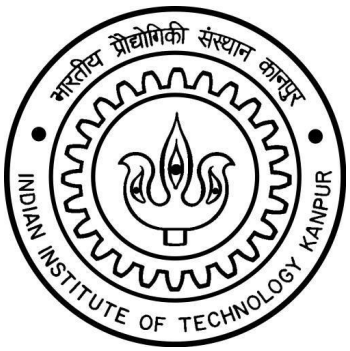


Science and Technology Council, Indian Institute of Technology Kanpur

Model Zoo

End Eval Project Report

July 28, 2021

Contents

1 Introduction

Deep Learning is a subfield of machine learning concerned with algorithms inspired by the structure and function of the brain called artificial neural networks.Deep learning excels on problem domains where the inputs (and even output) are analog. Meaning, they are not a few quantities in a tabular format but instead are images of pixel data, documents of text data or files of audio data. In deep learning, each level learns to transform its input data into a slightly more abstract and composite representation. In an image recognition application, the raw input may be a matrix of pixels; the first representational layer may abstract the pixels and encode edges; the second layer may compose and encode arrangements of edges; the third layer may encode a nose and eyes; and the fourth layer may recognize that the image contains a face. Importantly, a deep learning process can learn which features to optimally place in which level on its own. (Of course, this does not completely eliminate the need for hand-tuning; for example, varying numbers of layers and layer sizes can provide different degrees of abstraction.)

This project focuses on making a collection of Deep Learning models that can be applied to various fields ranging from image classification to super resolution of images.

Figure 1 below demonstrates a simple Neural Network.

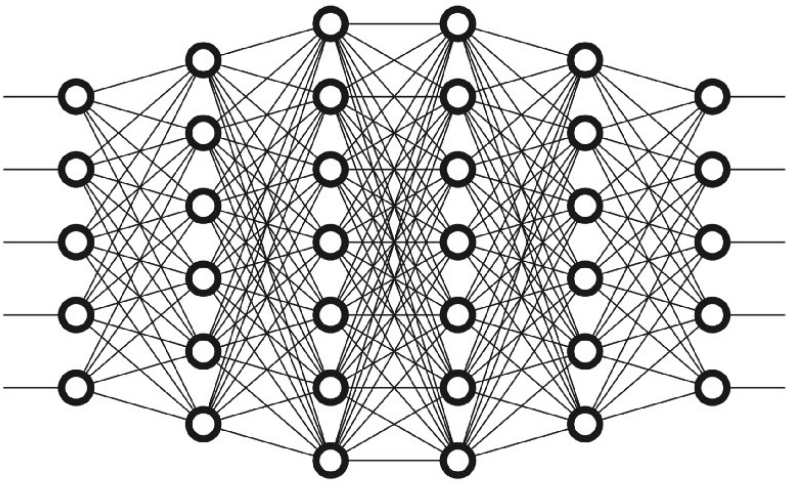


Figure 1: A simple neural network

2 Overview

We implemented 15 models in the Model-Zoo spanning many categories. We used both PyTorch and Tensorflow which are the most popular Deep Learning frameworks for implementation of the models. The project repository can be found on GitHub (`https://github.com/pclubiitk/model-zoo`). Each model is in a separate directory with a README including the usage and architecture details. All the models are additionally listed in the tables below :

Model	PyTorch	Tensorflow
Image Classification		
Inception-v3	✓	✗
Inception-v1	✓	✓
Xception	✗	✓
MobileNet-v1	✗	✓
RepVGG	✓	✓
MLP Mixer	✓	✓
I3D	✗	✓
Super Resolution		
Perceptual Losses	✓	✗

Model	PyTorch	Tensorflow
Object Detection		
Faster-RCNN	✗	✓
SSD	✗	✓
YOLO-v1	✗	✓
Medical Image Segmentation		
UNet	✓	✗
UNet++	✓	✗
Multimodal Models		
StackGAN++	✓	✓
Show and Tell	✓	✓

3 Categories

3.1 Classification

Classification has played an immense role in the field of deep learning. Deep Convolutional Neural Networks have led to a series of breakthroughs for image classification. Deep networks naturally integrate low/mid/high level features and classifiers in an end-to-end fashion. We have implemented seven models in this area :

- **Inception-v1** – Inception-v1 achieves improved utilization of computing resources by a design that allows increasing the depth and width of the network while keeping computational budget constant. It also introduces Inception module which are repeated many times in the model leading to an increased network depth.

- **Inception-v3** – Inception-v3 focuses on developing an efficient deep neural network architecture for computer vision. The most straightforward way of improving the performance of deep neural networks is by increasing their size. This includes both increasing the depth and width of the network. It explores ways to scale up networks in ways that aim at utilizing the added computation as efficiently as possible by suitably factorized convolutions and aggressive regularization.

- **Xception** – Xception is an extreme version of the Inception which has the same number of parameters as the standard Inception-V3 model but its improvement in performance is due to efficient use of model parameters. This model first uses a 1x1 (Pointwise Convolutions) to map cross-channel correlations and then would separately map the spatial correlations of every output channel. Thus the two mappings are entirely decoupled which is a stronger version of the hypothesis underlying the Inception architecture.

- **MobileNet-v1** – MobileNets are built primarily from depthwise separable convolutions initially introduced in and subsequently used in Inception models to reduce the computation in the first few layers. Flattened networks build a network out of fully factorized convolutions and showed the potential of extremely factorized networks. It is an efficient network architecture with two hyperparameters that can be used in mobile applications due to its small size and is efficient in comparison to the extremely dense networks.

- **RepVGG** – RepVGG is a simple but powerful architecture of convolutional neural network, which has a VGG-like inference time body composed of nothing but a stack of 3x3 convolution and ReLU, while the training-time model has a multi-branch topology. Such decoupling of the training time and inference-time architecture is realized by a structural re-parameterization technique so that the model is named RepVGG.

- **MLP Mixer** – Based exclusively on Multi-layer Perceptrons, MLP Mixer relies only on basic matrix multiplication, changes on data layout and scalar nonlinearities. It uses two types of MLP's, one for mixing per-location features and the other for mixing features from different spatial locations. It shows that although Convolutions and Attention are both sufficient for good performance, they are not necessary.

- **I3D** – It introduces a new Two-Stream Inflated 3D ConvNet (I3D) is based on 2D ConvNet inflation: filters and pooling kernels of very deep image classification ConvNets are expanded into 3D, making it possible to learn seamless spatio-temporal feature extractors from video while leveraging successful ImageNet architecture designs and even their parameters.This model serparately learns features from an RGB Stream and an Optical-Flow Stream and the outputs of the two streams are averaged to predict actions in the videos.

3.2 Object Detection

Goal – Detecting and Classifying any object present in a given image.
Object detection has applications in many areas of computer vision, including image retrieval and video surveillance.
This problem itself dates as back as 1960s. Several approaches are proposed to solve this like **You Only Look Once (YOLO)**, **Single Shot Detector (SSD)**, **Faster-RCNN**. They make real time object detection possible over large classes using Deep Neural Networks. We have successfully implemented YOLOv1, SSD, and Faster-RCNN. Some key points of them are as follows:

- **YOLOv1** – YOLO, a new approach to object detection in which we frame object detection as a regression problem to spatially separated bounding boxes and associated class probabilities. A single neural network predicts bounding boxes and class probabilities directly from full images in one evaluation making it very fast. Being a single neural network it can be optimized end-to-end directly on detection performance.

- **SSD** – We assign a number of default bounding boxes each of different aspect ratios and scale, accross various feature maps of different sizes. At prediction time, the network generates scores for the presence of each object category in each default box and produces adjustments to the box to better match the object shape. SSD completely eliminates proposal generation and subsequent pixel and encapsulates all computation in a single network.

- **Faster-RCNN** – Faster RCNN is a model which provides real time object detection with Region Proposal Networks which generates a number of region proposals or bounding boxes called Region of Interests (ROIs) that has high probability of containing any object. Detection Network takes input from both the Feature Network and RPN, and generates the final class and bounding box.

3.3 Medical Image Segmentation

Goal – Simplify an image to locate objects, making it easier to analyze for biomedical purposes.

- **UNet** – It works with very few training images and yields more precise segmentations. The main idea is to supplement a usual contracting network by successive layers, where pooling operators are replaced by upsampling operators. Hence, these layers increase the resolution of the output.

- **UNet++** – It is essentially a deeply-supervised encoder-decoder network where the encoder and decoder sub-networks are connected through a series of nested, dense skip pathways. The re-designed skip pathways aim at reducing the semantic gap between the feature maps of the encoder and decoder sub-networks.

3.4 Multimodal Models

Modality refers to the way in which something happens. Our experience of the world is multimodal — we see objects, hear sounds, feel the texture, smell odors, and taste flavors. In order for Artificial Intelligence to make progress in understanding it's surroundings, it needs to be able to interpret such multimodal signals together.

Deep Learning has made the field of Multimodal Learning efficient to explore for many reseachers and as a result several multimodal models are emerging dealing with a vast variety of Modals two of which we implemented are Image generation from text (**StackGAN++**), and Image captioning (**Show and Tell**).

- **StackGAN++** – StackGAN++ expands upon the idea of StackGAN to include more Discriminator and by producing images at different levels of Generators which go through different Discriminators at once generating conditional and unconditional losses. The major idea behind the presence of multiple Generators is to enhance the data collection from the Embeddings by introducing Embeddings at each stage.

- **Show and Tell** – It is a neural net which is fully trainable using stochastic gradient descent. It combines state-of-art sub-networks for vision and language models. These can be pre-trained on larger corpora and thus can take advantage of additional data.

3.5 Super Resolution

Goal – Reconstructing a high resolution photo-realistic image from its counterpart low resolution image.

- **Perceptual Losses for Super Resolution** – It proposes the use of perceptual loss functions for training feed-forward networks for image transformation tasks, thus combining the benefit of perceptual loss functions which can generate high resolution images and per-pixel loss functions used in training feed-forward convolutional neural networks for image transformation problems.

4 Images

This section contains some of the results of the Models.

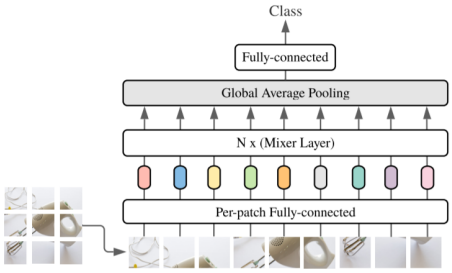


Figure 2: MLP Mixer architecture (Image Classification)

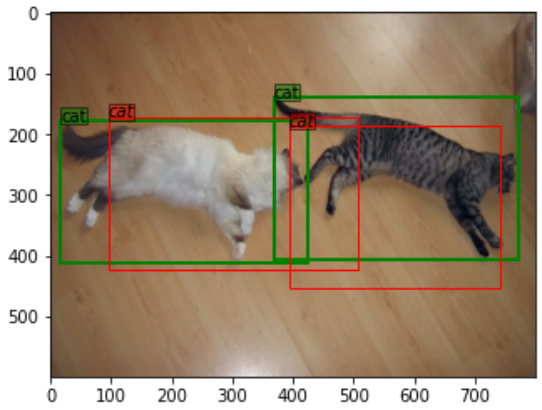


Figure 3: Faster RCNN (Object Detection)

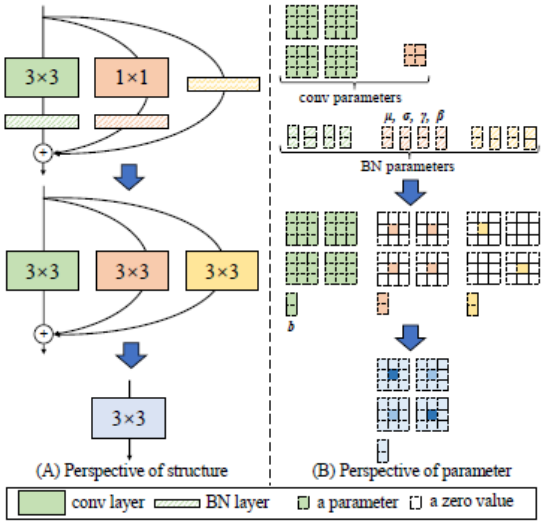
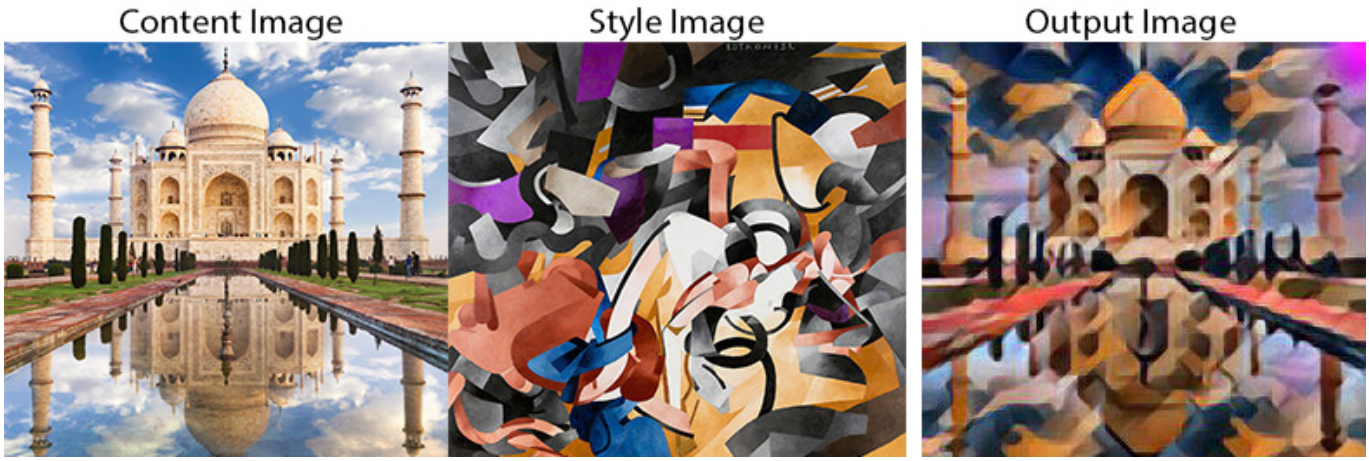


Figure 4: RepVGG (Image Classification)

Figure 5: "man wearing a red headband and a leather outfit standing outside"
-Predicted caption from Show and Tell



(a) Style Transfer result.

(b) Super Resolution result.

Figure 6: Perceptual Losses for Real-Time Style Transfer and Super-Resolution

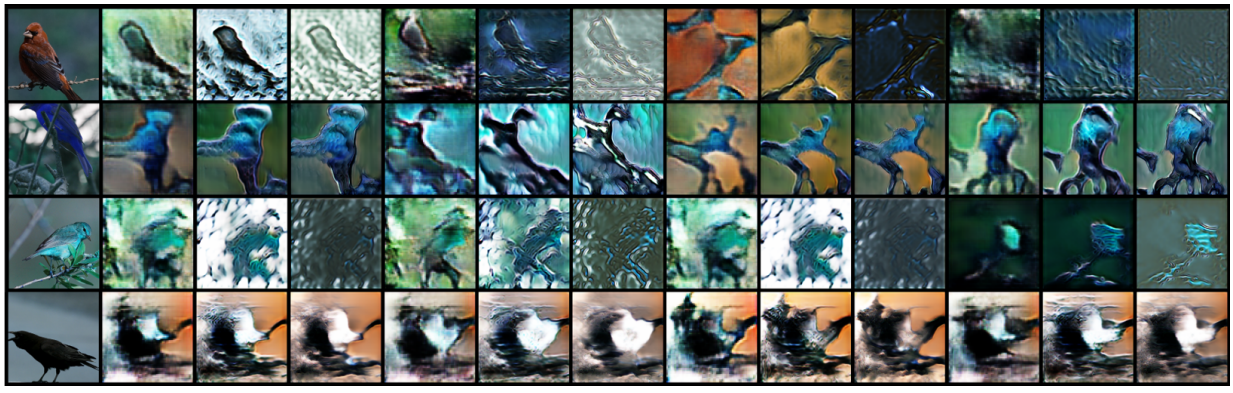


Figure 7: StackGAN++ (Multimodal Models)

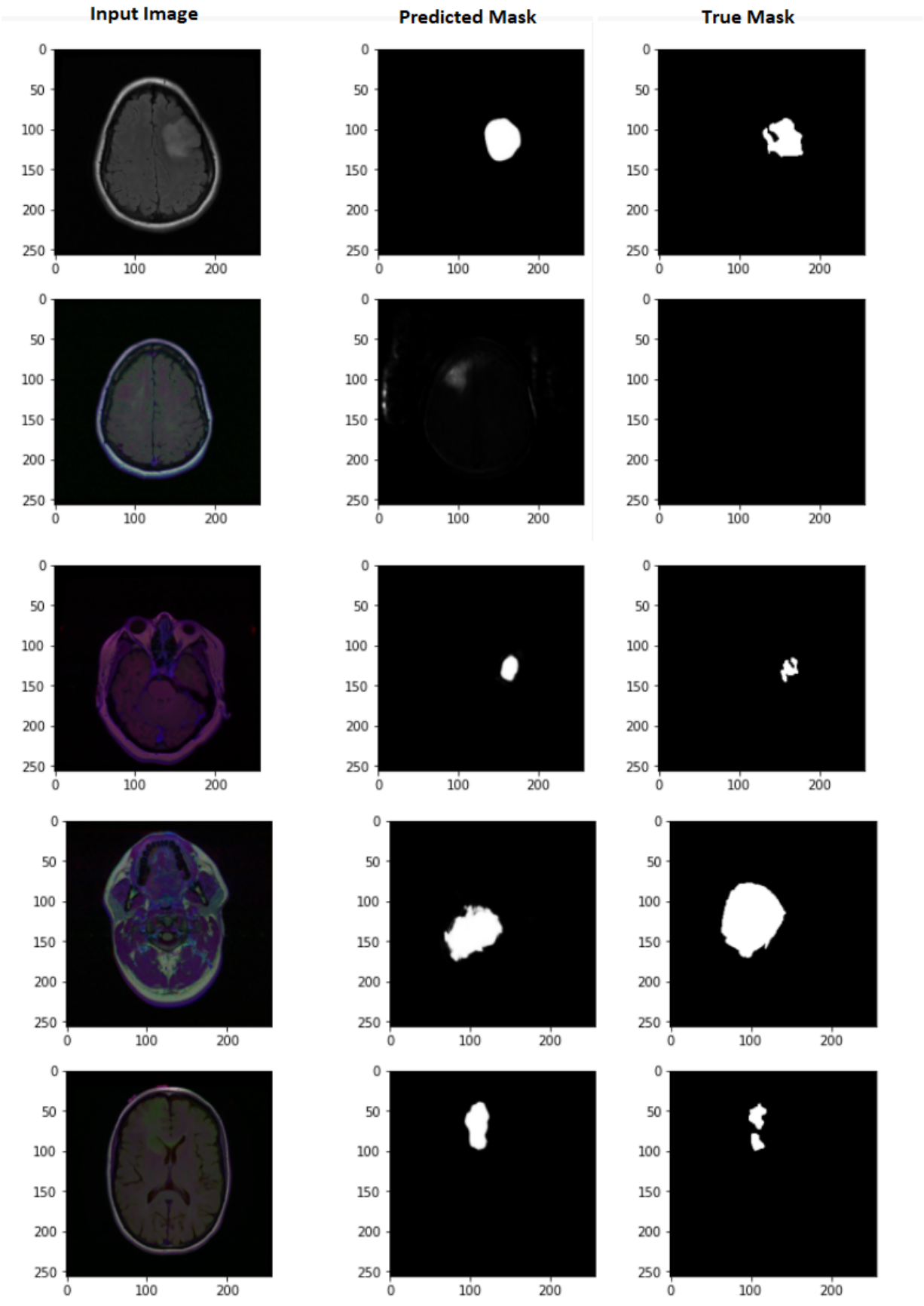


Figure 8: UNet (Medical Image Segmentation)

5 Team

5.1 Contributors

Aditya Tanwar	Ayush Kumar	Padam Sharma
Akhil Agrawal	Divyanshu Gangwar	Prem Bharwani
Aman Jain	Hitesh Anand	Rajarshi Dutta
Anirudha Brahma	Imad Khan	Rishav Bikarwar

5.2 Mentors

Atharv Singh Patlan	Rishabh Dugaye
Naman Gupta	Som Tambe