NATIONAL UNIVERSITY OF
COMPUTER AND EMERGING
SCIENCES, KARACHI

FYP-II REPORT

De-correlating Ensembles of Neural Networks

Author:
Ammar Adeel
(16k-3919)
,Mohammad Samran
Elahi (16k-3819)

Supervisor:
Dr. Tahir Q. Syed

A report submitted in fulfillment of the requirements for the degree of BS Computer Science

Department of
Computer Science

June 11, 2020

Declaration of Authorship

I, Ammar Adeel (16k-3919) ,Mohammad Samran Elahi (16k-3819) , declare that this thesis titled, "De-correlating Ensembles of Neural Networks" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

Plagiarism Undertaking

We solemnly declare that research work presented in the FYP Project titled "De-correlating Ensembles of Neural Networks" is solely our research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

We understand the zero tolerance policy of the HEC and University "National University of Computer and Emerging Sciences, Karachi" towards plagiarism. Therefore, we as Authors of the above FYP Report declare that no portion of our report has been plagiarized and any material used as reference is properly referred/cited.

We undertake that if we are found guilty of any formal plagiarism in the Project Report even after award of Bachelors degree, the University reserves the rights to withdraw/revoke our Bachelors degree and that HEC and the University has the right to publish our name on the HEC/University Website on which names of students are placed who submitted plagiarized thesis.

Author signature:

Name:

Abstract

Ensembles are a bunch of weak classifiers stacked together , individually these classifiers are of little use but when used together they outperform strong classifiers. Our goal was to create diversity between classifiers present in the ensemble. We hoped that by creating model diversity between classifiers we could increase the overall accuracy of the ensemble. Confidence estimation and model Diversity were the focus of our project , we hoped that by achieving this we could significantly increase the performance of the ensemble which would in turn make an impact in many different applications.

Acknowledgements

I would like to express my profound gratitude to my supervisor, Dr Tahir Syed and Co-supervisor Sir Behraj Khan for there thorough and productive support, and for his important time to guide on me. There wide knowledge and logical way of thinking have been of great value for me. I would also like to extend my gratitude to the jury members, Sir Danish, and Dr Rafi for their constructive critiques and discussions; that had helped me to improve the shortcomings of this work. Finally I would like to thank my Bharam Baloch for helping me with the implementation of the experiments.

Author
Ammar Adeel (16k-3919) ,Mohammad Samran Elahi (16k-3819)

Contents

Chapter 1

Introduction

Ensembles are weak classifiers aggregated in a way to achieve a better performance than any of the individual model involved in the ensemble.It has been observed that weak classifier when combined together efficiently through different ensembling techniques can outperform a single strong classifier.There are many ensembling techniques such bootstrap aggregation also known as Bagging which involves having each model in the ensemble vote with equal weight. Another common ensembling techniques is Boosting in which, training for each new model to we emphasize on the training instances that previous models mis-classified[1].Ensembles are an attractive choice because they have no theoretical limit on its accuracy given the models are independent in the error generation[2].Our problem is focused on De-correlating ensembles of neural networks.We are working to make each neural network in an ensemble more independent from each other so that feature re-learning is reduced.Although literature on this problem exist [3] for de-correlating models in ensembles,and various other approaches including genetic algorithm search based approaches,data augmentations, incremental building of the ensemble and using structural diversity, have been used but the biggest disadvantage of such approaches is wastage of computational resources.We wish to contribute to the field of model independence and propose an automated approach for ensembles of neural networks to learn rich and distinctive features resulting in better performing systems.Our aim is to implement various diversity metrics and incorporate them in a regularizer which introduces model diversity and independence to the system in an efficient manner and acts as a barrier against redundant feature learning.

Chapter 2

Literature Review

There are various approaches which have been used to improve diversity in the individual models to construct efficient ensembles which involve overproducing individual models and selection,using genetic algorithms for searching of diverse classifiers,Incremental approaches, using different architectures,random weight initialization,training on different subsets of data,train data augmentation, adding diversity measures in loss functions for minimization and different activation functions in the individual models.

The ADDEMUP algorithm proposed by david and jude[3] is an example of the overproducing and selection method used for diverse models.The approach involves generating neural networks for an initial population and defines a fitness function which incorporates both accuracy and diversity of the models.A new generation is then generated using mutation and crossover operations.The new population is then trained on the examples,giving emphasis to the misclassified examples from the previous population.New members are selected to add in the population by evaluating them on the fitness function devised.The new population is then pruned to the N most fit neural networks.A similar approach is used by Zhi-Hua Zhou [4] called GASEN.GASEN uses bootstrapping to generate N training samples.N neural networks are trained on the samples and then random weights are given to the individual trained networks.These weights are then evolved such that they minimize the generalization error of the ensemble.The weight assigned to the individual neural network can be used as a fitness criteria to join the population.The disadvantage of these approaches is that the overproduce and prune the pool of networks.The pruned networks result in wasted computational resources and training time.Also the evolving of weights and bootstrapping adds an overhead of computation on the training.

An example of incremental approaches is by mazurowski [5] which generates a pool of N models,and selects incrementally from those models using an add-if-better metric,i.e it adds the under evaluation model to the ensemble if the addition of the model improves the overall accuracy of the ensemble by a certain margin.M models are selected incrementally using this approach and an appropriate fusion method is chosen to aggregate the results of the individual models.A similar approach is used by jingyang [6] which uses

the same architecture of neural networks but different initial weights are generated and trained individually.From this pool of trained networks only the networks which are diverse are selected.The publication includes a novel diversity measure based on the sensitivity of the networks.The trained networks are classified according to their sensitivity and one network is selected from each class.These are then trained on a subset of the training data to determine the combination weights.Lastly the ensemble is constructed using the networks and the combination weights.

An ensemble of convolutional neural networks is devised by Gianluca Maguolo [7] to improve the accuracy of the ensemble by training it on different activation functions. The technique uses same architectures such as VGG and Resnet and train the networks with different activation functions.Some of the activation functions used are MeLU,Leaky Relu, Srelu and ENS.The hyperparameters for the respective activation functions are also given.The publication also involves a novel activation function.

Another technique to diversify the individual models in an ensemble is apply different augmentations on the training data and training the models on the different augmentations. Hydra [8] for instance uses random shifts,zooms and flips on the images on every epoch. Wen Chen [9] also use similar augmentations but also use a preprocessing step which involves applying functions on patches of the images,where the size of the patch is a hyper parameter. Wen uses WRN-28-10,WRN-28-20 and resnet in the ensemble,while Hydra is a novel architecture in itself.

NCL or Negative Correlation Learning proposed by Liu[10] uses a penalty term in the loss function of every individual classifier.N base classifiers are trained independently, the loss is calculated with respect to accuracy of the individual learners.The penalty term consists of two parts, the first involves the aggregated result of the whole ensemble while the second part calculates the summation of the difference of the individual classifiers from the aggregated result. The penalty term is added to the loss of the individual classifiers with a scaling factor and the new weights are calculated according to the loss using backpropagation.The scaling factor is a hyperparameter which is used to change the strength of the penalty term which is in the range [0,1].The new weights are then updated for each learner.The above process is done repeatedly and lastly the ensemble of neural networks is produced. Regularized negative correlation learning is a similar approach used by Huanhuan Chen [11] which adds another penalty term to the negative correlation framework which involves the individual weights of the base networks in the penalty term, these regularization parameters are then optimized using Bayesian inference.

Chapter 3

FYP-I approaches

The method we adopted to solve the problem is the addition of a regularizer in the loss function, and optimizing it accordingly. Our approach involves using N number of neural networks with the same architectures where N is odd.Due to the logical outcome after the aggregation N is kept odd such that if N-1 even results are tied the Nth result will result in a tie breaker.The architectures are kept constant with respect to datasets to eliminate dependency with respect to the neural network architecture.Different weight initialization is used on the networks using the Xavier weight initialization.The approach was used such that the starting point of exploration of each model is different and the probability of finding a good minima is increased.Activation and the general loss functions excluding the regularization term is chosen appropriately according to the nature of the dataset and are kept constant.During training the number networks are constant and do not change.There are two methods used for training the networks, one is Parallel training where the outputs of the softmax for the N neural networks are averaged during training time and the combined loss is calculated using the averaged softmax output.Hence the backward gradients flow in parallel together in all the neural networks in backpropagation.The second method used was individual training where loss is calculated on all the neural networks individually.

3.0.1 Pipeline and Results

We built a pipeline in FYP-I where we formalized the architecture and performed benchmarks on the architecture. The Datasetsthat we used to perform Benchmarks with were all numeric and were open source Datasets. We established two architectures and compared both of them on the open source Datasets.The following table shows our previous results :

Datasets	Ensemble with 3 NN		Ensemble with 5 NN	
	Individually	Parallel	Individually	Parallel
Mushroom	96.2%	93.5%	95.8%	95.4%
Heart	67.2%	53.4%	78.8%	73.2%
Genes	100%	84.8%	100%	88.5%
Australian	80.9%	79.3%	81.7%	69.9%
Thyroid	92.7%	92.8%	92.7%	92.7%

Chapter 4

FYP-II approaches

In the start of FYP-II we switched our approach to tackling the problem. Our initial plan was to use parallel training and incorporate the regularizer into the loss function,as a result decorrelating individual models in the ensemble to use less computational efficiency.The advantages obtained by this approach became obsolete as the computational resources used by simply training and selecting K best classifiers produced better results.Also,the new approaches such as snapshot ensembles, bayesian approaches,adding noise on forward pass and aggregating,and other similar approaches trained one neural network and produced an ensemble using this one trained network.As a result we researched the latest approaches yet again and found that confidence estimation was a problem, which these approaches failed to address and this problem was being approached using ensemble methods and hence we shifted our research direction towards confidence estimation.

4.1 Activation Based Feature Filtering

We wanted to find a method that would separate the data on which the model was not confident and was performing poorly so that we could use the low confidence subset of the data and train it on another model,penalizing it heavily on the non confident examples.The objective was to find a way to separate the data on which the model was not confident,train a new model on this data,and ensemble the results at the end hence increasing confidence of the ensemble. The solution of this problem has huge applications in safety critical applications such as health,self driving cars, surgery and most of all adversarial attacks,which are not yet properly handled and addressed in neural networks[12][13]. The following table shows our results:

Datasets	Model Accuracy	Accuracy on Threshold
Cifar10	69%	86%

Unfortunately this approach was not producing desired results therefore we had to transition towards the previous approach which

was to incorporate a regularizer into the loss function to increase model diversity.

4.2 Penalizing KL Divergence between individual models and aggregated ensemble decisions

A probabilistic measure constructed by Dr Tahir himself in some early literature was investigated and appropriately formulated which provides a distance between N individual classifier outputs and their aggregated ensemble results. Notation and details of the function are defined below.

4.2.1 Notation

C1 and *C2* are individual classifiers with outputs represented by *u1* and *u2*,There are *N* classes each represented by *Ci* where i goes to *N*. The output probability of a classifier *C1* given a class is represented by *p(u1 | Ci)*.The output probability of the aggregated result of the ensemble of two classifiers *C1* and *C2* given a class is given by *p(u1,u2 | Ci)*.

In general by conditional independence The probability of *p(A,B | C) = p(A | B,C)*p(B | C)* and if the the two variables *A* and *B* given *C* are independent, then *p(A,B | C) = p(A | C) * p(B | C)*

By this principle the conditional independence of two classifiers *C1* and *C2* given a class requires that *p(u1,u2 | Ci) = p(u1 | Ci) * p(u2 | Ci)*.

To enforce diversity between the classifiers we penalize the distance between the joint density *p(u1,u2 | Ci)* and the factorized density *p(u1 | Ci)p(u2 | Ci)* by adding the Kullback-Leibler divergence of the two densities in the loss function.The divergence is formulated as:

$$D(p(u1, u2|Ci)||p(u1|Ci)p(u2|Ci)) = \sum_{Ci} p(u1, u2|Ci) \log \frac{p(u1, u2|Ci)}{p(u1|Ci)p(u2|Ci)}$$

4.2.2 Diversity Measures

To assess the framework post hoc diversity measures were found such as the Qstatistics, the disagreement measure, and the double-fault measure[14]. These measures were used to measure diversity at train and test time to assess the functioning of the implemented regularizer.

4.2.3 Metric relation with Diversity

The details of the measures found and their relations found with diversity are as follows:

	D_k correct (1)	D_k wrong (0)
D_i correct (1)	N^{11}	N^{10}
D_i wrong (0)	N^{01}	N^{00}

Total, $N = N^{00} + N^{01} + N^{10} + N^{11}$.

where *Di* is the the output of the classifier *i*.

I) Q STATISTICS

Yule's Q statistic for two classifiers, Di and Dk,

$$Q(i,k) = \frac{N11N00 - N01N10}{N11N00 + N01 + N10}$$

For statistically independent classifiers, the expectation of Qi,k is 0[14].

II) DISAGREEMENT MEASURE

The disagreement measure is defined as:

$$Dis(i,k) = \frac{N01 + N10}{N11 + N10 + N01 + N00}$$

The diversity increases with the value of the disagreement measure[15].

III) DOUBLE FAULT MEASURE

The double fault measure is defined as:

$$Df(i,k) = \frac{N00}{N11 + N10 + N01 + N00}$$

The diversity decreases when the value of the double fault measure increases[15].

4.2.4 Experimental Setup

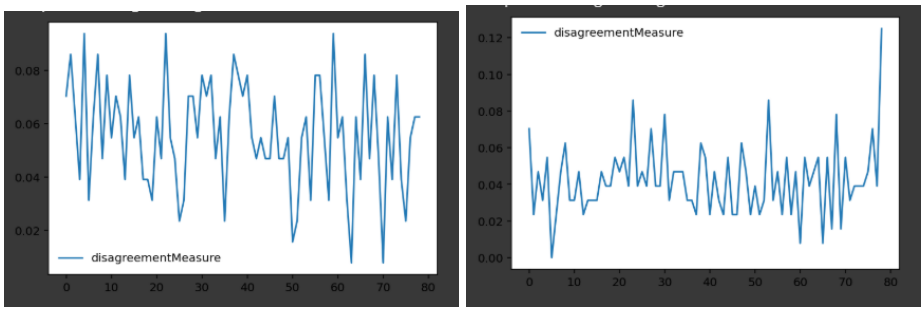
The base models of the ensemble are ANNs with 32*32*3 nodes in the input layer,1024 nodes in the first hidden layer and 256 in the second hidden layer with 10 output nodes for the 10 classes.The data set used for the experiment is CIFAR 10.

Two ensembles with exact architecture and setup are trained for 16 epochs in which all the above diversity metrics and the Kullback-Leibler divergence is recorded at train and test time, one with the regularizer and one without, with the regularizer scaled at 0.001.

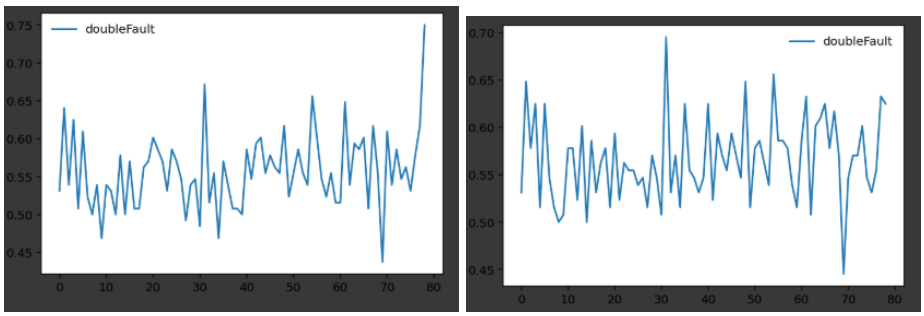
4.2.5 Findings

Reported below are the metrics and distances between the ensembles for comparison, where the left side shows the graphs with no regularizer and the right side with the regularizer in all measures respectively.

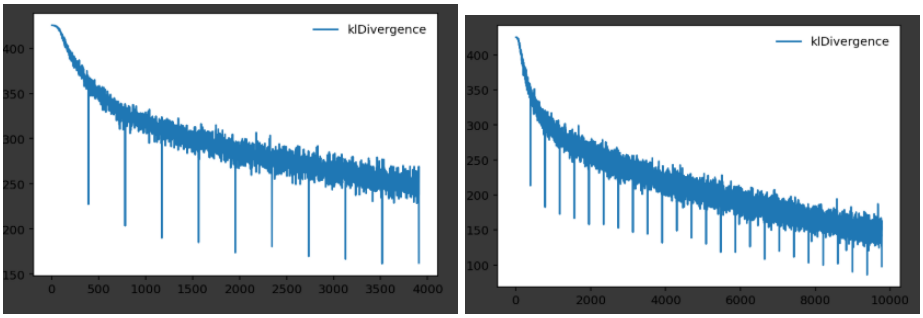
DISAGREEMENT MEASURE



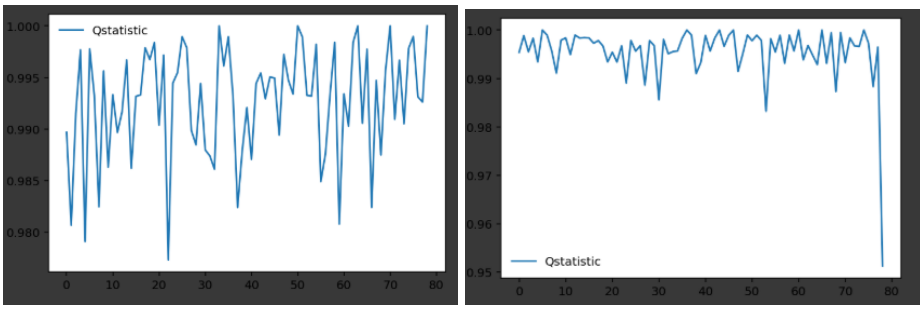
DOUBLEFAULT MEASURE



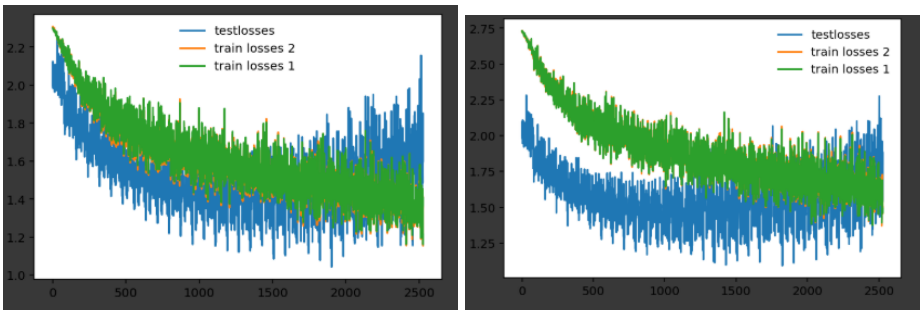
KULLBACK-LEIBLER DIVERGENCE AT TRAIN TIME



Q STATISTIC



TRAIN AND TEST LOSSES



The metrics clearly show that the regularizer improves the diversity in the inividual models, i.e the disagreement measure is higher,the double fault measure is lower and the Qstatistic is lower with the penalization, all of which are significantly better than the ensemble trained without the regularizer.

The Kullback-Leibler divergence curve is steeper with the regularizer and the train and test loss curves are closer than the curves without the penalized KL divergence measure.

The accuracy reported for the the both the ensemble setups is reported as 45 percent.

4.3 Conclusion and Future Work

We studied the the different methods to enforce diversity in individual models, the relationship between model independence and ensemble performance and found advantages and disadvantages of the different methods previously used.The main flaws with the approaches previously used were manual enforcement of the diversity measure,computational resources used in processing and searching,wastage of computational resources when pruning pools of models and a non existence of a generalized approach for enforcement.With a non linear probabilistic measure of independence we were successfully able to enforce model independence with the quality of generalization for any number of models independent of the architecture and kind of Neural network used.An approach to reporting individual model independence was defined with post hoc diversity measures which efficiently provide interpretation and comparison power between two ensemble setups. We hope the regularizer introduced will provide easy access to enforce diversity in ensemble methods resulting in better systems which can be used in various applications in the field.

The experiments conducted with the regularizer are with two ANN models and two setups,which can be improved by experimenting with different number of models,Convolutional neural networks and on different setups and datasets.But with the positive results,devised pipeline,and the experimentation and reporting setup, building better classifier teams in different setups would not be challenging.

4.4 References :

[1]Robert E. Schapire, boosting approach to machine learning an overview ,2003
[2] Hansen, Lars Kai and Salamon, Peter, Neural network ensembles, IEEE Transactions on Pattern Analysis Machine Intelligence, pp. 993-1001, 1990
[3] Opitz, D.W. and Shavlik, J.W., 1996. Generating accurate and diverse members of a neural-network ensemble. In Advances in neural information processing systems (pp. 535-541)
[4] Zhou, Z.H., Wu, J. and Tang, W., 2002. Ensembling neural networks: many could be better than all. Artificial intelligence, 137(1-2), pp.239-263.
[5] Mazurowski, M.A., Zurada, J.M. and Tourassi, G.D., 2009. An adaptive incremental approach to constructing ensemble classifiers: Application in an information-theoretic computer-aided decision system for detection of masses in mammograms. Medical physics, 36(7), pp.2976-2984
[7] Maguolo, G., Nanni, L. and Ghidoni, S., 2019. Ensemble of Convolutional Neural Networks Trained with Different Activation Functions. arXiv preprint arXiv:1905.02473
[8] Minetto, R., Segundo, M.P. and Sarkar, S., 2019. Hydra: an ensemble of convolutional neural networks for geospatial land classification. IEEE Transactions on Geoscience and Remote Sensing
[9] Chen, W., Gao, Y., Gao, L. and Li, X., 2018. A New Ensemble Approach based on Deep Convolutional Neural Networks for Steel Surface Defect classification. Procedia CIRP, 72, pp.1069-1072
[10] Liu, Y., Yao, X. and Higuchi, T., 2000. Evolutionary ensembles with negative correlation learning. IEEE Transactions on Evolutionary Computation, 4(4), pp.380-387
[11] Chen, H. and Yao, X., 2009. Regularized negative correlation learning for neural network ensembles. IEEE Transactions on Neural Networks, 20(12), pp.1962-1979
[12] Miyato, Takeru, et al. "Distributional smoothing with virtual adversarial training." arXiv preprint arXiv:1507.00677 (2015)
[13] Guo, Chuan, et al. "On calibration of modern neural networks." Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017.
[14]Kuncheva, Ludmila I., and Christopher J. Whitaker. "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy." Machine learning 51.2 (2003): 181-207.
[15]Tang, E. Ke, Ponnuthurai N. Suganthan, and Xin Yao. "An analysis of diversity measures." Machine learning 65.1 (2006): 247-271.

This page intentionally left blank.