




# Neural Networks as Artificial Specifications

I. S. Wishnu B. Prasetya<sup>(✉)</sup>  and Minh An Tran

Utrecht University, Utrecht, The Netherlands  
s.w.b.prasetya@uu.nl

**Abstract.** In theory, a neural network can be trained to act as an artificial specification for a program by showing it samples of the programs executions. In practice, the training turns out to be very hard. Programs often operate on discrete domains for which patterns are difficult to discern. Earlier experiments reported too much false positives. This paper revisits an experiment by Vanmali et al. by investigating several aspects that were uninvestigated in the original work: the impact of using different learning modes, aggressiveness levels, and abstraction functions. The results are quite promising.

**Keywords:** Neural network for software testing · Automated oracles

## 1 Introduction

Nowadays, many systems make use of external services or components to do some of their tasks, allowing services to be shared, hence reducing cost. However, we also need to take into account that third parties services may be updated on the fly as our system is running in production. If such an update introduces an error, this may affect the correctness of our system as well. One way to guard against this is by doing run time verification [2]: at the runtime the outputs of these services are checked against their formal specifications. Unfortunately, in practice it is hard to persuade developers to write formal specifications.

A more pragmatic idea is to use ‘artificial specifications’ generated by a computer. Another use case is automated testing. Tools like QuickCheck, Evosuite, and T3 [3, 6, 13] are able to generate test inputs, but if no specification is given, only common correctness conditions such as absence of crashes can be checked. Using artificial specifications would extend their range.

Although we cannot expect a computer to be able to on its own specify the intent of a program, it can still try to guess this intent. One way to do this is by observing some training executions to predict general properties of the program, e.g. in the form of ‘invariants’ (state properties) [5], finite state machine [12], or algebraic properties [4]. These approaches cannot however capture the full functionality of a program, e.g. [5] can only infer predefined families of predicates,

many are simple predicates such as  $o \neq \text{null}$  and  $x + y \geq 0$ . With respect to these approaches, neural networks offer an interesting alternative, since they can be trained to simulate a function [9].

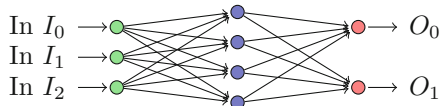
The trade off of using artificial specifications is the additional overhead in debugging. When a production-time execution violates such a specification, the failure may be either caused by an error triggered by the execution, or by an error in the training executions that were reflected in the predictions, or due to inaccuracy of the predictions. The first two cases expose errors (though the second case would take more effort to debug). However, the failure in the last case is a false alarm (false positive). Since we do not know upfront if a violation is a real error or a false positive, we will need to investigate it (debugging), which is quite labour intensive. If it turns out to be a false positive, the effort is wasted. Despite the potential, studies on the use of neural networks as artificial specifications are few: [1, 10, 11, 14]. They either reported unacceptably high rate of false positives, or do not address the issue.

In this paper we revisit an experiment by Vanmali et al. [14] that revealed  $\approx 16\%$  rate of false positives—a rate of above 5% is likely to render any approach unusable in practice. The challenge lies in the discrete nature of the program used as the experiment subject, making it very hard to train a neural network. This paper explores several aspects that were left uninvestigated in the original work, namely the influence of different learning modes, aggressiveness levels, and abstraction. The results are quite promising.

## 2 Neural Network as an Artificial Specification

Consider a program  $P$  that behaves as a function  $I \rightarrow O$ . An artificial *specification*  $\phi$  is a predicate  $I \times O \rightarrow \text{bool}$ ;  $\phi(x, P(x)) = \text{T}$  means that  $P$ 's output is judged as correct, and else incorrect. With respect to the intended specification  $\mathcal{G}$ ,  $\phi$ 's judgment is a *true positive* is when both  $\phi$  and  $\mathcal{G}$  judge a T, a *true negative* is when they agree on the judgement F, a *false positive* is when  $\phi$  judges F and  $\mathcal{G}$  judges T, and a *false negative* is when  $\phi$  judges T and  $\mathcal{G}$  judges F.

An *neural network* (NN) is a network of ‘neurons’ [9] that behaves as a function  $\mathbb{R}^M \rightarrow \mathbb{R}^N$ . We will restrict ourselves to *feed forward NNs* (FNNs) where the neurons are organized in linearly ordered layers [9]; an example is below:



The first layer is called the *input layer*, consisting of  $M$  neurons connected to the inputs. The last layer is the *output layer*, consisting of  $N$  neurons that produce the outputs. The layers in between are called *hidden layers*. An input neuron simply passes on its input, else it has  $k$  inputs and an additional input called ‘bias’ whose value is always 1 [9]. Each input connector has a weight  $w_i$ . The neuron’s output is the weighted sum of its inputs, followed by applying a so-called *activation*

function:  $out = f(\sum_{0 \leq i \leq k} w_i \cdot x_i)$ . A commonly used  $f$  is the logistic function, which we also use in our experiments.

Any continuous numeric function  $\mathbb{R}^M \rightarrow \mathbb{R}^N$ , restricted within any closed subset of  $\mathbb{R}^M$ , can be simulated with arbitrary accuracy by an FNN [7], which implies that an FNN can indeed act as an artificial specification for  $P$ , if  $P$  is injectable into such a numeric function. That is, there exists a continuous numeric function  $F: \mathbb{R}^M \rightarrow \mathbb{R}^N$  and injections  $\pi_I: I \rightarrow \mathbb{R}^M$  and  $\pi_O: O \rightarrow \mathbb{R}^N$  such that  $F$  encodes  $P$ : for all  $x \in I$ ,  $P(x) = \pi_O^{-1}(F(\pi_I(x)))$ . However, finding a right FNN is hard. A common technique to find one is by training an FNN using a set of sample inputs and outputs, e.g. using the back propagation [9] algorithm. It might be easier to train the NN to simulate  $\alpha \circ P$  instead, where  $\alpha$  is some chosen abstraction on  $P$ 's output values. The trade off is that we get a weaker specification.

Since an NN does not literally produce a `bool`, we couple its output vector  $\bar{z}' = NN(\pi_I(\bar{x}))$  to a so-called *comparator*  $\mathcal{C}: \mathbb{R}^N \rightarrow \mathbb{R}^N \rightarrow \text{bool}$  to calculate the judgement by comparing  $\bar{z}'$  with the observed output  $\bar{z} = \pi_O(\alpha(P(\bar{x})))$ . Basically, if their values are ‘far’ from each other, then the judgement is `F`, and else `T`. By adjusting what ‘far’ means we can tune the specification’s aggressiveness without having to tamper with the NN’s internals. In our experiments (below), the identity function  $id = (\lambda x . x)$  will be used as the injector  $\pi_I$  and  $\pi_O$ . Because  $id$  simply passes on its input, it will be omitted from the formulas.

### 3 Experiments

Figure 1 shows a credit approval program from the financial domain that was used as the experiment subject by Vanmali et al. [14]. The program takes 8 input parameters describing a customer. The output is a pair  $(b, y)$  where  $b$  is a boolean indicating whether the credit request is approved, and if so  $y$  specifies the maximum allowed credit. We will ignore  $b$  since [14] already shows that an FNN can accurately predict its value. Despite its size, the subject is quite challenging for an NN to simulate because it operates on a discrete domain (the numeric values are all integers). The whole input domain has 224000 possible values. We will use an FNN with 8 inputs (representing `approve`'s inputs) and a hidden layer with 24 neurons (adding more layers and neurons does not really improve the FNN's accuracy).

Five variations of the FNN will be used, as listed below, along with the used comparator  $\mathcal{C}$ .  $\mathcal{C}$  is parameterized with aggressiveness level  $A$  (integer 0 (least aggressive) ... 5) that determines  $\mathcal{C}$ 's policy to deal with non clear-cut cases.

1. The FNN direct has one output, which is trained to simulate  $y$ . Its comparator  $\mathcal{C}_A$  uses Euclidian distance, with sensitivity linearly scaled by  $A$ :  $\mathcal{C}_A(y, y') = |y - y'| < \epsilon_{max} - 0.01A$ , with  $\epsilon_{max} = 0.09$ .

```

1 approve(Citizenship , State , Region , Sex , Age , Marital , Dependents , Income) {
2   if (Region==5 || Region==6) Amount=0 ;
3   else if (Age<18) Amount=0 ;
4   else {
5     if (Citizenship==0) {
6       Amount = 5000+1000*Income ;
7       if (State==0)
8         if (Region==3 || Region==4) Amount = Amount*2 ;
9         else Amount = (int)(Amount*1.50) ;
10      else Amount = (int)(Amount*1.10) ;
11      if (Marital==0)
12        if (Dependents>0) Amount = Amount+200*Dependents ;
13        else Amount = Amount+500;
14      else Amount = Amount+1000 ;
15      if (Sex==0) Amount = Amount+500 ;
16      else Amount = Amount+1000;
17    }
18    else {
19      Amount = 1000 + 800 * Income;
20      if (Marital==0)
21        if (Dependents>2) Amount = Amount+100*Dependents ;
22        else Amount = Amount+100 ;
23      else Amount = Amount+300 ;
24      if (Sex==0) Amount = Amount+100 ;
25      else Amount = Amount+200 ;
26    }
27    if (Amount==0) Approved=F else Approved=T;
28    return (Approved , Amount); }

```

**Fig. 1.** The experiment subject: a credit approval program from [14].

2. The FNN  $\text{uni}_N$  has  $N$  outputs, trained to simulate  $\alpha_N \circ \text{approve}$ . The abstraction  $\alpha_N$  maps  $\text{approve}$ 's  $y$  output to a vector  $\bar{z} : [0.0..1.0]^N$  representing one of  $N$  uniform sized intervals in  $y$ 's range  $[0..18000]$ , such that the  $k$ -th interval is represented by a vector of 0's except a single 1 at the  $k$ -th position. If  $\bar{v} : [0.0..1.0]^N$ , let  $\text{winner}(\bar{v})$  be the index of the greatest element in  $\bar{v}$ . The comparator is more complicated. An obvious case is when  $\bar{z}' = \text{NN}(\bar{x})$  and  $\bar{z} = \alpha_{10}(\text{approve}(\bar{x}))$  report the same winner. If the NN's winner is confident of itself,  $\text{approve}$ 's output is judged as correct. When they produce different winners and the NN's winner is confident of itself, we judge  $\text{approve}$  to be incorrect. Other cases are non-clear-cut and judged depending on the aggressiveness level. The full definition of  $\mathcal{C}_A$  is shown below. The original work Vanmali et al. [14] only uses  $A = 3$  aggressiveness level.

```

function  $\mathcal{C}_A(\bar{z}, \bar{z}')$ 
   $k, j \leftarrow \text{winner}(\bar{z}), \text{winner}(\bar{z}')$  ;  $\text{agree} \leftarrow k = j$ 
  if  $\text{agree} \wedge |\text{agree} - \bar{z}'_j| < th_{low}$  then (obvious match) T
  else if  $\neg \text{agree} \wedge |\text{agree} - \bar{z}'_j| > th_{high}$  then (obvious mismatch) F
  else (non-clear-cut cases) case  $A$  of
    0 : (least aggressive: always accept) T
    1 : (reject when the NN contradicts agreement)  $\neg(\text{agree} \wedge |\text{T} - \bar{z}'_j| > th_{high})$ 
    2 : (always accept on agreement)  $\text{agree}$ 
    3 : (Vanmali et al. [14]: accept on conflicting results)  $\neg \text{agree} \vee |\text{T} - \bar{z}'_j| > th_{high}$ 
    4 : (only accept if NN's winner supports  $\bar{z}$ )  $|\text{agree} - \bar{z}'_j| < th_{low}$ 
    5 : (most aggressive: never accept) F
  end function

```

The thresholds  $th_{low}$  and  $th_{high}$  are set to 0.2/0.8.

3. The FNN  $\text{unimin}_N$  is a less presumptuous variant of  $\text{uni}$ , with  $th_{low}/th_{high}$  set to 0.1/0.9. This will cause more cases to be regarded as non-clear-cut.
4. The FNN  $\text{lower}_N$  is like  $\text{uni}_N$ , but trained to simulate  $\alpha_N \circ \text{low} \circ \text{approve}$ .  $\text{low}$  is used to ‘stretch’  $\alpha_N$  to divide  $y$  into finer intervals in the lower region of  $y$ ’s range, e.g. if we believe the region to be more error prone, and growing coarser towards the other end. We use the log function to do this:  $K * \log(1 + y/a)$  with  $K = 8000$  and  $a = 100$  controlling the steepness.
5. The FNN  $\text{center}_N$  is like  $\text{uni}_N$ , but trained to simulate  $\alpha_N \circ \text{ctr} \circ \text{approve}$ .  $\text{ctr}$  is used to ‘stretch’  $\alpha_N$  to divide  $y$  into finer intervals in the center region of  $y$ ’s range. We use logistic function  $\text{ctr}(y) = M/(1 + e^{-a(y-0.5M)})$  where  $M = 18000$  ( $y$ ’s maximum) and  $a = 0.0006$  control the function’s steepness.

**Training.** We randomly generate 500 distinct inputs (from the space of 224000 values) and collect the corresponding  $\text{approve}$ ’s outputs. This set of 500 pairs (input,output) forms the training data. For every type of FNN above and every aggressiveness level an FNN is trained.  $N$  controls the granularity of the used abstraction, so we also try various  $N$  (10..60). For each FNN, the connections’ weight is randomly initialized in  $[-0.5..0.5]$ . The training is done in a series of epochs using the back propagation algorithm [9]. We tried both the incremental learning mode [8, 9], where the FNN’s error is propagated back after each training input, and batch learning modes, where only the average error is propagated back, after the whole batch of training inputs (500 of them). Incremental learning is thus more sensitive to the influence of individual inputs.

**Evaluation.** To evaluate the FNNs’ ability to detect errors, we run them on 21 erroneous variations (mutants) of the subject as in [14]—due to limited space they are not shown here. For each mutant, 500 distinct random inputs are generated, whose outputs are ‘error exposing’ (distinguishable from the corresponding outputs of the correct subject). As an artificial specification, an FNN should ideally reject all these error exposing outputs. Each rejection is a true positive. We also generate 500 distinct random inputs and feed it to the (unmutated) subject. The FNN should accept the corresponding outputs—each rejection is a false positive.

Figure 2 shows some of the results. Except for  $\text{direct}$ , the training was done in 1500 epochs with learning rate 0.5. We can see that using abstraction improves the FNN’s performance: compare  $\text{direct}$  with  $\text{uni}_{30}$ . The latter obtains a true positive rate 68% on aggressiveness 2, implying that out of two erroneous executions,  $\text{uni}_{30}$  is likely to detect at least one, while when the aggressiveness level is set low, its rate of false positives is only around 2%. Abstraction also makes training easier: after 1500 epochs  $\text{uni}_{30}$  produces a mean square error (MSE) of  $\approx 0.0001$ , whereas the shown results for  $\text{direct}$  is obtained after 10000 epochs (incrementally) with 0.1 learning rate, yielding an  $\text{MSE} \approx 0.0004$ .

The experiment in [14] uses  $\text{unimin}_{10}$ . We believe [14] used batch learning because the reported MSE after 1500 epochs matches, namely  $\approx 0.05$ . However, as can be seen in Fig. 2, this leads to poor performance ( $\text{batched unimin}_{10}$ ). Incremental learning yields a much more accurate FNN ( $\approx 0.0001$  MSE), hence also better performance ( $\text{unimin}_{10}$ ). The performance of the FNN in [14] under our setup is indicated by the  $\text{vanmali}$ -markers in Fig. 2.

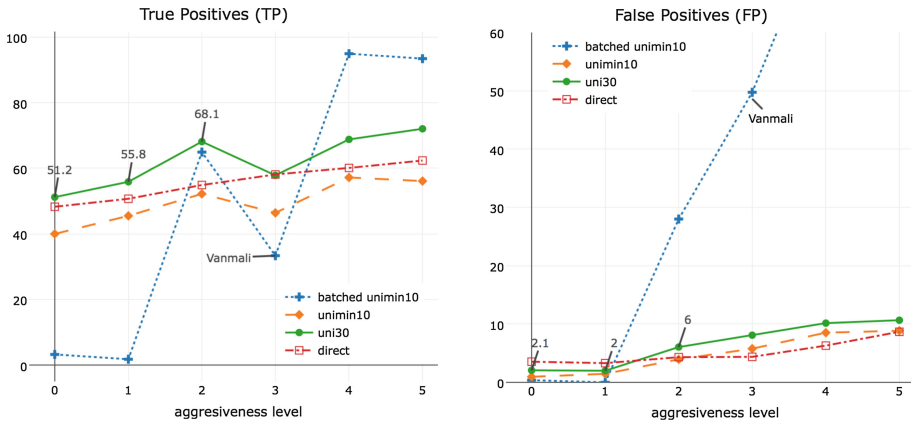


Fig. 2. The true positive and false positive rates (in %) of different FNNs.

The effect of using different abstractions and abstraction granularity (the  $N$  parameter) is shown in Fig. 3. Based on the results in Fig. 2, we now use the lowest aggressiveness level (0). The graph of uni shows that increasing  $N$  can greatly improve the FNN’s ability to detect error, while keeping the false positive rate below 5%. We also see  $\alpha_N$  and  $\alpha_N \circ low$  perform significantly better than  $\alpha_N \circ ctr$ , implying that the choice of the abstraction function matters. Compared to  $\alpha_N$ ,  $\alpha_N \circ low$  and  $\alpha_N \circ ctr$  introduce non-linear granularity. The results suggest that introducing more granularity in the region (of  $P$ ’s output) which is more error prone pays off.

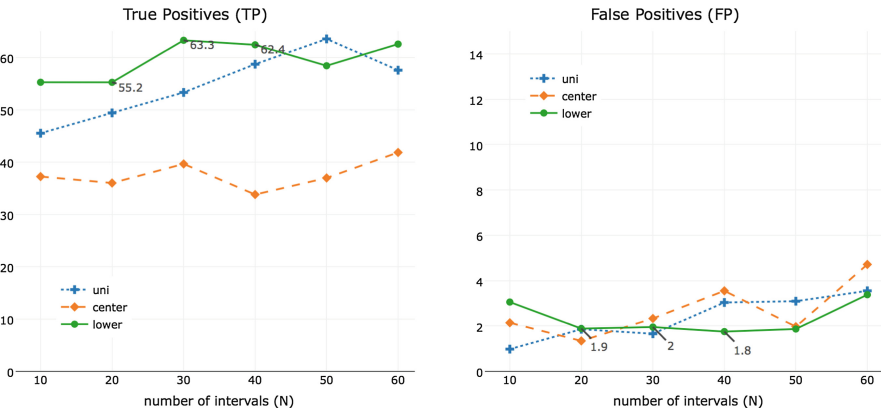


Fig. 3. The effect of different abstractions and the abstraction granularity ( $N$ ).

## 4 Conclusion

The experiment showed that, contrary to earlier attempts, it is possible to train Neural Networks, given an appropriate abstraction, to become an artificial specification for a non-trivial discrete-domain program with acceptable precision. As future work, more case studies are needed to see how this generalizes.

## References

1. Aggarwal, K., Singh, Y., Kaur, A., Sangwan, O.: A neural net based approach to test oracle. *ACM SIGSOFT Softw. Eng. Notes* **29**(3), 1–6 (2004)
2. Cao, T.D., Phan-Quang, T.T., Felix, P., Castanet, R.: Automated runtime verification for web services. In: *International Conference on Web Services (ICWS)*. IEEE (2010)
3. Claessen, K., Hughes, J.: QuickCheck: a lightweight tool for random testing of Haskell programs. In: *ACM SIGPLAN International Conference on Functional Programming* (2000)
4. Elyasov, A., Prasetya, W., Hage, J., Rueda, U., Vos, T., Condori-Fernández, N.: AB=BA: execution equivalence as a new type of testing oracle. In: *30th ACM Symposium on Applied Computing*. ACM (2015)
5. Ernst, M., et al.: The Daikon system for dynamic detection of likely invariants. *Sci. Comput. Program.* **69**(1), 35–45 (2007)
6. Fraser, G., Arcuri, A.: EvoSuite: automatic test suite generation for object-oriented software. In: *SIGSOFT FSE*, pp. 416–419 (2011)
7. Goodfellow, I., Bengio, Y., Courville, A.: *Deep Learning*. MIT Press, Cambridge (2016)
8. Joelfself: FANN C# NeuralNet float. <http://joelfself.github.io/FannCSharp>
9. Kriesel, D.: A brief Introduction on Neural Networks. [dkriesel.com](http://dkriesel.com) (2007)
10. Lu, Y., Ye, M.: Oracle model based on RBF neural networks for automated software testing. *Inf. Technol. J.* **6**(3), 469–474 (2007)
11. Mao, Y., Boqin, F., Li, Z., Yao, L.: Neural networks based automated test oracle for software testing. In: King, I., Wang, J., Chan, L.-W., Wang, D.L. (eds.) *ICONIP 2006 Part III*. LNCS, vol. 4234, pp. 498–507. Springer, Heidelberg (2006). [https://doi.org/10.1007/11893295\\_55](https://doi.org/10.1007/11893295_55)
12. Mariani, L., Pastore, F.: Automated identification of failure causes in system logs. In: *19th International Symposium on Software Reliability Engineering*. IEEE (2008)
13. Prasetya, I.S.W.B.: T3i: a tool for generating and querying test suites for Java. In: *10th Joint Meeting on Foundations of Software Engineering (FSE)*. ACM (2015)
14. Vanmali, M., Last, M., Kandel, A.: Using a neural network in the software testing process. *Int. J. Intell. Syst.* **17**(1), 45–62 (2002)