Herds of Dumb Models: A New Approach Towards Reliable and Safe AI

Nicholas Mc Guire*†, Imanol Allende*†, Carles Hernández ‡
*OpenTech EDV Research GmbH, Bullendorf, Austria
† Open Source Automation Development Lab (OSADL), Germany
‡ Universitat Politécnica de Valéncia, Spain

***Abstract*—In this article, we extrapolate the concept of a popular jury, characteristic of various judicial systems, to the field of Deep Learning with the aim of advancing towards a reliable and safe Artificial Intelligence (AI). Just as a jury composed of individuals without specific expertise reaches a decision, we propose the "Herds of Dumb" models, a method that is based on the use of multiple weak and unbiased models. The idea is to observe the degree of consensus among them in order to identify out-of-distribution (OOD) data when there is no unanimous agreement. This innovative approach opens new avenues for the creation of more robust and reliable AI systems.**

I. Introduction

The performance of individual Machine Learning (ML) models is commonly assessed through validation and testing, providing validation accuracy as an initial approximation of their correctness capabilities. However, implementing ML models in real scenarios presents challenges. Among others, they may face inputs that are not exactly the same as the ones trained for or even face untrained classes while at the same time having no output to indicate the *"non-fit"* case. As a result, this leads to an uncertainty in the behavior and response of the model. Models tend to identify similarities between feature sets, even if the class presented to them is unknown. On the one hand, they often classify inputs convincingly, even those to which they have not been trained. On the other hand, similarities between certain classes are not defined in detail. For instance, when does a car starts being a truck? Or the number 7 being a 1? To address these uncertainties, there are two fundamental strategies: i) enhance our understanding of AI to rectify the discrimination and feature uncertainty, or ii) construct the model and train it in such a way that a potentially large set of models yields predictable results on average, and aggregate these results using suitable statistical methods.

While efforts are underway to implement the first solution, it has an inherent limitation: the assurance for any single model will ultimately be qualitative. The second approach, based on a set of models, potentially allows for runtime detection of increased uncertainty and the reliable declaration of *"I don't know"* instead of producing an incorrect verdict. Implicitly, we are replacing correctness claims with a strong indicator i.e., consistency, although only if the model set is constructed with adequate independence and unbiasedness.

We introduce the concept of large sets of weak models, which we term as *"Herd of Dumb Models"* (HDM), and provide an overview of the theoretical basis for our claims. We also present preliminary empirical results derived from the well-known CIFAR-10/100 datasets. This approach offers a promising avenue for enhancing the functional safety of AI systems.

II. Problem statement

The issue with a single model is that we cannot determine whether any given input falls within the model's *"knowledge base"*. While we can employ anomaly or outlier detection techniques to ascertain if the input likely fits within the known discriminatory feature space, we cannot determine the level of assurance if it does fit. The fundamental problem lies in our inability to understand what the model's feature space represents. It is unlikely to align with the human feature space associated with any specific class, primarily because humans do not universally agree on the defining characteristics of a class, such as what constitutes a cat and when an input ceases to be a cat. In other words, our knowledge classes lack precise boundaries, and even if we establish such boundaries in certain cases, they may not align with others' perceptions. One potential solution is to formalize the knowledge or feature space to make it more comprehensible. The alternative is to randomize the feature relation structures so that there are (ideally) no common structures. Hence, it is possible to focus on the consistency of the emitted verdict of the different models as an indicator that there must be a common knowledge base even if this is not known to us. Otherwise, the models could not have emitted highly agreeable results. Statistical properties like the variance may then serve as an indicator of how strong this consensus actually is.

III. Herds of Dumb Models Concept

Our approach is based on the Condorcet Jury theorem [1], a political science theorem about the relative probability of a given group of independent individuals arriving at a correct decision. The theorem assumes that each member of the jury has an independent probability to vote for the correct verdict and are more likely to vote correctly (i.e., $p \geq 1/2$), then adding more voters increases the probability that the majority decision is correct.

Drawing parallels from this theorem, we propose a solution based on a group of unbiased models, or weakly biased. Given a sufficiently large models group, we can derive statistical

assurance on the aggregated verdict using social choice theory or social decision theory. Therefore, just as in popular juries several unbiased individuals with limited knowledge of the subject matter are employed to arrive at the correct verdict, we propose to use simple (i.e., untuned) models to evaluate how well they agree on the outcome. This approach, inspired by the Condorcet Jury theorem, allows us to leverage the wisdom of the crowd in the realm of Deep Learning.

What is the difference with ensemble methods? Just as in popular juries the individuals independent probabilities and, hence, they are unbiased, here the models have to be unbiased as well. This is achieved by employing diversity in model structure, training data set, and input augmentation (ensuring each model in the group is presented with similar but not identical input). Our proposed solution involves using a model generator that randomly selects layers, layer sequences, hyper-parameters, as well as the optimizer and its settings, within predefined sane bounds. Furthermore, just as in juries they are not experts in the subject matter, in this case the models are untuned. Hence, models are trained on a relatively small subset of the overall training dataset. This subset is chosen at random from a range common to all models, resulting in a slightly overlapping, rather than disjoint, training set.

Each selected model is subjected to a set of inputs derived from the original input, which is replicated and then subjected to image augmentation. The predictions provided by each model are then logically *"stacked"* into a results matrix, which is post-processed using simple statistical methods. From these statistics, we determine whether the models saw a credible input (based on strength and agreement), and hence whether the result is trustworthy or should be rejected with a declaration of *"I don't know"*. Note that from a functional safety perspective the declaration of *"I don't know"* is essential to go to a safe state.

The assurance level is expressed as the agreement level within and across the herd's predictions. Furthermore, following a common safety technique, we can add as a final layer of assurance M-out-of-N consensus. In other words, we can do several herds of random models to increase assurance.

IV. Preliminar results

The first results have been performed with convolutional neural network models trained with CIFAR-10. For this purpose, each of these models have been trained sampling 20 % of the training set. The evaluation has been performed by training 234 randoms models with mean accuracy of 80% (min 69% and max 82%) and built approximately with less than 200 thousand hyperparameters. From the total pool of models, we sampled randomly 128 models, grouped in 8 different herds (i.e., 16 models per herd). The aim is to evaluate the consensus of our herds with a 8-out-of-8 class consensus which allows to interpret the distribution of consensus development as an indication for the overall inference robustness. Ideally false-positive results would not achieve consensus at all. Even though in practice they do, they are due to spurious agreements and not as Common Cause Failures (CCF)s.

If we perform average ensembling with the training data, each herd provides an accuracy of 96% for the training data sets ($p \geq 1/2$ as required by Condorcet theorem). Note that these models where trained with 20% of the data. Nevertheless, if we apply the statistical filters to classify the outputs as classified or unknown, we see a cost in the classification performance. Table I shows the obtained results in each herd and the consensus (i.e., 8-out-of-8) in the last column.

TABLE I
ID: Cifar-10 with 300 images sampled from test data set

Herds	1	2	3	4	5	6	7	8	Consensus
Total probs	300	300	300	300	300	300	300	300	300
Classified (TP)	175	177	176	173	178	172	173	180	134
Misclassified (FP)	2	2	3	3	8	2	3	1	0
I don't know	123	121	121	124	114	126	124	119	74

The key evaluation for the current proposal is OOD data. For that, we use what is considered near-OOD [2], Cifar-100 in this case. Table II shows a high rejection rate, as can be expected. From the 300 images, only 15 images have the 8 herds matched resulting in a misclassification. However, after inspecting them manually, we can confirm the majority of them are collisions and not misclassifications, as Cifar-10 and Cifar-100 are not disjoint even if they do not share class labels (e.g., Cifar-10 automobiles/trucks and Cifar-100's vehicles1/vehicles2). To be precise, there are 4 misclassifications. Besides, the fact that we reject most but do identify (at least some of) the collisions correctly indicates that the approach works and can provide safe judgements on OOD inputs.

TABLE II
OOD: Cifar-100 with 300 sampled images

Herds	1	2	3	4	5	6	7	8	Consensus
Total probs	300	300	300	300	300	300	300	300	300
Classified (TP)	0	0	0	0	0	0	0	0	0
Misclassified (FP)	44	45	47	42	45	40	38	42	15
I don't know	256	455	254	258	255	260	262	258	200

V. Conclusions

While our current status is effective, it is not satisfactorily efficient; in other words, we still reject too many inputs, but those we do accept appear to be correct. It is clear that performance needs to be enhanced in order to be feasible to employ the method in a use case. Simultaneously, the preliminary results show promising signs for the development of trustworthy and safe AI, as the trend of reported false positives approaches zero.

References

[1] Marquis de Condorcet. Essay on the Application of Analysis to the Probability of Majority Decisions. *Paris: Imprimerie Royale*, page 1785, 1785.

[2] Debargha Ganguly and Debayan Gupta. Detecting Out-of-Distribution Data with Semi-supervised Graph "Feature" Networks.