# IoT ML HW1 Analytical Questions

1. **Consider a fully-connected network with 120 inputs, two hidden layers with 256 units each, and 10 output classes. Assume that all weights and activations are stored as 8-bit values.**
    1. *Find the total parameter storage required in bytes.*
        1. The number of weights for a given layer is equal to the product of the number of neurons in that layer and the number of neurons in the previous layer.
            1. $|W_i| = n_{i-1} * n_i$, where $W_i$ is the set of weights in layer $i$ and $n_i$ is the set of neurons in layer $i$
            2. Thus $|W| = \sum_{i=i}^{N} n_{i-1} * n_i$ where N is the number of layers
            3. The number of weights of this model is $120 * 256 + 256 * 256 + 256 * 10 = 98816$
        2. The number of biases is simply the number of hidden neurons and output neurons
            1. $|B| = \sum_{i=1}^{N} n_i = 256 + 256 + 10 = 522$
        3. The number of parameters is $|W| + |B| = 512 + 98816 = 99338$
        4. Since each parameter occupies one byte of space, the total parameter storage required is **99338 Bytes**
    2. *How many MACs (multiply-and-accumulate) are required to run one inference with this model?*
        1. The number of MACs required to perform inference is equal to the number of weights, which is **98816 MACs**
    3. *How much temporary storage (SRAM) is required to run this model? Remember you'll need to store the outputs from one layer and the outputs from the next layer at the same time. The*

*consecutive pair of layers with the largest combined requirements sets the amount of SRAM required.*

1. SRAM$= \max(n_i + n_{i+1}\ \forall i < N-1) = 256 + 256 =$ **512 Bytes SRAM**

2. **Consider a fully-connected network with 1280 inputs, two hidden layers with 512 units each, and 32 output classes. Assume that all weights and activations are stored as 32-bit values.**
   1. *How much parameter storage is required for this model?*
      1. First find parameter count
         1. $|P| = |W| + |B| = \sum_{i=1}^{N} n_{i-1} * n_i + n_i$
         2. $= (1280 * 512 + 512) + (512 * 512 + 512) + (512 * 32 + 32)$
         3. $= 934944$
      2. Then multiply by $\frac{32}{8} = 4$ bytes per parameter
         1. Storage = $934944 * 4 = 3739776$ **Bytes**
   2. *Running this model on an 80MHz processor that can compute 1 MAC every 4 cycles, how long will one inference take (answer in ms).*
      1. MACs = $|W| = \sum_{i=i}^{N} n_{i-1} * n_i = 1280 * 512 + 512 * 512 + 512 * 32 = 933888$
      2. Cycles = MACS x $\frac{\text{cycles}}{\text{MAC}} = 933888 * 4 = 3735552$ Cycles
      3. Time = $\text{Cycles} * \frac{\text{ms}}{\text{Cycle}} = 3735552 * \frac{1}{80 \times 10^3} = 46.6944$ **ms**
   3. *How much temporary storage (SRAM) is required to run this model?*
      1. SRAM$= 4 * \max(n_i + n_{i+1}\ \forall i < N-1) = 4 * (1280 + 512) =$ **7168 Bytes SRAM**