quant-scheme:QuantScheme.post_training_tf, round_mode=RoundingMode.ROUND_NEAREST,
bitwidth=4, enabled=True
          min:-4.319999967302595,  max=3.7799999713897705,  delta=0.5399999959128243,
offset=-8.0
  -------
  Output[0]: bw=4, encoding-present=True
```

```
    StaticGrid TensorQuantizer:
     quant-scheme:QuantScheme.post_training_tf, round_mode=RoundingMode.ROUND_NEAREST,
bitwidth=4, enabled=True
          min:-2.252160135904948,   max=1.9706401189168292,   delta=0.2815200169881185,
offset=-8.0
  -------
```

2.2. Analysis

Min (qmin): Numbers below these are clamped

Max (qmax): Numbers above these are clamped

Delta: Granularity of the fixed point numbers (is a function of the bit-width selected)

The Delta can be calculated using Min and Max using follwing equation:

$$\text{Delta} = \frac{\text{Max} - \text{Min}}{2^{\text{bitwidth}} - 1}$$