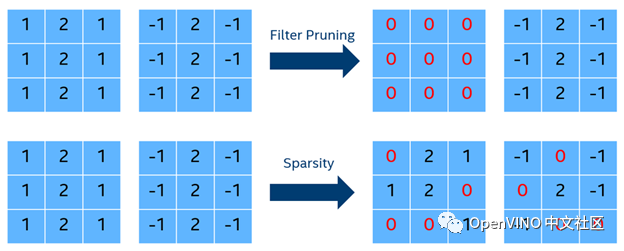
### 1.稀疏（Sparsity）

稀疏是另一种神经网络的压缩方法，目前主流的稀疏方法有以下两种：

· 结构化稀疏（或者叫剪枝），结构化稀疏以后的模型，会比原始模型的体积更小，因为其中一些layer间的通道或连接将被删除。

· 非结构化稀疏，非结构化稀疏后的模型，体积将和原始模型没有区别。但是权重值将被稀疏化（用更多的0来表示）。相较于结构化稀疏，通过非结构化，我们可以去除更多的敏感度不高权重值。



结构化稀疏是删除整层，非结构化稀疏是将该层部分weight置0（非结构化稀疏要依赖硬件进行稀疏矩阵的加速运算，否则无效）

稀疏的核心思想是**去除一些“参与度不高”的权重值**，神经网络中往往包含了远超单纯解决检测问题所需要的权重值，稀疏压缩就会根据每一个权重对于预测结果准确性的影响，进行权重压缩，保留“贡献大”的权重，删除“贡献小”的权重值。

Magnitude-based，此方法将**基于阈值**参数来删除对网络结果贡献较小的权重值（设为0），该方案优势在于它的模型压缩过程将比较快速，在稀疏之前，NNCF会使用权重衰减正则化方法，进一步减小权重值，然后再根据阈值，将其置零。

Regularization-based，此方法将**引入额外的loss function**来计算稀疏后的损失，以确保模型预测准确性下降的范围。在实际操作过程中，该方法将不会直接修改权重值，相反NNCF为再为每个权重值增加一个“重要性”的权重参数，并对其进行训练优化，这样做的好处，在不改变权重值的情况下，模型准确性的下降范围更低，并可以得到有效控制，但是该方法需要更多的训练周期来对模型进行压缩。

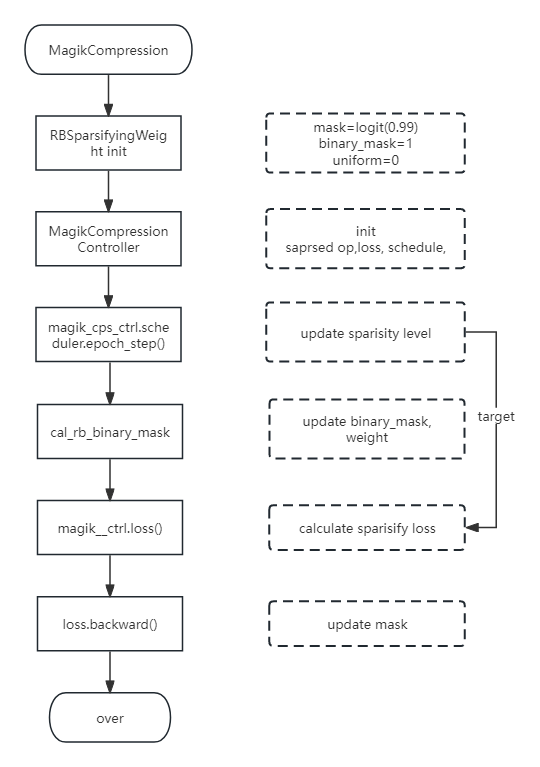
在这两种方法的选择上，如果你的训练样本量比较小，优先推荐Magnitude-based，训练样本量比较大，则可以使用Regularization-based获得更好的准确性。同时建议平滑地调整sparsity ratio，以平衡准确性和压缩率。

### 2.RB-Sparsity(Regularization-Based Sparsity )

为什么叫RB-Sparsity？这是因为：Loss+=SparseLoss-->update mask-->sparse model，而SparseLoss是L0损失函数，原始Loss加上了一个L0惩罚系数。

**原理：<https://github.com/openvinotoolkit/nncf/blob/develop/docs/compression_algorithms/Sparsity.md>**

#### 基本流程：



#### **关键代码：**

def epoch\_step(*self*, *next\_epoch*: Optional[int] = None) -> None:

super().epoch\_step(*next\_epoch*)

if *next\_epoch* is None:

*next\_epoch* = *self*.\_current\_epoch + 1

*self*.\_current\_epoch = *next\_epoch*

*self*.\_update\_sparsity\_level()

def \_update\_sparsity\_level(*self*) -> None:

if *self*.current\_epoch >= *self*.freeze\_epoch:

*self*.\_controller.freeze()

*self*.\_controller.set\_sparsity\_level(*self*.\_calculate\_sparsity\_level())

def \_calculate\_sparsity\_level(*self*) -> float:

current\_density = *self*.schedule(*self*.current\_epoch)

current\_level = 1.0 - current\_density

return min(current\_level, *self*.target\_level)

**ExponentialDecaySchedule. \_\_call\_\_()**

def \_\_call\_\_(*self*, *epoch*: int) -> float:

if *self*.target\_epoch == 0:

return *self*.target\_value

value = *self*.initial\_value \* np.power(*self*.decay\_rate, *epoch* / *self*.target\_epoch)

# decay\_rate = *target\_value* / *initial\_value*

*# 0.99,0.86,0.74,0.64,0.56,0.49(0.01,0.)*

return max(value, *self*.target\_value)

**RBSparsifyingWeight.forward()**

def forward(*self*, *weight*):

if is\_tracing\_state():

return *weight*.mul(*self*.binary\_mask)

tmp\_tensor = *self*.\_calc\_training\_binary\_mask(*weight*)

return apply\_binary\_mask\_impl(tmp\_tensor, *weight*)

def \_calc\_training\_binary\_mask(*self*, *weight*):

u = *self*.uniform if *self*.training and not *self*.frozen else None

if not *self*.frozen:

*self*.binary\_mask = binary\_mask(*self*.\_mask)

return calc\_rb\_binary\_mask(*self*.\_mask, u, *self*.eps)

def calc\_rb\_binary\_mask(*mask*, *uniform\_buffer*, *eps*):

if *uniform\_buffer* is not None:

*uniform\_buffer*.uniform\_()

*mask* = *mask* + logit(*uniform\_buffer*.clamp(*eps*, 1 - *eps*))

return binary\_mask(*mask*)

**SparseLoss.calculate()**

def calculate(*self*) -> torch.Tensor:

if *self*.disabled:

return 0 *## mask loss为0，--mask grad为0*

params = 0

loss = 0

sparse\_prob\_sum = 0

for sparse\_layer in *self*.\_sparse\_layers:

if not *self*.disabled and sparse\_layer.frozen:

raise AssertionError(

"Invalid state of SparseLoss and SparsifiedWeight: mask is frozen for enabled loss"

)

if not sparse\_layer.frozen:

sw\_loss = sparse\_layer.loss() *## 0-1 mask(sparse\_weight.binary\_mask)*

params = params + sw\_loss.view(-1).size(0) *## num of 0-1 mask*

loss = loss + sw\_loss.sum() *## sum(1)*

sparse\_prob\_sum += torch.sigmoid(sparse\_layer.mask).sum() *## mask*

*self*.mean\_sparse\_prob = (sparse\_prob\_sum / params).item() *## change with sparse\_layer.mask*

*self*.current\_sparsity = 1 - loss / params *## no change*

return ((loss / params - *self*.target) / *self*.p).pow(2) *# change with target with epoch\_step()*

**statistics()**

==> Statistics of the sparsified model:

+-----------------------------------------+-------+

| Statistic's name | Value |

+=========================================+=======+

| Sparsity level of the whole model | 0.488 |

+-----------------------------------------+-------+

| Sparsity level of all sparsified layers | 0.488 |

+-----------------------------------------+-------+

Statistics by sparsified layers:

+--------------------+-------------------+----------------+--------------------+

| Layer's name | Weight's shape | Sparsity level | Weight's |

| | | | percentage |

+====================+===================+================+====================+

| model.0.conv | [48, 3, 6, 6] | 0.059 | 0.025 |

+--------------------+-------------------+----------------+--------------------+

| model.1.conv | [96, 48, 3, 3] | 0.084 | 0.199 |

+--------------------+-------------------+----------------+--------------------+

| model.2.cv1.conv | [48, 96, 1, 1] | 0.021 | 0.022 |

+--------------------+-------------------+----------------+--------------------+

| model.2.cv2.conv | [48, 96, 1, 1] | 0.021 | 0.022 |

+--------------------+-------------------+----------------+--------------------+

| model.2.cv3.conv | [96, 96, 1, 1] | 0.030 | 0.044 |

+--------------------+-------------------+----------------+--------------------+

| model.2.m.0.cv1.co | [48, 48, 1, 1] | 0 | 0.011 |

| nv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.2.m.0.cv2.co | [48, 48, 3, 3] | 0 | 0.099 |

| nv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.2.m.1.cv1.co | [48, 48, 1, 1] | 0 | 0.011 |

| nv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.2.m.1.cv2.co | [48, 48, 3, 3] | 0.021 | 0.099 |

| nv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.3.conv | [192, 96, 3, 3] | 0.056 | 0.796 |

+--------------------+-------------------+----------------+--------------------+

| model.4.cv1.conv | [96, 192, 1, 1] | 0.052 | 0.088 |

+--------------------+-------------------+----------------+--------------------+

| model.4.cv2.conv | [96, 192, 1, 1] | 0.011 | 0.088 |

+--------------------+-------------------+----------------+--------------------+

| model.4.cv3.conv | [192, 192, 1, 1] | 0 | 0.177 |

+--------------------+-------------------+----------------+--------------------+

| model.4.m.0.cv1.co | [96, 96, 1, 1] | 0.059 | 0.044 |

| nv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.4.m.0.cv2.co | [96, 96, 3, 3] | 0.033 | 0.398 |

| nv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.4.m.1.cv1.co | [96, 96, 1, 1] | 0.010 | 0.044 |

| nv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.4.m.1.cv2.co | [96, 96, 3, 3] | 0.010 | 0.398 |

| nv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.4.m.2.cv1.co | [96, 96, 1, 1] | 0 | 0.044 |

| nv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.4.m.2.cv2.co | [96, 96, 3, 3] | 0.000 | 0.398 |

| nv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.4.m.3.cv1.co | [96, 96, 1, 1] | 0 | 0.044 |

| nv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.4.m.3.cv2.co | [96, 96, 3, 3] | 0.001 | 0.398 |

| nv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.5.conv | [384, 192, 3, 3] | 0.096 | 3.184 |

+--------------------+-------------------+----------------+--------------------+

| model.6.cv1.conv | [192, 384, 1, 1] | 0.006 | 0.354 |

+--------------------+-------------------+----------------+--------------------+

| model.6.cv2.conv | [192, 384, 1, 1] | 0.083 | 0.354 |

+--------------------+-------------------+----------------+--------------------+

| model.6.cv3.conv | [384, 384, 1, 1] | 0.068 | 0.707 |

+--------------------+-------------------+----------------+--------------------+

| model.6.m.0.cv1.co | [192, 192, 1, 1] | 0.007 | 0.177 |

| nv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.6.m.0.cv2.co | [192, 192, 3, 3] | 0.152 | 1.592 |

| nv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.6.m.1.cv1.co | [192, 192, 1, 1] | 0.005 | 0.177 |

| nv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.6.m.1.cv2.co | [192, 192, 3, 3] | 0.210 | 1.592 |

| nv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.6.m.2.cv1.co | [192, 192, 1, 1] | 0.006 | 0.177 |

| nv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.6.m.2.cv2.co | [192, 192, 3, 3] | 0.283 | 1.592 |

| nv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.6.m.3.cv1.co | [192, 192, 1, 1] | 0.012 | 0.177 |

| nv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.6.m.3.cv2.co | [192, 192, 3, 3] | 0.333 | 1.592 |

| nv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.6.m.4.cv1.co | [192, 192, 1, 1] | 0.023 | 0.177 |

| nv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.6.m.4.cv2.co | [192, 192, 3, 3] | 0.412 | 1.592 |

| nv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.6.m.5.cv1.co | [192, 192, 1, 1] | 0.028 | 0.177 |

| nv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.6.m.5.cv2.co | [192, 192, 3, 3] | 0.451 | 1.592 |

| nv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.7.conv | [768, 384, 3, 3] | 0.826 | 12.735 |

+--------------------+-------------------+----------------+--------------------+

| model.8.cv1.conv | [384, 768, 1, 1] | 0.369 | 1.415 |

+--------------------+-------------------+----------------+--------------------+

| model.8.cv2.conv | [384, 768, 1, 1] | 0.583 | 1.415 |

+--------------------+-------------------+----------------+--------------------+

| model.8.cv3.conv | [768, 768, 1, 1] | 0.411 | 2.830 |

+--------------------+-------------------+----------------+--------------------+

| model.8.m.0.cv1.co | [384, 384, 1, 1] | 0.358 | 0.707 |

| nv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.8.m.0.cv2.co | [384, 384, 3, 3] | 0.824 | 6.367 |

| nv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.8.m.1.cv1.co | [384, 384, 1, 1] | 0.326 | 0.707 |

| nv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.8.m.1.cv2.co | [384, 384, 3, 3] | 0.851 | 6.367 |

| nv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.9.cv1.conv | [384, 768, 1, 1] | 0.163 | 1.415 |

+--------------------+-------------------+----------------+--------------------+

| model.9.cv2.conv | [768, 1536, 1, 1] | 0.479 | 5.660 |

+--------------------+-------------------+----------------+--------------------+

| model.10.conv | [384, 768, 1, 1] | 0.199 | 1.415 |

+--------------------+-------------------+----------------+--------------------+

| model.13.cv1.conv | [192, 768, 1, 1] | 0.190 | 0.707 |

+--------------------+-------------------+----------------+--------------------+

| model.13.cv2.conv | [192, 768, 1, 1] | 0.120 | 0.707 |

+--------------------+-------------------+----------------+--------------------+

| model.13.cv3.conv | [384, 384, 1, 1] | 0.071 | 0.707 |

+--------------------+-------------------+----------------+--------------------+

| model.13.m.0.cv1.c | [192, 192, 1, 1] | 0.045 | 0.177 |

| onv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.13.m.0.cv2.c | [192, 192, 3, 3] | 0.469 | 1.592 |

| onv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.13.m.1.cv1.c | [192, 192, 1, 1] | 0.039 | 0.177 |

| onv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.13.m.1.cv2.c | [192, 192, 3, 3] | 0.457 | 1.592 |

| onv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.14.conv | [192, 384, 1, 1] | 0.032 | 0.354 |

+--------------------+-------------------+----------------+--------------------+

| model.17.cv1.conv | [96, 384, 1, 1] | 0.000 | 0.177 |

+--------------------+-------------------+----------------+--------------------+

| model.17.cv2.conv | [96, 384, 1, 1] | 0.020 | 0.177 |

+--------------------+-------------------+----------------+--------------------+

| model.17.cv3.conv | [192, 192, 1, 1] | 0.002 | 0.177 |

+--------------------+-------------------+----------------+--------------------+

| model.17.m.0.cv1.c | [96, 96, 1, 1] | 0 | 0.044 |

| onv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.17.m.0.cv2.c | [96, 96, 3, 3] | 0.001 | 0.398 |

| onv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.17.m.1.cv1.c | [96, 96, 1, 1] | 0 | 0.044 |

| onv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.17.m.1.cv2.c | [96, 96, 3, 3] | 0.001 | 0.398 |

| onv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.18.conv | [192, 192, 3, 3] | 0.221 | 1.592 |

+--------------------+-------------------+----------------+--------------------+

| model.20.cv1.conv | [192, 384, 1, 1] | 0.076 | 0.354 |

+--------------------+-------------------+----------------+--------------------+

| model.20.cv2.conv | [192, 384, 1, 1] | 0.052 | 0.354 |

+--------------------+-------------------+----------------+--------------------+

| model.20.cv3.conv | [384, 384, 1, 1] | 0.079 | 0.707 |

+--------------------+-------------------+----------------+--------------------+

| model.20.m.0.cv1.c | [192, 192, 1, 1] | 0.028 | 0.177 |

| onv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.20.m.0.cv2.c | [192, 192, 3, 3] | 0.313 | 1.592 |

| onv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.20.m.1.cv1.c | [192, 192, 1, 1] | 0.020 | 0.177 |

| onv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.20.m.1.cv2.c | [192, 192, 3, 3] | 0.302 | 1.592 |

| onv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.21.conv | [384, 384, 3, 3] | 0.632 | 6.367 |

+--------------------+-------------------+----------------+--------------------+

| model.23.cv1.conv | [384, 768, 1, 1] | 0.352 | 1.415 |

+--------------------+-------------------+----------------+--------------------+

| model.23.cv2.conv | [384, 768, 1, 1] | 0.223 | 1.415 |

+--------------------+-------------------+----------------+--------------------+

| model.23.cv3.conv | [768, 768, 1, 1] | 0.328 | 2.830 |

+--------------------+-------------------+----------------+--------------------+

| model.23.m.0.cv1.c | [384, 384, 1, 1] | 0.236 | 0.707 |

| onv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.23.m.0.cv2.c | [384, 384, 3, 3] | 0.707 | 6.367 |

| onv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.23.m.1.cv1.c | [384, 384, 1, 1] | 0.157 | 0.707 |

| onv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.23.m.1.cv2.c | [384, 384, 3, 3] | 0.664 | 6.367 |

| onv | | | |

+--------------------+-------------------+----------------+--------------------+

| model.24.m.0 | [24, 192, 1, 1] | 0.006 | 0.022 |

+--------------------+-------------------+----------------+--------------------+

| model.24.m.1 | [24, 384, 1, 1] | 0.084 | 0.044 |

+--------------------+-------------------+----------------+--------------------+

| model.24.m.2 | [24, 768, 1, 1] | 0.146 | 0.088 |

+--------------------+-------------------+----------------+--------------------+

Statistics of the RB-sparsity algorithm:

+----------------------------------------------------------------------+-------+

| Statistic's name | Value |

+======================================================================+=======+

| A target level of the sparsity for the algorithm for the current | 0.510 |

| epoch | |

+----------------------------------------------------------------------+-------+

| The probability that one weight will be zeroed | 0.480 |

+----------------------------------------------------------------------+-------+

### 补充：

1. 正则化

* L0正则化的值是模型参数中**非零参数**的个数。
* L1正则化表示各个参数**绝对值**之和。
* L2正则化标识各个参数的**平方和的开方值**。

1. 大数定律

这是因为根据大数定律，当我们从概率分布 Ps 中抽取足够多的样本时，这些样本的平均值会逐渐收敛到随机变量s的数学期望。换句话说，随着样本数量的增加，样本的平均值将越来越接近真实的数学期望。

这种现象是由大数定律保证的，它表明在一定条件下，随机变量序列的算术平均值会收敛到其数学期望。因此，通过从概率分布中采样一组值，并计算这些值的平均值，我们可以得到对数学期望的估计。

需要注意的是，这种估计方法依赖于样本数量的大小以及样本的独立性和同分布性。通常情况下，当样本数量足够大时，这种估计方法会提供一个较为准确的数学期望估计。