Distilling Knowledge

A Comprehensive Survey of Knowledge Distillation Techniques in Deep Learning

Team 3
Hitesh Goel (2020115003)
Devesh Marwah (2020115005)
Vithesh Reddy Adala (2020115002)

01 Problem Statement

Knowledge transfer in deep learning is resource-intensive.

02 Lit Review, Solution Proposed and Scope

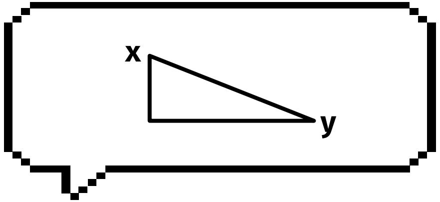
Transferring knowledge from a large model to a smaller model.

03 Implementation and Experiments

Compare with baseline models, conduct ablation studies.

04 Results and Contribution

Distillation improves the performance of smaller models



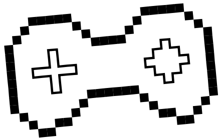
01

Problem Statement

Knowledge distillation is like making a concentrated version of your favorite drink - you boil down the essence of your teacher's knowledge and pour it into your own brain. Just be careful not to burn your neurons!

Motivation

The increasing demand for deploying deep learning models on low-level devices such as mobile phones necessitates the development of techniques that can **distill knowledge** from large models to small models with a significant reduction in model size and number of parameters, while maintaining a high level of accuracy.



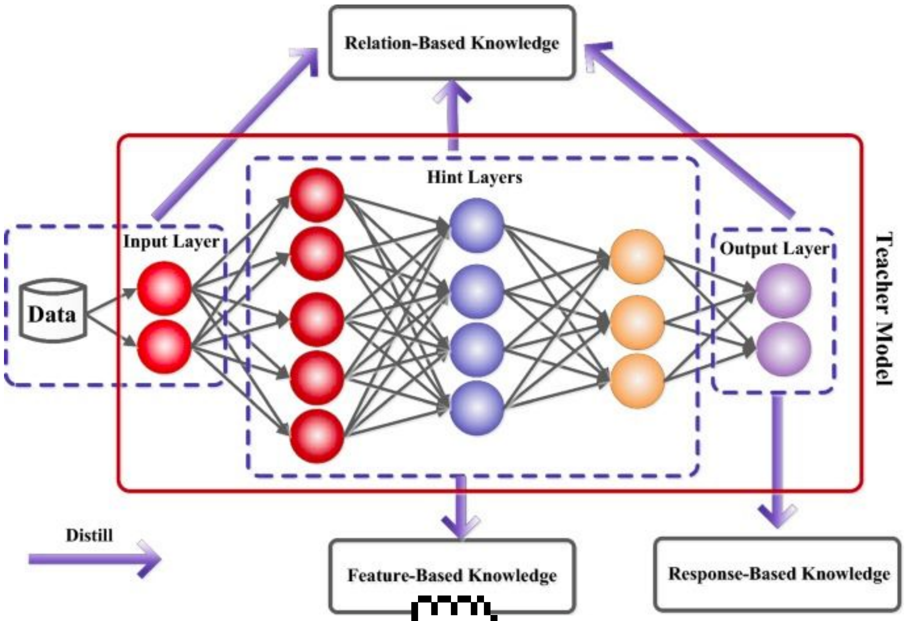
404 NOT FOUND?

02

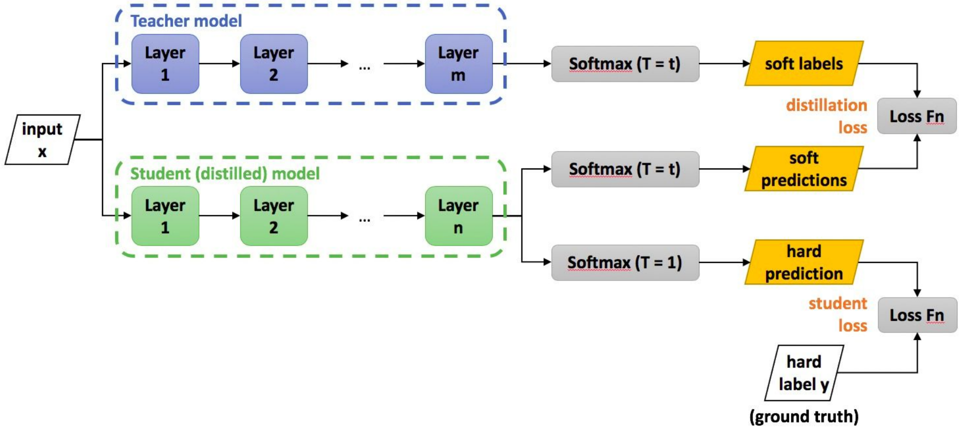
Literature Review, Solution Proposed & Scope

Knowledge distillation transfers knowledge from a large model to a smaller model.

Approaches for Knowledge Distillation
A Visual Overview



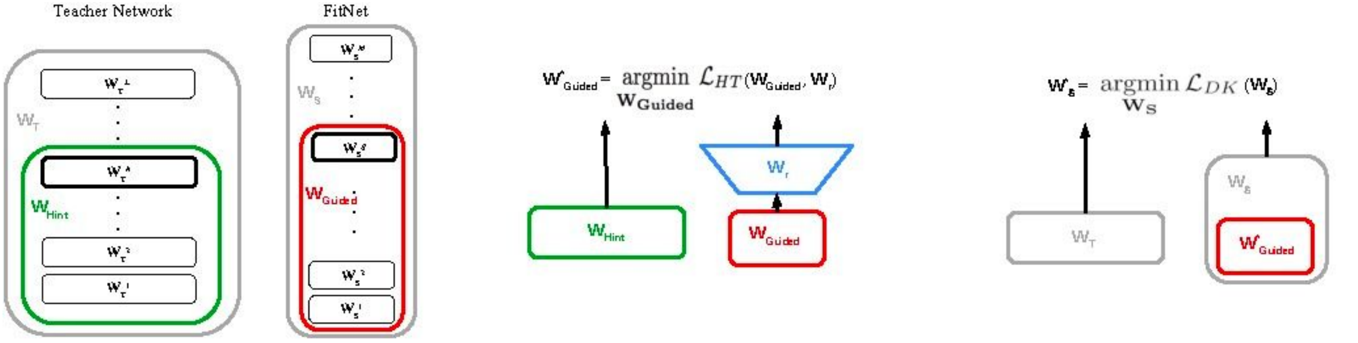
Paper 1: Hinton et al



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

$$\frac{\partial C}{\partial z_i} = \frac{1}{T}(q_i - p_i) = \frac{1}{T}\left(\frac{e^{z_i/T}}{\sum_j e^{z_j/T}} - \frac{e^{v_i/T}}{\sum_j e^{v_j/T}}\right)$$

Paper 2: FitNets



(a) Teacher and Student Networks (b) Hints Training (c) Knowledge Distillation

Hint-Based Learning Loss function:

$$\mathcal{L}_{HT}(\mathbf{W_{Guided}}, \mathbf{W_r}) = \frac{1}{2}||u_h(\mathbf{x}; \mathbf{W_{Hint}}) - r(v_g(\mathbf{x}; \mathbf{W_{Guided}}); \mathbf{W_r})||^2,$$

$$\mathcal{L}_{KD}(\mathbf{W_S}) = \mathcal{H}(\mathbf{y_{true}}, \mathrm{P_S}) + \lambda\mathcal{H}(\mathrm{P_T^\tau}, \mathrm{P_S^\tau}),$$

Paper 3: Relational Knowledge Distillation

$$\mathcal{L}_{\text{RKD}} = \sum_{(x_1,..,x_n)\in\mathcal{X}^N} l\big(\psi(t_1,..,t_n), \psi(s_1,..,s_n)\big),$$

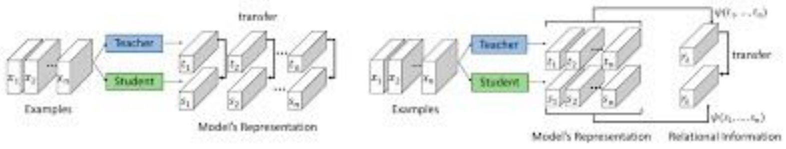
$$\psi_{\text{D}}(t_i, t_j) = \frac{1}{\mu}\,\|t_i - t_j\|_2\,,$$

$$\mathcal{L}_{\text{RKD-D}} = \sum_{(x_i,x_j)\in\mathcal{X}^2} l_\delta\big(\psi_{\text{D}}(t_i,t_j), \psi_{\text{D}}(s_i,s_j)\big),$$

$$l_\delta(x,y) = \begin{cases} \frac{1}{2}(x-y)^2 & \text{for } |x-y| \le 1, \\ |x-y| - \frac{1}{2}, & \text{otherwise.} \end{cases}$$

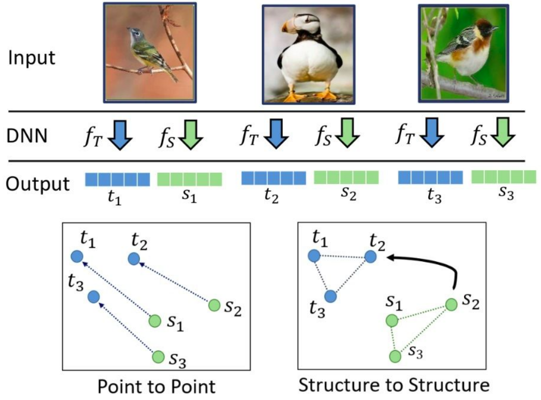
$$\psi_{\text{A}}(t_i,t_j,t_k) = \cos\angle t_i t_j t_k = \langle \mathbf{e}^{ij}, \mathbf{e}^{kj} \rangle$$

$$\text{where} \quad \mathbf{e}^{ij} = \frac{t_i - t_j}{\|t_i - t_j\|_2}, \mathbf{e}^{kj} = \frac{t_k - t_j}{\|t_k - t_j\|_2}.$$



Individual Knowledge Distillation

Relational Knowledge Distillation



Conventional KD

Relational KD

03

Implementation & Experiments

using PyTorch we compare the performance of novel techniques with predetermined baselines

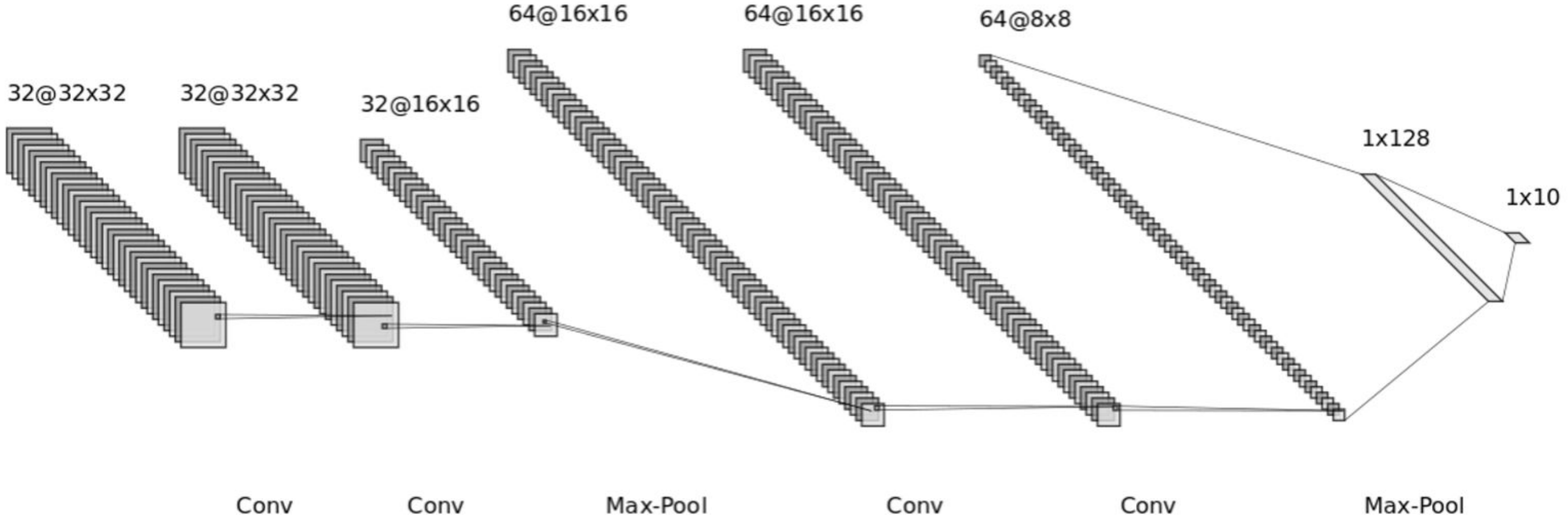
Data Description

Our Model is trained on CIFAR-10 dataset.
The CIFAR-10 dataset consists of 60000 32×32×3 colour images in 10 classes, with 6000 images per class.
There are 50000 training images and 10000 test images.

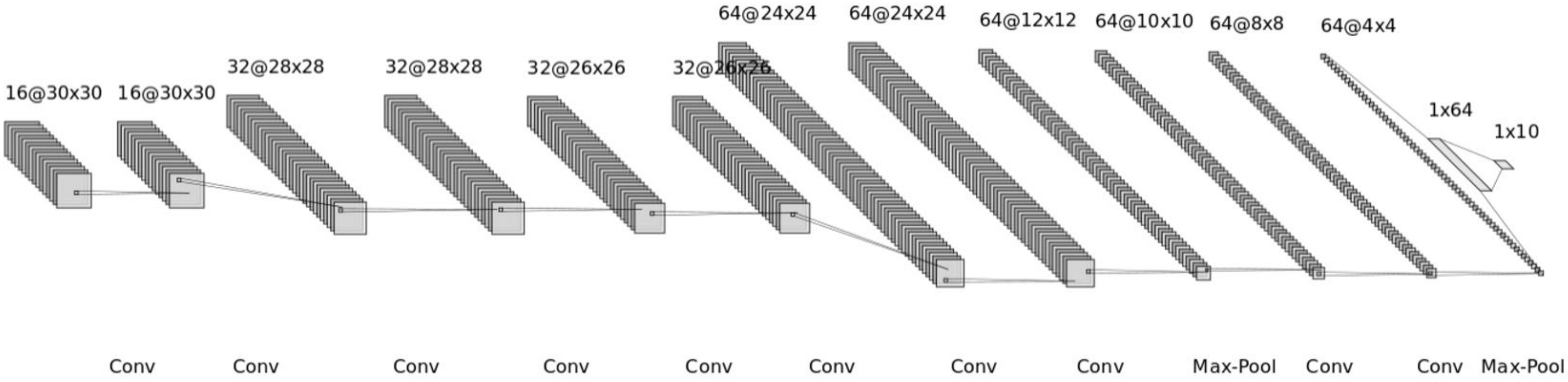
Models that we used

1. Pretrained ResNet
    a. used as the source of knowledge to transfer to a smaller student model.
    b. capable of achieving high accuracy on a variety of computer vision tasks.

2. Novel Teacher Model Architecture



3. Novel Student Model Architecture




After some intense coding sessions

04

Results & Contribution

Knowledge distillation can significantly reduce model size and number of parameters while maintaining high accuracy.

Baseline Results

	Score
ResNet110 Teacher	91%
ResNet20 Student	82%
Novel Teacher	75%
Novel Student	63%

Response Based (Hinton et al)

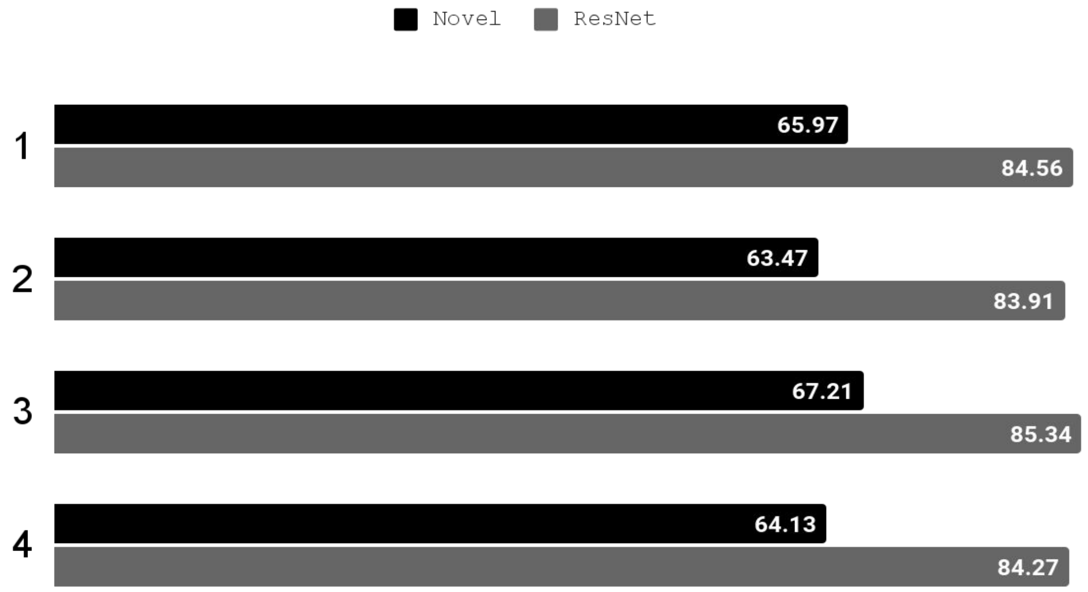
Novel Architecture

Temperature	Alpha	Score
10	0.2	65.97%
20	0.5	63.47%
5	0.2	67.21%
40	0.7	64.13%

Pretrained ResNet

Temperature	Alpha	Score
10	0.2	84.56%
20	0.5	83.91%
5	0.2	85.34%
40	0.7	84.27%

Response Based (Hinton et al)



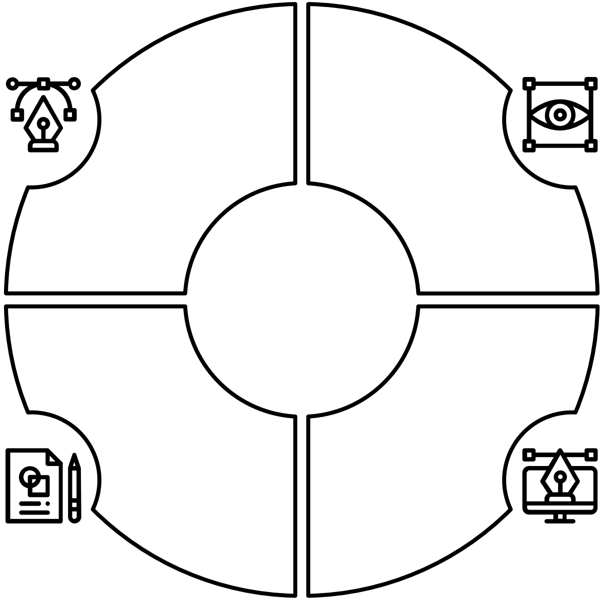
Feature Based (Romero et al)

Task	Novel Architecture Score	PreTrained ResNet
Mimicking middle layer	68.36%	85.57%
Mimicking final layer	54.06%	76.86%

Relation Based (devesh iska author likh dio)

w_dist	w_angle	Score
25	50	87.00
20	60	86,13

Team Contributions



Response-based

Devesh, Vithesh

Relation-based

Hitesh, Devesh

Feature-based

Hitesh, Vithesh

Presentation

All

Thank you