Specification Document:

**Hyper Parameters:**

1. **Patch** **Size** – The size of the estimator block. In the **baseline** case (Diagonal) for a value of 2 we would receive the matrix:

****

For a PS of 3 in the baseline case, this would amount to:

****

**Effects of Patch Size** –

* The number of combinatorial options to choose from. Notice that the maximum number of options for a patch size of **p** is:

****

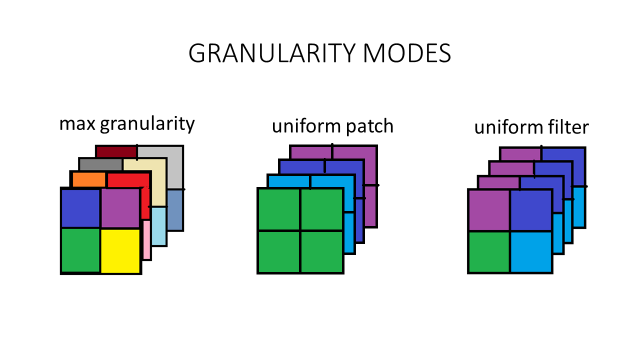
Remember that these options do take **less** space as p grows, seeing if the input data is 32x32, it fits  total patches for p=2, and  patches. But do to the fact that for the **Max-Granularity mode** we optimize to the “patch”, each different option of a patch would require a full forward run through the net. On ResNet18, there are 17 different **Spatial Layers** along with 17 Conv2d layers, RELU and more. Therefore, as p grows, we can expect an **exponential** increase in **runtime complexity.**

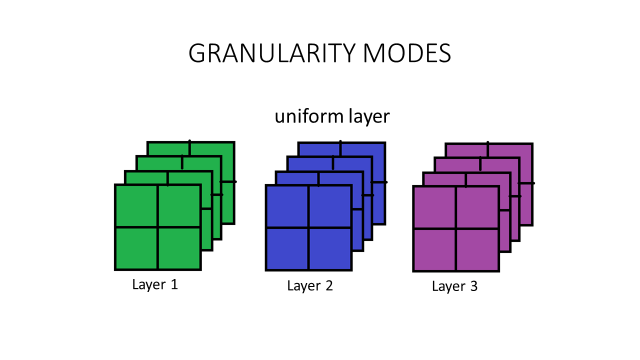
We can decimate some options with the **ones\_range** hyper-parameter – a range specifying which options are valid to us by the number of ones. For p=2, with a ones\_range of  we would receive: (All options for number of ones == 1, or number of ones == 2)

****

* **Spatial** **Locality** - The smaller the PS, the more spatial locality we can expect. We probably do not want to meddle with higher ps for that reason.

1. **Optimization Mode** – Four options:





Our algorithm would “**quantize**” the results. From the Max-Granularity that tests the effects of each patch on the entire net, we would then quantize over the channels, to find the best combinations that suit each channel. Lastly, we would quantize over the layers, finally reaching the best possibility feasible over the entire network.

Uniform Layer mode would simply run quantize over the layers, and then find the best possibility for the entire network ( A much easier and simpler optimization task )

1. **Granularity Threshold Parameter -** Fearing that the effects of a 2x2 patch on the entire network would not be substantial, we increase the spread of each mask to something that is more substantial, as we talked about today in the office.
2. **Neural Network -** A simple inheritance structure has been implemented, which allows for a simple installation of the Spatial Layers to any network you would like. Insertion of a new net is a very simple process. As of now, we can easily create any ResNet model immediately, and have been working on ResNet18 and ResNet30. Remember – more spatial layers, more optimization complexity.
3. **Dataset –** Currently working on CIFAR10. TEST\_SET\_SIZE is also a parameter we should consider. On CIFAR10 – we have 10000 images, of which we can choose to test only on 1000 for example, for much faster forward runs.

* I’ve looked at ImageNet today – it is **140GB** dataset. I’m not sure this is feasible (See next section)

**We have created timing approximations of all runs given the following parameters:**

* ResNet18
* Patch Size = 2
* Ones Range of [1,3)
* Granularity Threshold of 10 ( - the higher, the less effect this has)
* Test Set Size of 10000

**Total Runtime approximation:**

* **Max Granularity – 15.66 Days**
* **Uniform Filters – 1.13 Days**
* **Uniform Patch - 2.64 Days**
* **Uniform Layers – 0.75 Hours**

**If we drop the test set size to 1000:**

* **Max Granularity – 1.78 Days**
* **Uniform Filters – 3.05 Hours**
* **Uniform Patch - 7.27 Hours**
* **Uniform Layers – 0.08 Hours**

**Baseline Results for the diagonal option with Uniform Layer optimization:**

BASELINE FOR: ResNet18Spatial CIFAR10

PATCH SIZE: 2

TEST SET SIZE: 1000

----------------------------------------------------------------

operations saved: 22.953%

with accuracy of: 90.0% (Drop of 3.83%)

================================================================

================================================================

BASELINE FOR: ResNet18Spatial CIFAR10

PATCH SIZE: 2

TEST SET SIZE: 10000

----------------------------------------------------------------

operations saved: 23.241%

with accuracy of: 89.94% (Drop of 3.89%)

**Findings of MULTS in the different layers (After 1 forward run with a 100% “saved” ops – a mask of zeros on all Spatial Layers)**

Spatial Layer 0: Ops saved: 100.000 % [71368704 / 71368704] of [10.386 % / 10.386 %]

Spatial Layer 1: Ops saved: 100.000 % [71368704 / 71368704] of [10.386 % / 10.386 %]

Spatial Layer 2: Ops saved: 100.000 % [71368704 / 71368704] of [10.386 % / 10.386 %]

Spatial Layer 3: Ops saved: 100.000 % [71368704 / 71368704] of [10.386 % / 10.386 %]

Spatial Layer 4: Ops saved: 100.000 % [71368704 / 71368704] of [10.386 % / 10.386 %]

Spatial Layer 5: Ops saved: 100.000 % [42467328 / 42467328] of [6.180 % / 6.180 %]

Spatial Layer 6: Ops saved: 100.000 % [42467328 / 42467328] of [6.180 % / 6.180 %]

Spatial Layer 7: Ops saved: 100.000 % [42467328 / 42467328] of [6.180 % / 6.180 %]

Spatial Layer 8: Ops saved: 100.000 % [42467328 / 42467328] of [6.180 % / 6.180 %]

Spatial Layer 9: Ops saved: 100.000 % [21233664 / 21233664] of [3.090 % / 3.090 %]

Spatial Layer 10: Ops saved: 100.000 % [21233664 / 21233664] of [3.090 % / 3.090 %]

Spatial Layer 11: Ops saved: 100.000 % [21233664 / 21233664] of [3.090 % / 3.090 %]

Spatial Layer 12: Ops saved: 100.000 % [21233664 / 21233664] of [3.090 % / 3.090 %]

Spatial Layer 13: Ops saved: 100.000 % [18874368 / 18874368] of [2.747 % / 2.747 %]

Spatial Layer 14: Ops saved: 100.000 % [18874368 / 18874368] of [2.747 % / 2.747 %]

Spatial Layer 15: Ops saved: 100.000 % [18874368 / 18874368] of [2.747 % / 2.747 %]

Spatial Layer 16: Ops saved: 100.000 % [18874368 / 18874368] of [2.747 % / 2.747 %]

Grand total: 687144960/687144960 100.0 %

**Translation:**

1. Layer index ( Layer 16 is right after the 17th Conv2D layer, and right before Dense)
2. % of Ops saved from Total ops
3. Number of Ops Saved
4. Number of Total ops in layer
5. Percentage of Ops saved from ALL saved operations
6. Percentage of Ops from all operations

We the network gets deeper, we see less and less operations overall.

**Appendix**

================================================================

----------------------------------------------------------------

NET: ResNet18Spatial

DATASET: CIFAR10

PATCH SIZE: 2

ONES: 1-2

GRANULARITY: 10

TEST SET SIZE: 10000

----------------------------------------------------------------

==> Asserted test-acc of: 93.83 [9383/10000]

Mode.MAX\_GRANULARITY 15.66 [days]

----------------------------------------------------------------

iters 1st lvl iters 2nd lvl iters lQ

number 197760 39210 170

time 1127825 223615 1780

sec per iter

1st/2nd lvl: 5.703

lQ: 10.469

----------------------------------------------------------------

==> Asserted test-acc of: 93.83 [9383/10000]

Mode.UNIFORM\_FILTERS 1.13 [days]

----------------------------------------------------------------

iters 1st lvl iters 2nd lvl iters lQ

number 16160 170 170

time 94730 997 1780

sec per iter

1st/2nd lvl: 5.862

lQ: 10.469

----------------------------------------------------------------

==> Asserted test-acc of: 93.83 [9383/10000]

Mode.UNIFORM\_PATCH 2.64 [days]

----------------------------------------------------------------

iters 1st lvl iters 2nd lvl iters lQ

number 39040 170 170

time 224948 980 1780

sec per iter

1st/2nd lvl: 5.762

lQ: 10.469

----------------------------------------------------------------

==> Asserted test-acc of: 93.83 [9383/10000]

Mode.UNIFORM\_LAYER 0.75 [hours]

----------------------------------------------------------------

iters 1st lvl iters 2nd lvl iters lQ

number 170 0 170

time 936 0 1780

sec per iter

1st/2nd lvl: 5.505

lQ: 10.469

----------------------------------------------------------------

================================================================

================================================================

----------------------------------------------------------------

NET: ResNet18Spatial

DATASET: CIFAR10

PATCH SIZE: 2

ONES: 1-2

GRANULARITY: 10

TEST SET SIZE: 1000

----------------------------------------------------------------

==> Asserted test-acc of: 93.5 [935/1000]

Mode.MAX\_GRANULARITY 1.78 [days]

----------------------------------------------------------------

iters 1st lvl iters 2nd lvl iters lQ

number 197760 39210 170

time 128148 25408 186

sec per iter

1st/2nd lvl: 0.648

lQ: 1.096

----------------------------------------------------------------

==> Asserted test-acc of: 93.5 [935/1000]

Mode.UNIFORM\_FILTERS 3.05 [hours]

----------------------------------------------------------------

iters 1st lvl iters 2nd lvl iters lQ

number 16160 170 170

time 10698 113 186

sec per iter

1st/2nd lvl: 0.662

lQ: 1.096

----------------------------------------------------------------

==> Asserted test-acc of: 93.5 [935/1000]

Mode.UNIFORM\_PATCH 7.27 [hours]

----------------------------------------------------------------

iters 1st lvl iters 2nd lvl iters lQ

number 39040 170 170

time 25884 113 186

sec per iter

1st/2nd lvl: 0.663

lQ: 1.096

----------------------------------------------------------------

==> Asserted test-acc of: 93.5 [935/1000]

Mode.UNIFORM\_LAYER 0.08 [hours]

----------------------------------------------------------------

iters 1st lvl iters 2nd lvl iters lQ

number 170 0 170

time 109 0 186

sec per iter

1st/2nd lvl: 0.644

lQ: 1.096

----------------------------------------------------------------

================================================================