Compression Neural Nets

Under the guidance of,

Dr. Animesh Chaturvedi

Team Members

19bcs035 - Dasari Rishikesh

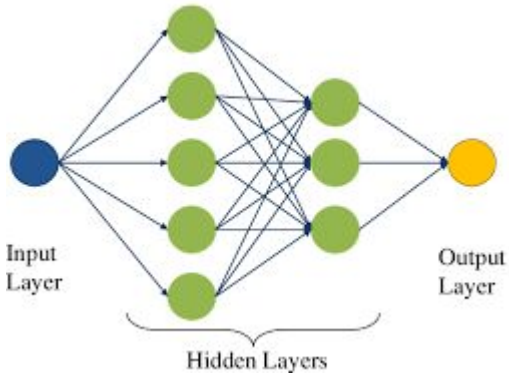
19bcs068 - Mathangi Sravan

19bcs101 - Sompalli Ajay kumar

Table of contents

Overview

- Deep Neural Networks has achieved great success in Computer vision tasks.
- It is generally optimised to get more accurate results.
- DNN have huge number of parameters sometimes in order of millions.

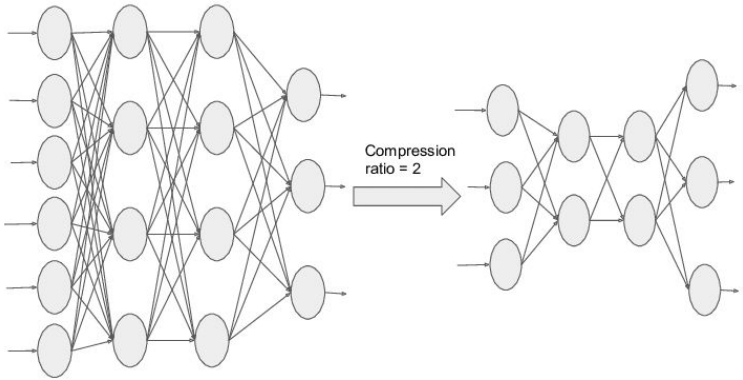


Motivation

- DNN are computationally expensive and memory intensive.
- The Deep Neural Network uses the high end GPUs and powerful processors.
- It is necessary for the devices like mobiles, UAVs(Unmanned Aerial Vehicles) and IoT devices needed for better privacy, less network bandwidth and real time processing.
- So goal to perform model compression and acceleration.

Problem Statement

Reduce the storage and energy required to run the interface of large networks so that they can be deployed on the portable devices with limited hardware resources.



Literature Review

- Compression with Pruning , Trained Quantization and Huffman coding by Song Han(Conference at ICLR, 2016)
- Knowledge distillation : A good teacher is patient and consistent by Lucas Beyer(CVPR, 2022)
- A Survey of Model Compression and Acceleration for Deep Neural Networks by Yu Cheng (IEEE, 2020)
- Combining Weight pruning and Knowledge distillation for CNN Compression by Nima Aghli (CVPR, 2021)

Categories

SUMMARIZATION OF DIFFERENT APPROACHES FOR MODEL COMPRESSION AND ACCELERATION.

Category Name	Description	Applications	More details
Parameter pruning and quantization	Reducing redundant parameters which are not sensitive to the performance	Convolutional layer and fully connected layer	Robust to various settings, can achieve good performance, can support both train from scratch and pre-trained model
Low-rank factorization	Using matrix/tensor decomposition to estimate the informative parameters	Convolutional layer and fully connected layer	Standardized pipeline, easily to be implemented, can support both train from scratch and pre-trained model
Transferred/compact convolutional filters	Designing special structural convolutional filters to save parameters	Convolutional layer only	Algorithms are dependent on applications, usually achieve good performance, only support train from scratch
Knowledge distillation	Training a compact neural network with distilled knowledge of a large model	Convolutional layer and fully connected layer	Model performances are sensitive to applications and network structure only support train from scratch

Methodology

- Initial approach with the parameter pruning and quantisation.

- Flaws exist when comes to complex architecture i.e., Residual Network models.

- Compression models doesn’t support change of model family and exist architecture-dependent challenges.

Deployed Environment:

Colab
TensorFlow

Datasets:

MNIST

CIFAR -10

Flowers 102

Network Architectures:

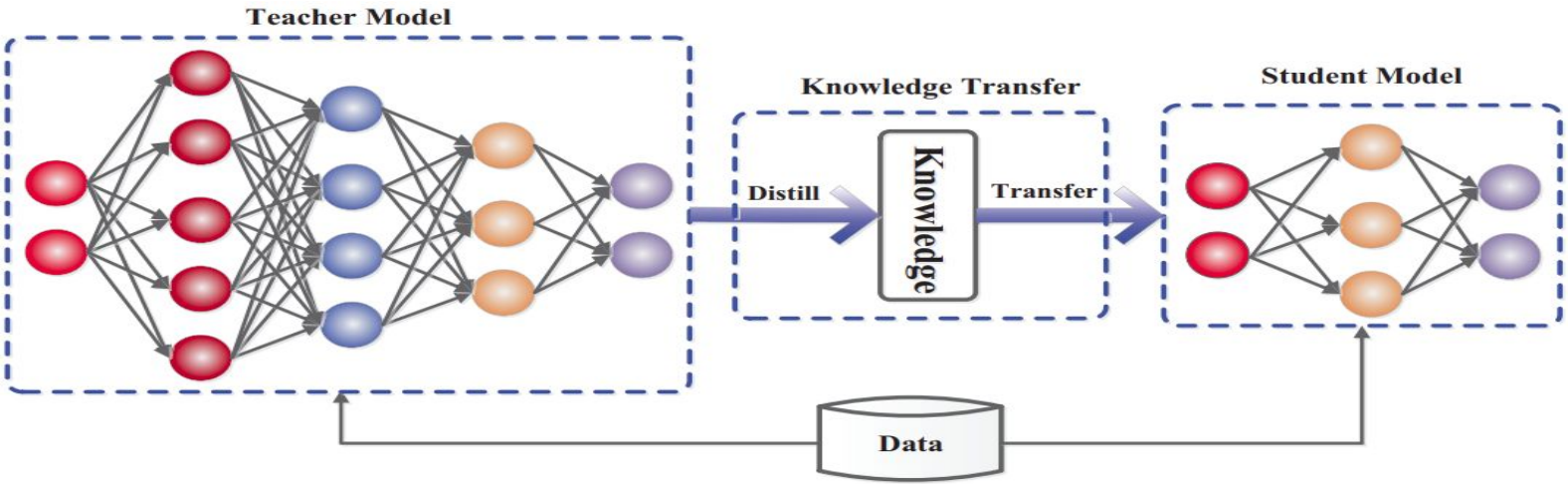
Convolution Neural Network

Residual Network

Knowledge distillation

- Knowledge distillation is technique of transferring knowledge from the larger into smaller models.
- The models are larger model are known as teacher and student model as student model.
- The compression model minimizes the loss function (distillation loss) and aimed at the matching the teacher logits as well as ground- truth labels.
- The logits are softened by applying a “temperature “ scaling function in the softmax, effectively smoothing out the probability distribution and revealing inter class relationships by teacher.

Illustration of KD



Ref: Knowledge Distillation: Principles, Algorithms, Applications - neptune.ai

Function Matching KD

- In, Function Matching, distillation approach yields student models that can actually match the performances of teacher models.

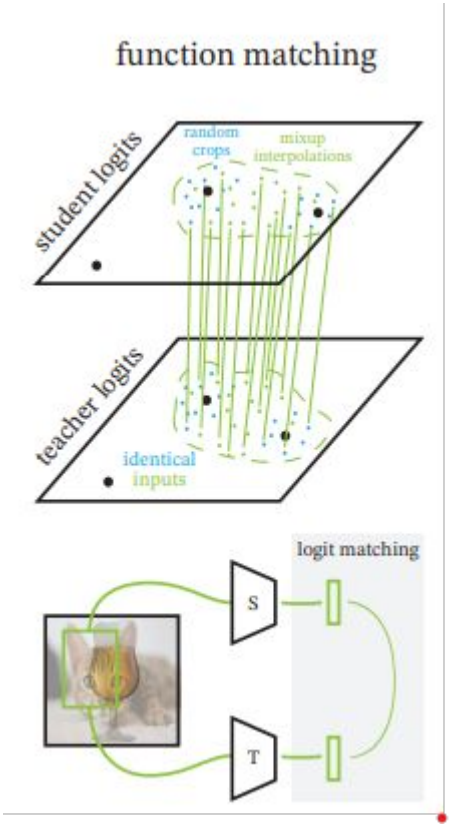
  The hypothesis has mainly three key takeaways:

- No use of ground-truth labels during the distillation process.

- Teacher and student should see the same input image views, i.e., same crop and augmentations.

Function Matching continued…

- The functions to match on a large number of support points to generalize them.
- Aggressive form of MixUp as the key augmentation recipe. The mix up is paired with “Inception style ” cropping.
- Long training schedules for distillation.



Ref: Knowledge distillation : A good teacher is patient and consistent by Lucas Beyer(CVPR, 2022)

Mathematical Interpretation

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

Loss 1(soft targets) = Cross Entropy(teacher_pred, student_pred)

Loss 2(Hard targets) = Cross Entropy(student_hardpred, hard labels)

Loss = α * Loss1 + (1-α) Loss 2

Ref : https://www.youtube.com/watch?v=gADXP5daZeM

Function Matching:

$$\mathrm{KL}(p_t||p_s) = \sum_{i \in \mathcal{C}} [-p_{t,i} \log p_{s,i} + p_{t,i} \log p_{t,i}]$$

Ref: Knowledge distillation : A good teacher is patient and consistent by Lucas Beyer(CVPR, 2022)

Results

- Knowledge distillation on MNIST Dataset

Datasets	Models	Architecture	Accuracy (sparse categorical accuracy)	Epochs
MNIST	CNN(Teacher)	Layers (256,512)	97.78%	25
MNIST	CNN(Student)	Layers (16,32)	97.50%	25
CIFAR-10	CNN(Teacher)	Layers (256,512)	69.72%	10
CIFAR-10	CNN(Student)	Layers (16,32)	58.55%	10

- Knowledge distillation using function matching hypothesis on Flowers-102 dataset

Dataset	Models	Accuracy (Top-1 Accuracy)
Flowers -102	BiT ResNet 101*3(Teacher)	98.18%
Flowers -102	BiT ResNet 50*1 (Student) for 1000 epochs	81.02%

Conclusion

In this project we have analysed application of compression models on different Network architectures. While using different models, we observed that the Pruning does not go well for all architectures, as because it changes the shapes of input in the architecture. Whereas Knowledge Distillation can be easily applied for different architectures without losing information. It provides close accuracy with the original model.

References

[1] Song Han, Huizi Mao, William J. Dally, "*Deep Compression: Compression Deep Neural Networks with pruning, Trained quantization and Huffman coding*", Conference at ICLR 2016

[2] Lucas Beyer, Xiaohua Zhai, Amelie Royer, Larisa Markeeva, *"Knowledge distillation: A good teacher is patient and consistent"*, CVPR 2022.

[3] J. Ba and R. Caruana, "Do deep nets really need to be deep?" in *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, 2014, pp. 2654-2662.

[4] Nima Aghli, Eraldo Ribeiro, *"Combining Weight Pruning and knowledge Distillation for CNN compression"*, CVPR 2021.

[5] Yu Cheng, Duo Wang, Pan Zhou, *"A survey of Model Compression and Acceleration for Deep Neural Networks"*, IEEE signal processing Magazine.

Thank You