

M.Tech Dissertation Report
on
**Modality Hallucination For Multi-Modal
Approaches To Visual Inference**

Submitted in the partial fulfillment of the requirements
for the award of degree

of
Master of Technology
in
Computer Science and Engineering

Submitted By:

NEHA PANDEY

17535016



Under the supervision of
Prof. R. BALASUBRAMANIAN

**Department of Computer Science and
Engineering**

Indian Institute of Technology Roorkee

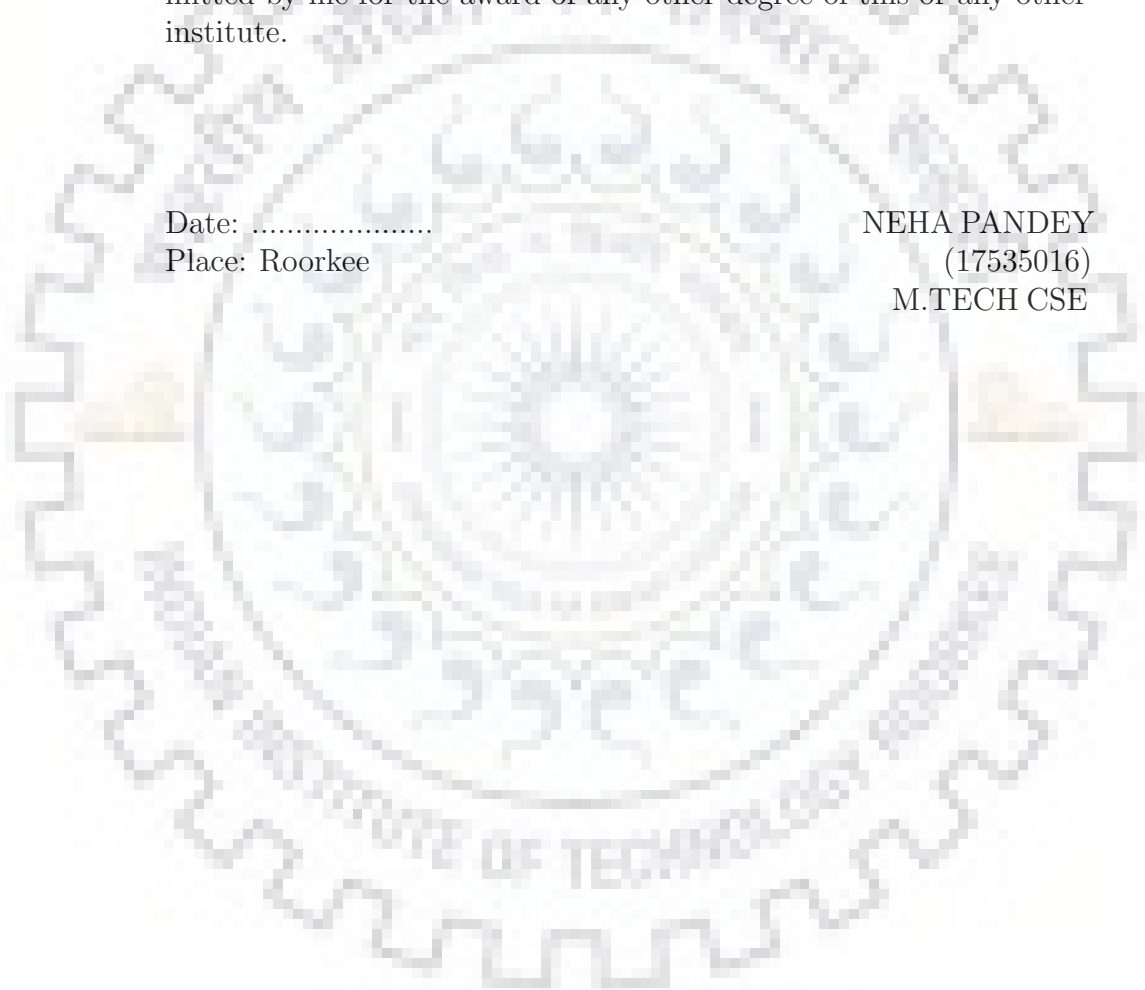
May, 2019

AUTHOR'S DECLARATION

I declare that the work presented in this dissertation with title **"Modality Hallucination For Multi-Modal Approaches To Visual Inference"** towards fulfillment of the requirement for the award of the degree of **Master of Technology in Computer Science & Engineering** submitted in the **Department of Computer Science & Engineering, Indian Institute of Technology Roorkee, India** is an authentic record of my own work carried out during the period of **May 2018 to May 2019** under the supervision of **Prof. R. BalaSubramanian**, Department of Computer Science and Engineering, Indian Institute of Technology Roorkee, Roorkee, India. The content of this dissertation has not been submitted by me for the award of any other degree of this or any other institute.

Date:
Place: Roorkee

NEHA PANDEY
(17535016)
M.TECH CSE



CERTIFICATE

This is to certify that Thesis Report entitled ”**Modality Hallucination For Multi-Modal Approaches To Visual Inference**” which is submitted by Neha Pandey (17535016), towards the fulfillment of the requirements for the award of the degree of **Master of Technology in Computer Science & Engineering** submitted in the **Department of Computer Science & Engineering, Indian Institute of Technology Roorkee, India** is carried out by her under my esteemed supervision and the statement made by the candidate in declaration is correct to the best of my knowledge and belief.

Date:

Place: Roorkee

Sign:

Prof. R. BalaSubramanian
Indian Institute of Technology Roorkee



ACKNOWLEDGEMENTS

I would first like to thank my thesis advisor Prof. R. BalaSubramanian for guiding me throughout my thesis work, helping me whenever needed and being a constant source of motivation. I am really grateful for having such wonderful and understanding mentor. I am also thankful to the Department of Computer Science Engineering and Tinkering Lab, IIT Roorkee for providing the valuable resources to aid my research.

I would like to thank my fellow classmates and friends for being there and helping me overcome the obstacles in my thesis work.

Last but not the least, I would like to thank my parents and siblings for their blessings and support without which I would not have reached this stage of my life.

NEHA PANDEY



ABSTRACT

The aim is to address the problem of exploiting multiple sources of information for object classification tasks when additional modalities that are present in the labeled training set are not available for inference. Considering the practicality of RGB-D object classifier, a modality hallucination architecture using multi-modal ConvNets has been proposed to incorporate depth information at training time. The modality hallucination network is trained to mimic mid-level features of depth images and learns a new RGB image representation. The single modality RGB test image is jointly processed using hallucination and RGB network and it outperforms the RGB model. As the deep networks based object classifiers require prohibitive runtimes to process images for real world applications, knowledge distillation framework has been proposed for the modality hallucination architecture with improved accuracy.



Contents

Author's Declaration	i
Certificate	ii
Acknowledgements	iii
Abstract	iv
1 Introduction	1
1.1 Background	1
1.2 Problem Statement	3
1.3 Contributions	3
1.4 Organization of Dissertation Report	4
2 Related Work	5
2.1 Literature Review	5
2.2 Research Gap	8
3 Proposed Solution	9
3.1 Hallucination Network Architecture	9
3.2 Knowledge Distillation Architecture	10
4 Experiment	13
4.1 Dataset	13
4.2 Hallucination Network	13
4.3 Knowledge Distillation	15
5 Result Analysis	16
6 Conclusion and Future Work	19
Bibliography	20

List of Figures

1.1.1 shows architecture of Fast R-CNN [5]	2
1.1.2 shows architecture for supervision transfer [12]	3
2.1.1 shows modality hallucination architecture during testing [3]	6
2.1.2 Object Detection on the NYUD2 test dataset where top scoring detection of RGB hallucination network for the image(green box) is correct and the top scoring detection of baseline RGB detector(red box) is incorrect [3]	6
2.1.3 Overview of system generating region proposals and classifies them in to object categories [4]	7
2.1.4 shows Fast R-CNN architecture for object detection [5]	7
2.1.5 shows training of student network using hints [10]	8
3.1.1 shows proposed solution for object classification using hallucination network.	10
3.2.1 shows teacher-student set up for knowledge distillation framework.	11
5.0.1 shows performance of teacher model at different temperatures for single object.	17
5.0.2 shows performance of all the models.	18

List of Tables

4.1	Input/Output for each network.	14
5.1	Accuracy of models.	16
5.2	Loss of models.	17



Chapter 1

Introduction

In object classification and recognition tasks, exploiting multiple sources of information can improve the performance significantly. The techniques assume that all the modalities (features) such as different image features or multiple modalities such as images and text are available during training and inference. These techniques are not suitable in dynamic scenarios like robotics applications when new modalities, for which no training data is available, are added during inference.

1.1 Background

Previously unseen features are leveraged under the assumption that, given the existing modalities, conditional distribution of new modalities is stationary. The unlabeled data learn a non-linear mapping from existing to novel modalities. This way missing modality is hallucinated from unlabeled data. This makes recognition system more effective in applications like personal robotics, where user cannot label set of examples when new sensor is added each time. There are many real world scenarios where unseen modalities are present during inference like unlabeled high resolution images are utilized to improve object detection for webcam, unlabeled color images are exploited for grayscale object detection, unlabeled text is used to improve visual classification[1].

For object recognition tasks, deep convolutional neural networks have made tremendous success. CNNs have fewer parameters and connections. They are easier to train and have large learning capacity[2].

Region Based Convolutional Neural network(R-CNN) is slow as it uses selective search first to extract the regions. For each extracted region, CNN is used to extract features. On the contrary, Fast Region based Convolutional neural network(Fast R-CNN) passes the image to ConvNet which generates Regions of Interest(ROIs)[5].

Performance of a machine learning algorithm can be improved by training different models on same dataset and then finding the average of their predictions. The predictions generated by ensemble of models are complicated and time consuming. The performance of

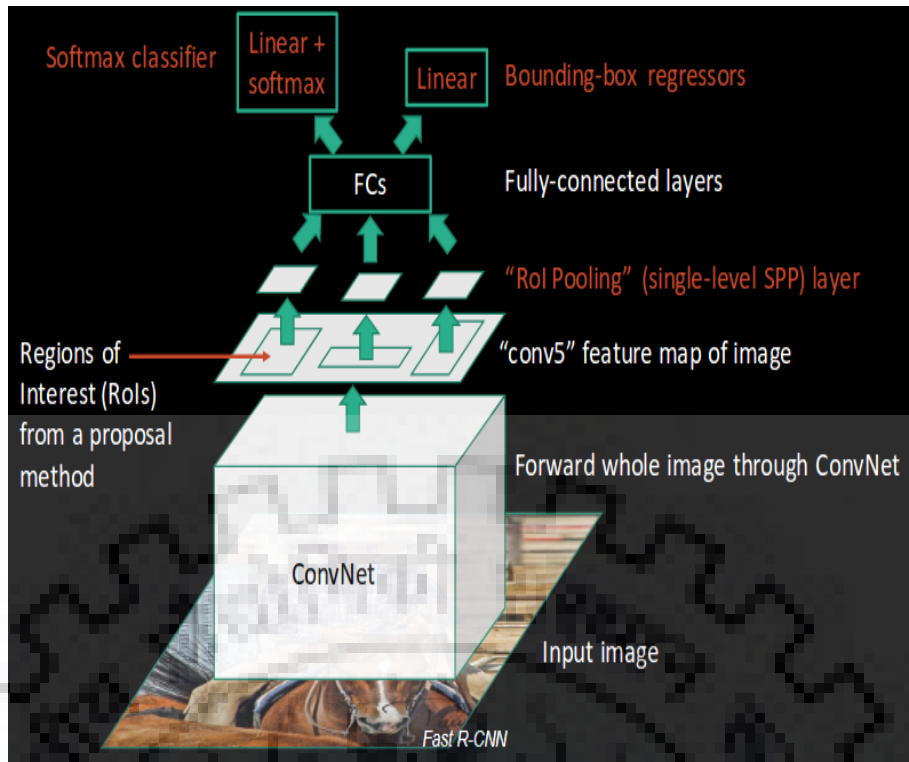


Figure 1.1.1: shows architecture of Fast R-CNN [5]

heavy models used in commercial system can be improved by distilling the knowledge of ensemble of heavy models to a single light weight model[7].

Deeper networks have high capacity and perform better under proper training. Compressed models are used to speed-up the complex models by decomposing the weights in each layer followed by fine-tuning or layer wise reconstruction to recover the accuracies. Knowledge distillation for multi-class object detection is challenging. For classification, knowledge distillation assumes that each class is equally important whereas for detection background class is far more prevalent. Detection is a complex task as it contains elements of both bounding box regression and classification[8].

For deep and wide networks, major drawback is that the result during inference is very time consuming. Also large memory is required as deep networks have more parameters. To reduce the computational burden at inference time, a novel approach has been proposed to train deep and thin networks, called FitNets, to compress to wide and shallower (but still deep) networks. The intermediate-level hints from the hidden layers of teacher are used to train the student so that the student network, also known as FitNet, learns an intermediate representation which is predictive of the intermediate representations of the teacher network[10].

The proposed method is used to transfer learned representations from one modality to another. The method uses paired images from the two modalities and utilizes the mid-level representations from the

labeled modality to supervise learning representations on the paired unlabeled modality. The method learns useful feature hierarchies in the unlabeled modality, which can be further improved with fine-tuning, and are still complementary to representations in the source modality [12].

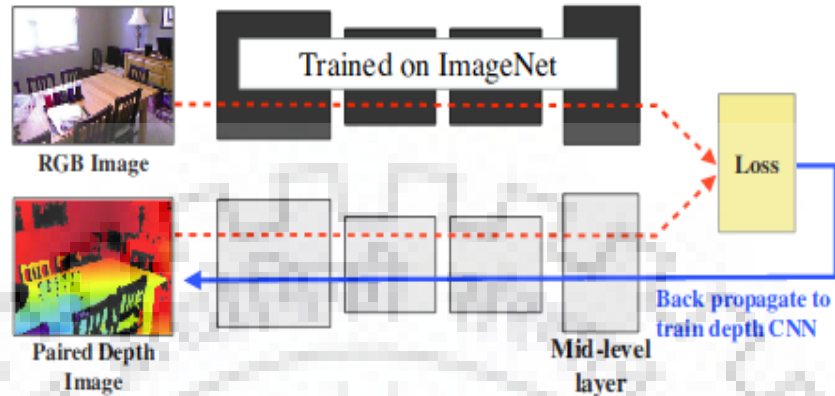


Figure 1.1.2: shows architecture for supervision transfer [12]

1.2 Problem Statement

- 1) Given a training set of RGB-D image pairs, the aim is to exploit multiple sources of information efficiently for object classification when single modality (RGB image) is present during inference.
- 2) For the above developed model, build a light weight model with improved accuracy so that the real time applications can process the images efficiently.
- 3) Further compare the performance of above distilled model and standalone model, with same structure as that of distilled model, trained on same dataset for object classification.

1.3 Contributions

The main contributions of this dissertation are:

- To develop an efficient object classification model by exploiting multiple sources of information.
- The model should be capable of identifying even small objects in absence of one or more modalities during inference.
- To develop the light weight model for above cumbersome model so that it require less computational runtimes for processing real time images efficiently with more accuracy.

- To compare the performance of complex model and the distilled model on the same test set for object classification.
- To compare the performance of distilled model and the standalone model, with same structure as that of distilled model, trained on same dataset for object classification.

1.4 Organization of Dissertation Report

Rest of the dissertation is organized as :

In Chapter 2, the related work done by other authors and their contributions have been discussed for object recognition and for distilling the knowledge of complex model. Also, the research gap, based on which we proposed the model to overcome these research challenges, has been presented.

In Chapter 3, the solution has been proposed to exploit multiple sources of information for object classification using the concept of hallucination of missing modalities. Further, knowledge distillation framework has been proposed for the hallucination model to decrease the runtime burden during inference.

In Chapter 4, the summary of hallucination network and knowledge distillation network and the dataset used for experiment has been presented.

In Chapter 5, the results have been based on the experiment performed which compares the accuracy of all the models for object classification.

In Chapter 6, we concluded the dissertation report and the future work.

Chapter 2

Related Work

Many techniques have been put forward to exploit multiple sources of information for performing object classification or detection. Mostly these techniques presume that, both during training and inference, all the modalities are present.

2.1 Literature Review

Probabilistic multiple kernel learning framework has been proposed to make use of multiple sources of information. Gaussian Processes (GPs) are used to learn mapping between existing and missing modalities. Missing modalities are used simultaneously with the old modalities in the proposed framework under the assumption that conditional distribution of new modalities is same at training and testing. For a new example, class label is assigned by finding the mean prediction of GP[1].

The modality hallucination architecture has been proposed to train RGB-D object recognition model which incorporates depth information during training. Hallucination network is trained to mimic depth level features and thus learns a new RGB representation. 3-channel model has been built using Fast R-CNN. Depth and RGB networks are independently trained using corresponding images. Depth network weights are then used to initialize hallucination network parameters followed by joint training of three networks.

At test time, RGB image is given as input and the final score is the softmax of the average of the predictions of two networks[3].

RGB-D object detection model has been put forward which uses R-CNN trained on RGB and depth image pair. The system detects contours and finds category for generated region proposals using Depth-CNN and RGB-CNN feature extraction networks[4].

Fast Region Based Convolutional Neural Network (Fast R-CNN) has been proposed for object recognition to increase the accuracy, training and testing speed. Fast R-CNN efficiently classifies object proposals and has higher detection quality. Training Fast R-CNN is a single stage and can update all the the layers. The entire image is passed to the ConvNet which generates RoIs and feature maps. The

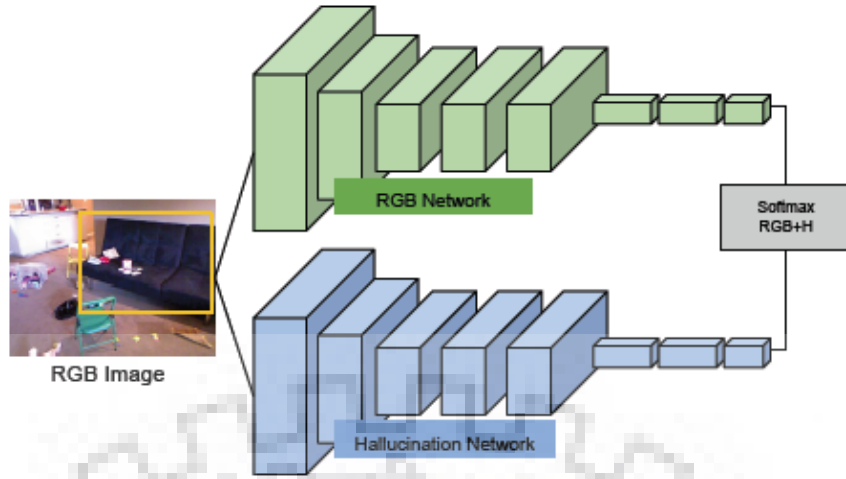


Figure 2.1.1: shows modality hallucination architecture during testing [3]

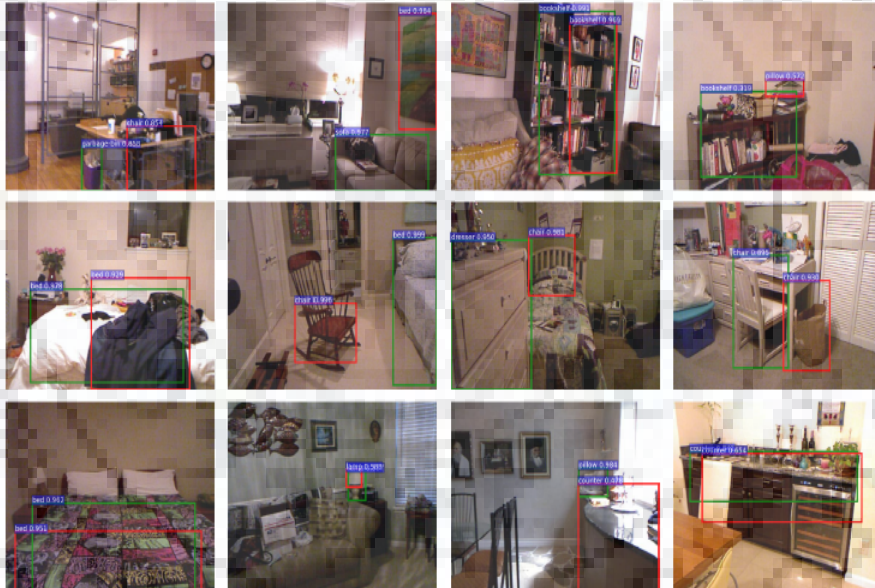


Figure 2.1.2: Object Detection on the NYUD2 test dataset where top scoring detection of RGB hallucination network for the image (green box) is correct and the top scoring detection of baseline RGB detector (red box) is incorrect [3]

model extracts features for each region, returns the bounding boxes and classifies each region. The architecture is trained end-to-end with a multi-task loss[5].

Knowledge distillation is used to transfer generalizations of complex models to a lighter model. A deep neural network is trained on large dataset with good regularization to increase the generalization ability for unseen data. A soft target distribution is produced by complex network at some high temperature. The distilled or lighter

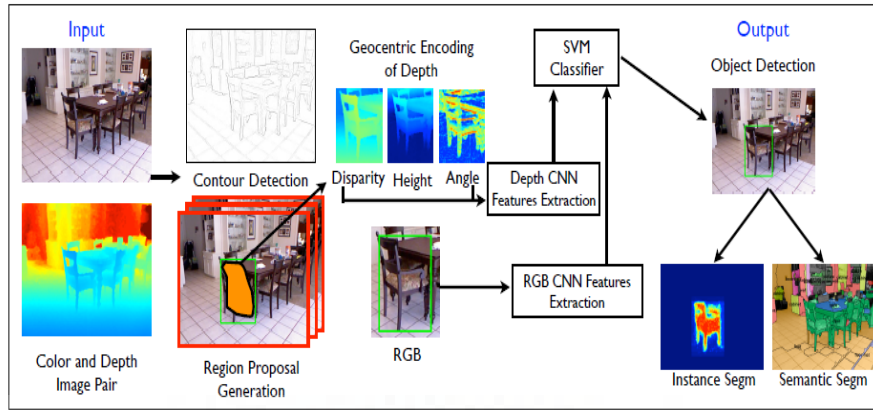


Figure 2.1.3: Overview of system generating region proposals and classifies them in to object categories [4]

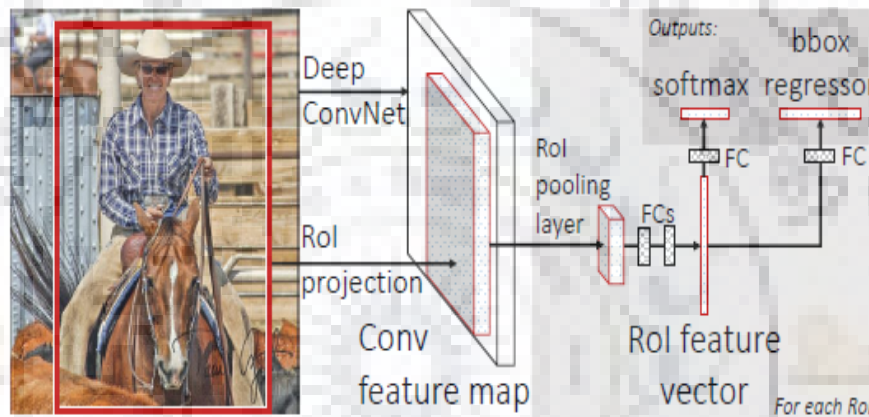


Figure 2.1.4: shows Fast R-CNN architecture for object detection [5]

model is trained at same temperature. The first objective is to compute cross entropy with soft target using softmax of lighter model at same temperature. The second objective is to calculate the cross entropy of correct labels at temperature $T = 1$, using logits in the softmax of distilled light weight model[7].

The framework has been proposed using knowledge distillation to learn Fast R-CNN based object detectors. The proposed loss function and hint-based learning helps in improving the performance. The compact models trained with the proposed framework execute significantly faster than complex network[8].

The student network captures the information provided by the true labels and also the finer structure learned by the teacher network. A hint is defined as the output of a teacher's hidden layer responsible for guiding the student's learning process[10].

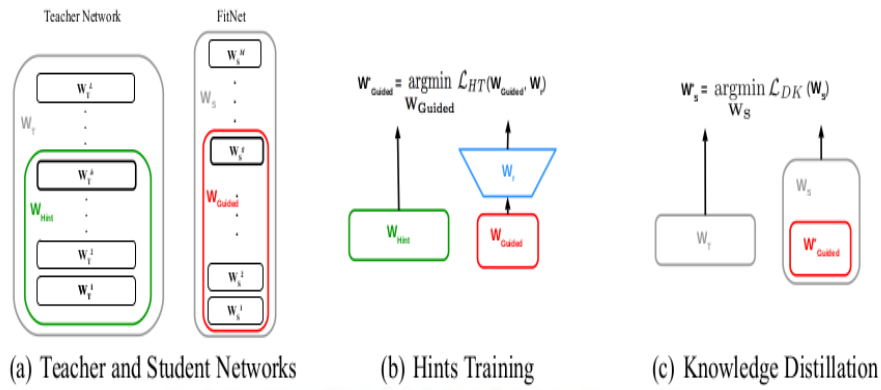


Figure 2.1.5: shows training of student network using hints [10]

2.2 Research Gap

The problem of exploiting multiple sources of information for object classification or recognition tasks when, additional modalities that are present in the labeled training set, are not available for inference needs to be tackled efficiently so that real time applications require less runtimes. This has been addressed by presenting a multi-modal CNN hallucination network. The efficiency of the object detection model can be improved using Faster R-CNN as it uses Region Proposal Network(RPN) instead of selective search for finding RoIs. Further Knowledge Distillation framework has been proposed to learn efficient object classification or detection models. As the above cumbersome model requires extensive runtime to process image for real time applications, Knowledge Distillation can be used with improved accuracy.

Chapter 3

Proposed Solution

The proposed solution is divided in to two parts:

- 1) For RGB-D object clasification, four networks comprising of multi-layer ConvNets have been used. To improve the classification performance when single modality is present during inference, hallucination network is build to incorporate the mid level depth features.
- 2) Knowledge Distillation framework is used to transfer the generalization ability of the above developed cumbersome model to distilled model using the weighted cross entropy loss of correct labels and the soft targets.

3.1 Hallucination Network Architecture

RGB-D object classification architecture consists of four channels, each made up of multi-layer ConvNets, as shown in the figure. RGB and depth imges are processed independently using RGB and depth network. The hallucination network takes RGB modality as input and learns corresponding depth features using depth network as depth modality is not present during inference. Through hallucination network, the depth modality shares information with the RGB modality . To achieve this, regression loss between paired depth and hallucination layers has been used. After some layer l , hallucination network needs to have similar features as depth network. So, hallucination loss encourages the network to mimic mid level depth features and is defined as below for any layer:

$$L_{halluciante}(l) = \|\sigma(A_l^{dNet}) - \sigma(A_l^{hnet})\|_2^2 \quad (3.1)$$

Initializing the hallucination network with depth parameters give the highest performance. The significant performance was achieved with hallucination loss on atleast middle activations [2].

The fourth network has been trained for classification of RGB-D objects using intermediate RGB features produced by RGB network and corresponding depth features predicted by hallucination network. During testing RGB image is given as input and the network predicts the class label for each object.

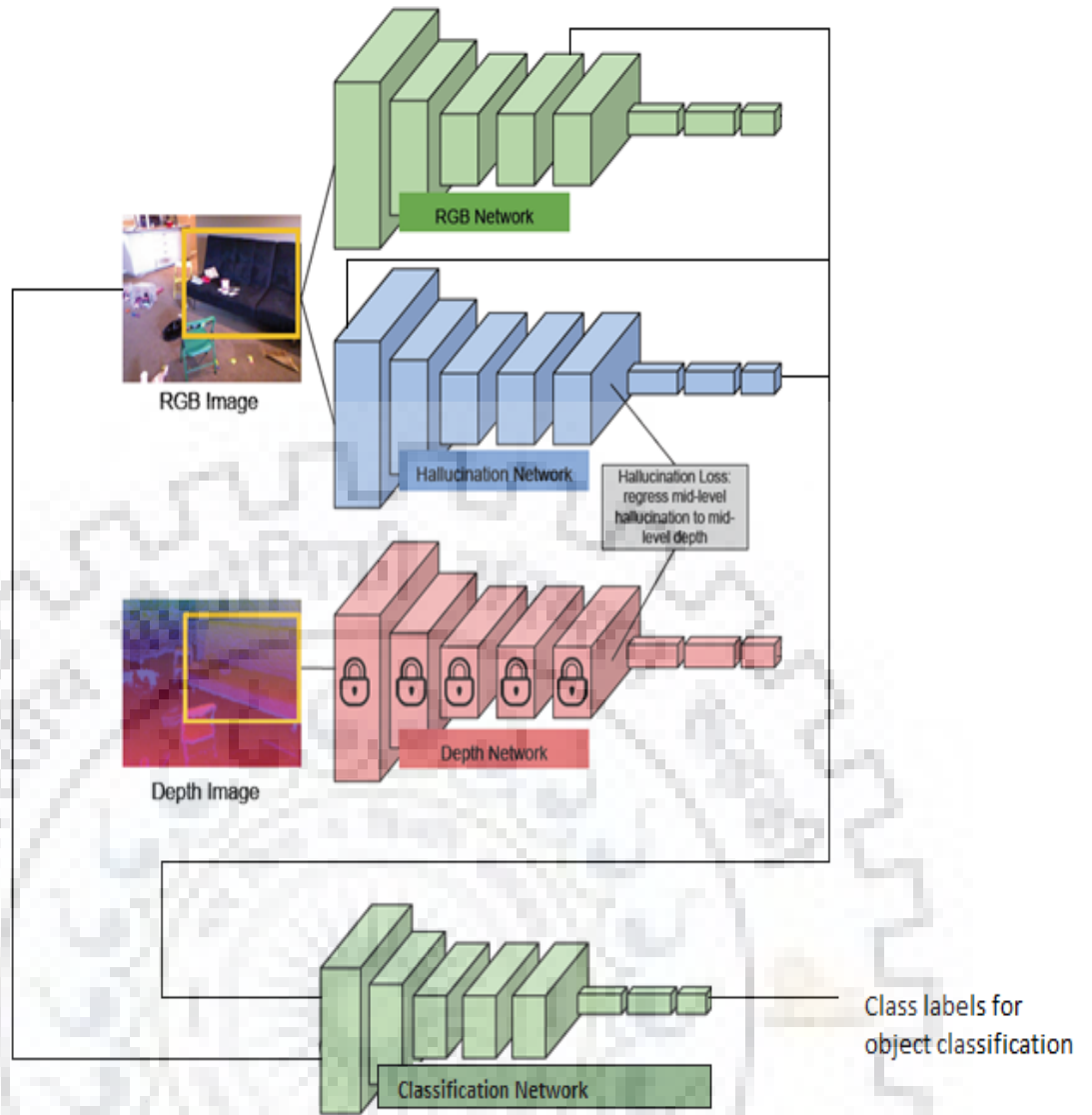


Figure 3.1.1: shows proposed solution for object classification using hallucination network.

3.2 Knowledge Distillation Architecture

The modality hallucination based RGB object classification model is cumbersome and requires computationally extensive time for running real applications. To overcome this limitation, light-weight distilled model can be built using knowledge distillation to generalize the ability of complex model. By using the knowledge of high capacity complex network, knowledge distillation learns efficient compact object classification network.

In teacher-student model, teacher is a deep and complex neural network trained on huge data with good regularization so that unseen data can be generalized well. A student network is light weight model trained by teacher model with the objective of learning most of the generalizations of teacher model. For quick predictions, the lighter or distilled model is preferred in productions with stringent production constraints.

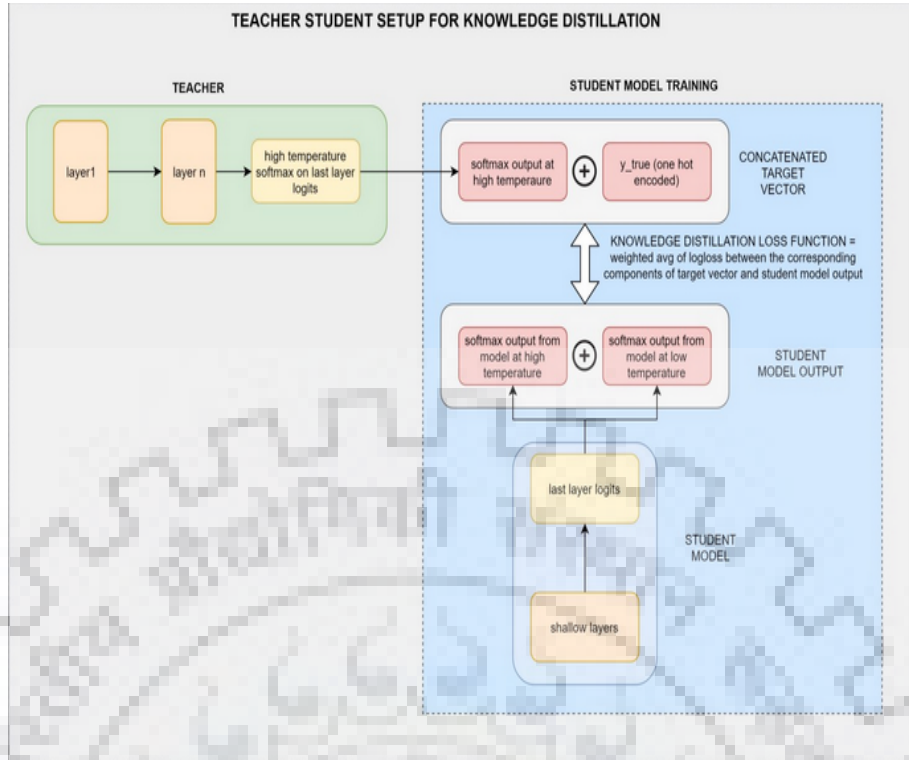


Figure 3.2.1: shows teacher-student set up for knowledge distillation framework.

Neural network uses softmax output layer to produce class probabilities for classification. The layer converts the logit, z_i , computed for each class into a probability, q_i , by comparing z_i with the other logits usually at temperature $T=1$.

$$q_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)} \quad (3.2)$$

In elementary form of distillation, knowledge is transferred to distilled model from complex model by training distilled model on transfer set and using soft target distribution in the transfer set that is produced by complex model at a high temperature in its softmax. While training, distilled model uses same high temperature and once it is trained, it uses $T=1$. When the labels for some or all of the transfer sets are known, the elementary form of distillation method can be made efficient by training distilled model to generate correct labels. To achieve this, weighted average of two objective functions can be used. The first is the cross entropy with soft targets. The high temperature used for producing soft targets from the complex model is used to determine the cross entropy. The second is the cross entropy with correct labels. This is determined using logits in softmax of distilled model at $T = 1$. The results are better when lower weight on second objective function is used. To ensure that relative contributions of soft and hard targets remain unchanged irrespective of the changes in temperature, the magnitudes of gradients produced

by soft targets which scale as $1/T^2$, are multiplied by T^2 .

With respect to each logit z_i of distilled model, each case in the transfer set contributes a cross-entropy gradient dC/dz_i . If the complex model has logits v_i which generate soft target probabilities p_i and transfer training is done at temperature T , then the gradient is as follows:

$$\frac{dC}{dz_i} = \frac{1}{T}(q_i - p_i) = \frac{1}{T} \left(\frac{e^{z_i/T}}{\sum_j e^{z_j/T}} - \frac{e^{v_i/T}}{\sum_j e^{v_j/T}} \right) \quad (3.3)$$

In comparison to the magnitude of logits, if temperature is high, then gradient can be approximated as follows:

$$\frac{dC}{dz_i} = \frac{1}{T} \left(\frac{1 + z_i/T}{N + \sum_j z_j/T} - \frac{1 + v_i/T}{N + \sum_j v_j/T} \right) \quad (3.4)$$

Further assuming that the logits have been zero-meaned separately for each transfer case so that

$$\sum_j z_j = \sum_j v_j = 0, \quad (3.5)$$

the gradient simplifies to:

$$\frac{dC}{dz_i} = \frac{1}{NT^2}(z_i - v_i) \quad (3.6)$$

Distillation, at lower temperatures, pays less attention to logits which are much more negative than the average. This is advantageous because these logits are completely unconstrained by cost function used for training the complex model so that they can be very noisy. On the other hand, the negative logits can convey useful information about the knowledge acquired by the complex model. So, intermediate temperatures give best performance suggesting that it can be beneficial to ignore very large negative logits.

Chapter 4

Experiment

The dataset and the architectures used for hallucination network and knowledge distillation have been discussed in the following sections.

4.1 Dataset

The developed model has been evaluated on standard RGB-D Kinect object dataset. The dataset consists of common household objects which have been organized in to different categories. For each object, video sequences were captured for one rotation. For experiment purpose, 10 categories of objects were taken. Each category contains around 1000 RGB and depth images. The dataset used is split into 10182 training images, 2100 validation images and 2000 testing images. The model has been trained using training set and is evaluated on validation set. The results have been presented on the test set.

4.2 Hallucination Network

For experiment, base network architecture for each of the four networks consists of convolutional, pooling, dropout, RELU and fully connected layers.

First network (RGB network) is a CNN classifier with RGB images as input. The network consists of 2 convolutional layers with RELU activation, 1 max pooling layer, 2 dropout layers, 1 RELU layer, 1 fully connected layer followed by a softmax layer. The loss used is categorical cross entropy with stochastic gradient descent optimizer.

Second network (Depth network) is a CNN classifier with depth images as input. The network consists of 2 convolutional layers with RELU activation, 1 max pooling layer, 2 dropout layers, 1 RELU layer, 1 fully connected layer followed by a softmax layer. The loss used is categorical cross entropy with stochastic gradient descent optimizer.

The third network (Hallucination network) is a regression model which hallucinates the mid level features of depth network. The model takes RGB images and the corresponding intermediate RGB features produced by RGB network as inputs and predicts the cor-

Table 4.1: Input/Output for each network.

Network	Model	Input	Output
RGB	CNN classifier	RGB images	Class label
Depth	CNN classifier	Depth images	Class label
Hallucination	CNN regressor	RGB images, Corresponding Intermediate RGB features produced by RGB network	Depth features
Fourth	CNN classifier	RGB images, Corresponding Intermediate RGB features produced by RGB network, Corresponding depth features predicted by hallucination network	Class label

responding intermediate depth features. The network consists of 1 convolutional layer with RELU activation, 1 max pooling layer, 2 dropout layers, 2 RELU layers, 1 fully connected layer followed by a linear activation layer. The loss used is mean squared error with adam optimizer. The learning of hallucination network is guided by the hallucination loss which is the euclidean distance between the activations of depth network and hallucination network. Any layer of the depth network can be hallucinated but the performance achieved with middle layer was highest and uniformly distributed across all categories.

The fourth network is a CNN classifier which takes following as inputs:

- 1) RGB images.
- 2) Corresponding intermediate RGB features produced by RGB network.
- 3) Corresponding depth features predicted by hallucination network.

The fourth network consists of 1 convolutional layer with RELU activation, 1 max pooling layer, 1 dropout layer, 1 RELU layer, 1 fully connected layer followed by a softmax layer. The loss used is categorical cross entropy with stochastic gradient descent optimizer. The network thus classifies the RGB images into different categories.

4.3 Knowledge Distillation

The experiment further consists of building a light weight model (student model) for the above complex model (teacher model). The generalizations of teacher model is transferred to student model. While training student model, softened probabilities, which are the outputs collected by applying high temperature softmax are taken as target instead of making one hot encoding as hard targets. The student architecture is a light weight CNN classifier made up of fully connected and softmax layers. The distillation loss used is the weighted average of logloss between the corresponding components of concatenated vector of soft targets + hard targets.

The standalone model is a CNN classifier made up of fully connected and softmax layers same as that of student model. The standalone model is trained and evaluated on same dataset.



Chapter 5

Result Analysis

The results obtained by evaluating the proposed solution on standard RGB-D Kinect object dataset have been shown in the table below. RGB model has comparatively higher accuracy than depth model. The loss achieved by hallucination model is less and thus outperforms baseline RGB model for object classification.

The student model has better performance than teacher model and requires less computational time. Hence, the student model can be used in stringent production environment with better accuracy for object classification.

The performance of student model is much better than standalone model, with same structure as that of student model, trained on same dataset for object classification. This shows the knowledge transferred from teacher is better than training distilled network on same dataset. As the depth of the teacher network increases, the student becomes more informative. The student model can thus be used in scenarios where training data is less for distilled model.

Table 5.1: Accuracy of models.

Model	Training Accuracy (%)	Validation Accuracy (%)
RGB	97	95
Depth	87	96
Teacher	92	95
Student	97	96
Standalone	74	77

The probability distribution for all classes at different temperatures for the teacher model has been shown in the figure. The soft predictions of the model is good at high temperature.

Table 5.2: Loss of models.

Model	Loss Function	Training Loss (%)	Validation Loss (%)
RGB	Categorical cross-entropy	0.1120	0.0260
Depth	Categorical cross-entropy	0.3527	0.1952
Hallucination	Mean Squared Error	0.00032	0.00071
Teacher	Categorical cross-entropy	0.5416	0.1057
Student	Categorical cross-entropy	0.2185	0.2206

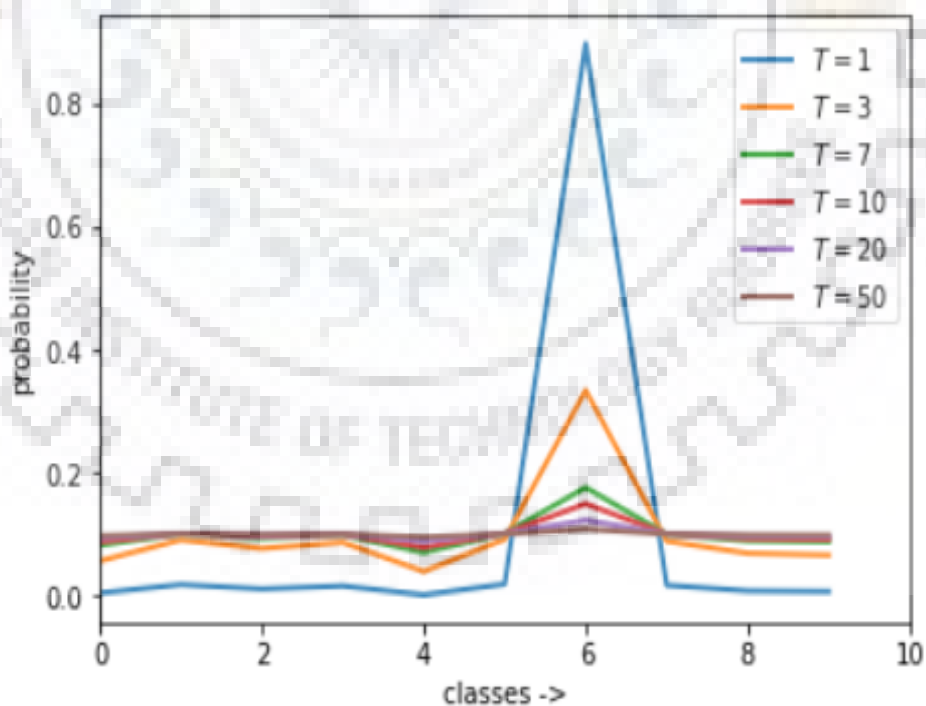


Figure 5.0.1: shows performance of teacher model at different temperatures for single object.

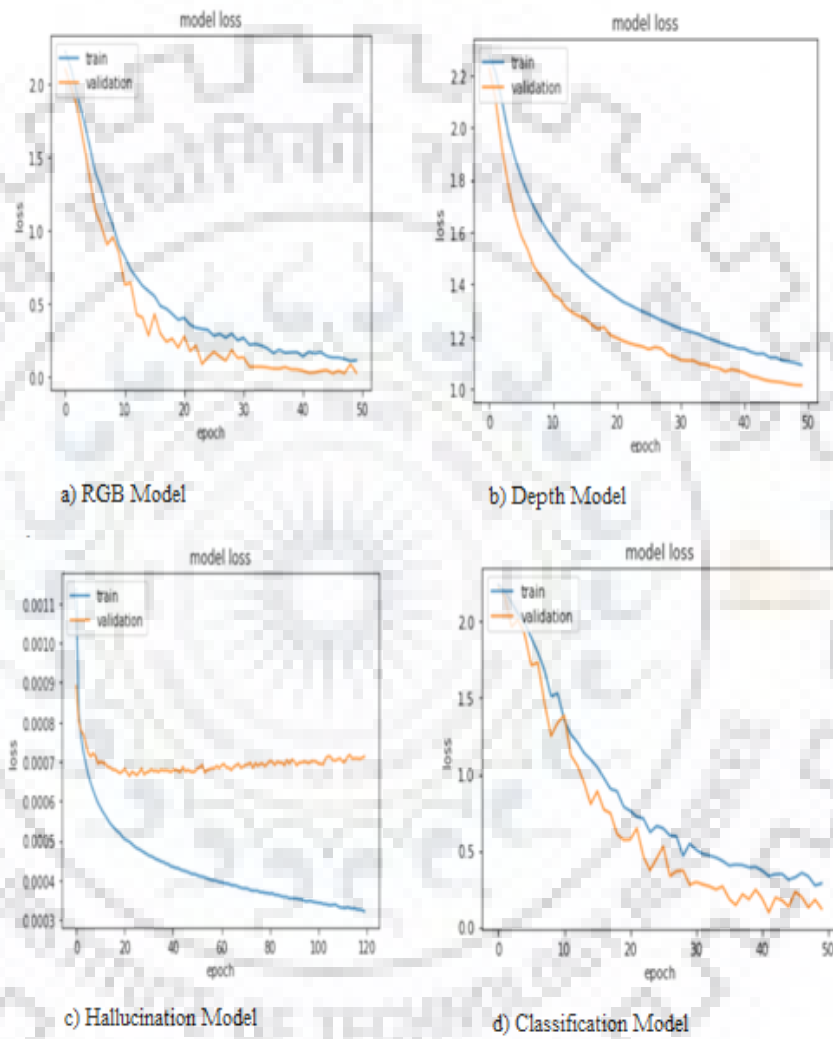
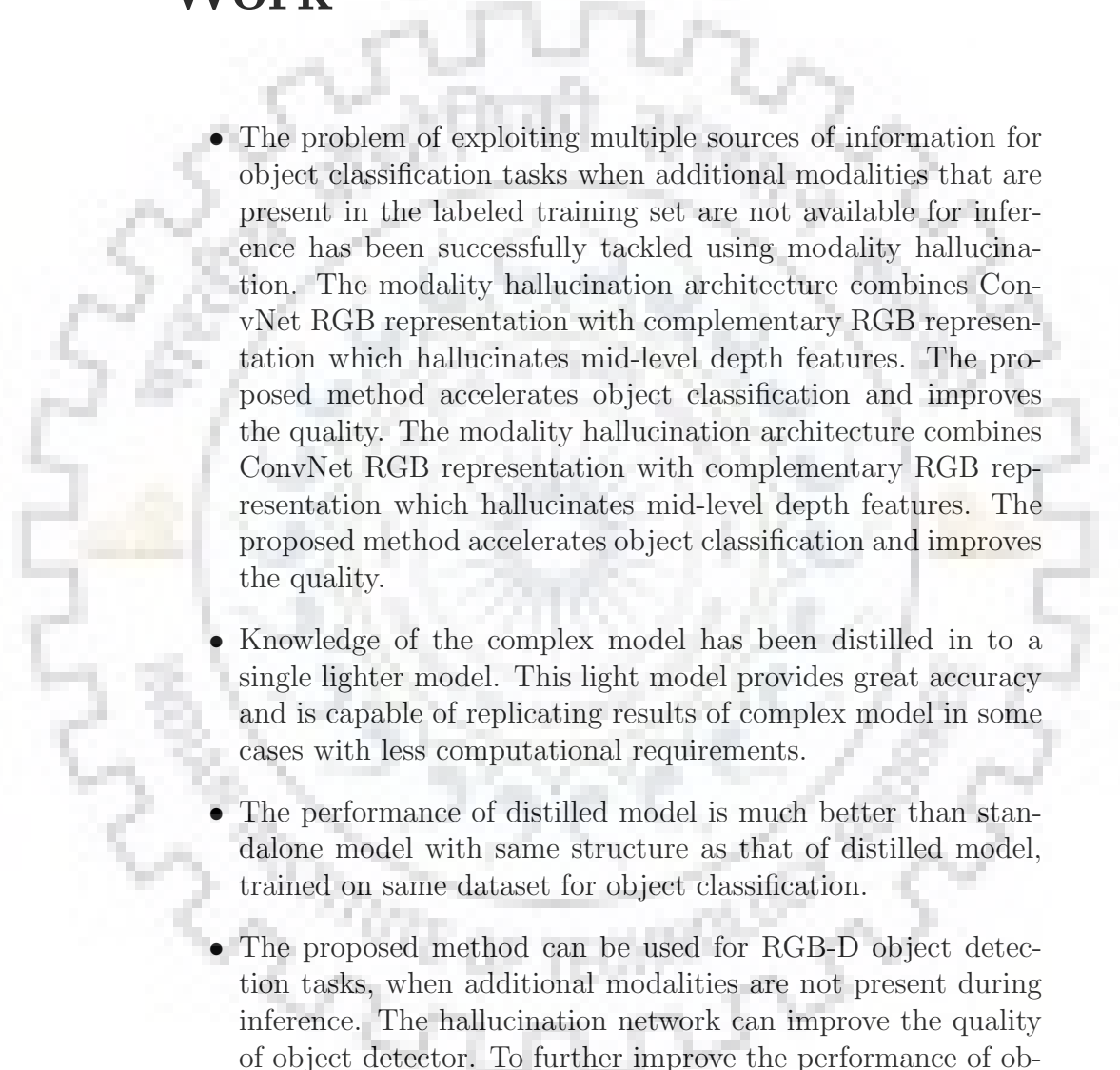


Figure 5.0.2: shows performance of all the models.

Chapter 6

Conclusion and Future Work

- 
- The problem of exploiting multiple sources of information for object classification tasks when additional modalities that are present in the labeled training set are not available for inference has been successfully tackled using modality hallucination. The modality hallucination architecture combines ConvNet RGB representation with complementary RGB representation which hallucinates mid-level depth features. The proposed method accelerates object classification and improves the quality. The modality hallucination architecture combines ConvNet RGB representation with complementary RGB representation which hallucinates mid-level depth features. The proposed method accelerates object classification and improves the quality.
 - Knowledge of the complex model has been distilled in to a single lighter model. This light model provides great accuracy and is capable of replicating results of complex model in some cases with less computational requirements.
 - The performance of distilled model is much better than standalone model with same structure as that of distilled model, trained on same dataset for object classification.
 - The proposed method can be used for RGB-D object detection tasks, when additional modalities are not present during inference. The hallucination network can improve the quality of object detector. To further improve the performance of object detector, Faster R-CNN can be used. Instead of selective search, Faster R-CNN uses Region Proposal Network(RPN) which takes feature maps, generated using ConvNets as inputs.
 - An application using GAN(Generative Adversarial Network) can be build for hallucinating missing modalities to exploit multiple sources of information. Further the accuracy and efficiency of the proposed distilled network and GAN can be compared.

Bibliography

- [1] C. M. Christoudias, R. Urtasun, M. Salzmann, and T. Darrell, “Learning to recognize objects from unseen modalities,” in *European Conference on Computer Vision*, pp. 677–691, Springer, 2010.
- [2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, pp. 1097–1105, 2012.
- [3] J. Hoffman, S. Gupta, and T. Darrell, “Learning with side information through modality hallucination,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 826–834, 2016.
- [4] S. Gupta, R. Girshick, P. Arbel aez, and J. Malik, “Learning rich features from rgb-d images for object detection and segmentation,” in *European Conference on Computer Vision*, pp. 345–360, Springer, 2014.
- [5] R. Girshick, “Fast r-cnn,” in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.
- [6] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in *Advances in neural information processing systems*, pp. 91–99, 2015.
- [7] H. Geoffrey, V. Oriol, and D. Jeff, “Distilling the knowledge in a neural network,” in *Proc. NIPS workshop*, 2014.
- [8] G. Chen, W. Choi, X. Yu, T. Han, and M. Chandraker, “Learning efficient object detection models with knowledge distillation,” in *Advances in Neural Information Processing Systems*, pp. 742–751, 2017.
- [9] W. Choi, M. Chandraker, G. Chen, and X. Yu, “Learning efficient object detection models with knowledge distillation,” Sept. 20 2018. US Patent App. 15/908,870.
- [10] R. Adriana, B. Nicolas, K. S. Ebrahimi, C. Antoine, G. Carlo, and B. Yoshua, “Fitnets: Hints for thin deep nets,” *Proc. ICLR*, 2015.
- [11] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, “Fitnets: Hints for thin deep nets,” *arXiv preprint arXiv:1412.6550*, 2014.

- [12] S. Gupta, J. Hoffman, and J. Malik, “Cross modal distillation for supervision transfer,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2827–2836, 2016.
- [13] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [14] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.
- [15] K.-H. Kim, S. Hong, B. Roh, Y. Cheon, and M. Park, “Pvanet: Deep but lightweight neural networks for real-time object detection,” *arXiv preprint arXiv:1608.08021*, 2016.

