Justifying Neural Network Ensembles

Table of Contents

1. Why ensemble over a single network?

The motivation to create ensembles comes from the human behavior analogy, that expects that M heads will think better than a single one.

If we consider a single network, we expect a distribution of the errors that it will make to be based on the training data set and the random initialization of the weights. This distribution will have a mean called the Expectation value. The Figure 1 shows 4 equivalent networks trained with different initialization weights, and the crosses are their E.

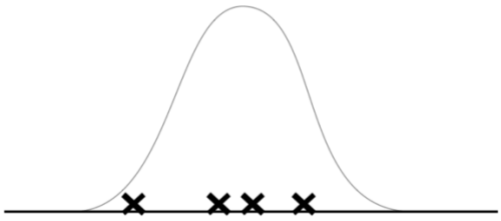

Figure 1. Error distribution of an unbiased estimator

We can see each network as a member of an ensemble. The more networks in the ensemble, the more we are encouraging the mean of their outputs to approximate the real value. This can be used to understand that if we increase the diversity of the elements of the ensemble, the better approximator it will be.

Now we can talk more in detail about the diversity.

2. What is Diversity?

We can think of diversity as the common meaning. It will refer to how different each member of an ensemble acts. If we have very different ways to treat the same input, then we will say our ensemble is diverse, whereas if all the members yield to the same conclusion after seen the same input, the lack of diversity and there is no point in ensemble, since we could have a really close result avoiding all the extra thinking.

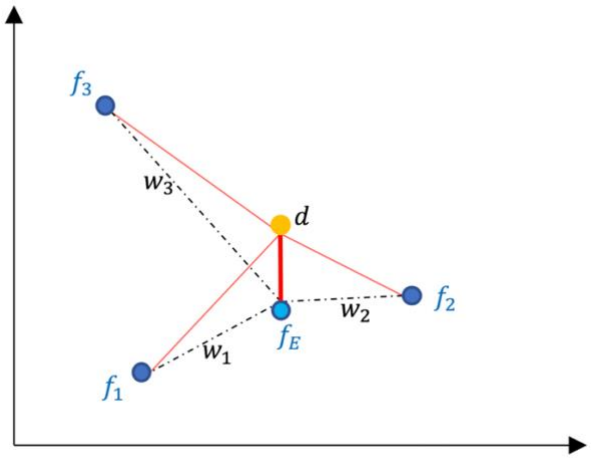

Figure 2. Ambiguity visualization

Ambiguity Decomposition

"If we have any given set of predictors, the square error of the convex-combined ensemble will be less than or equal to the average squared error of the individuals"

Being the result given by a weighted average (), the above sentence is expressed as ():

$$f_E = \sum w_i \cdot f_i \qquad \text{(Eq. 1)}$$

$$(f_E - d)^2 = \sum w_i \cdot (f_i - d)^2 - \sum w_i \cdot (f_i - f_E)^2 \qquad \text{(Eq. 2)}$$

Where in Eq.2, and using error for quadratic error:

- $(f_E - d)^2$ Error of the Ensemble
- $\sum w_i \cdot (f_i - d)^2$ Weighted average of the error of the individuals
- $\sum w_i \cdot (f_i - f_E)^2$ *Ambiguity term*

The Ambiguity term measures the amount of variability among the ensemble individual answers for this pattern.

Since this term is always positive, it is guaranteed that the ensemble error is lower than the average individual error. The larger this term, the larger the ensemble error reduction. However, as the variability of the individuals increase, so does the weighted average of the errors of those individuals. **This shows that diversity itself is not enough**.

2.1. Classification Error Diversity

Unfortunately, we cannot have an expression that similarly decomposes the classification error rate into the error rates of the individuals and a term that quantify their diversity.

2.2. How to create diversity?

When constructing an ensemble, we may choose to take information about diversity into account or not. Therefore, a distinction can be made into *explicit* and *implicit* diversity methods. To understand this:

- Bagging is an implicit method, since diversity is created when randomly samples the training patterns but at no point is a measurement taken to ensure diversity
- Boosting is an explicit method, because it ensures diversity when selecting the training patters for the incoming individuals of the ensemble

3. What decisions we have to make when building an ensemble?

Building an ensemble does not only consists on create and run several individual learners and train them on the same dataset.

Several decisions affect the intrinsic behavior of the ensemble as a model.

During Training

I divide these decisions into 3 groups depending on the part of the network we are looking at, as shown in :

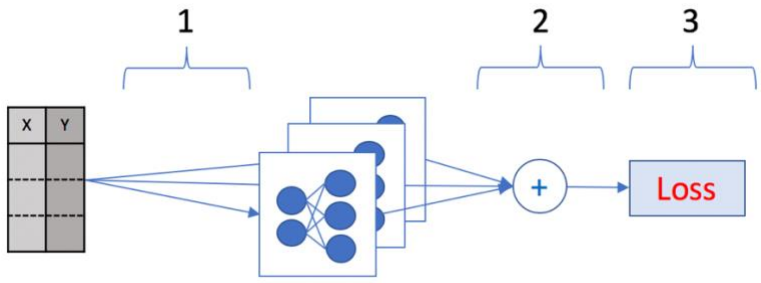

Figure 3. Scheme of a neural network ensemble

1 – How should the data be distributed between the ensembles?

- Bagging:
- Boosting:
- Random initialization:
- …

2 – How the output from the individual learners should be aggregated?

- Naïve averaging
- Majority Voting
- Super Learner
- Mixture of Experts
- …

3 – How should the loss be defined?

- On the scores - Before the Softmax
- On the probabilities – After the Softmax
- Ensemble aware – without averaging the gradients
- …

Post Training

4 – How the inference can be improved?

- How to reduce memory consumption?
- How to increase forward pass speed?
- …

All these concepts are covered in next corresponding works.

4. Ensemble Techniques

There are 3 categories upon the majority of neural network ensemble techniques can be placed:

4.1. Starting point in the Hypothesis Space

The hypothesis space is the space where the function approximator can follows a trajectory while training.

This approach then means that each network will start with a different random initialization, increasing the probability of taking different trajectories.

It has been proved that random initialization of the weights of a layer is the less effective method for generating diversity after network type, training set structure and number of hidden units.

4.2. Set of Accessible Hypotheses

Given that for a fixed network architecture, hypotheses are accessible or inaccessible depending on the training subset, these techniques vary the training data, or the architecture employed for different ensemble members.

4.2.1. Manipulation Training Data

It has been proven that combining the results of one type of classifier on different feature sets is far more effective than combining the results of different classifiers on one feature set.

4.2.2. Manipulation of Network Architectures

It has been proven that the variation in the number of hidden nodes is (after weight initialization) the least useful method of creating diversity in NN ensembles, due to the methodological similarities in the supervised learning algorithms (however the experiment they do is very limited and indicates there is some work to be done here).

4.2.3. Hybrid Ensembles

Examples of hybrid ensembles is to combine a random forest with neural networks or combines DTs with NNs in ensemble and use genetic programming to evolve a suitable combination rule.

Conclusions from the studies on these ensembles indicate that they will produce estimators with differing specialties and accuracies in different regions of the space. This specialization implies that with hybrid ensembles, *selection* of a single estimator rather than *fusion* of the outputs of all estimators may be more effective.

4.3. Hypothesis Space Traversal

Given a particular search space, which Is defined by the architecture of the network and the data, we could occupy any point in that space to give us a particular hypothesis. How we choose to traverse the space of possible hypotheses determines the final ensemble.

4.3.1. Regularization Methods

Use the same principle of regularization term in the loss function, to control a trade-off between two objectives: individual accuracy and ensemble accuracy.

4.3.2. Evolutionary Methods

The goal of evolutionary algorithm is to maintain the diversity in the population of individuals you are evolving to ensure you explore a large area of the search space. The trade-off is between exploration-exploitation to find the best individuals the fastest possible. The downside is that these methods don't combine the best individuals as ensembles pretend, but they choose the best individuals.

Evolutionary methods can be used to evolve a population of neural networks, using fitness sharing to encourage diversity, and then combines the entire population as an ensemble.

5. Ensemble of deep networks proof of improvement

In order to give some context in how ensembling improves a single network, here it is shown two examples on famous deep convolutional network architectures: ResNet20.

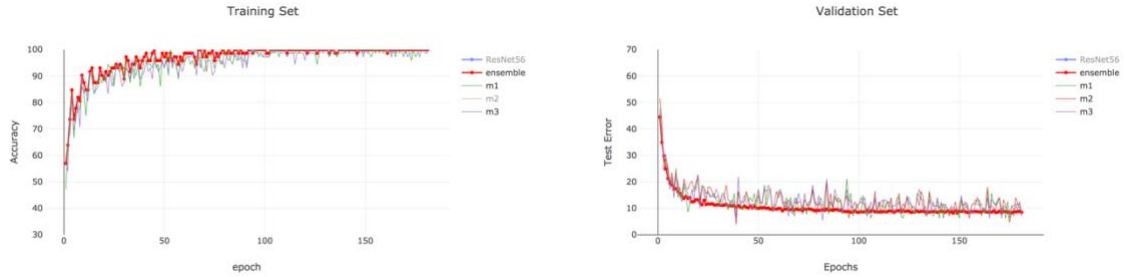

Figure 4. 3 single ResNet20 against an ensemble of 3 ResNet20

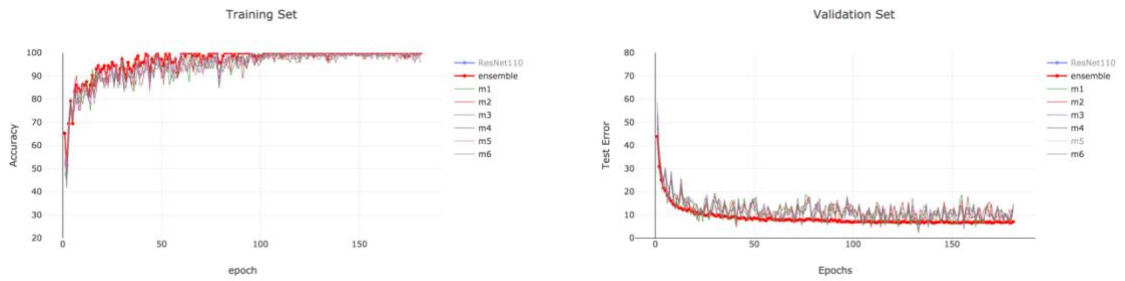

Figure 5. 6 single ResNet20 against an ensemble of 6 ResNet20

Both Figure 4 and Figure 5 shows a comparison between the performance of single networks ResNet20, against and ensemble of 3 and 6 ResNet20 respectively.

We can see how in both, training and testing, the ensemble performance is better than the individual models on their own.

Note that this has been simple averaging of the outputs of the individual models to provide the ensemble output. As shown in section 2, there are more intelligent strategies to do this aggregation and get even more profit from ensembling approaches.