Automatic model selection for neural networks

David Laredo

***Abstract*—Neural networks and deep learning are changing the way that artificial intelligence is being done. Efficiently choosing a suitable model (including hyperparameters) for a specific problem is a time-consuming task. Choosing among the many different combinations of neural networks available gives rise to a staggering number of possible alternatives overall. Here we address this problem by proposing a fully automated framework for efficiently selecting a neural network model given a specific problem (whether it is classification or regression). Our proposal focuses on a distributed decision-making algorithm for keeping the most promising models among a pool of possible models for three of the major neural network architectures, namely Convolutional Neural Networks (CNNs), Multi-Layer Perceptrons (MLPs) and Recurrent Neural Networks (RNNs). This work develops AutoNN, a new micro genetic algorithm (along with a new representation for the neural network as a genotype and new crossover and mutation operator) that automatically and efficiently finds the most suitable neural network model for a problem specified by the user. Our evaluation on four different datasets show that the AutoNN effectively finds suitable neural network models while being efficient in terms of the computational burden, our results are compared against other state of the art methods such as Auto-Weka and SkLearn.**

***Index Terms*—artificial neural networks, model selection, hyperparameter tuning, distributed computing, evolutionary algorithms.**

I. Introduction

MACHINE learning studies automatic algorithms that improve themselves through experience. Given the large amounts of data currently available in many fields such as engineering, biomedical, finance, etc, and the increasingly computing power available machine learning is now practiced by people with very diverse backgrounds. Increasingly, users of machine learning tools are non-experts who require off-the-shelf solutions. The machine learning community has aided these users by making available a variety of easy to use learning algorithms and feature selection methods such as WEKA [1], PyBrain [2] or MLLib [3]. Nevertheless, the user still needs to make some choices which not may be obvious or intuitive (selecting a learning algorithm, hyperparameters, features, etc) thus leading to the selection of non optimal models.

Recently, deep neural networks have gained a lot of attention due to the newer models (CNN, RNN, Deep Learning, etc.) and their flexibility and generality for solving a large number of problems: regression, classification, natural language processing, recommendation systems, just to mention a few. Furthermore, there are a lot software libraries which makes their implementations easy to use (TensorFlow [4], Keras [5], Caffe [6], CNTK [7], etc). Nevertheless, the task of picking the right neural network model (hyperparameters included) can be even more complicated and time consuming than that of other algorithms. Given the popularity of neural networks, specially among non computer scientist we will restrict our efforts in this study to them and leave other machine learning algorithms for future work.

D. Laredo is with the Department of Mechanical Engineering, University of California, Merced, CA, 95340 USA e-mail: dlaredorazo@ucmerced.edu.

Usually, the process of selecting a suitable machine learning model for a particular problem is done in an iterative manner. First, an input dataset must be transformed from a domain specific format to features which are predictive of the field of interest, once features have been engineered users must pick a learning setting appropriate to their problem, e.g. regression, classification or recommendation. Next users must pick an appropriate model, such as support vector machines (SVM), logistic regression or any flavor of neural networks (NNs). Each model family has a number of hyperparameters, such as regularization degree, learning rate, number of neurons, etc, and each of these must be tuned to achieve optimal results. Finally, users must pick a software package that can train their model, configure one or more machines to execute the training and evaluate the model's quality. It can be challenging to make the right choice when faced with so many degrees of freedom, leaving many users to select a model based on intuition or randomness and/or leave hyperparameters set to default, this approach will usually yield suboptimal results.

This suggests a natural challenge for machine learning: given a dataset, to automatically and simultaneously chose a learning algorithm and set its hyperparameters to optimize performance. As mentioned in [1] the combined space of learning algorithm and hyperparemeters is very challenging to search: the response function is noisy and the space is high dimensional involving both, categorical and continuous choices and containing hierarchical dependencies (e.g. hyperparameters of the algorithm are only meaningful if that algorithm is chosen). Thus, identifying a high quality model is typically costly (in the sense that entails a lot of computational effort) and time consuming.

Distributed and cloud computing provide a compelling way to accelerate this process, but also present additional challenges. Though parallel storage and processing techniques enable users to train models on massive datasets and accelerate the search process by training multiple models at once, the distributed setting forces several more decisions upon users: what parallel execution strategy to use, how big a cluster to provision, how to efficiently distribute computation across it, and what machine learning framework to use. These decisions are onerous, particularly for users who are experts in their own field but inexperienced in machine learning and distributed systems.

To address this challenges we propose AutoNN a flexible and scalable system to automate the process of selecting

artificial neural network models. The key contributions of this paper include: 1) a new way to encode neural networks as genotypes for evolutionary computation algorithms, 2) new crossover and mutation operators to generate valid neural networks models from an evolutionary algorithm, 3) defining a way of measuring the similarity between two neural networks, 4) all these components together make a new evolutionary algorithm, which we call AutoNN, which can be used to find an optimal neural network architecture for a given problem.

The remainder of this paper is organized as follows: Section II formally introduces the model selection problem (also referred as CASH problem). The related work is briefly reviewed in Section III. Next the AutoNN algorithm and all of its components are described in detail in Section IV, experiments to test the algorithm and comparison against other state of the art methods are presented in Section ... . Finally conclusions and future work are discussed in Section ... .

II. The CASH problem

In this section we introduce and formally describe the model selection problem, for this section we borrow the definitions given in [8]. This work focuses on supervised learning: learning a function $f : \mathcal{X} \mapsto \mathcal{Y}$ with finite $\mathcal{Y}$. A learning algorithm A maps a set $\{d_1, \ldots, d_n\}$ of training data points $d_i = (\mathbf{x_i}, \mathbf{y_i}) \in \mathcal{X} \times \mathcal{Y}$ to such a function. Most learning algorithms A further expose hyperparameters $\lambda \in \mathbf{\Lambda}$, which change the way the learning algorithm A_λ works. One example of hyperparameters is the number of neurons in a hidden layer of an ANN, another common example is the learning rate α of a neural network. These hyperparameters are typically optmized in an "outer loop" that evaluates the performance of each hyperparameter configuration using cross-validation.

A. Model selection

Given a set of learning algorithms $\mathcal{A}$ and a limited amount of training data $\mathcal{D} = \{(\mathbf{x_1}, \mathbf{y_1}, \ldots \mathbf{x_n}, \mathbf{y_n})\}$, the goal of model selection is to determine the algorithm $A^* \in \mathcal{A}$ with optimal generalization performance. Generalization performance is estimated by splitting $\mathcal{D}$ into disjoint training and validation sets $\mathcal{D}_t$ and $\mathcal{D}_v$ respectively, learning function f by applying A^* to $\mathcal{D}_t$, and evaluating the predictive performance of this function on $\mathcal{D}_v$. Using k-fold validation, which splits the data into k equal sized partitions $\mathcal{D}_v^1, \ldots, \mathcal{D}_v^k$ and sets $\mathcal{D}_t^i = D \backslash D_v^i$ for $i = 1, \ldots, k$ the model selection problem is written as:

$$A^* \in \underset{A \in \mathcal{A}}{\operatorname{argmin}} \frac{1}{k} \sum_{i=1}^{k} \mathcal{L}\left(A, \mathcal{D}_t^i, \mathcal{D}_v^i\right), \tag{1}$$

where $\mathcal{L}\left(A, \mathcal{D}_t^i, \mathcal{D}_v^i\right)$ is the loss achieved by A when trained on $\mathcal{D}_t^i$ and evaluated on $\mathcal{D}_v^i$.

B. Hyperparameter optimization

The problem of optimizing the hyperparameters $\lambda \in \mathbf{\Lambda}$ of a given learning algorithm A is conceptually similar to that of model selection. Some key differences are that hyperparameters are often continuous, that hyperparameter spaces are often high dimensional, and that we can exploit correlation structure between different hyperparameter settings $\lambda_1, \lambda_2 \in \mathbf{\Lambda}$. Given n hyperparameters $\lambda_1, \ldots, \lambda_n$ with domains $\Lambda_1, \ldots, \Lambda_n$, the hyperparameter space $\mathbf{\Lambda}$ is a subset of the crossproduct of these domains: $\mathbf{\Lambda} \subset \Lambda_1 \ldots \Lambda_n$. This subset is often strict, such as when certain settings of one hyperparameter render other hyperparameters inactive. For example, the parameters determining the specifics of the third layer of a deep belief network are not relevant if the network depth is set to one or two. More formally, following [9], we say that a hyperparameter λ_i is conditional on another hyperparameter λ_j, if λ_i is only active if hyperparameter λ_j takes values from a given set $V_i(j) \subsetneq \Lambda_j$; in this case we call λ_j a parent of λ_i. Conditional hyperparameters can in turn be parents of other conditional hyperparameters, giving rise to a tree-structured space [10] or, in some cases, a directed acyclic graph (DAG) [9]. Given such a structured space $\mathbf{\Lambda}$, the (hierarchical) hyperparameter optimization problem can be written as:

$$\lambda^* \in \underset{\lambda \in \mathbf{\Lambda}}{\operatorname{argmin}} \frac{1}{k} \sum_{i=1}^{k} \mathcal{L}\left(A_\lambda, \mathcal{D}_t^i, \mathcal{D}_v^i\right), \tag{2}$$

In this study we consider the more general combined algorithm selection and hyperparameter optimization (CASH). That is we intend to optimize both problems at the same time.

III. Literature Review

Automatic model selection has been of research interest since the uprising of deep learning. This is no surprise since selecting an effective combination of algorithm and hyperparameter values is currently a challenging task requiring both deep machine learning knowledge and repeated trials. This is not only beyond the capability of layman users with limited computing expertise, but also often a non-trivial task even for machine learning experts [11].

To make machine learning accessible to non-expert users, researchers have proposed various automatic selection methods for machine learning algorithms and/or hyperparameter values for a given supervised machine learning problem. These methods' goal is to find, withind a pre-specified resource limit (usually specified in terms of time, number of algorithms and/or combintations of hyperparameter values), an effective algorithm and/or combination of hyperparameter values that maximize the accuracy measure on the given machine learning problem and data set. Using an automatic selection method method, the machine learning practitioner can skip the manual and iterative process of selecting and efficient combination of of hyperparameter values and neural network model, which is high labor intensive and requires a high skill set in machine learning.

In the recent years a number of tools have been made available for users to automate the model selection and/or hyperparameter tuning, in the following we present a brief survey of the most popular methods.

A. AutoWEKA

Auto-WEKA [12] is a system designed to help machine learning users by automatically searching through the joint

space of WEKA's learning algorithms and their respective hyperparameter settings to maximize performance using a state-of-the-art Bayesian optimization method. AutoWEKA addresses the CASH problem by treating all of WEKA as a single, highly parametric machine learning framework, and using Bayesian optimization to find a strong instantiation for a given dataset. AutoWEKA also natively supports parallel runs (on a single machine) to find good configurations faster and save the N best configurations of each run instead of just the single best. AutoWEKA is tightly integrated with WEKA and does provide support for Multilayer Perceptrons (MLP).

B. Auto-sklearn

Auto-sklearn [13] is Auto-WEKA's sister package, it uses the same Bayesian optimizer but comprises a smaller space of models and hyperparameters, however it includes additional meta-learning techniques.

C. TuPAQ

TuPAQ [11] is a system designed to efficiently and scalably automate the process of training predictive models. One of its main features is a planning algorithm which decides on an efficient parallel execution strategy during model training while identifying new hyperparameter configurations and proactively eliminating models which are unlikely to provide good results. TuPAQ is aimed at large scale machine learning, it builds on top of the well know Apache Spark. TuPaq only focuses on classification problems and considers only three model families (Support Vector Machines, Logistic Regression and nonlinear SVMs), each with several hyperparameters. TuPAQ performs batching to train multiple models simultaneously and deploys bandit resource allocation to allocate more resources to the most promising models. TuPAQ does not provide support for neural networks.

IV. An evolutionary framework for the CASH problem

While there is a number of methods for automatic model selection and hyperparameter tuning, the most popular ones still have room for improvement. In the case of AutoWEKA and Auto-sklearn they do not provide good support for large machine learning problems, nor provide support for distributed computing. TuPAQ on the other hand, provides wide support for distributed computing, maximizing the use of computational resources through the use of sophisticated optimizations, nevertheless its restricted to only classification problems and does not provide support for neural networks.

We propose to implement a system for automatically selecting the most fitting neural network architecture (only fully connected networks in the first stage) for a given problem, whether it is classification or regression. Furthermore, we plan that the system should be scalable and should be able to be used in distributed computing environments, allowing it to be usable for large datasets and complex models. To achieve the latter we propose to build our system using Ray [14] which is a distributed system designed with large scale distributed machine learning in mind.

For this work we will consider three major architectures of neural networks, namely multilayer percpetrons (MLPs) [15], convolutional neural networks (CNNs) [16] and recurrent neural networks (RNNs) [17]. Each one of these architectures can be built by stacking together a *valid* combination of any of the four following layers: fully connected layers, recurrent layers, convolutional layers and pooling layers.

We say that a neural network architecture is *valid* if it complies with the following set of rules, which we derived empirically from our practice in the field:

- A fully connected layer can only be followed by another fully connected layer
- A convolutional layer can be followed by a pooling layer, a recurrent layer, a fully connected layer or another convolutional layer
- A recurrent layer can be followed by another recurrent layer or a fully connected layer.
- The first layer is user defined according to the type of architecture chosen (MLP, CNN or RNN)
- The last layer is always a fully connected layer with either a softmax activation function for classification or a linear activation function for regression problems

A. The fitness function

In order to steer the search in the most promising search directions, a carefully designed fitness function is required. The framework's goals are to generate a neural network architecture with good predictive power for the class of problem at hand while keeping the complexity of the network as low as possible. Measuring the predictive power of the network is straightforward; having a valid neural network we can assess its predictive power by training it on the set $\mathcal{D}_t$ and then evaluating the predictions using the set $\mathcal{D}_v$. A more robust approach would be performing a k-fold cross-validation which splits the data into k equal sized partitions $\mathcal{D}_v^1, \ldots, \mathcal{D}_v^k$ and sets $\mathcal{D}_t^i = D \backslash D_v^i$ for $i = 1, \ldots, k$.

Let $A \in \mathcal{A}$ be a certain neural network architecture trained on set $\mathcal{D}_t$, let also $\mathcal{P}_A(\mathcal{D}_v)$ represent the performance of the neural network A when tested using validation set $\mathcal{D}_v$ and the user-defined performance indicator p, where p is usually any of the metrics listed in Table III . Using k-fold cross-validation the average performance of the algorithm can be written as

$$p = \frac{1}{k} \sum_{i=1}^{k} \mathcal{P}_A(\mathcal{D}_v^i) \tag{3}$$

For measuring the complexity of the architecture we consider the number of trainable weights w of the neural network which is a good indicator of how complex the architecture is.

Using p and w we propose the following fitness function

$$f = p + \lambda w, \tag{4}$$

where $\lambda \in [0, 1]$ is a scaling factor that indicates how much does the number of trainable weights w affects the overall fitness of the neural network. By setting $\lambda = 1$ the preference is given to very compact architectures, while $\lambda = 0$ will only care about architectures that find the best possible value for p regardless of their complexity.

B. Evolutionary algorithms

The main part of the framework consists of an evolutionary algorithm, evolutionary algorithms (EAs) are a family of methods for optimization problems. The methods do not make any assumptions about the problem, treating it as a black box that merely provides a measure of quality given a candidate solution. Furthermore, EAs do not require the gradient when searching for optimal solutions, making them very suitable for applications such as neural networks.

In the following we describe the very basics of evolutionary algorithms as an introduction for the reader.

Every evolutionary algorithm consists of a population of individuals (sometimes EAs are also referred as population based algorithms). Each individual in the population is indeed a potential solution to the optimization problem. Individuals are generally encoded, this encoded solution is often called a genotype while the actual representation of the genotype in the domain of the problem is referred as a fenotype, for our application the fenotype represents the neural network architecture while the genotype will be defined later on. Each solution is evaluated using a so-called fitness function, where the function represents how does the individual performs with respect to a certain metric.

At every iteration a new generation of solutions is generated by using crossover and mutation operators. Crossover operator is an evolutionary operator used to combine the information of two parents to generate new offspring while the mutation operator is used to maintain genetic diversity from one generation of the population to the next.

The basic template for an evolutionary algorithm is the following

Algorithm 1 Basic Evolutionary Algorithm

Let $t = 0$ be the generation counter
Create and initialize an n_x-dimensional population, $\mathcal{C}(0)$, to consist of n_s individuals
while stopping condition not true **do**
  Evaluate the fitness, $f(\mathbf{x}_i(t))$, of each individual, $\mathbf{x}_i(t)$
  Perform reproduction to create offspring
  Select the new population, $\mathcal{C}(t+1)$
  Advance to the new generation, i.e. $t = t + 1$
end while

Among the many different choices for evolutionary algorithms three major trends currently lead the way, we refer to the genetic algorithms (GAs), evolutionary strategies (ES) and genetic programming (GP) [15].

One of the major drawbacks of EAs is the time penalty involved in evaluating the fitness function. If the computation of the fitness function is computationally expensive, as in this case, then using any flavor of EA may be very computationally expensive and in some instances unfeasible. Micro-genetic algorithms [18] are one variant of GAs whose main difference is the use of small populations (less than 10 individuals per population). In this work we will follow general principles of micro-GA in order to reduce the computational burden of the algorithm.

Specifically speaking and taking inspiration from the micro-GA the pseudocode for our proposed algorithm is described in Algorithm 2. Let C_p and M_p be the crossover and mutation probabilities respectively, let also G_{max} be the maximum number of allowed generations and E_{max} the maximum number of repetitions for the micro-GA.

Algorithm 2 Neural Network Evolution

Let $t_e = 0$ be the experiments counter
while $t_e < E_{max}$ **do**
  Let $t_g = 0$ be the generation counter
  Create and initialize an n_x-dimensional population, $\mathcal{C}(0)$, to consist of n_s individuals, where $n_s <= 10$. See section IV-D
  while $t_g < G_{max}$ or nominal convergence not reached **do**
    Check for nominal convergence in $\mathcal{C}(t)$. See section ...
    Evaluate the fitness, $f(\mathbf{x}_i(t))$, of each individual, $\mathbf{x}_i(t)$. See section ...
    Identify best and worst individuals of $\mathcal{C}(t)$
    Replace worst individual in $\mathcal{C}(t)$ with best from $\mathcal{C}(t-1)$
    Perform selection. See section ...
    Perform crossover of individuals in $\mathcal{C}(t)$ with $C_p = 1$. Let $\mathcal{O}(t)$ be the offspring population. See section ...
    For each individual in $\mathcal{O}(t)$ perform mutation with M_p probability. See section ...
    Make $\mathcal{C}(t+1) = \mathcal{O}(t)$
    $t_g = t_g + 1$
  end while
  Append best solution from previous run to $\mathcal{B}$
  $t_e = t_e + 1$
end while
Final Solution is best existing solution in $\mathcal{B}$

C. Encoding neural networks as genotypes

In order to perform the optimization of neural network architectures a suitable encoding for the neural networks is needed. A good encoding has to be flexible enough to represent neural network architectures of variable length while also making it easy to verify the *validity* of a proposed neural network architecture.

While array based encodings are quite popular for numerical problems, they often use a fixed-length genotype. While it is possible to use an array based representation for encoding a neural network, this would require the use of very large arrays, furthermore verifying the validity of the encoded neural network is hard to achieve. Three-based representation as those used in genetic programming [15] enables more flexibility when it comes to the length of the genotype, nevertheless imposing constraints for building a valid neural network requires traversing the entire tree or making use of complex data structures every time a new layer is to be stacked in the model.

For this work, the chosen encoding is list-based, that is, the genotype is represented as a list of arrays, where the length of

the list can be arbitrary. Each array within the list represents the details of a given layer as described in Table IV. A visual depiction of the array is presented in Table I.

Layer type	Neuron number	Activation function	CNN filter size	CNN kernel size	CNN stride	Pooling size	Dropout rate

TABLE I: Visual representation of a neural network layer as an array.

Let us illustrate the proposed encoding with an example, let S_e be a model composed of several stacked layers as those shown in Table I.

$$S_e = [[1, 264, 2, 0, 0, 0, 0, 0], [5, 0, 0, 0, 0, 0, 0, 0, 0.65], [1, 464, 2, 0, 0, 0, 0, 0], [5, 0, 0, 0, 0, 0, 0, 0, 0.35], [1, 872, 2, 0, 0, 0, 0, 0], [1, 10, 3, 0, 0, 0, 0, 0]]$$

The neural network representation of the model just presented is described in Table II

Layer type	Neurons	Activation Function	Dropout Ratio
Fully connected	264	ReLU	n/a
Dropout	n/a	n/a	0.65
Fully Connected	464	ReLU	n/a
Dropout	n/a	n/a	0.35
Fully Connected	872	ReLU	n/a
Fully Connected	10	Softmax	n/a

TABLE II: Neural network model.

Encoding the neural network as a list of arrays presents two big advantages over other representations. First, the number of layers that can be stacked is, in principle, arbitrary. Second, the validity of an architecture can be verified, in constant time, every time a new layer is to be stacked to the model, this is due to the fact that in order to stack a layer in between the model one just needs to verify the previous layer and the layer ahead to check for compatibility. The rules for stacking layers together are described in Table V. As can be observed, the ability of stacking layers dynamically and verifying its correctness as a new layer is stacked allows for a powerful representation that can build several kinds of neural networks such as fully connected, convolutional and recursive.

D. Generating valid models

Generating valid models is straightforward. An initial layer type has to be specified by the user, the initial layer type can be FullyConnected, Convolutional or Recurrent. As it can be seen, defining the initial layer type effectively defines the type of architectures that can effectively be generated by the algorithm, i.e. if the user chooses FullyConected as the initial layer, all the generated models will be fully connected, if the user chooses Convolutional as initial layer the algorithm will generate Convolutional models only and so on.

Just as the initial layer type has to be defined in advance, the final/output layer is also defined in advance, in fact, all of the generated models share the same output layer. The output layer is always a FullyConnected layer, furthermore, it is generated based on the type of problem to solve (classification or regression). In the case of classification problems the number of neurons is defined by the number of classes in the problem and the softmax function is used as activation function. For regression problems the number of neurons is one and the activation function used is the linear function.

Having defined the architecture type and the output layer generating an initial model is an iterative process of stacking new layers that comply with the rules in Table V. A user defined parameter m_l is used to stop inserting new layers, every time a new layer is stacked in the model a random number $n_r \in [0, 1]$ is generated, if $n_r < m_l$ and if the current layer is compatible with the last layer (according to Table V) then no more layers are inserted. With regards to layers that have an activation function, even though in principle any valid combination is possible, for this application we choose to keep all the activations for similar layers the same across the model since this is usually the common practice.

E. Selection

In order to generate n_s offsprings $2n_s$ parents are required. The parents are chosen using a selection mechanism which takes the population $\mathcal{C}(t)$ at generation the current generation and returns a list of parents for crossover. For our application, the selection mechanism used is based on the binary tournament selection [15], [18]. A description of the mechanism is given next:

- Select n_p parents at random where $n_p < n_s$.
- Compare the selected elements in a pair-wise manner and return the most fit individuals.
- Repeat the procedure until $2n_s$ parents are selected.

It is important to note in the above procedure that the larger n_p is, the more the probable that the best individual in the population is chosen as one of the parents, this is not a desirable behavior, thus we warn the users to keep n_p small. Also, recall from Algorithm 2 that our approach uses elitism, therefore the best individual of a current generation goes unchanged in the next generation.

F. Crossover operator

Since the encoding chosen for this task is rather peculiar, the existing operators are not suitable for our encoding. In this section we describe in detail the used crossover operator. Our operator is based on the two point crossover operator for genetic algorithms [19] in the sense that two points are selected for each parent, nevertheless our operator is more restrictive in order to ensure the generation of valid architectures. The selection mechanism is described in Algorithm 3. The following algorithm will be executed for n_s times, where n_s is a user defined parameter, at most or until a valid offspring is generated. Nevertheless, based on our experience with the algorithm it usually takes only 1 attempt to successfully generate a valid offspring. Finally we would like to note that although this is the implementation we used, it may not be the only one to achieve the expected results.

In Algorithm 3 when we mean compatibility between two points we mean that such two points can be interchanged

Algorithm 3 Crossover Method

Let S_1, S_2 be the arrays containing the stacked layers of a neural network model in parents 1 and 2 respectively.
Take two random points (r_1, r_2) from S_1 where $r_1 <= r_2$
if $r_1 = r_2$ **then**
  $r_2 = len(S_1 - 1)$
else
  pass
end if
Find all the compatible pairs of points $(r_3, r_4)_i$ in S_2 that are compatible with (r_1, r_2) where $r_3 < r_4$ and $r_4 - r_3 < l_{max}$
Randomly pick any of the pairs $(r_3, r_4)_i$
Replace the layers in S_1 between r_1, r_2 inclusive with the layers in S_2 between r_3, r_4 inclusive. Label the new model as S_3
Rectify the activation functions of S_3 to match the activation functions of S_1

and still comply with the building rules stated in Table V. Let us illustrate Algorithm 3 with an example. Consider the following models

$$S_1 = [[1, 264, 2, 0, 0, 0, 0, 0], [5, 0, 0, 0, 0, 0, 0, 0.65], [1, 464, 2, 0, 0, 0, 0, 0], [5, 0, 0, 0, 0, 0, 0, 0.35], [1, 872, 2, 0, 0, 0, 0, 0], [1, 10, 3, 0, 0, 0, 0, 0]]$$

$$S_2 = [[1, 56, 0, 0, 0, 0, 0, 0], [5, 0, 0, 0, 0, 0, 0, 0.25], [1, 360, 0, 0, 0, 0, 0, 0], [1, 480, 0, 0, 0, 0, 0, 0] [1, 88, 0, 0, 0, 0, 0, 0], [5, 0, 0, 0, 0, 0, 0, 0.2], [1, 10, 3, 0, 0, 0, 0, 0]]$$

Lets take $r_1 = 1$ and $r_2 = 3$, since these points are going to be removed from the model then we need to find the compatible layers with $S_1[r_1 - 1]$ and $S_1[r_2]$ according to the rules described in Table V. Note however that if $r_1 = 0$, i.e. the initial layer, then only a layer whose layer type is equal to the layer type of $S_1[0]$ is compatible. Thus, for this example the compatible pairs of points $(r_3, r_4)_i$ are:

$$[(0,0), (0,2), (0,4), (0,5), (1,2), (1,4), (1,5), (2,2), (2,4), (2,5), (4,4), (4,5), (5,5)]$$

Now assume we pick at random the pair $(2, 4)$, thus the offspring which we will for simplicity call S_3 looks like:

$$S_3 = [[1, 264, 2, 0, 0, 0, 0, 0], [1, 360, 2, 0, 0, 0, 0, 0], [1, 480, 2, 0, 0, 0, 0, 0], [1, 88, 2, 0, 0, 0, 0, 0], [1, 872, 2, 0, 0, 0, 0, 0], [1, 10, 3, 0, 0, 0, 0, 0]]$$

which is indeed a valid model, the reader can check the actual model representations for each of the models in this example in Tables VII to IX. Notice though that all the activation functions of the same layer types are changed to match the activation functions of the first parent S_1, this is what we call activation function rectification, which basically means changing all the activation functions of the layers that share the same layer type between S_1 and S_3 to the activation functions used in S_1.

We would like to highlight one important feature of this crossover operator, namely that it has the ability to generate neural network models of different sizes, i.e. it can shrink or increase the size of one of the parents. This is a desirable behavior as in real life, machine learning experts will often try various sizes of neural network models when trying to find the one that has the best inference capabilities.

G. Mutation operator

The mutation operator is used to induce small changes to some of the models generated through the crossover mechanism. In the realm of evolutionary computation these subtle changes tend to improve the exploration properties of the current population (genetic diversity) by injecting random noise to the current solutions. Although according to [18] mutation is not needed in the micro-GA, we believe some sort of mutation is needed in our application in order to get more diverse models which could potentially lead to better inference abilities, nevertheless, our mutation approach will be less disruptive in order to mitigate its effect. Following the same ideas found on the literature we developed a mutation operator to handle neural network models.

As stated above our mutation approach is less disruptive than the common mutation operators [15], this decision follows two main reasons: First, is the fact that usually micro genetic algorithms don't make use of the mutation algorithm since the crossover operator has already induce significant genetic diversity in the population. The second reason is related to the way neural networks are usually built by human experts, commonly experts try a number of models and then make subtle changes to each of them in order to try to improve the inference ability of them, such changes usually involve changing the parameters in a layer, adding or removing a layer, adding regularization or changing the activation functions.

Based on the principles described above, our mutation process randomly chooses one layer of the model and the proceeds to make one of the following operations:

- Change a parameter of the layer chosen for a value complying the values stated in Table IV.
- Change the activation function of the layer. This would involve rectifying the entire model (described in section IV-F).
- Add a dropout layer if the chosen layer is compatible.

This operations together provide a rich set of possibilities for performing efficient mutation while still keeping valid models after mutation is performed.

H. Determining nominal convergence

Nominal convergence is one of the criteria used for early stopping of the evolutionary procedure of our algorithm. Some

literature defines the convergence in terms of the fitness of the individuals [], while in [] the convergence is defined in terms of the genotype or phenotype of the individuals. Although convergence based on the actual fitness of the individuals may be easier to asses given that the fitness is already calculated, we believe that an assessment of convergence based on the actual genotype of the individuals suits our needs better, this follows the following reasoning.

Since neural networks are stochastic in nature, we expect some variations in the fitness of the individuals at every different run, furthermore since we are running the training process for only few epochs (in order to avoid a high computational burden) the performance of the networks can be quite different and would not be a reliable indicator of convergence. Instead, to assess convergence we look at the genotype and compute the similarities between the different individuals.

To compute the similarities between different individuals we take the following approach. Let S_1, S_2 where $len(S_2) >= len(S_1)$ be the genotype representing two different models, let also $S_1 - S_2$ represent the layer-wise difference between each model and $S_i[j]$ be the array representation of the $j-th$ layer of model i, $S_2 - S_1$ is defined in Algorithm 4.

Algorithm 4 $S_2 - S_1$ Method

Let $d \in \mathbb{R}$ be the distance between the two models. Make $d = 0$
for Each layer i in S_1 except last layer **do**
  $d = d + norm_2(S_2[i] - S_1[i])$
end for
for Each remaining layer i in S_2 except last layer **do**
  $d = d + norm_2(S_2[i])$
end for
Return d

REFERENCES

[1] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[2] T. Schaul, J. Bayer, D. Wierstra, Y. Sun, M. Felder, F. Sehnke, T. Rucksties, and J. Schmidhuber. Pybrain. *JMLR*, 11:743–746, 2010.

[3] Xiangrui Meng, Joseph Bradley, Burak Yavuz, Evan Sparks, Shivaram Venkataraman, Davies Liu, Jeremy Freeman, DB Tsai, Manish Amde, Sean Owen, Doris Xin, Reynold Xin, Michael J. Franklin, Reza Zadeh, Matei Zaharia, and Ameet Talwalkar. Mllib: Machine learning in apache spark. *Journal of Machine Learning Research*, 17(34):1–7, 2016.

[4] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[5] François Chollet et al. Keras. https://github.com/fchollet/keras, 2015.

[6] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[7] Frank Seide and Amit Agarwal. Cntk: Microsoft's open-source deep-learning toolkit. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 2135–2135, New York, NY, USA, 2016. ACM.

[8] Thornton C., Hutter F., Hoos H., and Leyton-Brown K. Auto-weka: Combined selection and hyperparameter optimization of classification algorithms. In *KDD*, 2013.

[9] Hutter F., Hoos H., Leyton-Brown K, and Stutzle T. Paramils: and automatic algorithm configuration framework. *JAIR*, 36(1):267–306, 2009.

[10] Bergstra J., Bardenet R., Bengio Y., and Kegl B. Algorithms for hyper-parameter optimization. In *NIPS*, 2011.

[11] Sparks ER., Talwalkar A., Smith V., Kottalam J., Pan X, and Gonzales JE. Automated model search for large scale machine learning. In *SoCC*, pages 368–380, 2015.

[12] Thornton C., Hutter F., Hoos H., Leyton-Brown K, and Kotthoff L. Auto-weka 2.0: Automatic model selection and hyperparameter optimization in weka. *JMLR*, 2016.

[13] Feurer M., Klein A., Eggensperger K., Springenberg J., Blum M., and Hutter F. Efficient and robust automated machine learning. In *NIPS*, volume 17, pages 1–5, 2015.

[14] Philipp Moritz, Robert Nishihara, Stephanie Wang, Alexey Tumanov, Richard Liaw, Eric Liang, William Paul, Michael I. Jordan, and Ion Stoica. Ray: A distributed framework for emerging AI applications. *CoRR*, abs/1712.05889, 2017.

[15] Douglas P. Engelbrecht. *Computational Intelligence. An Introduction*. Wiley, 2007.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[17] Zachary Chase Lipton. A critical review of recurrent neural networks for sequence learning. *CoRR*, abs/1506.00019, 2015.

[18] Krishnakumar K. Micro-genetic algorithms for stationary and non-stationary function optimization. In *SPIE Proceedings: Intelligent Control and Adaptive Systems*, pages 289–296, 1989.

[19] Holland J. *Adaptation in Natural and Artificial Systems*. MIT Press, 1992.

Metric name	Definition
Root Mean Squared Error	Regression
Accuracy	Classification
Precision	Classification
Recall	Classification
F1	Classification

TABLE III: Common performance metrics for neural networks

Cell number	Cell name	Data Type	Represents	Applicable to	Values
0	Layer type	Integer	The type of layer. See table V	MLP/RNN/CNN	$x \in \{1, \ldots, 5\}$
1	Neuron number	Integer	Number of neurons/units in the layer	MLP/RNN	$8 * x$ where $x \in \{1, \ldots, 128\}$
2	Activation function	Integer	Type of activation function. See table VI	MLP/RNN/CNN	$x \in \{1, \ldots, 4\}$
3	Filter size	Integer	Number of filters generated by the layer	CNN	$8 * x$ where $x \in \{1, \ldots 64\}$
4	Kernel size	Integer	Size of the kernel used for convolutions	CNN	3^x where $x \in \{1, \ldots, 6\}$
5	Stride	Integer	Stride used for convolutions	CNN	$x \in \{1, \ldots, 6\}$
6	Pooling size	Integer	Size for the pooling operator	CNN	2^x where $x \in \{1, \ldots 6\}$
7	Dropout rate	Float	The dropout rate applied to the following layer	MLP/RNN/CNN	$x \in [0, 1]$

TABLE IV: Details of the representation of a neural network layer as an array.

Layer type	Layer name	Can be followed by
1	Fully connected	$[1, 5]$
2	Convolutional	$[1, 2, 3, 5]$
3	Pooling	$[1, 2]$
4	Recurrent	$[1, 4]$
5	Dropout	$[1, 2, 4]$

TABLE V: Neural network stacking/building rules.

Index	Activation function
0	Sigmoid
1	Hyperbolic tangent
2	ReLU

TABLE VI: Available activation functions.

Layer type	Neurons	Activation Function	Dropout Ratio
Fully connected	264	ReLU	n/a
Dropout	n/a	n/a	0.65
Fully Connected	464	ReLU	n/a
Dropout	n/a	n/a	0.35
Fully Connected	872	ReLU	n/a
Fully Connected	10	Softmax	n/a

TABLE VII: Neural network model corresponding to S_1.

Layer type	Neurons	Activation Function	Dropout Ratio
Fully connected	56	Sigmoid	n/a
Dropout	n/a	n/a	0.25
Fully Connected	360	Sigmoid	n/a
Fully Connected	480	Sigmoid	n/a
Fully Connected	80	Sigmoid	n/a
Dropout	n/a	n/a	0.20
Fully Connected	10	Softmax	n/a

TABLE VIII: Neural network model corresponding to S_2.

Layer type	Neurons	Activation Function	Dropout Ratio
Fully connected	264	ReLU	n/a
Fully Connected	360	ReLU	n/a
Fully Connected	480	ReLU	n/a
Fully Connected	88	ReLU	n/a
Fully Connected	872	ReLU	n/a
Fully Connected	10	Softmax	n/a

TABLE IX: Neural network model corresponding to S_3.