Medical Image Classification with Normal , Reversed and Defective Knowledge Based Training Strategies

A Project / Dissertation as a Course requirement for

Master of Sciences in Data Science and Computing

A Manickavela

19226

SRI SATHYA SAI INSTITUTE OF HIGHER LEARNING
(Deemed to be University)

Department of Mathematics and Computer Science

Muddenahalli Campus

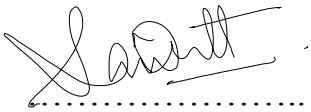
April 2020

SRI SATHYA SAI INSTITUTE OF HIGHER LEARNING
(Deemed to be University)

Dept. of **Mathematics & Computer Science**
Muddenahalli Campus

CERTIFICATE

This is to certify that this Project / Dissertation titled **Medical Image Classification with Normal , Reversed and Defective Knowledge Based Training Strategies** submitted by **A Manickavela, 19226**, Department of Mathematics and Computer Science, Muddenahalli Campus is a bonafide record of the original work done under my/our supervision as a Course requirement for the Degree of Master of Science in Data Science and Computing.



……..

Dr Sampath Lonka

Project Supervisor

Countersigned by

……...

Dr. Rita Gupta
Head of the Department

Place: Muddenahalli

Date: 20-04-2021

DECLARATION

The Project / Dissertation titled **Medical Image Classification with Normal , Reversed and Defective Knowledge Based Training Strategies** was carried out by me under the supervision of Dr Sampath Lonka, Department of Mathematics and Computer Science , Muddenahalli campus as a Course requirement for the Degree of Master of Science in Data Science and Computing and has not formed the basis for the award of any degree, diploma or any other such title by this or any other University.

.....

A Manickavela

19226

II Msc

Muddenahalli Campus

Place : Muddenahalli

Date : 20-04-2021

ACKNOWLEDGEMENTS

This project work has involved a lot of unseen hands without which it wouldn't have been possible, I thank all those who have extended their helping hand for bringing in Support , Guidance and Motivation throughout this work.

Foremost I would like to thank **Bhagawan Sri Sathya Sai Baba** for constantly being there to instill motivation and desire for work with great self-confidence. If not for his constant support and reminders this work would no where be where it is.

I thank my Professor guide , **Shri Dr Sampath Lonka**, for his constant support,guidance ,trust and the confidence he had in me which had kept me going.

I am also grateful to **Shri Sai Natarajan**,MSCA Researcher , Barcelona for coming up with the idea , for his constant guidance and pointers at each stage of the work and inspiration. I thank him for his timely help and advice in the technological front.

I would also like to thank all the faculty members at the Department of Mathematics and Computer Science, Sri Sathya Sai Institute of Higher Learning, for their timely support and feedback.

I also thank all my classmates for their motivation, and perceptions.

I thank my parents and sister for their constant support and belief in me which has made me pick myself at each fall.

Table of Contents

1.Objective

To explore , experiment and analyse the 3 of many different Knowledge Distillation strategies, namely, Normal Knowledge Distillation, Reverse Knowledge Distillation and Defective Knowledge Distillation performance on Classification of Medical Image data.

2.Motivation

With Deep learning ubiquitous in vastly varied fields and domains , several industries are trying to incorporate them with several researchers. Computing power which is required to Deploy or train them has somewhat become a bottleneck. Knowledge Distillation is one such method which is used to get over this issue. This was the motivation to take up on this topic and explore further.

3.Introduction

Medical Image classification is a prominent problem in the domain of image recognition. It aims to classify images into different categories that would further help the Doctors for diagnosis or further research. Doctors usually use their experience gained over time to extract out features and then classify them to different classes , this task is difficult , boring , tedious and this approach is also insatiable and leads to unrepeatable outcomes.

The advent of CNN has led to several SOTA architectures being proposed and there are numerous published works which have shown very good results with respect to Medical Image Classification.

Knowledge Distillation is a well tested knowledge transfer concept in Neural networks. The black box of knowledge hidden in the teacher model's(complex and sophisticated) logit layer is used to improve the student models(small). After the normal(i.e vanilla) Knowledge distillation was proposed for neural networks by Hinton. Several new strategies have come up on the same lines.Two such that we will explore are Reverse Knowledge Distillation and Defective Knowledge Distillation.

4. Methods and Algorithms Used

4.1. Neural Networks

Neural Networks are a class of algorithms that are inspired by biological neurons , they try to mimic the operations of the human brain to recognize the underlying relationship and patterns between the vast amount of data.

An ANN is made up of cells that are connected layerly and work together to produce a desired result. A cost function is defined for the network which estimates how well the algorithm is doing for an assigned task.

Deep Neural Networks is an ANN with multiple layers between the input and output layers ,these multiple layers. DNN can model complex non-linear relationships

4.2. CNN Architectures

Convolutional Neural Networks are a class of Deep Neural Network that recognizes and classifies the spatial features from images , they are widely used for analysis of Images. But with few preprocessing they can be applied to other forms of data too.

CNN architectures generally apply the convolutional layers successively to the input data , continuously downsampling the spatial dimension while increasing the number of feature maps.

It is these architectures that act feature extractors for image classification , segmentation , object detection and many more

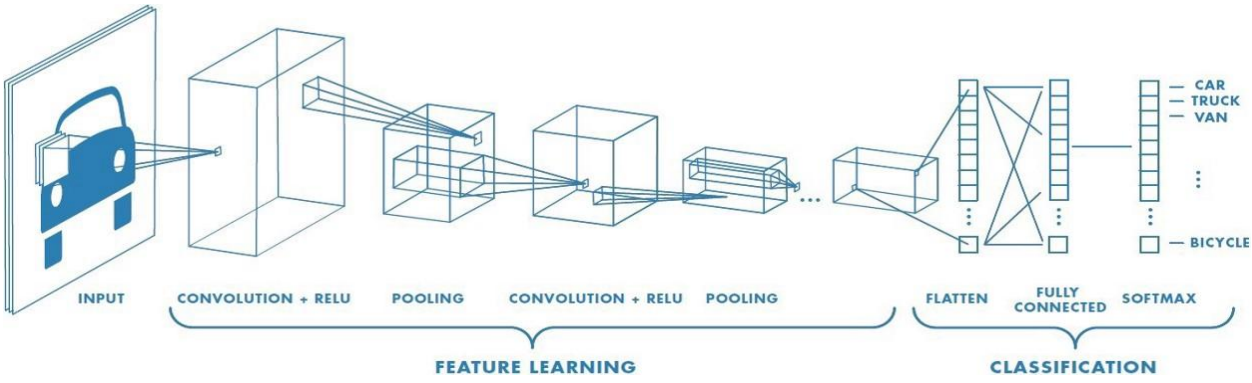


Image Source :1

4.3. Residual Network (ResNet)

Increase in depth of the very Deep NN led to the problem of Vanishing gradients. It was overcome by using Residual Networks . It uses a technique called skip connection. This skip connection as the name suggests skips the training over a few layers and connects directly to the output of a layer ahead of it.

Residual Block becomes the basic block of residual network.

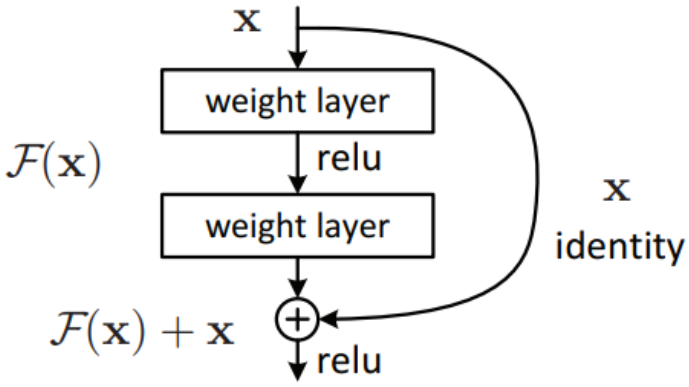


Image Source : 2

Residual network takes inspiration for various aspects from its predecessors like Alexnet, LeNet ,VGG and etc. But the network is built much deeper and it delivers exceptionally good performance.

ResNet has many variants to it. It varies with the number of layers that makes the network deeper.

Well known Resnet variants are Resnet-18, Resnet-34, Resnet-50, Resnet-101

Resnet110 and Resnet-152.

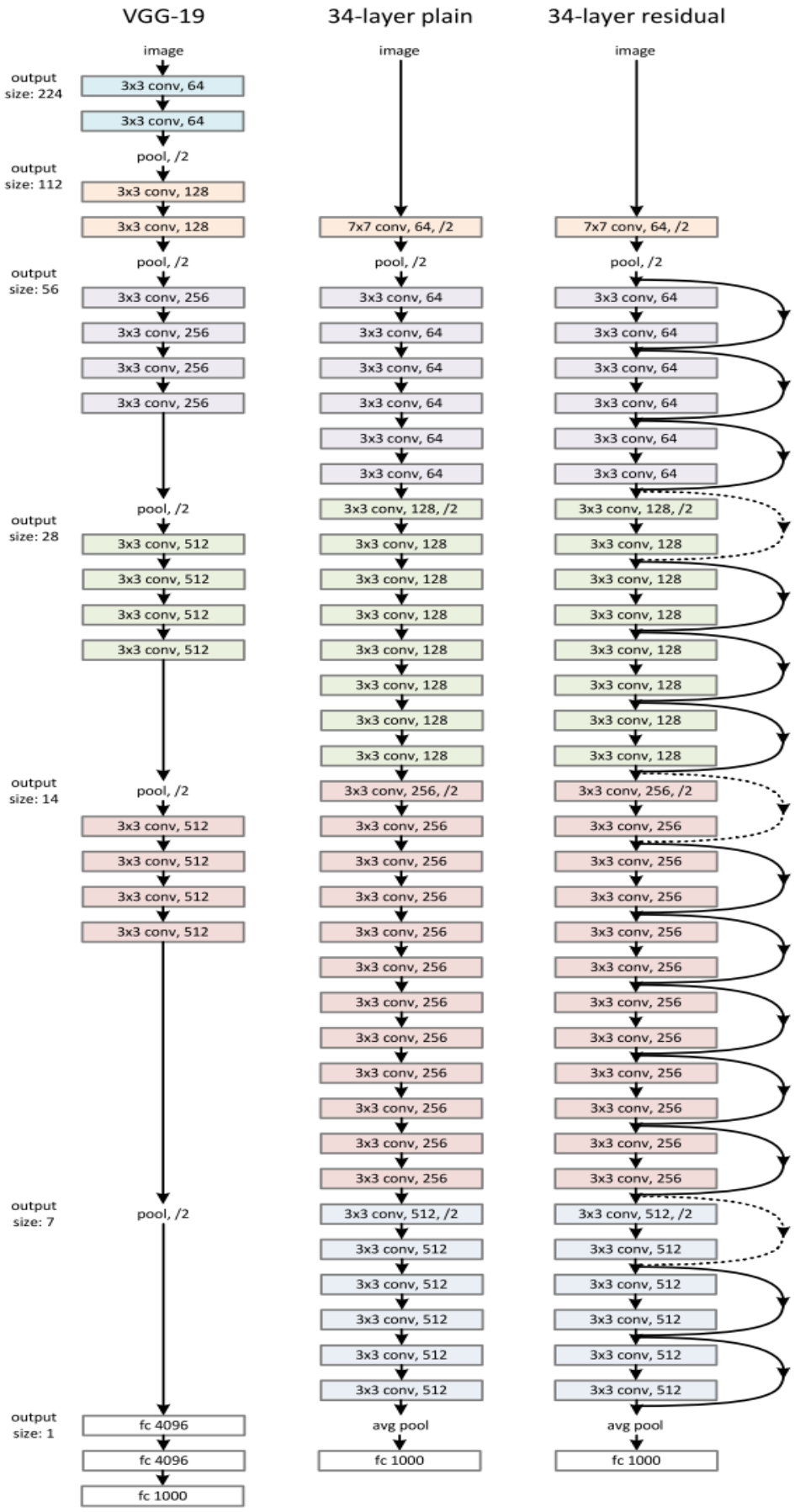


Image Source : 3

4.4. Squeeze and Excitation network (SeNet)

This architecture breaks itself from the monotony of its predecessor's habit of increasing the number of layers to improve the performance. One important characteristic that has been left out in the previous architectures is missing out on channel dependencies. Convolution aggregates all the channel information. They have introduced an new architectural design to capture the channel relationships. This architectural unit is called Squeeze-and-Excitation(SE) block.It Selectively emphasizes informative features and suppresses less useful ones.

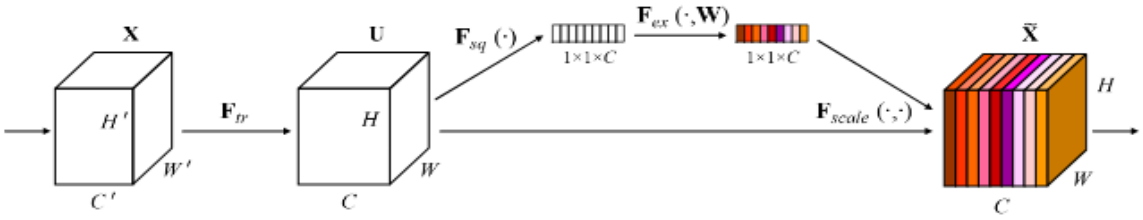


Image Source : 4

SE block can be used with any standard architectures, ResNet is one such.

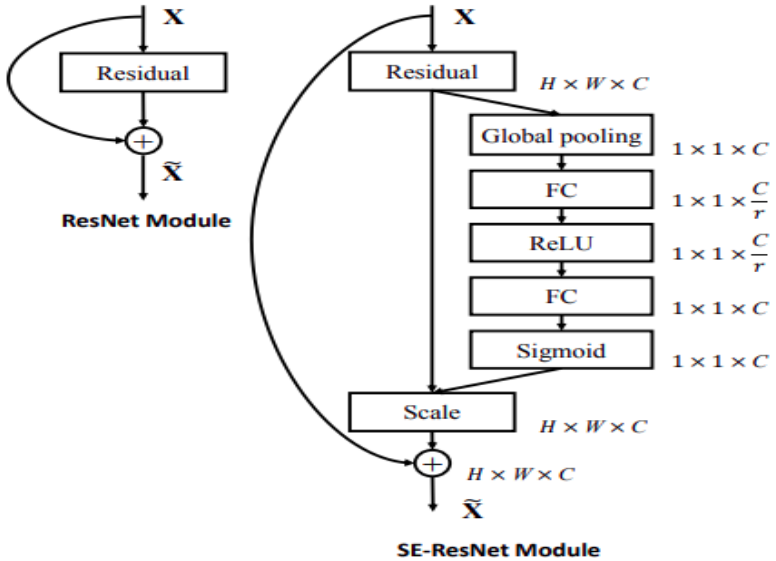


Image Source : 5

With SE block as the basic block for ResNet it is termed SeResnet.

All the Resnet variants take up the same form but with SeNet block with the name Sersnet. Example Resnet18 -> SeResNet18.

4.4. VGG-11

VGG stands for Visual Geometry Group. This architecture is an inspiration from AlexNet which incorporated ReLu activation function in place of commonly used tanh. It was also an relatively deeper network compared to AlexNet. It was stacked with convolution and pooling layers.

5. Knowledge Distillation

Knowledge Distillation is a technology through which learned knowledge is transferred from a large cumbersome model to a smaller model, these 2 models are termed as Teacher and Student model.

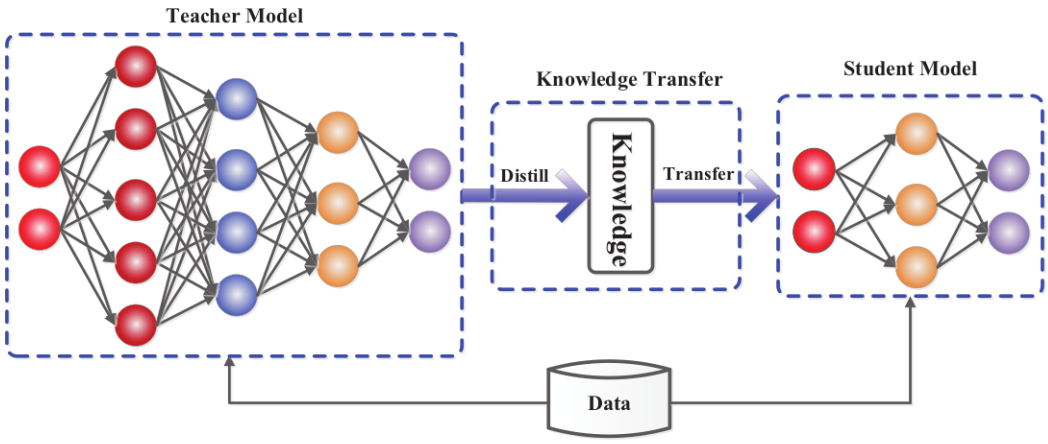


Image Source : 5

Several models fail miserably at the time of deployment as they still require a fair amount of GPU or other compute support. This issue becomes much bigger when the model is for incorporation with an Edge or Iot device as they can't meet the required hardware specificities. Knowledge Distillation helps us overcome this problem by transferring the knowledge specifically learnt by the Teacher model to the small Student model, a light weight model which is better fit for deployment..

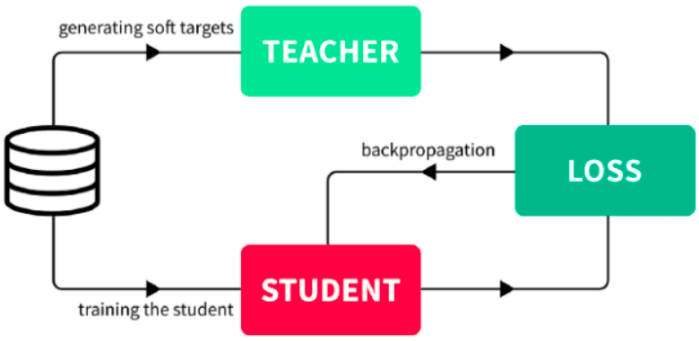


Image Source :6

The Teacher model could either be a very cumbersome model or an ensemble of separately trained models, once the Teacher model is trained, it's learnt knowledge is transferred to a small Student model that is more suitable for the deployment. This knowledge transfer is called 'Distillation'.

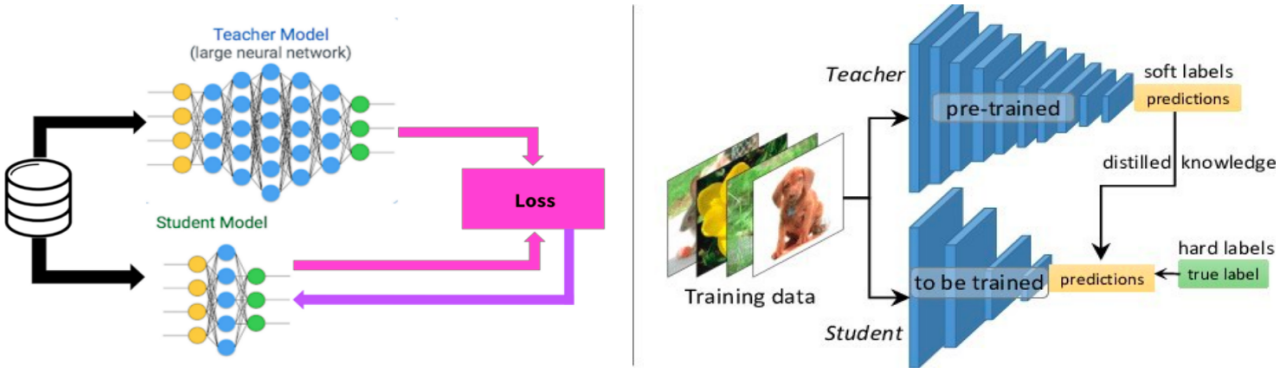


Image Source : 7

5.1 Objective function and generalization

Objective functions used for training are generally designed to reflect the true objective of the users intentions. But during training the models are optimized towards the training data whereas the real objective is to generalise well to new data.

If the Complex Teacher model that is trained normally generalizes well, then the Student model that is trained from the Teacher model generalizes much better then the student model trained directly.

5.2 Knowledge Distillation method and Loss function

Teacher model is generally trained using cross entropy loss on the ground truth labels of the problem . The class probabilities (Soft targets) of the Teacher model is then used to train the Student model.

From Regularly trained teacher network, convert the pre-softmax logits z_i computed for each class into a probability q_i with the below given equation.

$$q_i = \frac{exp(z_i/T)}{\Sigma exp(z_j/T)}$$

T is called the temperature (T >= 1) , for the standard softmax function it is set to 1. The above softmax operator converts the logit's values z_i to pseudo-probabilities . Higher the value of temperature T, it will give probability distribution which is much softer over the classes. Softened distribution might reveal the hidden information for incorrect classes.

The knowledge from the teacher is transferred to the student by a minimizing the knowledge distillation loss (L_{KD})

$$L_{KD} = \alpha T^2 * \text{Crossenthropy}(Q_s^\tau, Q_T^\tau) + (1 - \alpha) * \text{Crossenthropy}(Q_s, y_{true})$$

Q_s^τ, Q_T^τ are the softened targets of the teacher and student using the same temperature T and α is another hyperparameter that tunes the weighted average between the 2 components of the loss.

First component of the loss function forces the optimization towards the similar softened distribution for the student.

Second component forces the optimization towards approximating the ground truth label as usual.

6. Other KD strategies

Several Knowledge Distillation strategies have been proposed over time, we will be exploring the 2 strategies namely, Reverse KD and Defective KD. Other works are mentioned in the 10.Conclusion and Future Scope section.

Beyond the well proven facts about Knowledge distillation that the Teacher can improve the student, the student can also improve and enhance the Teacher by reversing the KD procedure.

A poorly trained Teacher , that is a model with lower accuracy than the students can still improve the Student significantly, this strategy is termed as Defective KD.

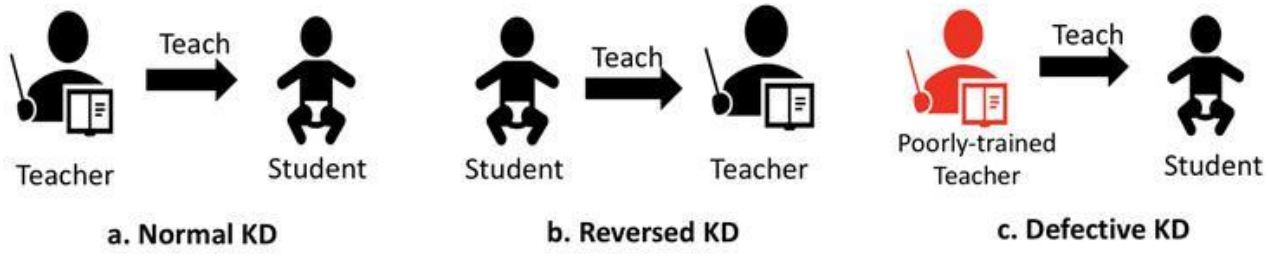


Image Source : 8

7. Tools and Technologies

7.1 Developing Environment

- Processor : Intel i7 - 10510U CPU @ 1.80GHz
- Memory: 8 GB primary memory
- Operating System: Ubuntu 20 .04 LTS
- Hard Drive : 512GB Solid State Drive

7.2 Technologies

Programming Environment

- Kaggle Notebooks

Python Packages

- Pytorch
- Torchvision
- Albumentation
- Pretrainedmodels
- Timm
- Numpy
- Pandas
- Scikit-learn
- Matplotlib
- Seaborn
- Neptune

7.3 Neptune.ai

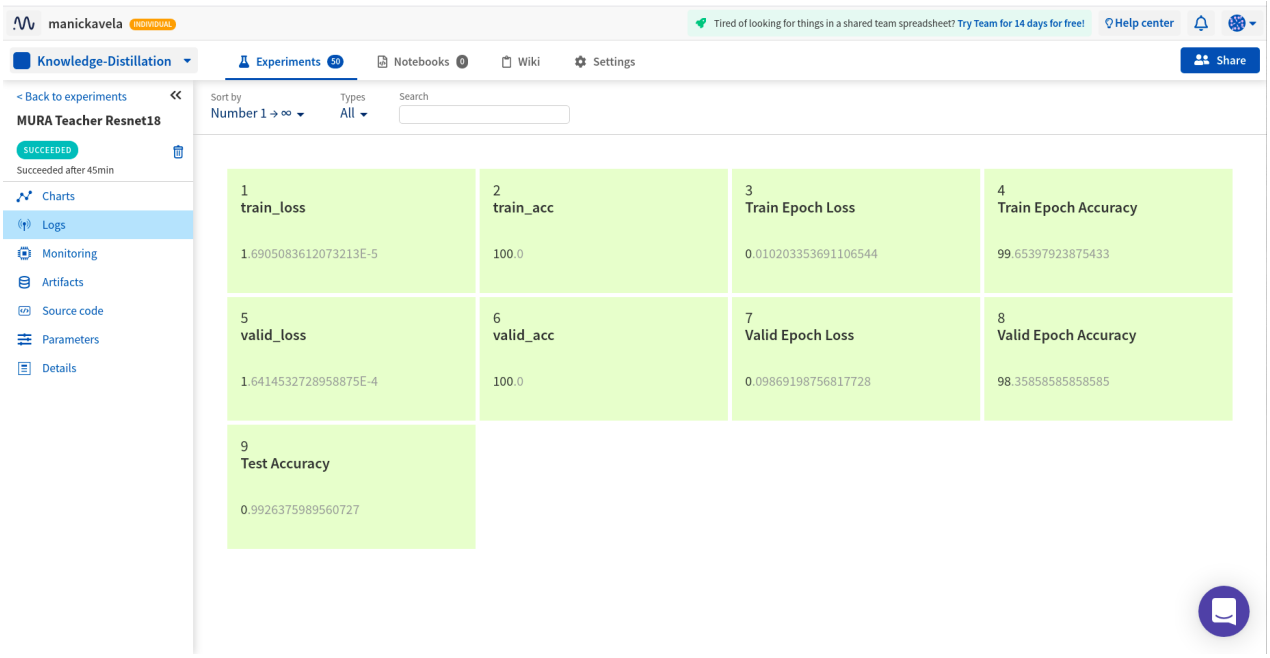
Neptune brings organisation and collaboration to data science projects. Everything is secured and backed-up in an organised knowledge repository. Neptune makes it much easier to log, store, display, organise, and query all meta data generated during the training.

Any of the below can be logged with neptune.

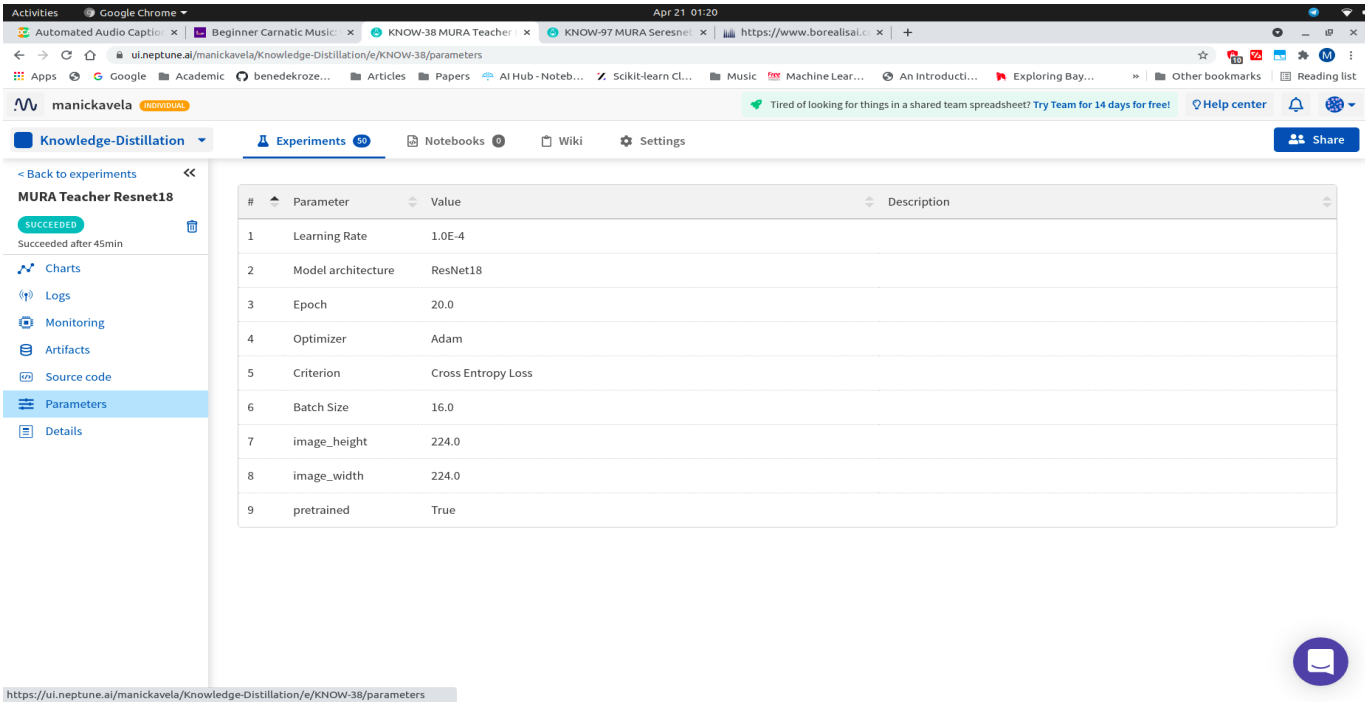
1. Metrics
2. Hyperparameters
3. Learning curves
4. Training code and configuration files
5. Predictions (images , tables and etc)
6. Diagnostic charts
7. Console logs
8. Hardware logs and more

Automatic interactive charts will be generated for the logs.With all experiments logged with neptune.

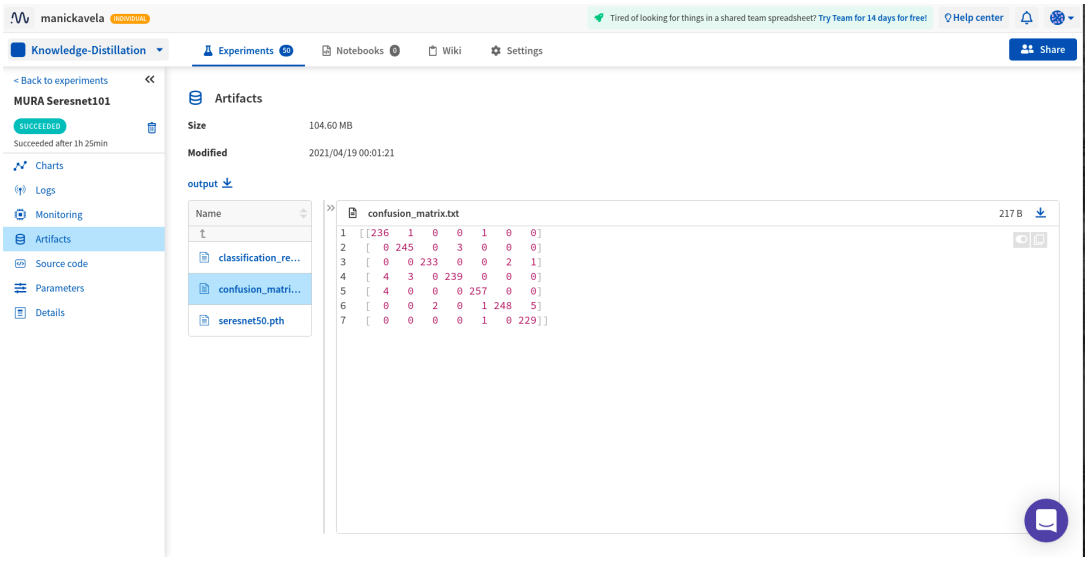
Experiments can be filtered by metric and parameters. Easily compare several different logs. Since this is an experimental work, Neptune has been of great help in managing and visualising the logs and metrics. Model weights too are stored in the experiments in neptune along with their confusion matrix and classification report with the classification metrics details.



With all the logged metrics which where measured during training.



All the model parameters in one place .



Artifacts of Model weights , confusion matrix and classification report with metrics.

8. Implementation

The implementation of the Neural networks is completely done with the Pytorch Framework.

Student and Teacher Networks are implemented in reference to their own right full authors implementation. These implementations are stored in the **model.py** utility file .

The data loaders and other support classes are stored in the utility file by name **mura_dataloader.py**

The experiments are done by importing these utility files into the Kaggle notebooks with appropriate classes. At few instances **timm** library has been used which is a python package for SOTA architectures with both trained and untrained networks.

All the logs and relevant metrics along with the models have been stored in the neptune.ai website for easy visualisation and to compare the results between several models.

9. Experiments and Discussions

9.1 MURA Dataset

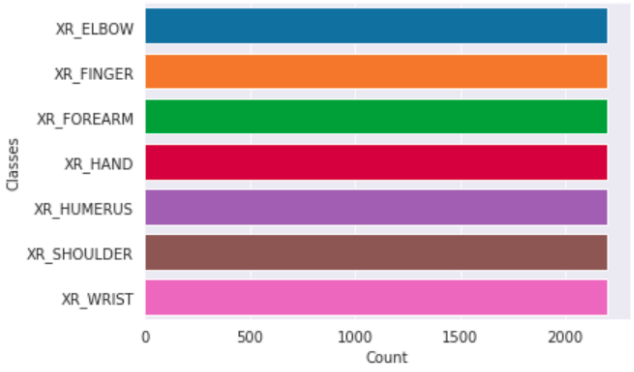
MURA dataset is a collection of images of Musculoskeletal radiographs images which were collected and processed with the purpose of improving the medical image technology to the level of an expert with professional experience. It was collected by the stanford ML group.

MURA contains multi-view radiographic images of seven body tissues Elbow, Finger, Hand, Humerus, Forearm, Shoulder and Wrist.

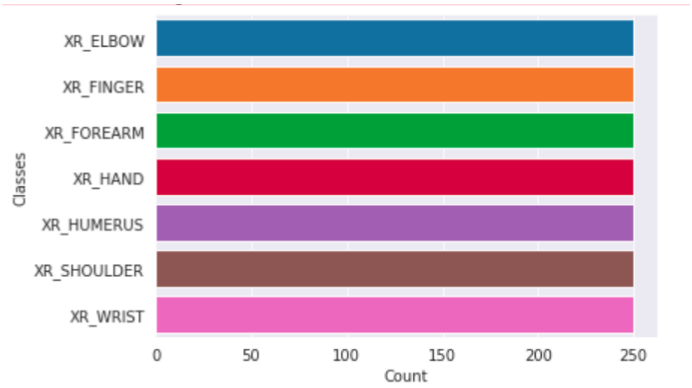
Actual MURA dataset problem was to classify an image to Normal and Abnormal but we haven't taken up the exact problem and have rather moved to another sub-problem with the dataset of classifying the images into multi class classification problem.

The Dataset is already divided into Train and Valid set.Train Set has 2201 images for each class, summing up to 15,407 images totally , whereas the Valid set has 250 images in each class summing up to 1750 images.

The classes are already well balanced , the dataset is chosen as such to focus more on Knowledge Distillation and its strategies.



Train Set



Valid Set

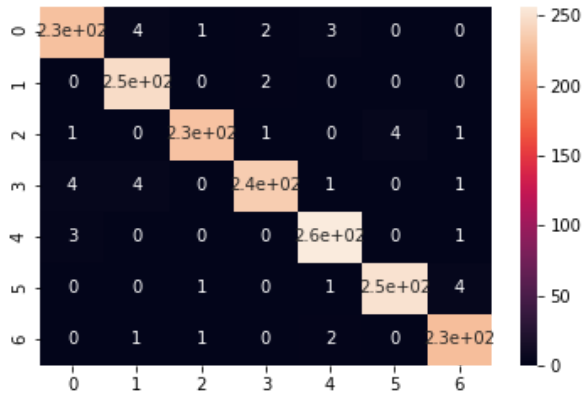
9.2 Detailed classification model results

Confusion Matrix

It is a table depicting the performance of an classification algorithm on an unseen data generally , letting us understand more about the models generality.It is also known as error matrix.

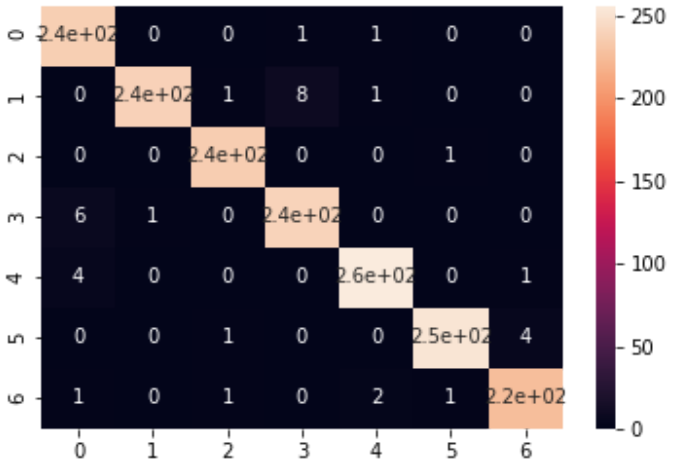
9.2.1 Teacher Networks

VGG11



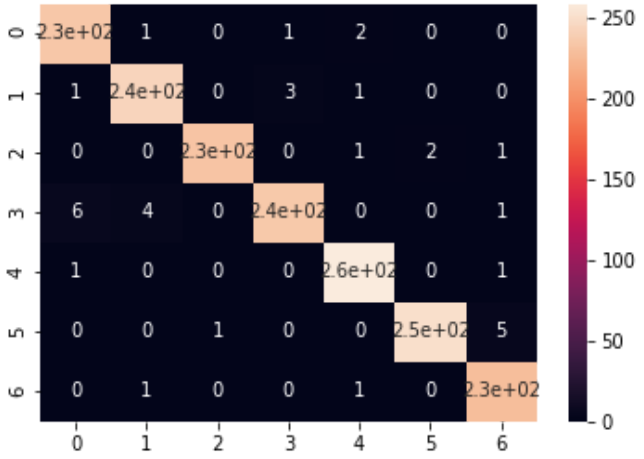
	precision	recall	f1-score	support
0	0.97	0.96	0.96	238
1	0.96	0.99	0.98	248
2	0.99	0.97	0.98	236
3	0.98	0.96	0.97	246
4	0.97	0.98	0.98	261
5	0.98	0.98	0.98	256
6	0.97	0.98	0.98	230
accuracy			0.97	1715
macro avg	0.97	0.97	0.97	1715
weighted avg	0.98	0.97	0.97	1715

Resnet18



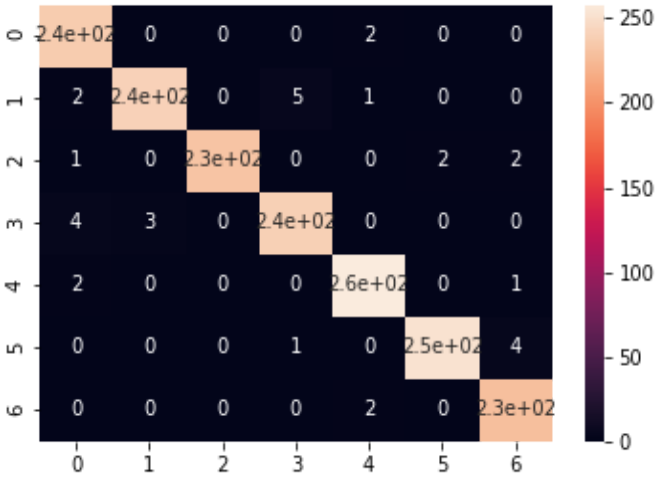
	precision	recall	f1-score	support
0	0.96	0.99	0.97	238
1	1.00	0.96	0.98	248
2	0.99	1.00	0.99	236
3	0.96	0.97	0.97	246
4	0.98	0.98	0.98	261
5	0.99	0.98	0.99	256
6	0.98	0.98	0.98	230
accuracy			0.98	1715
macro avg	0.98	0.98	0.98	1715
weighted avg	0.98	0.98	0.98	1715

Resnet34



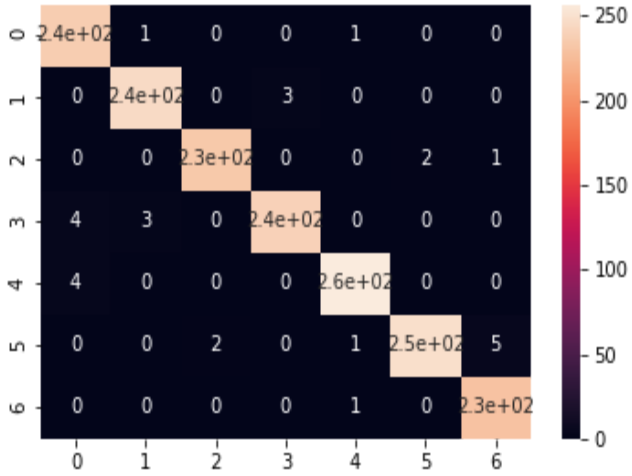
	precision	recall	f1-score	support
0	0.97	0.98	0.97	238
1	0.98	0.98	0.98	248
2	1.00	0.98	0.99	236
3	0.98	0.96	0.97	246
4	0.98	0.99	0.99	261
5	0.99	0.98	0.98	256
6	0.97	0.99	0.98	230
accuracy			0.98	1715
macro avg	0.98	0.98	0.98	1715
weighted avg	0.98	0.98	0.98	1715

Resnet50



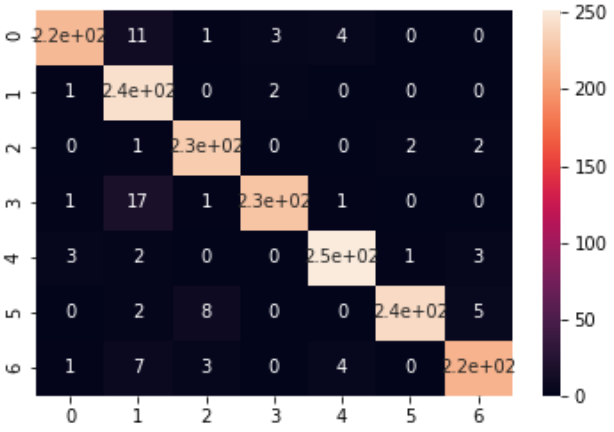
	precision	recall	f1-score	support
0	0.96	0.99	0.98	238
1	0.99	0.97	0.98	248
2	1.00	0.98	0.99	236
3	0.98	0.97	0.97	246
4	0.98	0.99	0.98	261
5	0.99	0.98	0.99	256
6	0.97	0.99	0.98	230
accuracy			0.98	1715
macro avg	0.98	0.98	0.98	1715
weighted avg	0.98	0.98	0.98	1715

Seresnet50



	precision	recall	f1-score	support
0	0.97	0.99	0.98	238
1	0.98	0.99	0.99	248
2	0.99	0.99	0.99	236
3	0.99	0.97	0.98	246
4	0.99	0.98	0.99	261
5	0.99	0.97	0.98	256
6	0.97	1.00	0.98	230
accuracy			0.98	1715
macro avg	0.98	0.98	0.98	1715
weighted avg	0.98	0.98	0.98	1715

Seresnet101



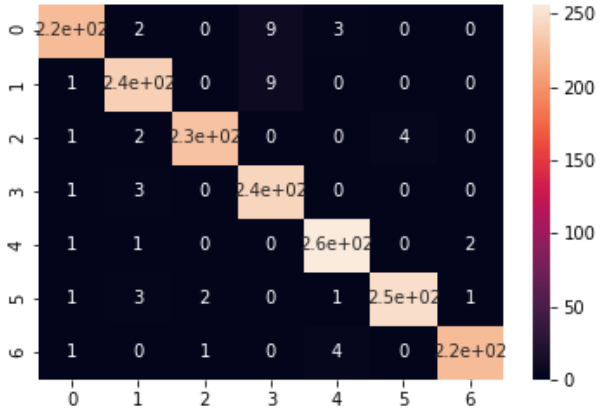
	precision	recall	f1-score	support
0	0.97	0.92	0.95	238
1	0.86	0.99	0.92	248
2	0.95	0.98	0.96	236
3	0.98	0.92	0.95	246
4	0.97	0.97	0.97	261
5	0.99	0.94	0.96	256
6	0.96	0.93	0.95	230
accuracy			0.95	1715
macro avg	0.95	0.95	0.95	1715
weighted avg	0.95	0.95	0.95	1715

9.2.2 Normal Knowledge Distillation

Teachers - *Seresnet50 , Seresnet101*

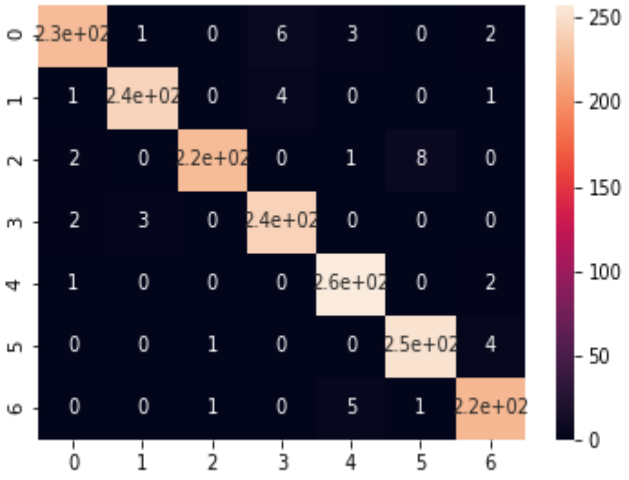
Student - *Resnet18 , Resnet34 , Resnet50*

Seresnet50 -> VGG11



	precision	recall	f1-score	support
0	0.97	0.94	0.96	238
1	0.96	0.96	0.96	248
2	0.99	0.97	0.98	236
3	0.93	0.98	0.96	246
4	0.97	0.98	0.98	261
5	0.98	0.97	0.98	256
6	0.99	0.97	0.98	230
accuracy			0.97	1715
macro avg	0.97	0.97	0.97	1715
weighted avg	0.97	0.97	0.97	1715

Seresnet101 -> VGG11

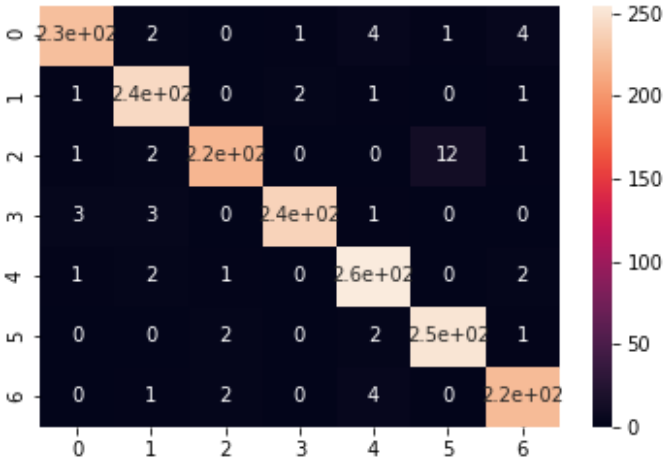


	precision	recall	f1-score	support
0	0.97	0.95	0.96	238
1	0.98	0.98	0.98	248
2	0.99	0.95	0.97	236
3	0.96	0.98	0.97	246
4	0.97	0.99	0.98	261
5	0.97	0.98	0.97	256
6	0.96	0.97	0.97	230
accuracy			0.97	1715
macro avg	0.97	0.97	0.97	1715
weighted avg	0.97	0.97	0.97	1715

Seresnet50 -> Resnet18

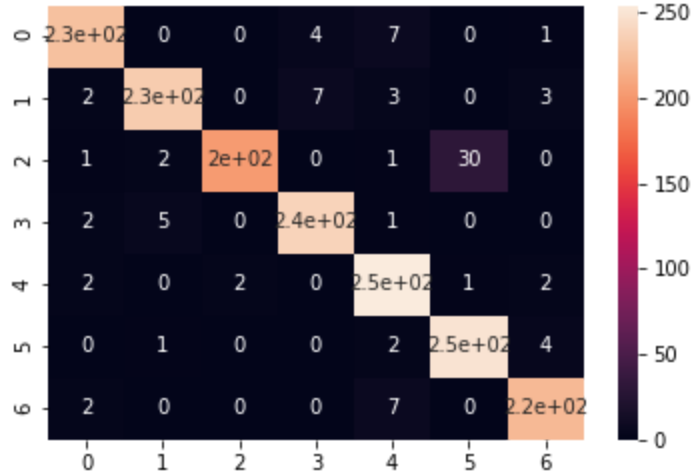
	precision	recall	f1-score	support
0	0.98	0.95	0.96	238
1	0.98	0.98	0.98	248
2	0.98	0.97	0.98	236
3	0.97	0.98	0.97	246
4	0.93	0.99	0.96	261
5	0.97	0.97	0.97	256
6	0.98	0.95	0.97	230
accuracy			0.97	1715
macro avg	0.97	0.97	0.97	1715
weighted avg	0.97	0.97	0.97	1715

Seresnet101 -> Resnet18



	precision	recall	f1-score	support
0	0.97	0.95	0.96	238
1	0.96	0.98	0.97	248
2	0.98	0.93	0.95	236
3	0.99	0.97	0.98	246
4	0.96	0.98	0.97	261
5	0.95	0.98	0.97	256
6	0.96	0.97	0.97	230
accuracy			0.97	1715
macro avg	0.97	0.97	0.97	1715
weighted avg	0.97	0.97	0.97	1715

Seresnet50 -> Resnet34

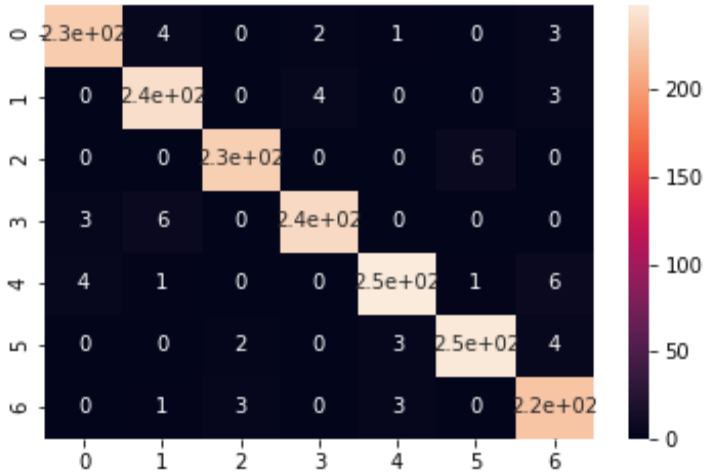


	precision	recall	f1-score	support
0	0.96	0.95	0.96	238
1	0.97	0.94	0.95	248
2	0.99	0.86	0.92	236
3	0.96	0.97	0.96	246
4	0.92	0.97	0.95	261
5	0.89	0.97	0.93	256
6	0.96	0.96	0.96	230
accuracy			0.95	1715
macro avg	0.95	0.95	0.95	1715
weighted avg	0.95	0.95	0.95	1715

Seresnet101 -> Resnet34

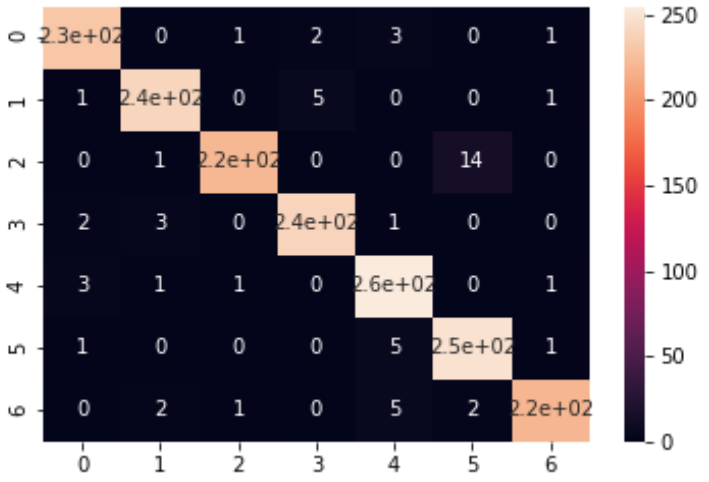
	precision	recall	f1-score	support
0	0.93	1.00	0.96	238
1	0.98	0.97	0.98	248
2	0.94	0.96	0.95	236
3	0.98	0.98	0.98	246
4	0.98	0.95	0.97	261
5	0.95	0.95	0.95	256
6	0.98	0.94	0.96	230
accuracy			0.96	1715
macro avg	0.96	0.96	0.96	1715
weighted avg	0.96	0.96	0.96	1715

Seresnet50 -> Resnet50



	precision	recall	f1-score	support
0	0.97	0.96	0.96	238
1	0.95	0.97	0.96	248
2	0.98	0.97	0.98	236
3	0.98	0.96	0.97	246
4	0.97	0.95	0.96	261
5	0.97	0.96	0.97	256
6	0.93	0.97	0.95	230
accuracy			0.97	1715
macro avg	0.96	0.97	0.96	1715
weighted avg	0.97	0.97	0.97	1715

Seresnet101 -> Resnet50



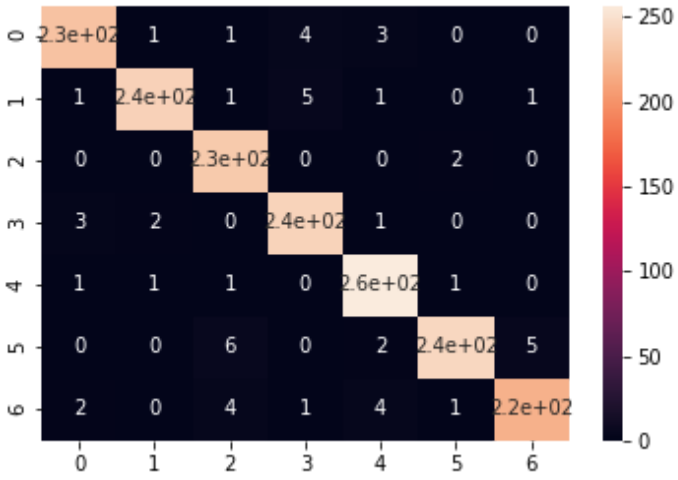
	precision	recall	f1-score	support
0	0.97	0.97	0.97	238
1	0.97	0.97	0.97	248
2	0.99	0.94	0.96	236
3	0.97	0.98	0.97	246
4	0.95	0.98	0.96	261
5	0.94	0.97	0.96	256
6	0.98	0.96	0.97	230
accuracy			0.97	1715
macro avg	0.97	0.97	0.97	1715
weighted avg	0.97	0.97	0.97	1715

9.2.2 Reverse Knowledge Distillation

Teachers - Resnet18 , Resnet34 , Resnet50

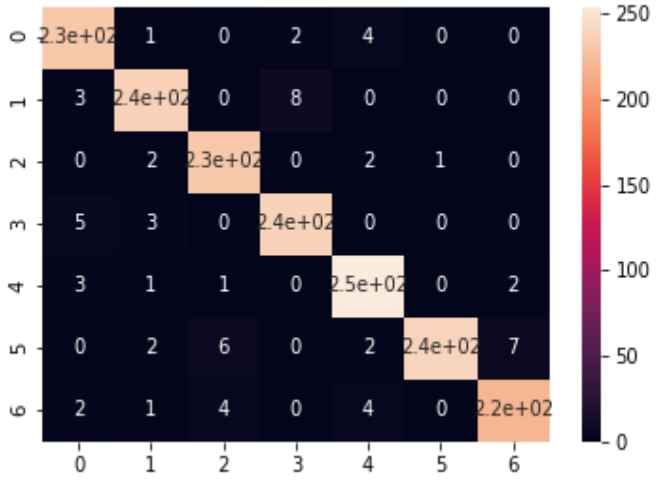
Student - Seresnet50 , Seresnet10

VGG11 -> Seresnet50



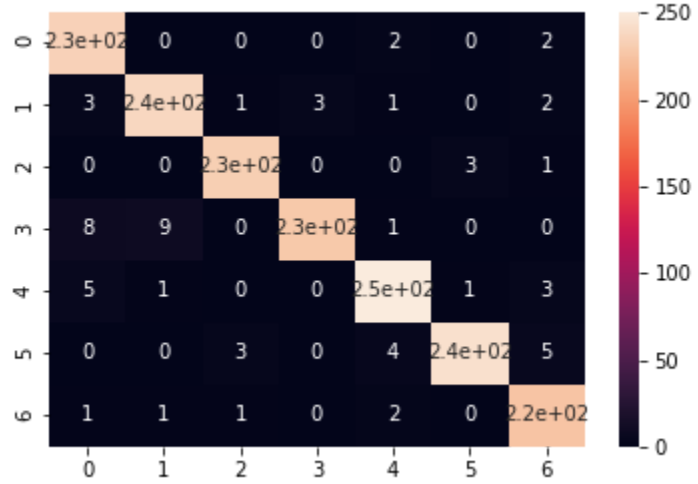
	precision	recall	f1-score	support
0	0.97	0.96	0.97	238
1	0.98	0.96	0.97	248
2	0.95	0.99	0.97	236
3	0.96	0.98	0.97	246
4	0.96	0.98	0.97	261
5	0.98	0.95	0.97	256
6	0.97	0.95	0.96	230
accuracy			0.97	1715
macro avg	0.97	0.97	0.97	1715
weighted avg	0.97	0.97	0.97	1715

VGG11 -> Seresnet101



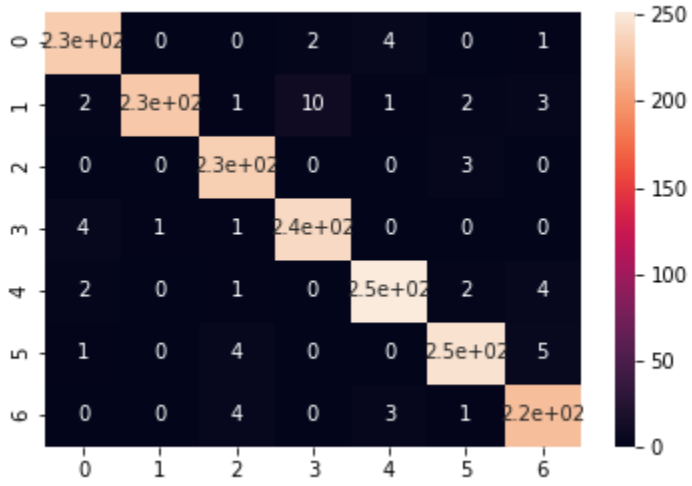
	precision	recall	f1-score	support
0	0.95	0.97	0.96	238
1	0.96	0.96	0.96	248
2	0.95	0.98	0.97	236
3	0.96	0.97	0.96	246
4	0.95	0.97	0.96	261
5	1.00	0.93	0.96	256
6	0.96	0.95	0.96	230
accuracy			0.96	1715
macro avg	0.96	0.96	0.96	1715
weighted avg	0.96	0.96	0.96	1715

Resnet18 -> Seresnet50



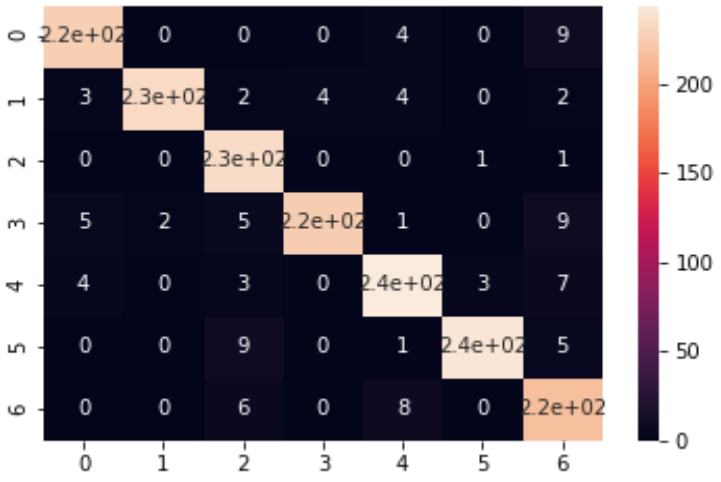
	precision	recall	f1-score	support
0	0.93	0.98	0.96	238
1	0.96	0.96	0.96	248
2	0.98	0.98	0.98	236
3	0.99	0.93	0.96	246
4	0.96	0.96	0.96	261
5	0.98	0.95	0.97	256
6	0.95	0.98	0.96	230
accuracy			0.96	1715
macro avg	0.96	0.96	0.96	1715
weighted avg	0.96	0.96	0.96	1715

Resnet18 -> Seresnet101



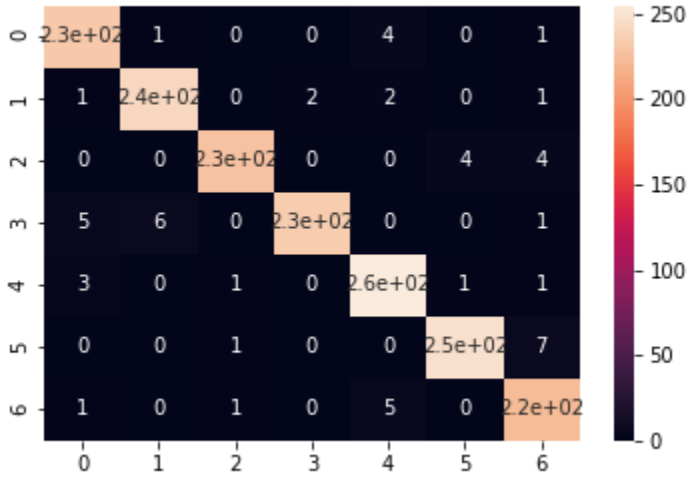
	precision	recall	f1-score	support
0	0.96	0.97	0.97	238
1	1.00	0.92	0.96	248
2	0.95	0.99	0.97	236
3	0.95	0.98	0.96	246
4	0.97	0.97	0.97	261
5	0.97	0.96	0.96	256
6	0.94	0.97	0.95	230
accuracy			0.96	1715
macro avg	0.96	0.96	0.96	1715
weighted avg	0.96	0.96	0.96	1715

Resnet34 -> Seresnet50



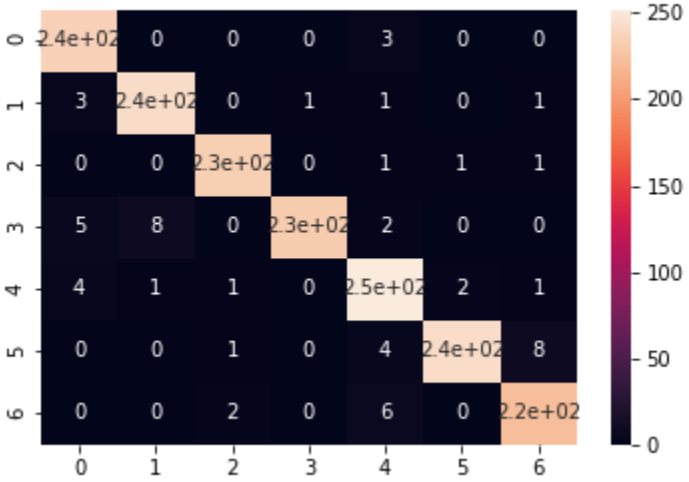
	precision	recall	f1-score	support
0	0.95	0.95	0.95	238
1	0.99	0.94	0.96	248
2	0.90	0.99	0.95	236
3	0.98	0.91	0.95	246
4	0.93	0.93	0.93	261
5	0.98	0.94	0.96	256
6	0.87	0.94	0.90	230
accuracy			0.94	1715
macro avg	0.94	0.94	0.94	1715
weighted avg	0.95	0.94	0.94	1715

Resnet34 -> Seresnet101



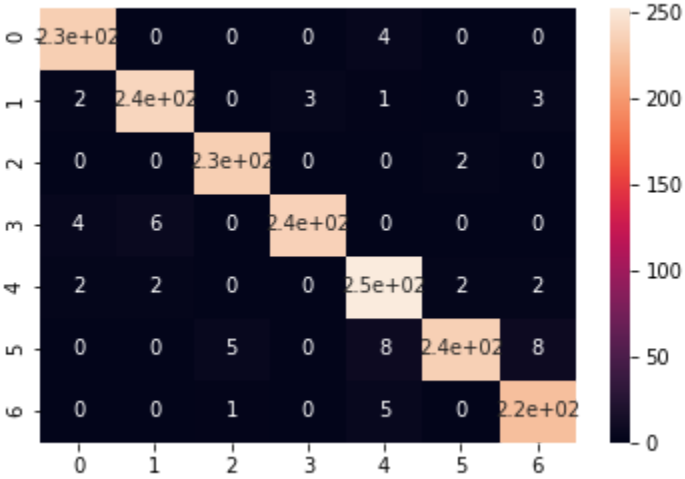
	precision	recall	f1-score	support
0	0.96	0.97	0.97	238
1	0.97	0.98	0.97	248
2	0.99	0.97	0.98	236
3	0.99	0.95	0.97	246
4	0.96	0.98	0.97	261
5	0.98	0.97	0.97	256
6	0.94	0.97	0.95	230
accuracy			0.97	1715
macro avg	0.97	0.97	0.97	1715
weighted avg	0.97	0.97	0.97	1715

Resnet50 -> Seresnet50



	precision	recall	f1-score	support
0	0.95	0.99	0.97	238
1	0.96	0.98	0.97	248
2	0.98	0.99	0.99	236
3	1.00	0.94	0.97	246
4	0.94	0.97	0.95	261
5	0.99	0.95	0.97	256
6	0.95	0.97	0.96	230
accuracy			0.97	1715
macro avg	0.97	0.97	0.97	1715
weighted avg	0.97	0.97	0.97	1715

Resnet50 -> Seresnet101



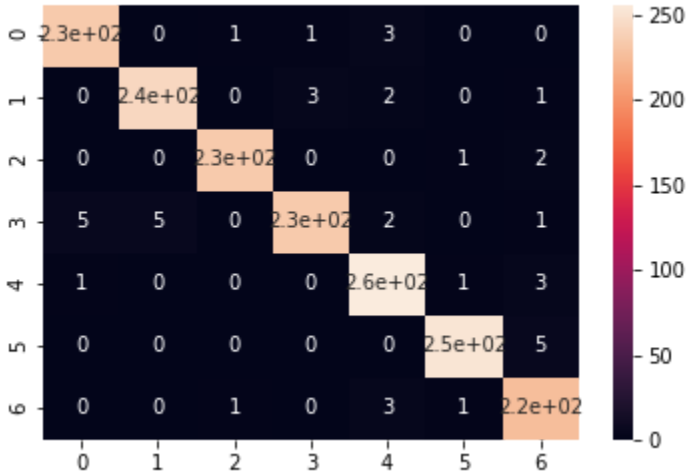
	precision	recall	f1-score	support
0	0.97	0.98	0.97	238
1	0.97	0.96	0.97	248
2	0.97	0.99	0.98	236
3	0.99	0.96	0.97	246
4	0.93	0.97	0.95	261
5	0.98	0.92	0.95	256
6	0.95	0.97	0.96	230
accuracy			0.97	1715
macro avg	0.97	0.97	0.97	1715
weighted avg	0.97	0.97	0.96	1715

9.2.3 Defective Knowledge Distillation

Teachers - *Seresnet50 , Seresnet101*

Student - *Resnet18 , Resnet34 , Resnet50*

Seresnet50 -> Resnet18

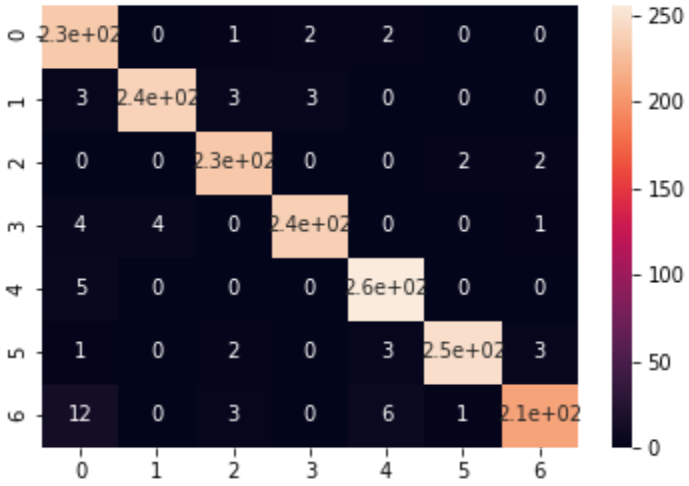


	precision	recall	f1-score	support
0	0.97	0.98	0.98	238
1	0.98	0.98	0.98	248
2	0.99	0.99	0.99	236
3	0.98	0.95	0.96	246
4	0.96	0.98	0.97	261
5	0.99	0.98	0.98	256
6	0.95	0.98	0.96	230
accuracy			0.98	1715
macro avg	0.98	0.98	0.98	1715
weighted avg	0.98	0.98	0.98	1715

Seresnet101 -> Resnet18

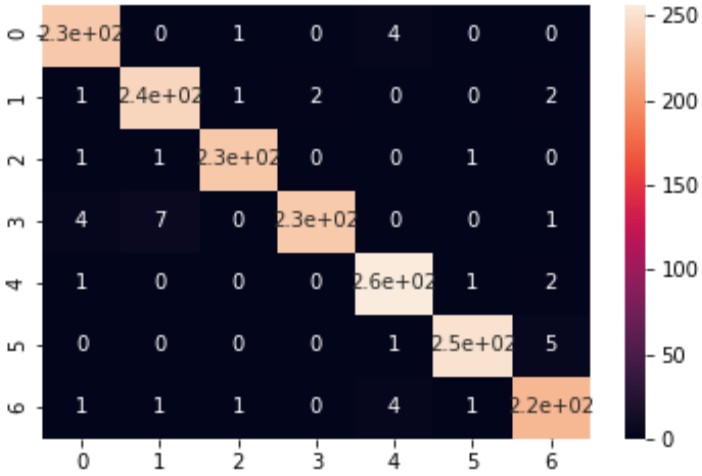
	precision	recall	f1-score	support
0	0.96	0.98	0.97	238
1	0.98	0.98	0.98	248
2	0.99	0.98	0.98	236
3	0.99	0.96	0.97	246
4	0.96	0.98	0.97	261
5	0.97	0.98	0.97	256
6	0.96	0.93	0.94	230
accuracy			0.97	1715
macro avg	0.97	0.97	0.97	1715
weighted avg	0.97	0.97	0.97	1715

Seresnet50 -> Resnet34



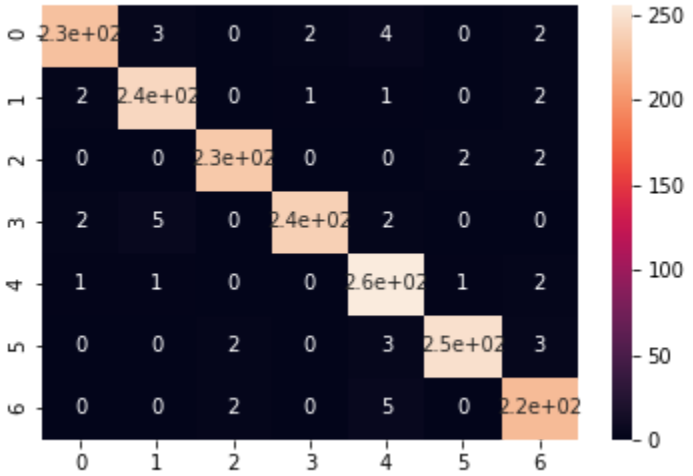
	precision	recall	f1-score	support
0	0.90	0.98	0.94	238
1	0.98	0.96	0.97	248
2	0.96	0.98	0.97	236
3	0.98	0.96	0.97	246
4	0.96	0.98	0.97	261
5	0.99	0.96	0.98	256
6	0.97	0.90	0.94	230
accuracy			0.96	1715
macro avg	0.96	0.96	0.96	1715
weighted avg	0.96	0.96	0.96	1715

Seresnet101 -> Resnet34



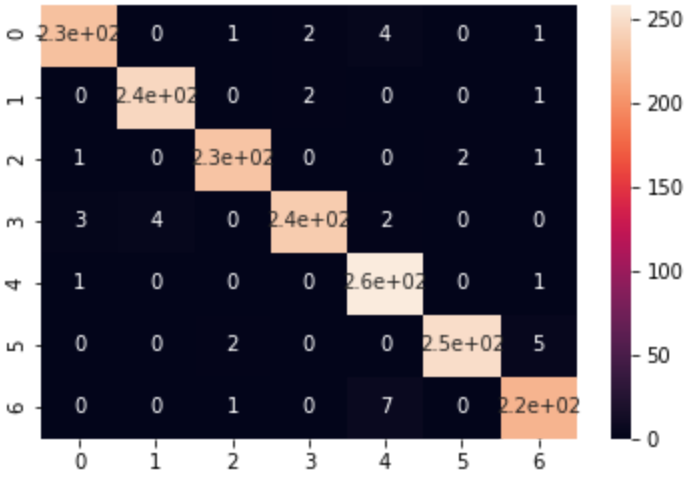
	precision	recall	f1-score	support
0	0.97	0.98	0.97	238
1	0.96	0.98	0.97	248
2	0.99	0.99	0.99	236
3	0.99	0.95	0.97	246
4	0.97	0.98	0.98	261
5	0.99	0.98	0.98	256
6	0.96	0.97	0.96	230
accuracy			0.97	1715
macro avg	0.97	0.97	0.97	1715
weighted avg	0.97	0.97	0.97	1715

Seresnet50 -> Resnet50



	precision	recall	f1-score	support
0	0.98	0.95	0.97	238
1	0.96	0.98	0.97	248
2	0.98	0.98	0.98	236
3	0.99	0.96	0.98	246
4	0.94	0.98	0.96	261
5	0.99	0.97	0.98	256
6	0.95	0.97	0.96	230
accuracy			0.97	1715
macro avg	0.97	0.97	0.97	1715
weighted avg	0.97	0.97	0.97	1715

Seresnet101 -> Resnet50



	precision	recall	f1-score	support
0	0.98	0.97	0.97	238
1	0.98	0.99	0.99	248
2	0.98	0.98	0.98	236
3	0.98	0.96	0.97	246
4	0.95	0.99	0.97	261
5	0.99	0.97	0.98	256
6	0.96	0.97	0.96	230
accuracy			0.98	1715
macro avg	0.98	0.98	0.98	1715
weighted avg	0.98	0.98	0.98	1715

9.3 Accuracy Study over the models

9.3.1 Normal Knowledge Distillation

In our experiments 2 SeResnet architecture variant is taken as Teacher networks , SeResnet50 and SeResnet101,

4 models taken for students are VGG11, Resnet18 , Resnet34 and Resnet50.

Teacher Networks

NORMAL KD		
Seresnet50	Train_Acc	0.9992
	Validation_Acc	0.988
	Test_Acc	0.9817
Seresnet101	Train_Acc	0.9882
	Validation_Acc	0.9524
	Test_Acc	0.8138

Student Networks

		SeresNet50	Seresnet101
Resnet18	Train_acc	0.9801	0.9798
	Validation_acc	0.9564	0.9539
	Test_acc	0.9704	0.9664
Resnet34	Train_acc	0.9795	0.9748
	Validation_acc	0.9457	0.9621
	Test_acc	0.9467	0.9629
Resnet50	Train_acc	0.9753	0.9749
	Validation_acc	0.9486	0.9532
	Test_acc	0.9652	0.9664
VGG-11	Train_acc	0.993	0.9938
	Validation_acc	0.964	0.9608
	Test_acc	0.9693	0.9716

It is observed that the more complex and cumbersome the teacher network is , it is generalizing well. Generalisation is also dependent on the architecture of the student model. Resnet34 and Resnet50 seem to generalize well to the validation set but Resnet18 seems to have little overfit

to the train set data as its test accuracy has come down with both the teachers Seresnet50 and Seresnet101.

9.3.2 Reverse Knowledge Distillation

Teacher Network

Reverse KD		
Resnet18	Train_Acc	99.653979
	Validation_acc	98.358586
	Test_Acc	0.992638
Resnet34	Train_Acc	99.524221
	Validation_acc	98.232323
	Test_Acc	0.875289
Resnet50	Train_Acc	99.437716
	Validation_acc	98.106061
	Test_Acc	0.88614

Reverse KD on the chosen KD models results are in the above table. Only Resnet18 seems to fit and generalize well to the dataset. Both Resnet34 and Resnet50 seems to be overfitting to the train data and perform very poorly on the unseen test data

Student Network

Student ⇣ \ Teacher ⇢		Resnet18	Resnet34	Resnet50	VGG-11
Seresnet50	Train_Acc	0.9702	0.9684	0.9704	0.9761
	Validation_acc	0.964	0.9362	0.9728	0.9671
	Test_acc	0.9635	0.9432	0.967	0.9681
Seresnet101	Train_Acc	0.9735	0.972	0.9719	0.9788
	Validation_acc	0.9646	0.9589	0.9684	0.9722
	Test_acc	0.9641	0.9693	0.9652	0.9618

The overfitting that was observed in the Teachers of Reverse KD doesn't seems to overfit the student models. The test accuracy on all the student models are well balanced and it has generalised well to the data. But the credit could also be asserted to the Student network itself as it trains its complexness with the data and Teacher model.

9.3.3 Defective Knowledge Distillation

Teacher Network

The teacher networks are trained with the condition of validation accuracy being above 50 , it is to experiment how well the students generalise with a poorly learned teacher model

Defective KD		
Seresnet50	Train_Acc	73.724704
	Validation_Acc	53.75
	Test_Acc	0.522441
Seresnet101	Train_Acc	50.699906
	Validation_Acc	62.5
	Test_Acc	0.609246

Student Network

The overall performance of all the student networks seems to have generalized well with respect to the previous models. Train accuracy and validation are much better than any other student models. Test accuracy is giving the best results too.

Especially Resnet50 along with the distilled knowledge and its own complex learning seems to perform extremely well with the unknown seen data.

Teacher ↓ \ Student →		**Seresnet50**	**Seresnet101**
Resnet18	Train_acc	0.9921	0.99
	Validation_acc	0.9678	0.9682
	Test_acc	0.9731	0.9699
Resnet34	Train_acc	0.9913	0.9916
	Validation_acc	0.9663	0.9703
	Test_acc	0.9635	0.9745
Resnet50	Train_acc	0.9917	0.9939
	Validation_acc	0.9633	0.9722
	Test_acc	0.971	0.9762
Resnet34	Train_acc	0.9985	0.9997
	Validation_acc	0.9678	0.964
	Test_acc	0.9733	0.9641

10. Conclusion and Future Works

It is a general observation that resnet34 seems to perform a little worse than its 18 and 50 variants. Generalisation of the models are visible with several student instances , out of those Students trained with Defective KD seems to stand out exceptionally.

VGG-11 being the smaller model seems to perform well and generalise well , further proving the fact that smaller model learns to generalise well.

The concept of Knowledge distillation works definitely while training the less cumbersome Student models from the Teachers models. Few models generalise very well with Teachers Knowledge and by its own learning and perform exceptionally well with the unseen data.

Major issue that is coming up with the medical image domain is that well trained models fail miserably to new similar data with a different meta-data like another radiographic machine or etc. Therefore making it impossible to use these models.

One proposed idea is it ensemble these trained models and training a Student Model which could perform well on the unseen data

There is so much future scope and exploration with Knowledge Distillation , as KD with different strategies is only one aspect of Teacher-Student architecture . There are several more aspects like , knowledge types , distillation schemes , distillation algorithms and applications.

This paper titled Knowledge Distillation : A Survey gathers up several of these aspects and experiment them

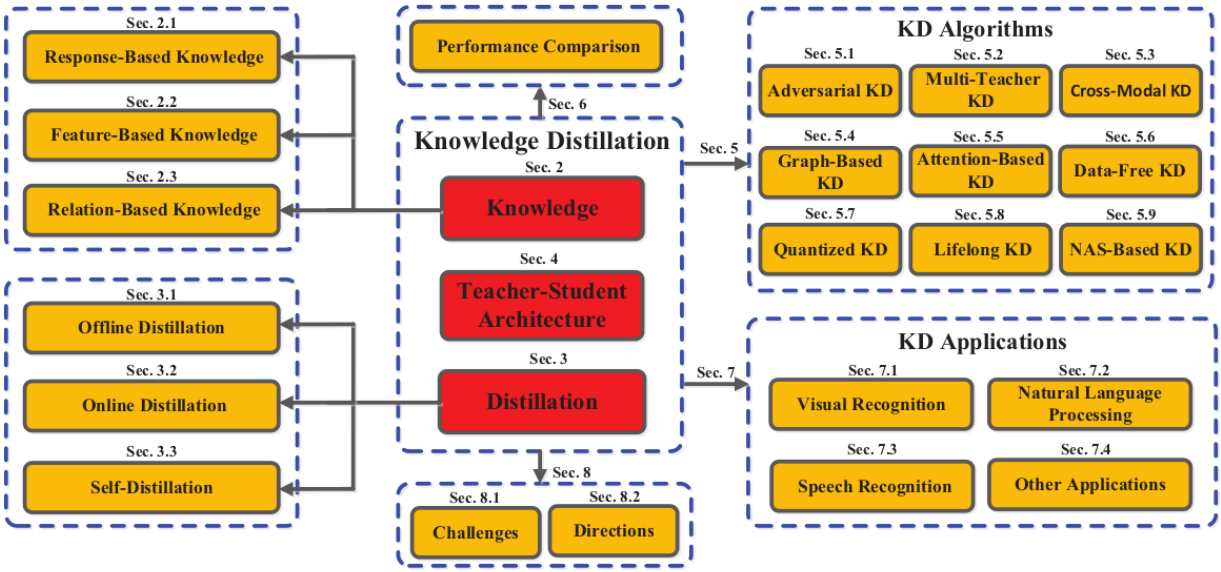


Image Source : 8

11. Bibliography

- Geoffrey Hinton,Oriol Vinyals,Jeff Dean Distilling the knowledge in Neural Networks. arXiv:1503.02531 .
- Xiaozheng Xie, Jianwei Niu , Xuefeng Liu, Zhengsu Chen, Shaojie Tang and Shui Yu(2021). A Survey on Incorporating Domain Knowledge into Deep Learning for Medical Image Analysis. arXiv:2004.12150.
- Liyang Chen ,Yongquan Chen, Juntong Xi, Xinyi Le (2020).Knowledge from the original network: restore a better pruned network with knowledge distillation. https://doi.org/10.1007/s40747-020-00248-y
- Exploring Knowledge Distillation of Deep Neural Networks for efficient Hardware Solutions.Haitong Li
- Alex Krizhevsky , Ilya Sutskever , Geoffrey E. Hinton . Imagenet classification with Deep Convolutional Neural Networks.
- Li Yuan ,Francis EH Tay ,Guilin Li , Tao Wang ,Jiashi Feng Revisiting Knowledge Distillation via Label Smoothing Regularization.
- How Knowledge distillation compresses Neural Networks, Tivadar Danka
- Knowledge Distillation Simplified, Prakhar Ganesh
- Squeeze and Excitation Network, Paul Louis Prove
- Squeeze and Excitation Network, Rachel Draelos

Image Source

1. Slide 12, Introduction to CNN (Stanford University, 2018)
2. Deep Residual Learning for Image Recognition ,arXiv:1512.03385v1
3. Squeeze and Excitation network , arXiv:1709.01507v4
4. Knowledge Distillation: A Survey, arXiv:2006.05525
5. can a neural network train other networks , Towards Data Science
6. Towards Data Science , Knowledge distillation simplified
7. Revisiting Knowledge Distillation via Label Smoothing Regularization, arXiv:1909.11723
8. Knowledge Distillation: A Survey, arXiv:2006.05525.