FLOATING-POINT NUMBER PROCESSING METHOD AND APPARATUS, COMPUTE DEVICE, AND STORAGE MEDIUM

ABSTRACT

This disclosure, related to the field of computer technologies, discloses a floating-point number processing method and apparatus, a compute device, and a storage medium. The compute device obtains a first floating-point number, and decodes the first floating-point number to obtain a sign, an exponent, and a mantissa. The first floating-point number may include a first sign field, a bit width indication field, a first exponent field, and a first mantissa field. The bit width indication field indicates a bit width D occupied by the first exponent field in a total bit width N of the first floating-point number. When the total bit width is fixed, the bit width of the first exponent field and a bit width of the first mantissa field may dynamically change with a value indicated by the bit width indication field. The bit width indication field may further indicate that the first mantissa field represents the mantissa or a bias of the exponent. Reusing the first mantissa field can ensure that precision of the floating-point number is slightly affected, and can increase coding space of the exponent, thereby achieving a larger dynamic range of a value represented by the floating-point number, and satisfying requirements of AI training and inference.

FIG. 6