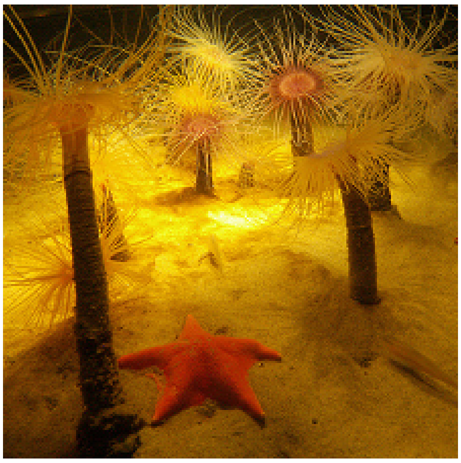
Ensembles of neural networks

René Fabricius, Faculty of management science and informatics, University of Zilina
fabricius@stud.uniza.sk

The goal of the project was to design and test nonparametric methods for combining outputs of neural networks. These neural networks were employed in the classification task on the validation set of ImageNet dataset. ImageNet validation set consists of 50000 images belonging to 1000 different classes. Outputs of the networks were posteriors for each image in the validation set. An example of the classification is shown below for an image from the ImageNet.



Correct class: sea anemone, anemone

5 most probable classes as classified by the most accurate of the used networks:
1. starfish, sea star (86.113%)
2. goldfish, Carassius auratus (7.665%)
3. sea anemone, anemone (2.777%)
4. axolotl, mud puppy, Ambystoma mexicanum (0.901%)
5. sea cucumber, holothurian (0.869%)

For combining the posteriors, we used two coupling methods formulated in [1]. These methods were originally designed for combining probabilistic outputs of binary classifiers in order to obtain multi class posteriors. In accordance with their origin, the input to these coupling methods are pairwise probability estimations for every pair of considered classes. For an attribute vector $\mathbf{x}$ and a pair of classes $i, j \in \{1, ..., k\}$, r_{ij} is an estimation of pairwise probability $\mu_{ij} = P(y = i | y = i \vee y = j, \mathbf{x})$, where y is the class of observation $\mathbf{x}$. Since we need these pairwise probability estimations for every pair of classes $i, j \in \{1, ..., k\}$, we can represent the input as a matrix $R = (r_{ij})$.

Output of a coupling method is a vector of posteriors $\mathbf{p}$ whose element p_i estimates the probability of an observation $\mathbf{x}$ belonging to a class i: $p_i \simeq P(y = i|\mathbf{x}), i = 1, ..., k$.

Each of the neural networks which we combine outputs a vector of posteriors $\mathbf{q}$. From this vector, we computed for each neural network a matrix of pairwise probabilities Q given by:

$$Q_{ij} \equiv \begin{cases} q_i/(q_i + q_j), & \text{if } i \neq j \\ 0, & \text{otherwise} \end{cases}.$$

We averaged these matrices and obtained a matrix R, which we then used as an input for used coupling methods. For averaging the matrices Q, we used two aggregation operators: mean and median. Both operators were applied elementwise. Four **new classification algorithms** resulted from combining two mentioned aggregation operators with two coupling methods. We refer to these algorithms as: m1_avg, m1_median, m2_avg, m2_median.

As customary with ImageNet, we evaluated the accuracy of classification by two metrics. These metrics are referred to as top1 and top5. Metric top1 considers the classification correct only if the class with the highest posterior is the correct one. Whereas metric top5 is satisfied if the correct class is among the 5 classes predicted as most probable. Reasoning behind the use of metric top5 lies in the fact, that in many ImageNet images, there are multiple objects displayed in a single image. This can be seen also in the image above - on the seabed, there are both anemones and a starfish.

All classification algorithms were implemented in programming language Python with the use of libraries numpy and pytorch. Experiment computations lasted approximately two and a half weeks and were executed on a faculty server equipped with graphic card RTX 2080 Ti.

Experiments were performed with 15 neural networks. We created 1260 different subsets of these networks and for each subset, we tested all four new classification algorithms. Histograms of accuracy of these algorithms according to top5 metric are displayed in the Figure 1. Horizontal axes represent top5 accuracy - portion of images that were classified correctly according to this metric. On vertical axes are counts of subsets which fall into each bin. Red dashed line represents the median accuracy of each classification algorithm over all tested subsets. Orange dashed line represents the accuracy for a subset combining all 15 networks. Accuracies of individual networks are displayed by green dashed lines. The most accurate classification algorithm according to the median of top5 accuracies over all tested subsets of networks is the algorithm m2_median. Highest accuracy of a single model was achieved by classification algorithm m2_avg applied to a subset of four networks: DenseNet169, DenseNet201, ResNet152 a Xception.

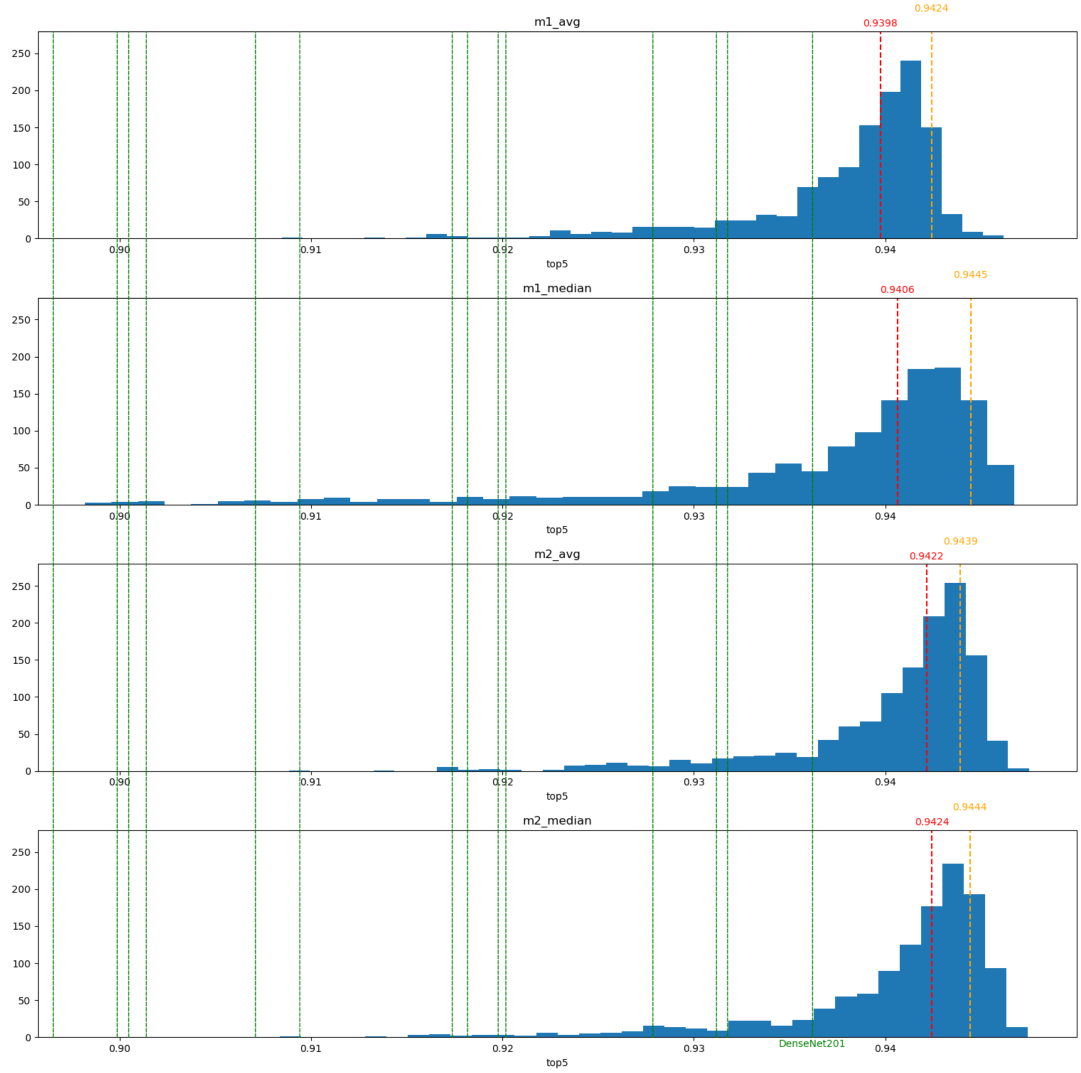


Figure 1: Histograms of accuracies of classification algorithms over all tested neural network subsets.

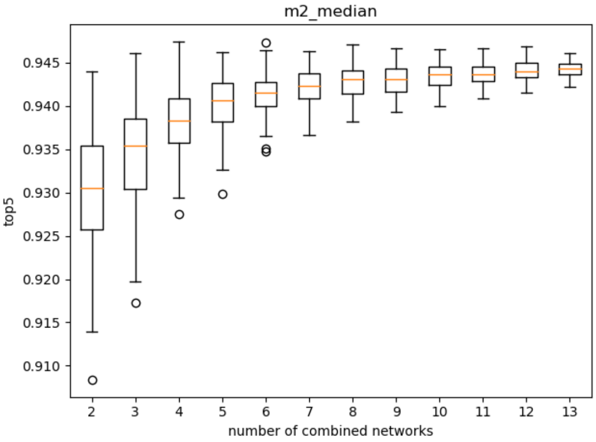


Figure 2: Relation between top5 accuracy of classification algorithm m2_median and number of combined networks.

We examined the relation between top5 accuracy of classification algorithm m2_median and the number of combined networks. This relation is displayed by the use of boxplots in the Figure 2. On horizontal axis is displayed the number of combined networks. On vertical axis is top5 accuracy. As can be seen in the figure, the accuracy of the classification algorithm increases with increasing number of combined networks. It can also be observed that variance of accuracies is smaller for subsets consisting of more networks.

Acknowledgement. This research was carried out from 4.2. 2020 to 7.5. 2020 and was financed by Faculty of management science and informatics, University of Zilina. Implemented algorithms can be found in [2].

References

[1] T.-F. W. et al, *JMLR*, 2004. [Online]. Available: http://www.jmlr.org/papers/volume5/wu04a/wu04a.pdf

[2] R. Fabricius. (2020). [Online]. Available: https://github.com/ReneFabricius/NNEnsembles