Quantisation

# Model weights.

* The larger the weight the greater the accuracy but increased required resources.
* Normally floating point
* Quantisation rescales weights and/or activations.
* Quantisation often uses INT8, INT4
* Mapping function
  + Map high precision to lower precision.
  + Per layer
  + Q(r) = round((r/S)+Z)
    - S = scaling factor
    - Z = bias value to map a zero
  + Map the actual range of values from the larger range not the possible range.
    - Sensitive to outliers
    - Different values could be mapped to the same value.
    - Use percentiles instead.
    - Use a calibration dataset – Kullback-Leibler divergence.
* Post train quantisation.
  + Dynamic
    - Convert weights ahead of time.
    - Resale activations on the fly
  + Static
    - Convert weights ahead of time.
    - Use a calibration dataset to convert activations ahead of time
* Quantisation aware training
  + Resource expensive on large LLM’s
* Dequantization process performed at inference time