Examples of Quiz 1 questions

* Give the binary representation (bit string) for the decimal number -12.6 in the fixed-point s8.3 format. Use rounding towards nearest value.

* For a neural network, number -2.8 is encoded asymmetrically using 8-bit unsigned integers. What is that unsigned integer and corresponding bit string if representation's slope *s* is 0.2 and the zero-point *z* is 100?

* Determine the value represented by the bit string **1 10101 1010000000** in the case of IEEE754 binary16 floating-point format, which has word length 16, exponent length 5, and exponent bias 15.

* Determine the binary representation (bit string) for the decimal number 101.5, when it is encoded in IEEE754 binary16 floating-point format (has word length 16, exponent length 5, exponent bias 15).

* What is the result if non-saturating 2's complement addition 15+25 is done with 6-bit input/output word length?

* A FIR filter, which has three coefficients -3, 8, and -3, is implemented on a MAC unit, which has 6-bit two's complement signal input. What is the range of the filter output and therefore how many bits are sufficient for the MAC accumulator register?

* Give reasons why a floating-point DSP implementation can be a better option than a fixed-point one.

* A DSP platform offers an integer MAC unit, which takes in 8-bit coefficient/signal values, and the accumulator has word length 16 bits. Saturation is not implemented in the MAC unit. You are implementing a FIR filter, which has 50 coefficients. What would you do to guarantee that accumulator overflows (wrap-ups) are not possible during computation?

* Explain what the normal and subnormal modes are in IEEE754 standard floating-point formats.

* What is truncation? What is its disadvantage in DSP implementation?

* How can overflow problems be avoided in DSP implementation?