Exploring Ensembling in deep learning

Antonio Bruno[a,*], Massimo Martinelli[a,], Davide Moroni[c,***]**

[a] Institute of Information Science and Technologies, National Research Council of Italy, via G. Moruzzi 1, 56124 Pisa, Italy

* e-mail: antonio.bruno@isti.cnr.it
** e-mail: massimo.martinelli@isti.cnr.it
*** e-mail: davide.moroni@isti.cnr.it

Antonio Bruno, Massimo Martinelli and Davide Moroni shares the first authorship

Abstract— Ensembling is a very well known strategy consisting in fusing several different models to achieve a new model for classification or regression tasks. Ensembling has been proven to provide superior performance in various contexts related to pattern recognition and artificial intelligence. The winners of public challenges in image analysis often adopt solutions based on Ensembling. The idea of Ensembling has also provided suggestions for introducing recent deep learning architectures with multiple layer connections that mimic ensembling approaches. However, the full potential offered by Ensembling is not yet fully exploited. This paper aims to explore possible research directions and define new fusion approaches. Preliminary experimental tests show favourable results with an increment in accuracy regarding the number of operations needed in training and inference.

Keywords: Ensembling, machine learning, deep learning, image classification, convolutional neural networks

INTRODUCTION

Representation learning and deep learning have achieved amazing results in the last decades, obtaining unparalleled performance under challenging tasks such as image classification and recognition [1]. After the first impressive leap, many works have been of incremental nature in the previous years, often focused on architecture engineering for achieving minor improvements over a sensible increase in complexity. Indeed, the performance gain versus the computational increment ratio has become less attractive. Therefore, research has moved to find optimal tradeoffs between accuracy and computational load [2]. This is also motivated by the widespread adoption of deep learning paradigms that calls for the sustainable use of artificial intelligence (AI). AI has operational costs directly measurable in energy consumption, therefore having a relevant environmental footprint [3].

Ensembling is a well-known approach that permits using a collection of different models (e.g. classifiers or regressors) to obtain a new model having other (and hopefully superior) performance with respect to predefined metrics. Ensembling has a long history that dates back even before the birth of machine learning. Indeed, it is customary to state that the first application of Ensembling is majority voting in statistics as in the claim of the theorem by Marquis de Condorcet: he proved in 1785 that if the probability of each voter being correct is more than one half and the voters are independent, then the addition of more voters increases the likelihood of the majority vote being correct. Long after that, Ensembling has been used to turn weak models into superior models showing encouraging results in several domains. It has been reported that ensemble models often become in the first place in public competitions such as those promoted by Kaggle.

The relationship between deep learning and Ensembling is at least twofold. From one side, basic constructions of Ensembling known as bagging, boosting and stacking have somehow influenced architectures commonly used in deep learning and the way they are trained. For instance, residual networks behave like ensembles of relatively shallow networks [4]. On the other side, thanks to Ensembling strategies, deep learning models can be used as basic models to build more complex models. This paper focuses on this second aspect by recapping Ensembling and its role in deep learning, exploring several directions. Preliminary results are then announced on a relevant dataset.

RELATED WORKS

Ensembling generally refers to machine learning approaches in which a set of *weak learners* (or basic models) is turned into a *strong learner* (or ensemble model). The set of weak learners might consist of homogenous models (i.e. they are all from the same family or architecture) or might be heterogeneous, i.e. the basic models belong to different machine learning paradigms. The basic example is to put together multiple models trained for solving the same classification or regression task and then combine them together in some fashion, e.g. by performing majority voting in the case of classification or averaging in the case of regression. The scope of performing Ensembling is generally related to the desire to reduce the bias or variance that affect a machine learning task [6].

As it is well known, very simple model might have a great error in achieving good performance on a dataset, even during training. This is generally linked to the low representation capabilities of simple models that cannot capture all the complex patterns in the training datasets. Such error during training is referred to as the bias of the model. By converse, very complex models have many degrees of freedom to adhere to the training dataset completely and obtain excellent performance during training. However, they capture not only the relevant features of the problem but also learn insignificant features of the training dataset. This results in relatively poor performance during test and validation: the model is overfitted to the training dataset and does not reach good general results, having scarce generalization capabilities. We refer to this issue, saying that the model has high *variance*.

The three basic approaches to performing ensemble are bagging, boosting, and stacking.

In general, bagging reduces the variance among the base classifiers, while boosting-based ensembles lead to bias and variance reduction. Stacking is commonly used as a bias reducing technique.

In more detail, bagging is performed by subdividing the training datasets into different subsets according to some criteria, e.g., balancing class distributions inside each subset or other forms of equalization. Then, each subset of the training set is used to train a weal classifier. Any such classifier ideally has a low bias on the training set but possibly high variance. Using a fusion layer, the outputs of the single classifiers are combined by performing (weighted) voting or performing a weighted average. The model given by the fusion of the weak classifiers is called a strong classifier, and it potentially exhibits lower variance. Notice that the weak classifiers might be trained independently and in parallel. Very often, such classifiers share the same architecture.

In boosting instead, weak classifiers are very simple and low complexity but are trained cleverly, for example, using cascading. Adaboost [7] is one of the most popular approaches in which each classifier is trained so as to properly deal with the examples in the training set on which previous weak classifiers have failed. The boosting concept is also known to be the backbone behind well-known architectures like Deep Residual networks [5].

Finally, stacking often considers heterogeneous weak learners. Training is performed in parallel, while a final combination is obtained by training a meta-model to output a prediction based on the different weak models' predictions. Deep convex nets (DCN) [8] are recognized to be a deep learning architecture composed of a variable number of modules stacked together to form the deep architecture.

In general, all of these approaches have been used in conjunction with deep learning models. The review [9] presents some recent literature on the subject systematically.

METHODOLOGY

As seen in the previous section, ensemble and deep learning have a twofold relationship. This section aims to briefly report some experiments on Ensembling that is worthwile exploring for optimizing deep learning models.

A) Varying the number of classifiers. After having fixed a deep learning architecture, such architecture can be regarded as a weak model. Different training runs starting with random weights might result in different classifiers. Majority voting can be applied in this case. The dependence of the performance with respect to the number of classifiers might be studied.

B) Sampling strategies and balancing. Besides performing training of all the weak learners on the full training set as described before, procedures for sampling can be applied. For instance, using disjoint datasets for each weak classifier helps have a set of indipendnet classifiers. In addition, stratification can be applied to keep the same class frequencies in each subset; conversely, it might be interesting to explore the possibilities given by training each classifier to make it specialized in addressing a special class.

C) Control size and model complexity. The ensembling approaches can be performed by keeping track of the model size and model complexity; this can help understand heuristics for the optimal choice of weak learners' dimensions and ensemble size.
D) Stacking at the deep feature levels. In many cases, the first layers of a deep network perform feature extraction while the final layers, usually fully connected, perform classification/regression. A possibility in ensembling is given by stacking weak models by removing the final classification layers from each one and training an ad hoc meta classifier.
E) Learning strategies. Given the trained ensemble, it is still possible to perform fine-tuning of model parameters suitably training the model by freezing or not some of the overall network layers.

CONCLUSIONS

This short paper has explored several directions for introducing Ensembling in the deep learning context. The approaches and the involved ideas are well-grounded in previous knowledge and guarantees connected to Ensembling in machine learning, yet there are many possible pathways and combinations to explore. In some preliminary experiments, we studied an adaptive enemsbling based on bagging, making it possible to achieve 100% accuracy on a know dataset in agricultural applications [10]. Other expeirments are under active development, and the results will be reported in the future.

COMPLIANCE WITH ETHICAL STANDARDS

This article is a completely original work of its authors; it has not been published before and will not be sent to other publications until the PRIA editorial board decides not to accept it for publication"

REFERENCES

1. Pouyanfar, Samira, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, and Sundaraja S. Iyengar. "A survey on deep learning: Algorithms, techniques, and applications." ACM Computing Surveys (CSUR) 51, no. 5 (2018): 1-36.
2. Tan, Mingxing, and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." In International conference on machine learning, pp. 6105-6114. PMLR, 2019.
3. Schwartz, Roy, Jesse Dodge, Noah A. Smith, and Oren Etzioni. "Green ai." Communications of the ACM 63, no. 12 (2020): 54-63.
4. Veit, Andreas, Michael J. Wilber, and Serge Belongie. "Residual networks behave like ensembles of relatively shallow networks." Advances in neural information processing systems 29 (2016).
5. He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.
6. Dong, Xibin, Zhiwen Yu, Wenming Cao, Yifan Shi, and Qianli Ma. "A survey on ensemble learning." Frontiers of Computer Science 14, no. 2 (2020): 241-258.
7. Schapire, Robert E. "Explaining adaboost." In Empirical inference, pp. 37-52. Springer, Berlin, Heidelberg, 2013.
8. Deng, Li, and Dong Yu. "Deep convex net: A scalable architecture for speech pattern classification." In Twelfth annual conference of the international speech communication association. 2011.
9. Ganaie, Mudasir A., and Minghui Hu. "Ensemble deep learning: A review." arXiv preprint arXiv:2104.02395 (2021).
10. Mohanty, Sharada P., David P. Hughes, and Marcel Salathé. "Using deep learning for image-based plant disease detection." Frontiers in plant science 7 (2016): 1419.

Authors profile *(including Biography and Photo)*

ANTONIO BRUNO received the Master Degree in Computer Science from the University of Pisa. He received a research grant from the Signal and Images Lab (SI-Lab) of the Institute of Information Science and Technologies (ISTI) for collaborating on the Barilla AGROSAT Plus project. His main research interests include deep learning for structured domain (e.g sequences, trees, images).

MASSIMO MARTINELLI is member of the Signals & Images research laboratory at ISTI-CNR since 2000, at CNR since 1987.
Head of the "Software Technologies and Frameworks Area" at SI-Lab since 2016.
He was member of the W3C Multimedia Semantics Incubator Group (2006-2007)
He is currently the Principal Investigator of the projects TiAssisto (Tuscany region), CloudPathology (industrial), Barilla Agrosatplus (industrial), RadIoPoGe (collaboration).
Leader of the of the Scientific Collaboration Agreements with the UO Otolaryngology, audiology and phoniatrics (UNIPI), and with the Italian Mountain Medicine Society.
Member of the Doseteam4you group of the Department of Diagnostics and Interventional Radiology of the University Hospital of Pisa. Member of the Topic Board of the Sensors MDPI Journal, Topic Editor of "Machine Learning and Biomedical Sensors" of the Sensors MDPI Journal, Guest Editor of the special Issues "Intelligent Sensors for Monitoring Physical Activities" and "Wearable Sensors and Internet of Things for Biomedical Monitoring" of the same Journal.
Guest Editor of "Artificial Intelligence in Point of Care Diagnostics" of the Frontiers Journal.
His main scientific interests include Computer Vision, Deep Learning, Decision Support Systems, Web technologies.

DAVIDE MORONI received the M.Sc. degree (Hons.) in mathematics from the University of Pisa, in 2001, the Diploma from the Scuola Normale Superiore of Pisa, in 2002, and the Ph.D. degree in mathematics from the University of Rome La Sapienza, in 2006. He is a Researcher with the Institute of Information Science and Technologies (ISTI), National Research Council, Italy, Pisa. He is currently the Head of the Signals and Images Lab, ISTI. He is the Chair of the MUSCLE working group of the European Consortium for Informatics and Mathematics. Since 2018, he serves as the Chair of the Technical Committee 16 on Algebraic and Discrete Mathematical Techniques in Pattern Recognition and Image Analysis of the International Association for Pattern Recognition (IAPR). He is an Associate Editor of IET Image Processing. His main research interests include geometric modeling, computational topology, image processing, computer vision, and medical imaging. At the moment, he is leading the ISTI-CNR team in the National Project PON MIUR S4E, working on maritime safety and security, and in the regional Project IRIDE addressing AR technologies and computer vision of Industry 4.0.