ITUEvents

ITU-ML5G-PS-018

Compression of Deep Learning models

ERICSSON

ITU
AI/ML in 5G
Challenge

Applying machine learning in communication networks

ai5gchallenge@itu.int

Gold Sponsor

TRA هيئة تنظيم الاتصالات
TELECOMMUNICATIONS REGULATORY AUTHORITY

Bronze sponsors

CISCO ZTE

Organized by

ITU

ITU-ML5G-PS-018T

Multi-Context based Knowledge Distillation

Team Name: GAN torrents

Thrivikram G L
Intern, Ericsson Research, Chennai

Vidya G
Intern, Ericsson Research, Chennai

Sethuraman T V
Indian Institute of Technology, Madras

Satheesh Kumar Perepu
Mentor, Senior Researcher, EGI
Ericsson Research & Development, Chennai

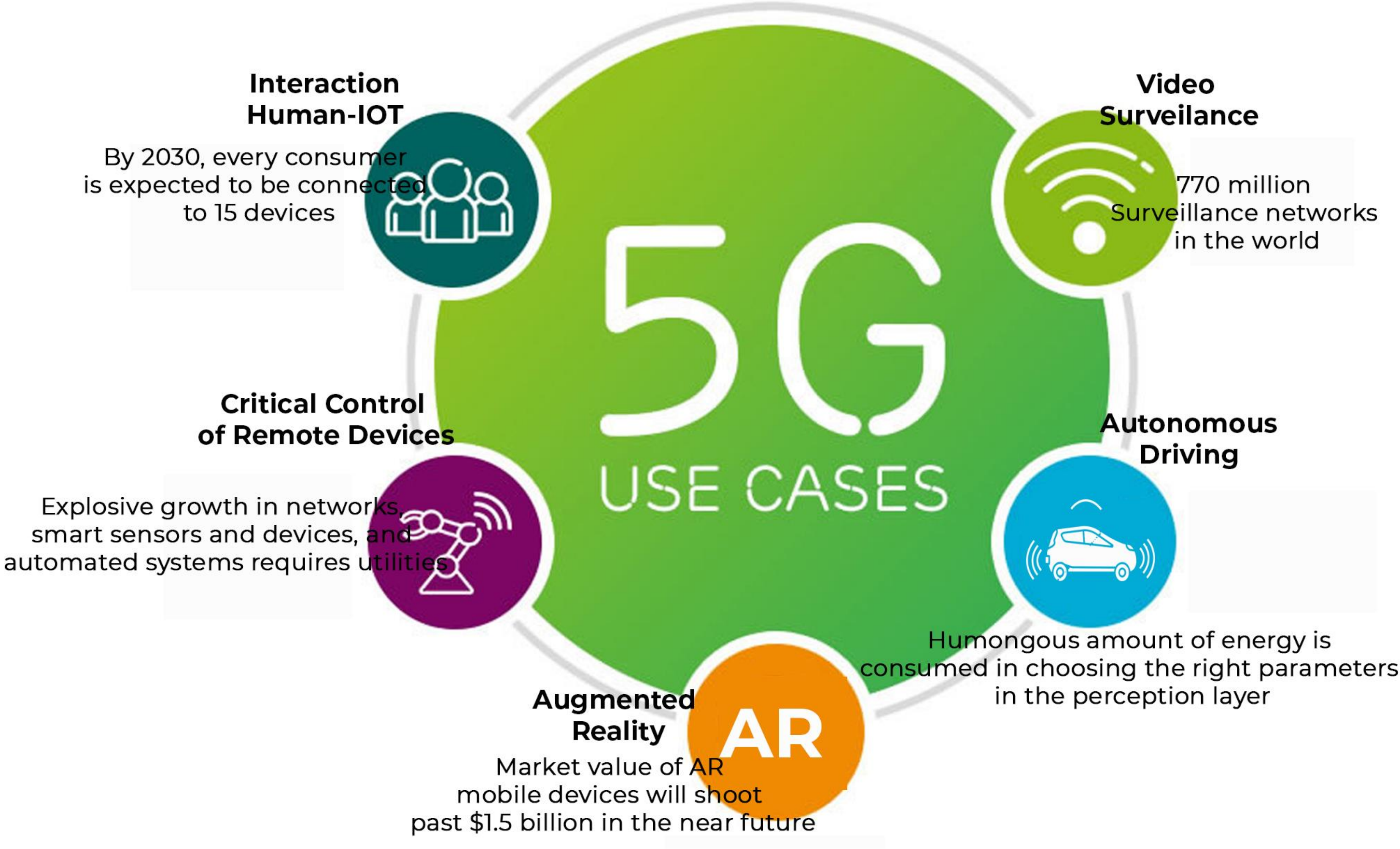
Saravanan Mohan
Mentor, Principal Researcher
Ericsson Research & Development, Chennai

BACKGROUND

Deep Learning Models Today :

During the last few years, deep learning has been the basis of many successes in artificial intelligence, including a variety of applications in computer vision, reinforcement learning, and natural language processing. But it has the following challenges:

- Millions (and even billions) of parameters
- Demands heavy computation power
- Too large to be deployed on edge devices.
- Difficult to be operated in real time.
- Large latency



Interaction Human-IOT

By 2030, every consumer is expected to be connected to 15 devices

Video Surveilance

770 million Surveillance networks in the world

Critical Control of Remote Devices

Explosive growth in networks, smart sensors and devices, and automated systems requires utilities

Autonomous Driving

Humongous amount of energy is consumed in choosing the right parameters in the perception layer

Augmented Reality

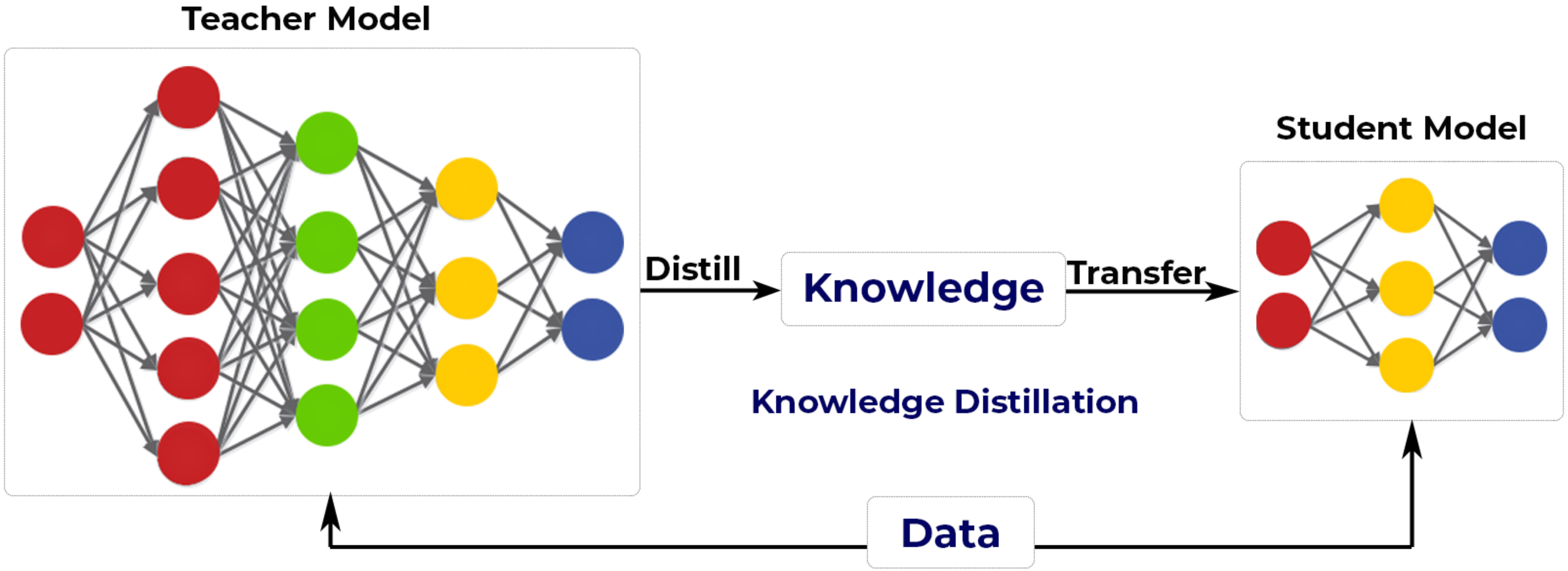
Market value of AR mobile devices will shoot past $1.5 billion in the near future

PRIOR ARTS

Model Pruning	Model Quantization	Knowledge Distillation
A model optimization technique that involves eliminating unnecessary values in the weight tensor	The process of reducing the number of bits that represent a number (the format has so far been 32-bit floating point, or FP32).	An effective technique to transfer information from one network to another network whilst training constructively.
Challenges • Very difficult to train from scratch. • Suffers from some loss of accuracy. • Difficult to generalise.	**Challenges** • Reduced memory footprint but not much increase in computing efficiency. • Difficult to generalise	**Challenges** • A teacher can't effectively distill it's knowledge to students for all the data distribution. • Not much insight on best student teacher combination.

Hardware Constraints: TensorFlow, TFLite, MXNet, and PyTorch enable developers to quantize models but they are not well suited to execute on a variety of hardware platforms. Eg. **TFLite** is optimized to run inference on ARM CPU edge devices but it **does not have efficient support for Intel CPUs and Nvidia GPUs.** **

Knowledge Distillation (KD) : (S-T) learning framework



A large (teacher) pre-trained network is used to train a smaller (student) network. However, different student architectures can perform better on different distributions data. A teacher can’t effectively distill it’s knowledge to students for all the data distribution. To alleviate this shortcoming, **we introduce multi-student knowledge distillation, which employs a multiple student model to bridge the gap between the data distribution and the student meta architecture.** To the best of our knowledge we are the first group to attempt multi-student KD framework.

RESOURCES

CIFAR-10 dataset

The CIFAR-10 dataset consists of 60000 32x32 colour images in 10 classes, with 6000 images per class.

GPU: Nvidia P100

RAM: 16GB

LIBRARIES:

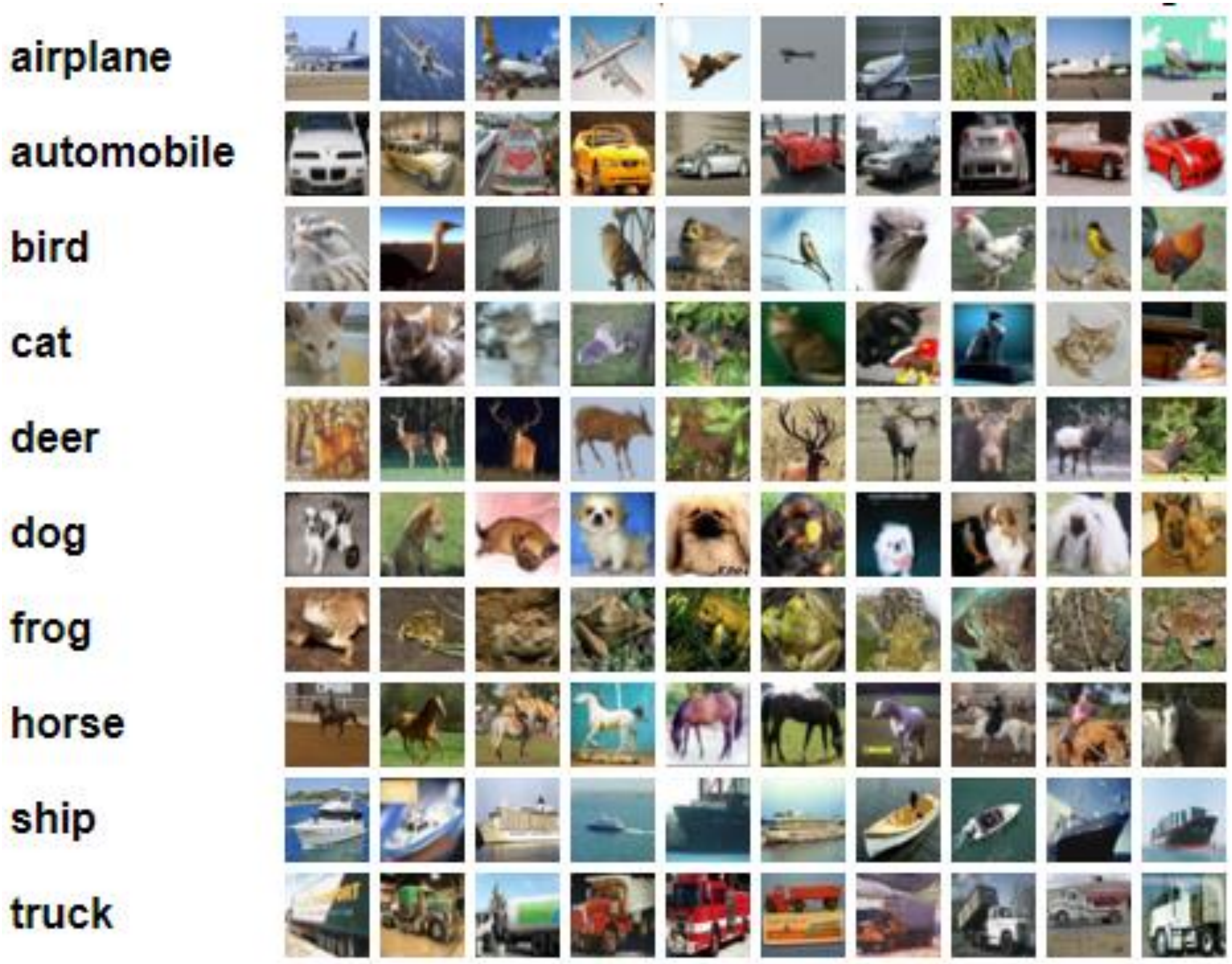
CUDA 10.0

CUDNN 8.0

Pytorch 1.6.0

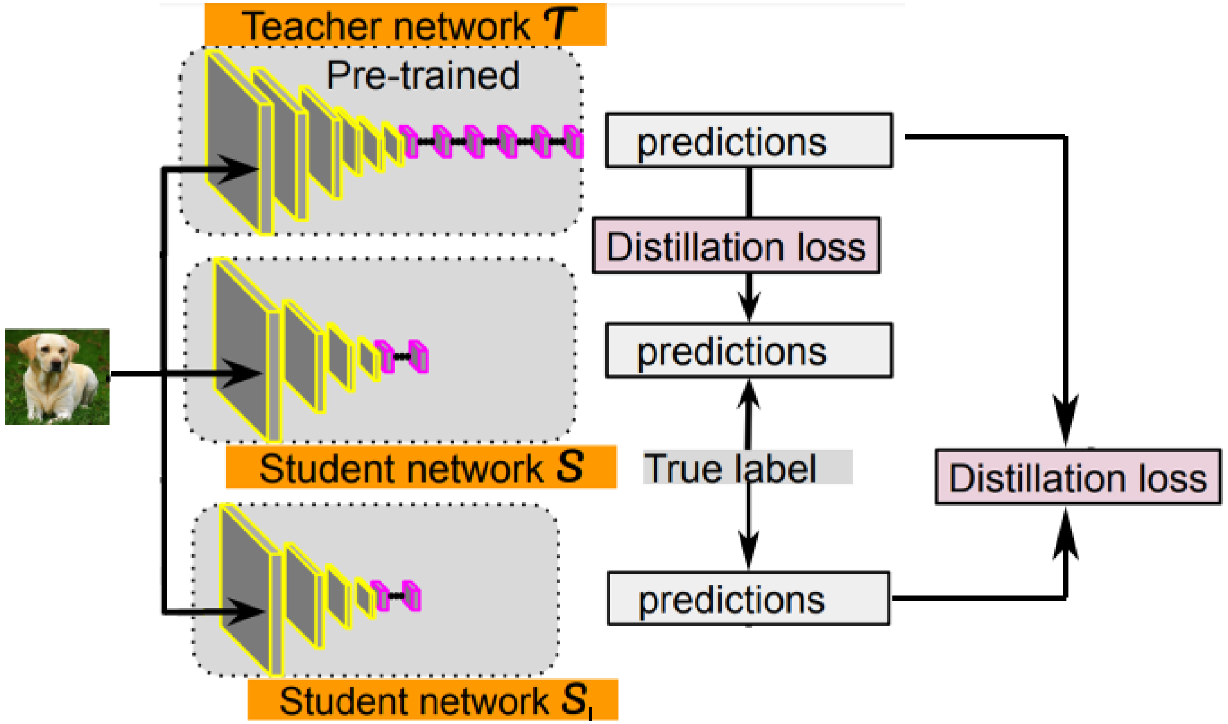
Python 3

Torchvision

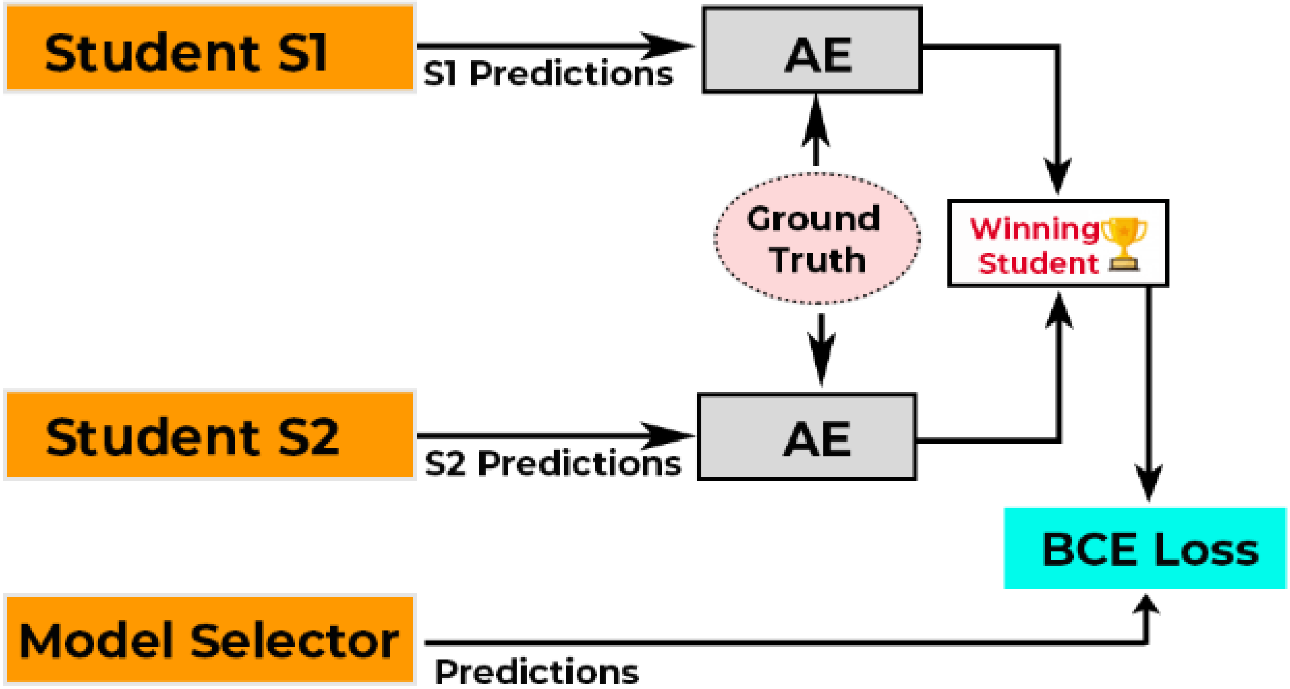


HIGH LEVEL OVERVIEW OF OUR APPROACH - TRAINING

MULTI-STUDENT KD



MODEL SELECTION



Teacher network **T** : ResNet50

Student network **S1** : DenseNet121

Student network **S2** : GoogleNet

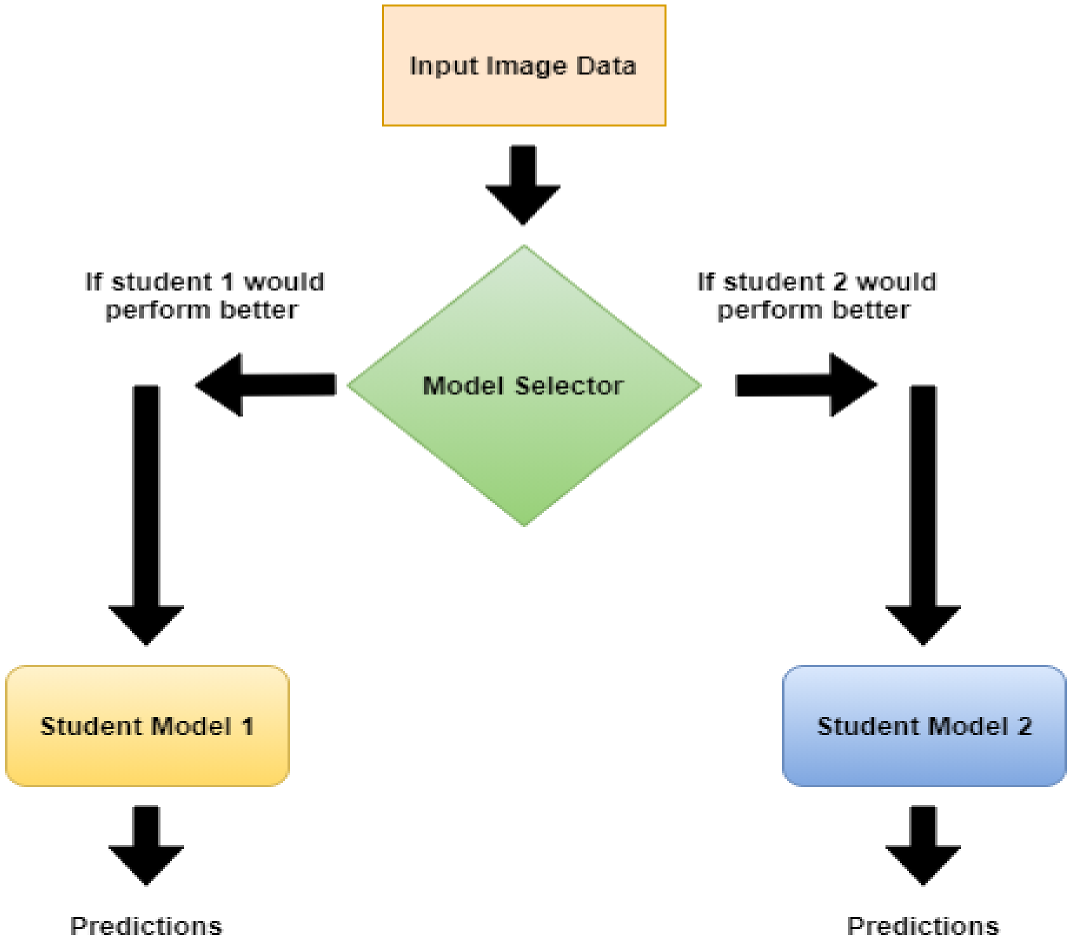
AE : Absolute Error

BCE Loss: Binary Cross Entropy Loss

Distillation Loss =KD Loss+ BCE Loss

Final Stand Alone Architecture – Inference Phase

Proposed Architecture for Inference

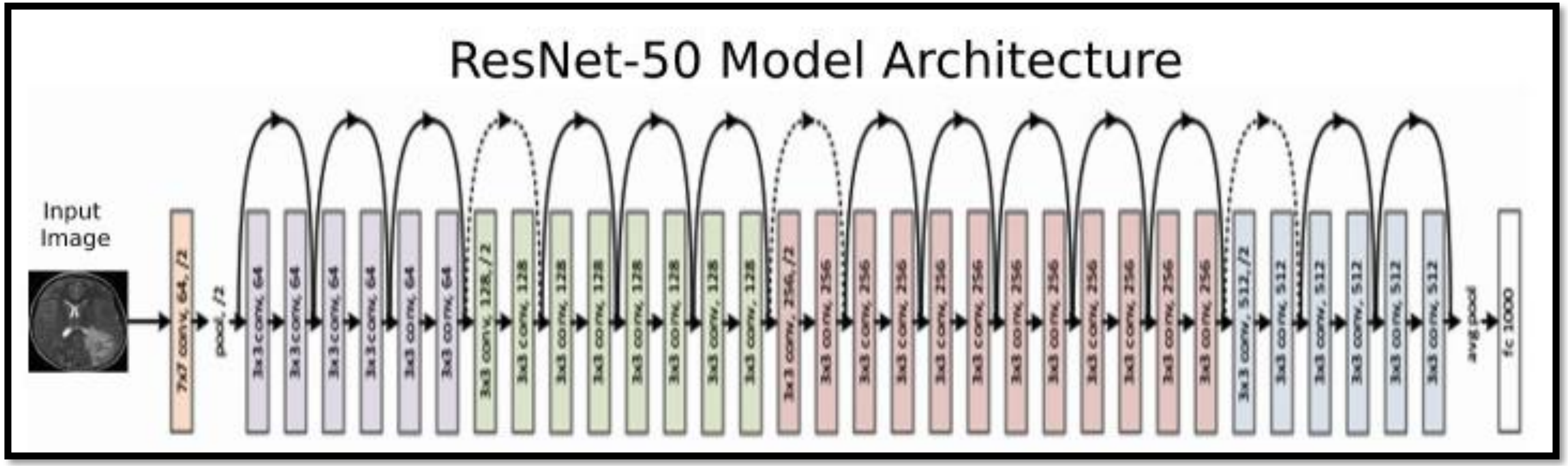


Student network **S1** : DenseNet121

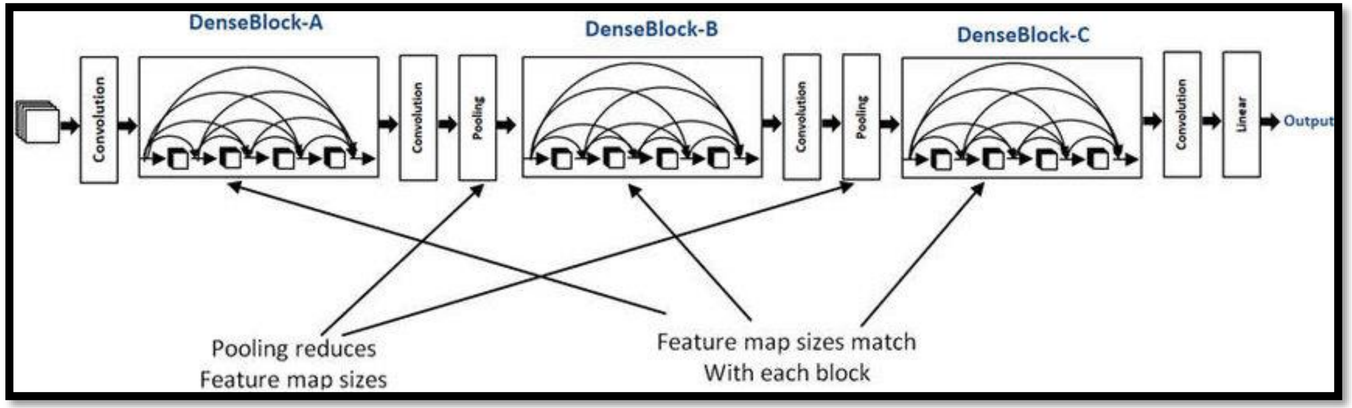
Student network **S2** : GoogleNet

Here the Model Selector extracts the features of the input data and estimates the student model that would work better on that input data and route the data to the corresponding student model to perform the desired task.

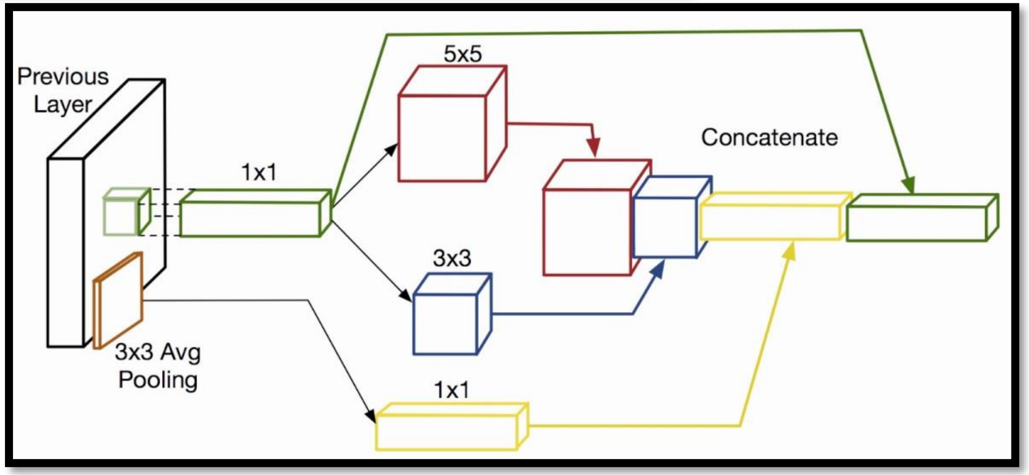
META ARCHITECTURE



Teacher Model- Resnet 50



DenseNet121- Student Model 1



Inception GoogleNet – Student model 2

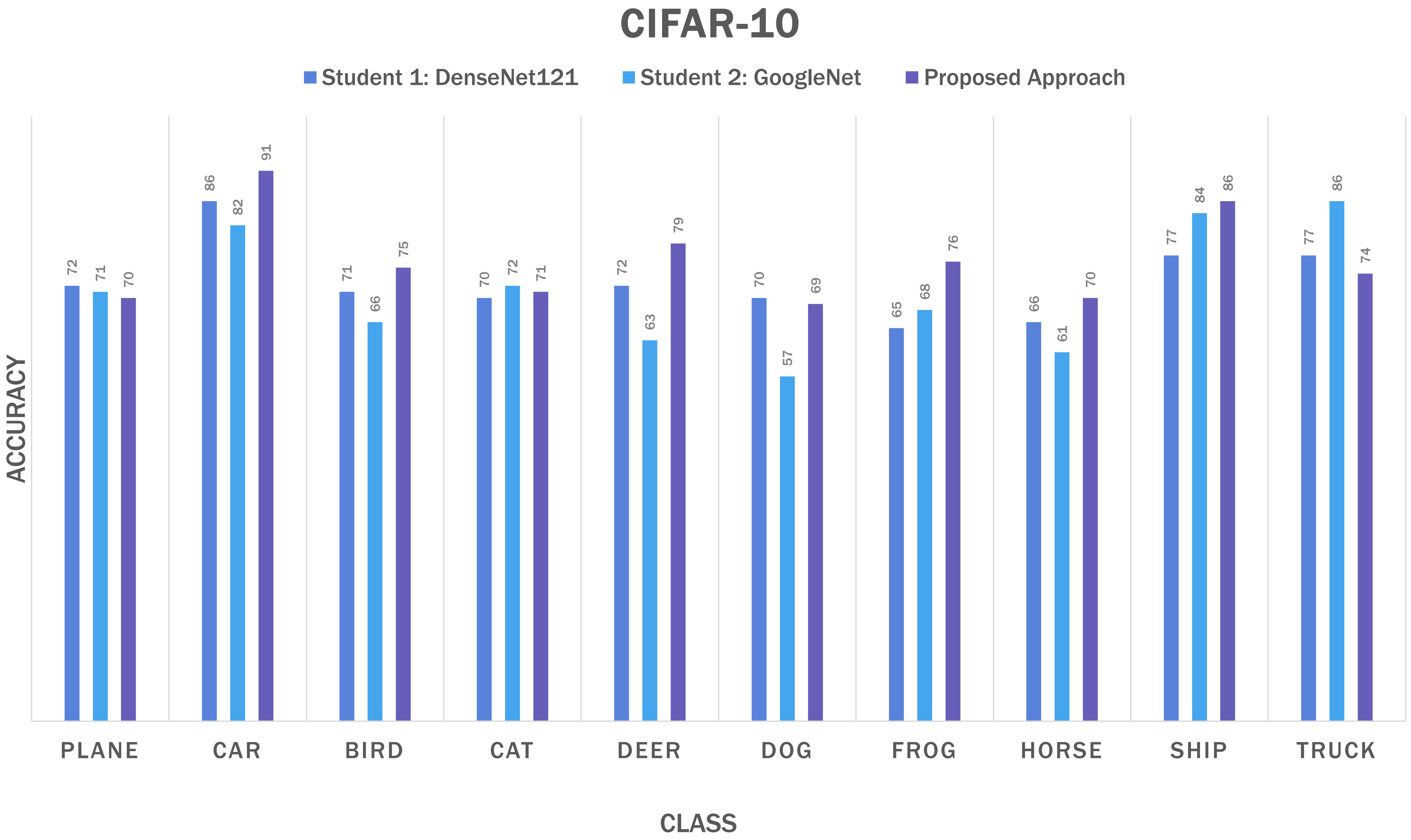
Curtesy: Google Images

MODEL ATTRIBUTES

Model Attribute	Teacher	Student 1	Student 2	Model Selector
Model	ResNet50	DenseNet121	GoogleNet	CNN (3 Layers)
No. of Parameters	25.6M	3.27M	6.07M	0.55M
Parameter size	98MB	31MB	25MB	2.21MB
No. of layers	50	11	22	3

WORKFLOW:

1. Training the teacher model.
2. Using the softmax outputs of teacher to train both the student networks.
3. Optimising both Binary Cross Entropy loss and knowledge distillation loss.
4. Using a model selector network to learn the attributes of the input image data and map it to the corresponding student model. (by learning the attributes of the model also)
5. Optimising the model selector loss over the period of time.
6. Model selector becomes better at predicting the corresponding student model for a given data that would give better predictions.


CIFAR-10
Student 1: DenseNet121
Student 2: GoogleNet
Proposed Approach
ACCURACY
72
71
70
PLANE
86
82
91
CAR
71
66
75
BIRD
70
72
71
CAT
72
63
79
DEER
70
57
69
DOG
65
68
76
FROG
66
61
70
HORSE
77
84
86
SHIP
77
86
74
TRUCK
CLASS

COMPRESSION

Compression with respect to Teacher			
Compression of Individual Student Models	**No. of Parameters**	**Space consumed**	**MAC**
DenseNet-121	**87.30%**	**68.40%**	**21%**
GoogleNet	**76.30%**	**74.49%**	**68.10%**

Compression with our proposed approach	No. of Parameters	Space consumed
DenseNet-121+ Model Selector	**85.17%**	**66%**
GoogleNet+Model Selector	**74.15%**	**72.23%**

Overall compression	No. of Parameters	Space consumed
Teacher(ResNet-50)+DenseNet-121+GoogleNet+ Model Selector	**61.40%**	**40.61%**

SELECTION CRITERIA

WHY DensNet121 and GoogleNet as Student Models??

- We observed that both model’s behaviour was complementary to each other across all the classes unlike other combinations.

- Also, the chosen models had very less number of parameters (3M and 6.07M) and provided nearly the same accuracy as the teacher.

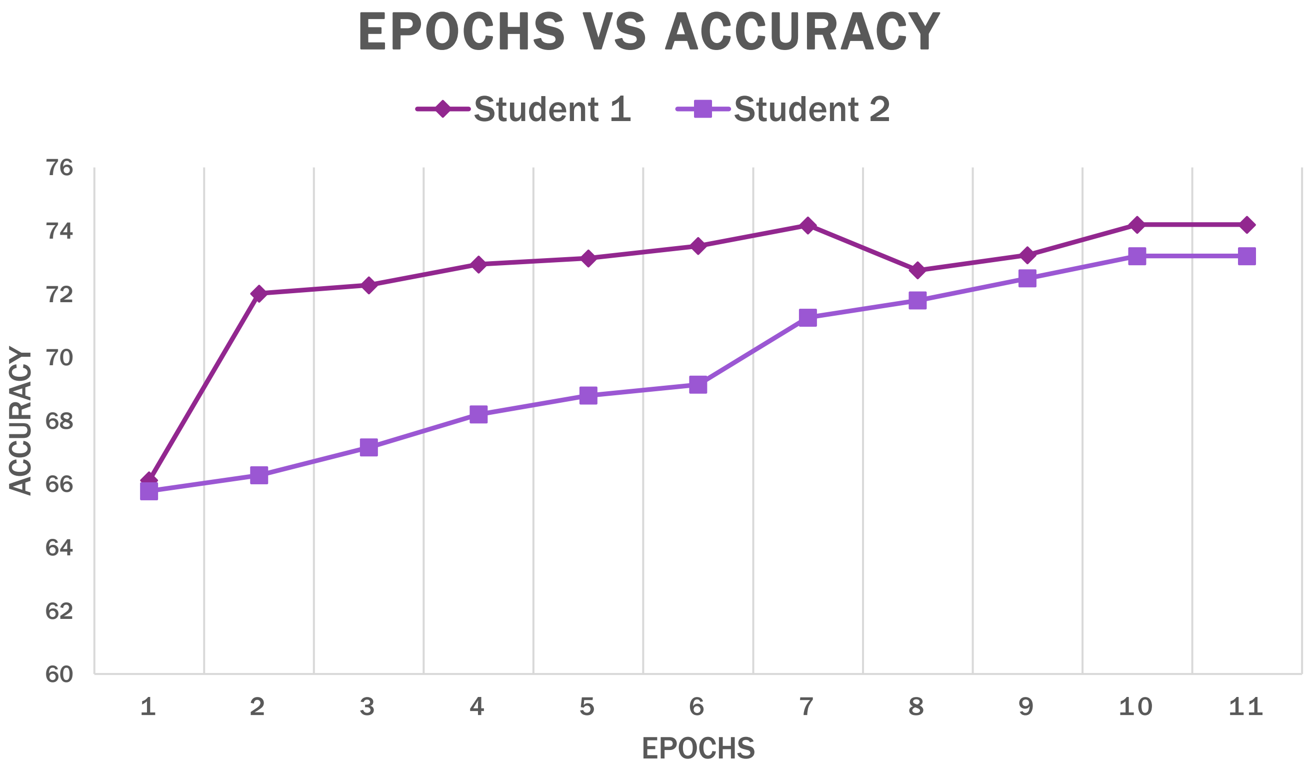
Model Selector

- Model Selector plays a fundamental role in identifying which student model will give the best accuracy for a given image data. So, the role of the model selector is to extract the features from the input image data and select the corresponding student model which will provide the best estimation.

CNN Model	Batch Norm	Without Batch Norm
2 Layer	Poor performance	Poor performance
3 Layer	Good Performance	Poor performance
4 Layer	Poor generalisation	Poor generalisation

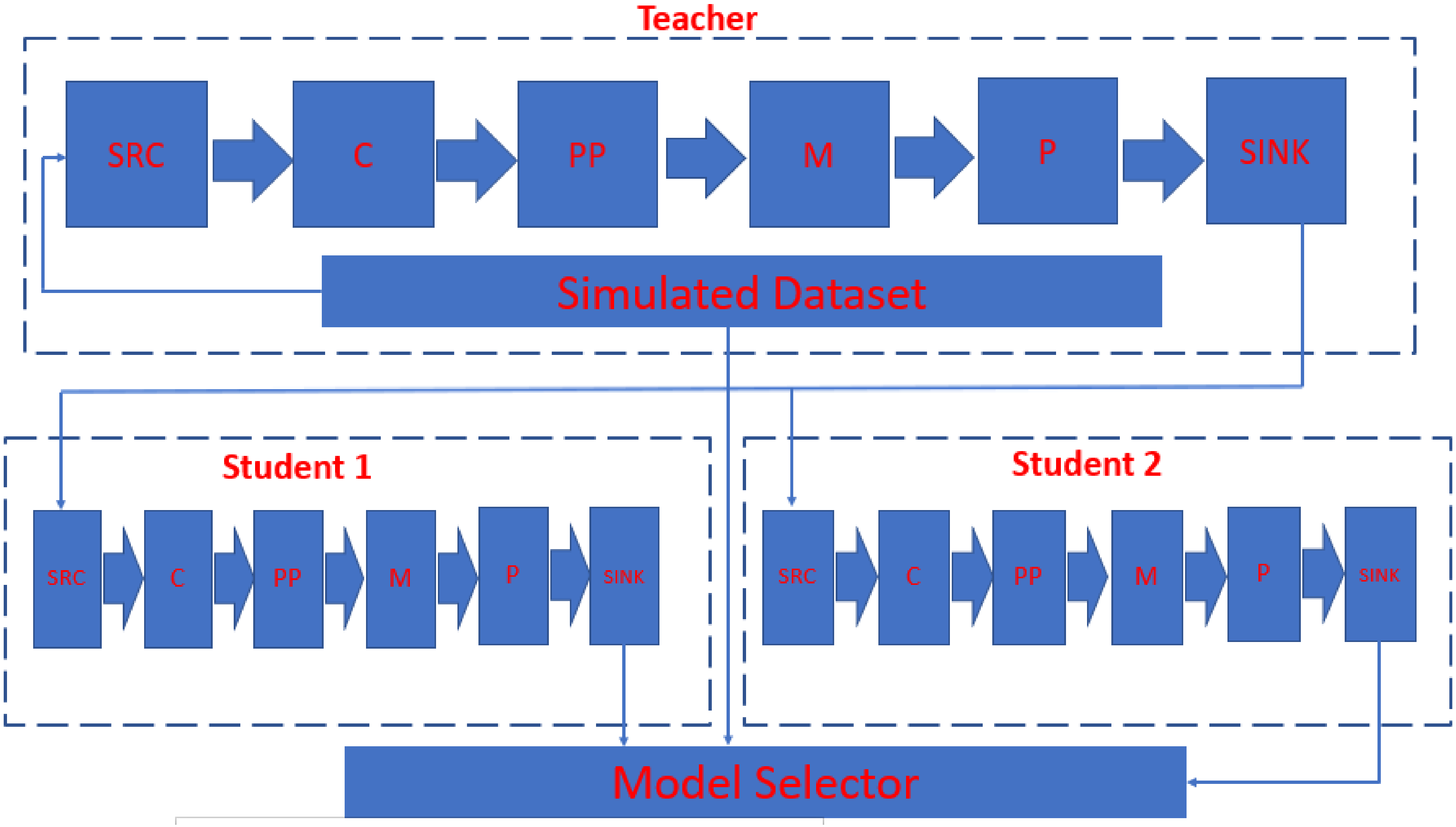
Model	Accuracy
Teacher: ResNet50	81%
Student1:DenseNet121	74.19%
Student2: GoogleNet	73.21%
Model Selector + Students(Both 1 and 2)	81.79%

It can be inferred from the table that both the student model perform much better when augmented with a model selector.



We observe that the individual student model's accuracy and the accuracy of the proposed approach has not been compromised and in fact exceeded the teacher models accuracy. We are able to gain(achieve) this accuracy with a huge margin of compression . (Refer compression Table)

ML ARCHITECTURE FOR TRAINING – ITU FORMAT



ML ARCHITECTURE FOR INFERENCING ITU FORMAT

SRC: Data Source

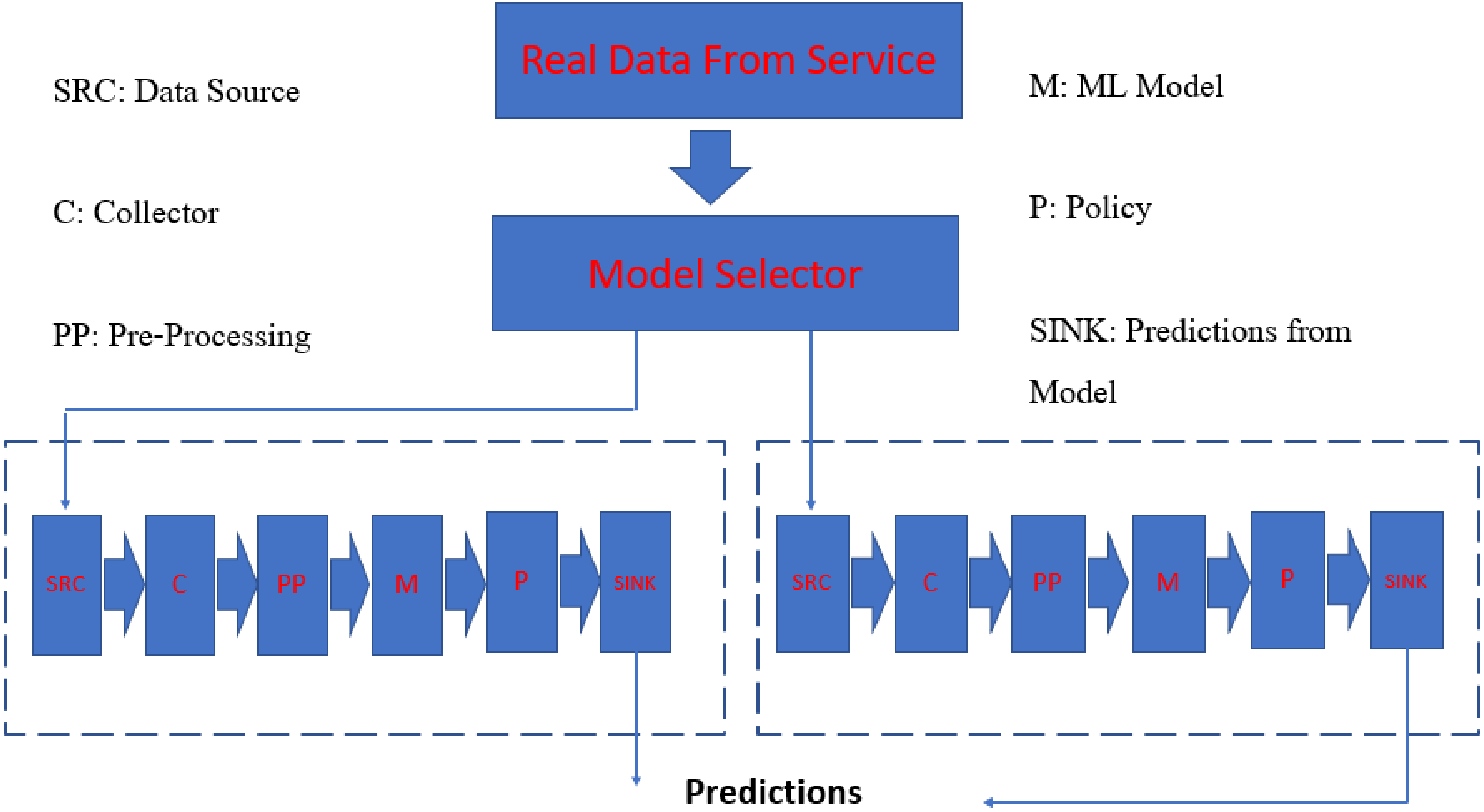
C: Collector

PP: Pre-Processing

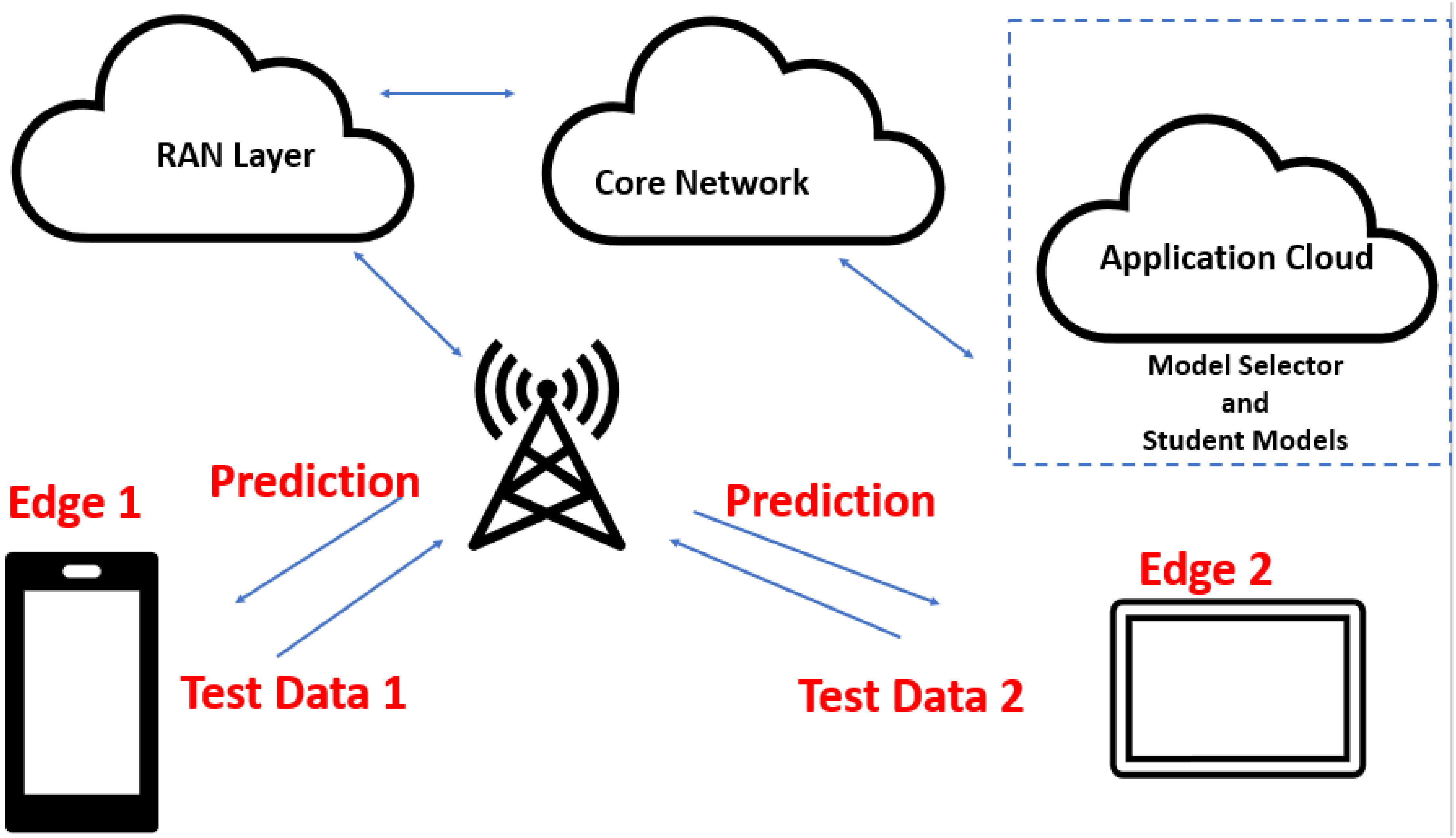
M: ML Model

P: Policy

SINK: Predictions from Model



EDGE IMPLEMENTATION OVER NETWORK



CONCLUSION

In this work, we observe that the underlying data geometry affects the models predictions and convergence to a great extent. Our key finding is that, knowledge Distillation (KD) is not a panacea for inferencing on all kinds of data.

Since the world is moving more towards an active learning paradigm, one student model on an edge device will not suffice to accommodate all the variations in the attributes. So, we proposed a multi-student Architecture augmented with a highly compressed model selector network which can understand the attributes of an input data and route it to the corresponding student model which can give better results by utilising least amount of computing resources.

To prove our hypothesis we have shown the benefits of this approach on CIFAR-10 Dataset by carrying out extensive empirical study. The results reveal that, the proposed Multi-context Aware architecture perform better than independent student models.

The observed results demand a pressing need for understanding the nuances of context aware meta architectures which can be deployed in 5G Edge computing scenarios, as we believe this can succeed as a general and a practical approach