

Algorithm 1 Custom floating-point quantization.

Input: $MODEL$ as the CNN.

Input: E_{size} as the target exponent bit size.

Input: M_{size} as the target mantissa bits size.

Input: $STDM_{size}$ as the IEEE 754 mantissa bit size.

Output: $MODEL$ as the quantized CNN.

```
1: for  $layer$  in  $MODEL$  do
2:   if  $layer$  is  $Conv2D$  or  $SeparableConv2D$  then
3:      $filter, bias \leftarrow GetWeights(layer)$ 
4:     for  $x$  in  $filter$  and  $bias$  do
5:        $sign \leftarrow GetSign(x)$ 
6:        $exp \leftarrow GetExponent(x)$ 
7:        $fullexp \leftarrow 2^{E_{size}-1} - 1$  ▷ Get full range value
8:        $cman \leftarrow GetCustomMantissa(x, M_{size})$ 
9:        $leftman \leftarrow GetLeftoverMantissa(x, M_{size})$ 
10:      if  $exp < -fullexp$  then
11:         $x \leftarrow 0$ 
12:      else if  $exp > fullexp$  then
13:         $x \leftarrow (-1)^{sign} \cdot 2^{fullexp} \cdot (1 + (1 - 2^{-M_{size}}))$ 
14:      else
15:        if  $2^{STDM_{size}-M_{size}-1} - 1 < leftman$  then
16:           $cman \leftarrow cman + 1$  ▷ Above halfway
17:        if  $2^{M_{size}} - 1 < cman$  then
18:           $cman \leftarrow 0$  ▷ Correct mantissa overflow
19:           $exp \leftarrow exp + 1$ 
20:        end if
21:      end if
22:       $x \leftarrow (-1)^{sign} \cdot 2^{exp} \cdot (1 + cman \cdot 2^{-M_{size}})$ 
23:    end if
24:  end for
25:   $SetWeights(layer, filter, bias)$ 
26: end if
27: end for
```