

Word spotting

Project Report

Submitted by

Pratham Nagpure
(112001054)

and

Priyanshu Gupta
(112001033)

under the guidance of

Dr. Narayanan C Krishnan



INDIAN INSTITUTE
OF TECHNOLOGY
PALAKKAD

Contents

List of Figures	3
List of Tables	4
1 Introduction	6
2 Literature Review	7
3 Methodology	9
3.1 Large model (teacher model).....	9
3.2 Small model (student model).....	10
4 Results and Discussions	11
5 Conclusion and Future Work	12
References	13

List of Figures

1.1 Teacher, student models for knowledge distillation.....	7
1.2 Different types of knowledge in teacher model.....	8
4.1 Example image and ground truth.....	11

List of Tables

3.1 CNN architecture of teacher model..... 9

3.2 CNN architecture of student model.....10

4.1 Results of teacher model..... 11

4.2 Results of student model..... 11

Abstract

In this paper we develop a large artificial neural network model and attempt to make a lightweight model from that larger model for word spotting.

Chapter 1

Introduction

Getting the word labels from the images of handwritten texts is a challenging task. A model with artificial neural network requires many layers which have large amount of parameters. Large neural nets are difficult to train as they have a high amount of features and also difficult to deploy on end devices, if the size is reduced then accuracy is lost. To get a model of small size as well as good accuracy we need some model compression method. One such method is knowledge distillation.

We develop a large model which is an artificial neural network which performs well and then we will create a student model and train it using the method of knowledge distillation and try to get the accuracy like the teacher model.

Chapter 2

Literature Review

For compression of the model various methods are used such as parameter compression, knowledge distillation. Knowledge distillation is the process of moving information from a large, cumbersome model or set of models to a single, more manageable model that may be used in real-world applications as we can see in the figure 1.1 teacher model has many parameters and the student model has less parameters. Knowledge distillation is frequently used with neural networks that have complicated architecture with numerous layers and parameters. As for speech recognition, image recognition, and natural language processing deep learning is used so knowledge distillation is also used in real life purposes. As edge devices have less memory and computational capacity it is difficult to use deep neural networks on them. To handle this situation model compression should be performed and a method was proposed to transfer the knowledge of a large training model to a smaller model without serious loss of performance.

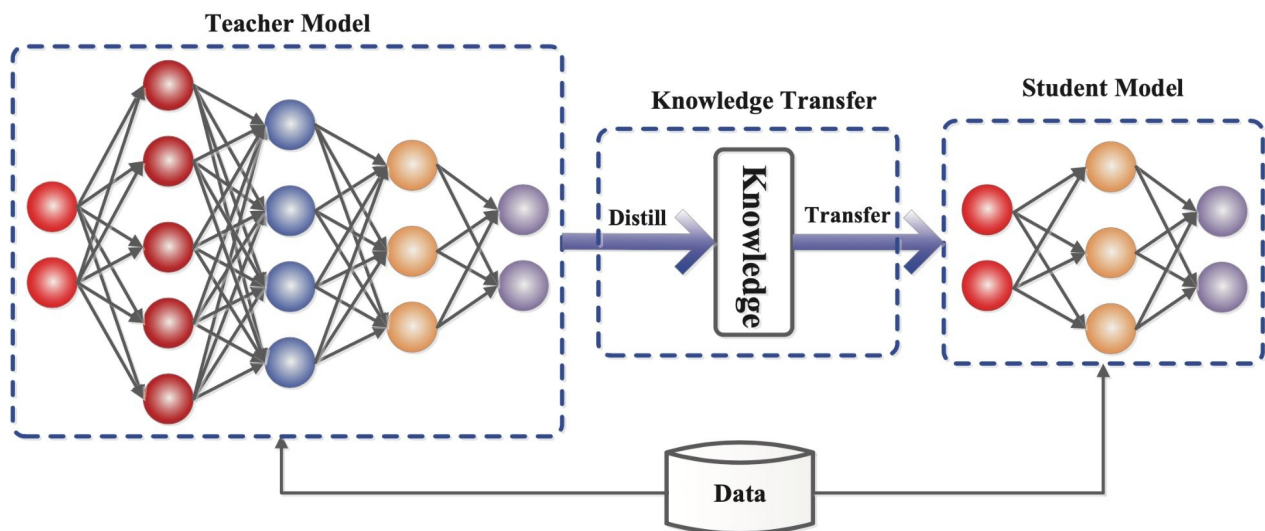


Figure 1.1 Teacher, student models for knowledge distillation | Source [1]

Knowledge in a neural network usually refers to the weights and biases that have been learnt. In addition, a huge deep neural network has a wide variety of information sources. While some knowledge distillations concentrate on the weights or activations of intermediary layers, the logits are often used as the source of teacher information. The connection between various neuronal activations and activation types, as well as the model's parameters, are other essential pieces of information. There are three types of knowledge response-based, feature-based relationship-based knowledge we can visualize them in figure 1.2

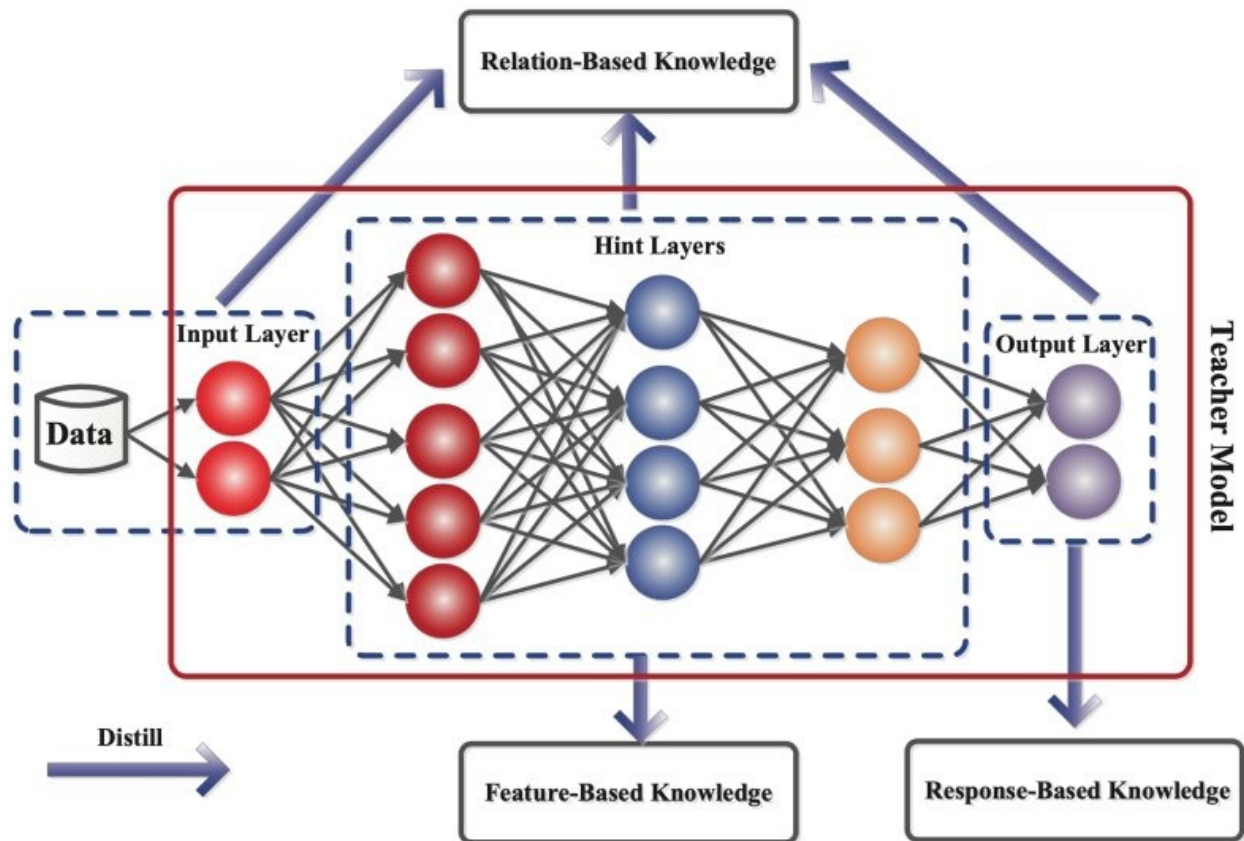


Figure 1.2 Different types of knowledge in teacher model | Source [1]

Response based knowledge: It focuses on only the output layer of the teacher model with the hypothesis that the student will learn to mimic the teacher and the loss is calculated as distillation loss.

Feature based knowledge: It also captures the data in the intermediate layers and this knowledge is used to train the student model which is incorporated in distillation loss

Relation based knowledge: It uses the relationship between the feature maps, similarity matrix and probabilistic distributions

We use Response based knowledge in our work

Chapter 3

Methodology

The first objective is to develop a teacher model which is a large model which gives good performance

3.1 Large model (Teacher model)

The teacher model is based on the model in [2], It consists of three components, first is a convolutional neural network (CNN), CNN is a class of artificial neural network (ANN) which is used for visual image analysis. Second is recurrent neural network (RNN) which is also a class of ANN which is commonly used for tasks such as handwriting recognition. The RNN consists of 2 Bi-directional Stacked long short-term memory (LSTM) cells. Finally there is connectionist temporal classification (CTC). The output from the RNN is fed into the CTC. CTC helps us to train the neural network (NN) by giving us a loss called as CTC-loss and also helps in decoding the text.

CNN architecture is based on the CNN in [3] which consists of residual neural network (ResNet) which is also a type of ANN which contains skip connections and shortcuts to jump over some layers. Maxpool is used to reduce the size of the output from the convolutional layers.

The CNN architecture:

Count	Layer Type	Dimension	Filter count
1	conv	7x7	32
1	maxpool	2x2	-
2	resblock	3x3	64
1	maxpool	2x2	-
4	resblock	3x3	128
1	maxpool	2x2	-
4	resblock	3x3	256
1	Column maxpool	-	-
1	Conv	5	256

Table 3.1 CNN architecture of teacher model

The final output of the CNN is fed into the RNN.

3.2 Small model (Student model)

The student model also consists of the three components but they are reduced in size. The CNN architecture is a reduced in size version of the Pho(SC)Net architecture in [2]. Relu activation layer is used after CNN

The CNN architecture:

Count	Layer type	Dimension	Filter count
1	conv	3x3	8
1	maxpool	1x2	-
1	conv	3x3	8
1	maxpool	2x2	-
4	conv maxpool	3x3 1x2	16 -

Table 3.2 CNN architecture of student model

The neural network is trained over the CTC loss of the ground truth as well as the distillation loss. Here we use response based knowledge distillation so the distillation loss is concerned with the final output of the teacher network. The distillation loss is the CTC-loss between the prediction of the teacher network and the prediction of the student network so the final loss L is

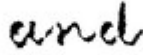
$$L = \text{CTCloss}(\text{ground truth}, \text{student prediction}) + \text{distillation loss}$$

$$\text{distillation loss} = \text{CTCloss}(\text{student prediction}, \text{teacher prediction})$$

As both the loss consists of CTC loss the weights of both the losses is 1 and none of them dominate each other. While training the student network tries to minimize this loss.

Chapter 4

Results and Discussions



ground truth: and

Figure 4.1 Example image and ground truth

Here we use the George Washington (GW) dataset. It is a collection of homogeneous word labels from the letters of George Washington and his affiliates. This data set is split in to 4 splits. Figure 4.1 shows an example of the dataset.

CER is the character error rate.

It is the sum of [(edit distance between predicted text and actual text)/length of actual text] over all data points

Splits	Accuracy (seen)	CER (seen)
Split 1	93.31	2.22
Split 3	93.83	2.55
Average	93.57	2.385

Table 4.1 Results of teacher model

Table 4.1 shows the accuracy of the Teacher model on the splits and from that we can infer that the teacher model is good on the datasets

Splits	Accuracy (seen)	CER (seen)
Split 1	73.73	12.8
Split 2	74.79	11.49
Split 3	58.62	23.29
Split 4	61.84	20.89
Average	67.245	17.1175

Table 4.2 Results of student model

Table 4.2 shows the accuracy of the student model on the splits. The average accuracy has 28% decrease in the student model than the teacher model. This shows that student model is weaker than the teacher model, in the distillation loss of the student model the final layer of teacher model was considered which is not sufficient to maintain the accuracy. Possibly to achieve higher accuracy the intermediate layers should also be included to calculate the distillation loss like in feature based knowledge distillation and also relational based knowledge distillation which might give different results.

Chapter 5

Conclusion and Future Work

Our teacher model which had the three components CNN, RNN, CTC and ResNet based CNN for teacher model has performed well over the dataset as seen in the results, the student model which is light weight than the teacher is not as high performing as the teacher.

We used response based knowledge distillation which included only the last layer of the teacher model to calculate the distillation loss which is not enough to achieve high accuracy so possibly incorporating other methods like feature and similarity based knowledge distillation could give different results.

References

- [1] Jianping Gou, Baosheng Yu , Stephen J. Maybank , Dacheng Tao "Knowledge Distillation: A Survey" pp 1-6
- [2] Ravi Bhatt "Word Recognition in Historical Documents" pp 7-12
- [3] George Retsinas, Giorgos Sfikas, Basilis Gatos & Christophoros Nikou "On-the-Fly Deformations for Keyword Spotting", DAS 2022 pp 338–351