ML Repro Report

Scope of reproducibility:

The main thing I was trying to reproduce was their result on ILSVRC, i.e. top1 and top5 accuracy on classification task into 1000 classes of imagenet data.

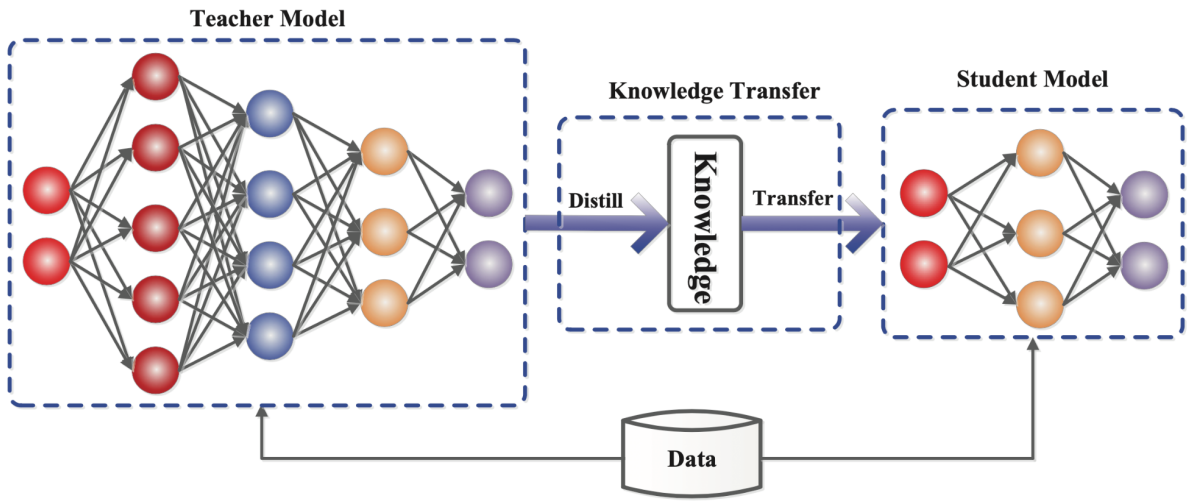
Introduction:

paper link:

https://arxiv.org/pdf/2305.19412.pdf

The paper's main aim was to carry out a first-of-its-kind study unveiling that modern large-kernel ConvNets are remarkably more effective teachers for small-kernel ConvNets due to more similar architectures than vision transformers.

Here, the technique referred to is knowledge distillation, which transfers knowledge from a large, unwieldy model or set of models to a single, smaller model that can be practically deployed under real-world constraints.



In the paper, the teacher model used was SLaK which was introduced recently in this paper→

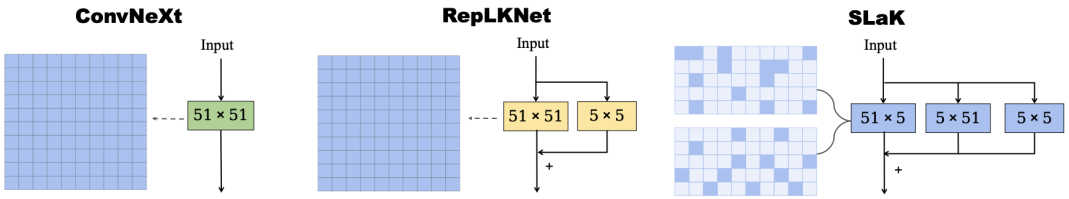
https://arxiv.org/pdf/2207.03620.pdf

SLak is inspired by a large kernel convent ‘RepLKnet’, introduced recently with improved performances with a kernel size 31*31. But performance tends to saturate in bigger kernels in RepLKnet.

SLaK applies extremely large kernels from the perspective of sparsity, which can smoothly scale up kernels to 61×61 with better performance.

FAILURES OF EXISTING APPROACHES TO GO BEYOND 31×31 KERNELS:

One plausible explanation is that although the receptive field may be enlarged by using extremely large kernels, it might fail to maintain the desirable property of locality.



Also the student model used was ConvNext which is a CNN with kernel size 7*7.

I trained my own slak model and used pretrained weights for convnext.

Code implementation:

All the code was implemented on kaggle notebooks using the P100 GPU.

The code could be rerun with just one modification: changing the paths where needed. Rest all the modifications to the code have been committed to forked repos of authors by me.

Code to install all dependencies is given in the ipynb files.

Results on SLaK:

Authors results: Based on imagenet-1k dataset trained for 300 epochs

model	image size	parameters	top1 acc	top5 acc
SLaK-t	224*224	30M	82.5	-
SLaK-b	224*224	55M	83.8	-
SLaK-s	224*224	91M	84.0	-

Reproduced results: Based on CIFAR-100 dataset

model	image size	parameters	top1 acc	top5 acc
SLaK-t(300 epochs)	32*32	30M	69.4	89.29
SLaK-b(150 epochs)	32*32	55M	58.05	85.64
SLaK-s	32*32	-	-	-

unable to train SLaK-S due to limited GPU access.

SLak-t: It took almost 2 hours to train on Kaggle.

SLak-b: It took almost 9 hours to train on Kaggle.

SLaK-t

```
Test:  [  0/105]  eta: 0:01:43  loss: 1.2752 (1.2752)  acc1: 73.9583 (73.9583)  acc5: 89.5833 (89.5833)
time: 0.9860  data: 0.4085  max mem: 3258
Test:  [ 10/105]  eta: 0:00:10  loss: 1.2752 (1.3414)  acc1: 70.8333 (70.7386)  acc5: 89.5833 (88.8258)
time: 0.1090  data: 0.0373  max mem: 3258
Test:  [ 20/105]  eta: 0:00:05  loss: 1.3222 (1.3170)  acc1: 68.7500 (70.5853)  acc5: 89.5833 (89.4345)
time: 0.0216  data: 0.0001  max mem: 3258
Test:  [ 30/105]  eta: 0:00:03  loss: 1.3222 (1.3303)  acc1: 67.7083 (69.9261)  acc5: 88.5417 (89.0457)
time: 0.0210  data: 0.0002  max mem: 3258
Test:  [ 40/105]  eta: 0:00:02  loss: 1.3762 (1.3520)  acc1: 66.6667 (69.1311)  acc5: 87.5000 (89.0244)
time: 0.0213  data: 0.0002  max mem: 3258
Test:  [ 50/105]  eta: 0:00:02  loss: 1.4296 (1.3695)  acc1: 65.6250 (68.6683)  acc5: 87.5000 (88.8276)
time: 0.0217  data: 0.0002  max mem: 3258
Test:  [ 60/105]  eta: 0:00:01  loss: 1.3617 (1.3554)  acc1: 67.7083 (68.9037)  acc5: 87.5000 (89.0027)
time: 0.0213  data: 0.0002  max mem: 3258
Test:  [ 70/105]  eta: 0:00:01  loss: 1.3007 (1.3606)  acc1: 69.7917 (68.8967)  acc5: 88.5417 (88.9231)
time: 0.0211  data: 0.0002  max mem: 3258
Test:  [ 80/105]  eta: 0:00:00  loss: 1.3484 (1.3615)  acc1: 68.7500 (68.8915)  acc5: 90.6250 (89.0046)
time: 0.0219  data: 0.0002  max mem: 3258
Test:  [ 90/105]  eta: 0:00:00  loss: 1.2944 (1.3491)  acc1: 70.8333 (69.1850)  acc5: 90.6250 (89.1598)
time: 0.0204  data: 0.0001  max mem: 3258
Test:  [100/105]  eta: 0:00:00  loss: 1.1760 (1.3418)  acc1: 70.8333 (69.3379)  acc5: 89.5833 (89.2327)
time: 0.0175  data: 0.0001  max mem: 3258
Test:  [104/105]  eta: 0:00:00  loss: 1.1760 (1.3391)  acc1: 71.8750 (69.4000)  acc5: 91.6667 (89.2900)
time: 0.0200  data: 0.0001  max mem: 3258
Test: Total time: 0:00:03 (0.0310 s / it)
* Acc@1 69.400 Acc@5 89.290 loss 1.339
Accuracy of the network on 10000 test images: 69.40000%
```

Notebook editor cells

SLaK-b

```
Test:  [  0/105]  eta: 0:01:58  loss: 1.5693 (1.5693)  acc1: 61.4583 (61.4583)  acc5: 87.5000 (87.5000)
time: 1.1276  data: 0.4827  max mem: 6294
Test:  [ 10/105]  eta: 0:00:13  loss: 1.6251 (1.6817)  acc1: 61.4583 (58.9962)  acc5: 87.5000 (86.0795)
time: 0.1453  data: 0.0440  max mem: 6294
Test:  [ 20/105]  eta: 0:00:08  loss: 1.6294 (1.6753)  acc1: 60.4167 (58.7302)  acc5: 86.4583 (86.1607)
time: 0.0460  data: 0.0002  max mem: 6294
Test:  [ 30/105]  eta: 0:00:06  loss: 1.6584 (1.6863)  acc1: 58.3333 (58.5013)  acc5: 86.4583 (85.7863)
time: 0.0449  data: 0.0002  max mem: 6294
Test:  [ 40/105]  eta: 0:00:04  loss: 1.6598 (1.6943)  acc1: 57.2917 (57.7236)  acc5: 84.3750 (85.5945)
time: 0.0448  data: 0.0002  max mem: 6294
Test:  [ 50/105]  eta: 0:00:03  loss: 1.6722 (1.7032)  acc1: 56.2500 (57.3325)  acc5: 84.3750 (85.4167)
time: 0.0447  data: 0.0001  max mem: 6294
Test:  [ 60/105]  eta: 0:00:02  loss: 1.6627 (1.6878)  acc1: 58.3333 (57.9064)  acc5: 86.4583 (85.8094)
time: 0.0447  data: 0.0001  max mem: 6294
Test:  [ 70/105]  eta: 0:00:02  loss: 1.6627 (1.6936)  acc1: 57.2917 (57.7612)  acc5: 87.5000 (85.7835)
time: 0.0448  data: 0.0001  max mem: 6294
Test:  [ 80/105]  eta: 0:00:01  loss: 1.7497 (1.7022)  acc1: 55.2083 (57.7161)  acc5: 84.3750 (85.5581)
time: 0.0448  data: 0.0001  max mem: 6294
Test:  [ 90/105]  eta: 0:00:00  loss: 1.7033 (1.6956)  acc1: 59.3750 (57.9441)  acc5: 85.4167 (85.5769)
time: 0.0446  data: 0.0001  max mem: 6294
Test:  [100/105]  eta: 0:00:00  loss: 1.6650 (1.6944)  acc1: 59.3750 (58.0136)  acc5: 86.4583 (85.6229)
time: 0.0444  data: 0.0001  max mem: 6294
Test:  [104/105]  eta: 0:00:00  loss: 1.6467 (1.6905)  acc1: 59.3750 (58.0500)  acc5: 86.4583 (85.6400)
time: 0.0470  data: 0.0001  max mem: 6294
Test: Total time: 0:00:05 (0.0567 s / it)
* Acc@1 58.050 Acc@5 85.640 loss 1.690
Accuracy of the network on 10000 test images: 58.05000%
```

- Unfortunately, I could not distil the above models with pre-trained convNext.
- So, no reproduced results after distillation.