


# *Efficient Computation of Robustness of Con- volutional Neural Net- works*

Paper report.

---

Wasim Essbai



# Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Background and basic definitions</b>	<b>2</b>
2.1	The limits of the uniform sampling approach . . . . .	3
<b>3</b>	<b>Adaptive Ssampling by Parabolic Estimation</b>	<b>3</b>
3.1	Maximum Error Estimation of the Computed Robustness . . . .	4

# 1 Introduction

Validation of CNNs is based mainly on their robustness, i.e., their ability to correctly classify perturbed input data. The idea is then to check CNN’s accuracy over different datasets, obtained from the original one by perturbing it. The problem is that this approach is time consuming, since data can be altered in many ways. The described paper aims to present an efficient technique to compute robustness of a CNN by selecting only the relevant alteration levels.

Moreover, the alteration of data should be done considering real alterations, coming from the domain knowledge, and not by exploiting the internal structure of the network. However, for a correct robustness estimation, input data have to be perturbed at different level, and this is time and resource computing. The presented method, ASAP (Adaptive Sampling by Parabolic Estimation), aims to do it efficiently. It’s clear that there is some trade off between the robustness estimated and computation time. ASAP tries to estimate the accuracy curve using a parabolic approximation. By doing this it’s possible to trade off the precision of the computed robustness against the time required to compute it.

# 2 Background and basic definitions

**Definition 1:** Alteration. An alteration of type  $A$  of an input  $t$  is a transformation of  $t$  that mimics the possible effect on  $t$  when a problem during its acquisition, or in its elaboration, occurs in reality. The altered data are denoted with  $P^{A_l}$ , that is the set obtained altering the original data in  $P$  with alteration of type  $A$  and level  $l$ .

Intuitively, given an alteration interval  $[L_A, U_A]$  the robustness is the portion of that interval in which the CNN still has an acceptable accuracy.

**Definition 2:** Robustness. Let be  $\Theta$  a threshold representing the minimum accepted accuracy. The robustness of a CNN  $C$  w.r.t. alteration of type  $A$  in the range  $[L_A, U_A]$  is:

$$rob_A(C, P) = \frac{\int_{L_A}^{U_A} H(acc(C, P^{A_l}) - \Theta) dl}{L_A - U_A}$$

$$\text{where } H(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}.$$

Computing the robustness using the above formula is clearly hard, and sometimes also not feasible. A naive solution is to uniformly sample in  $[L_A, U_A]$  and compute the accuracy for the sampled points. More formally:

**Definition 3:** (Uniform robustness). Given  $n$  equidistributed points  $SP = \{l_1, \dots, l_n\}$  sampled in the interval  $[L_A, U_A]$ , the uniform robustness is defined as:

$$rob_A(C, P) = \frac{n_{acc}}{n} = \frac{|\{l \in SP | acc(C, P^{A_l}) \geq \Theta\}|}{|SP|}$$

Numerical integration approaches are based also on points sampling, but the user has no control over the sampled points. This may result in oversampling not relevant areas and/or undersampling relevant ones.

## 2.1 The limits of the uniform sampling approach

Following this approach the total time required to perform robustness analysis is  $t_{tot} = n \cdot k \cdot t_A$ , where  $n$  is the number of alteration levels,  $n$  the inputs number and  $t_A$  the time required for applying the alteration  $A$  to single input. In general, it's possible to act on  $n$  to reduce the computational effort, but this has some drawbacks, especially when analyzing networks whose accuracy varies a lot.

Thus, a way to solve these problems is to adaptively select the points to be sampled, and not only uniformly, following the usual approach done for input values in software testing.

## 3 Adaptive Ssampling by Parabolic Estimation

The ASAP algorithm to automatically selects the points where to evaluate the accuracy. The best points to select would be those in which the accuracy curve intersects the threshold  $\Theta$ , to detect the bounds of the various intervals. However, the analytical form of the curve is not known a priori, and it is not possible to compute these intersections, so the idea is to select points as close as possible to  $\Theta$ .

The method is based on the assumption that, once computed the accuracy for two alteration levels  $A$  and  $B$ , the real curve between  $A$  and  $B$  is included in the area between two parabolas passing through the points  $A$  and  $B$ , and having concavity depth respectively  $+\hat{a}$  and  $-\hat{a}$ . If there is an intersection between that area and the threshold, and the distance between  $A$  and  $B$  is sufficiently large, then it's possible to compute the accuracy in the middle point  $M$  between  $A$  and  $B$ , and add it to the sample set. Now it's possible to apply the same procedure recursively on the intervals  $[A, M]$  and  $[M, B]$ . In this way, the number of evaluated points is adaptively determined and depends on the value of the parameter  $\hat{a}$ . Intuitively, the higher is the value of  $\hat{a}$  the higher is the number of alteration levels evaluated by the algorithm, that means that it must be chosen based on the robustness estimation accuracy needed and available resources.

At the end of the algorithm execution there is a set  $RES = \{\langle l_1, acc_1 \rangle, \dots, \langle l_n, acc_n \rangle\}$ , starting which it's possible to compute the robustness generalizing Def. 3 as follows:

$$rob_A(C, P) = \frac{\sum_{j=2}^n H(acc_j - \Theta)(l_j - l_{j-1})}{U_A - L_A}$$

It's possible to show that ASAP uses fewer alteration levels than the uniform sampling, so it saves time, but it still provides an accurate approximation of the robustness.

### 3.1 Maximum Error Estimation of the Computed Robustness

With the presented method it's possible to quantify the maximum error in the robustness estimation. Let be  $IP = \{(l_j, l_{j+1}) | \langle l_j, acc_j \rangle, \langle l_{j+1}, acc_{j+1} \rangle \in RES \wedge (parabIntsct(l_j, acc_j, l_{j+1}, acc_{j+1}, \hat{a}, \Theta) \vee parabIntsct(l_j, acc_j, l_{j+1}, acc_{j+1}, -\hat{a}, \Theta))\}$  the pairs set of two consecutive points from RES such that the parabolas passing from them with concavity depth  $\pm \hat{a}$  intersect the threshold  $\Theta$ . Intuitevely, this means that the real curve may intersect, but ASAP has quit sampling because the two points have alteration levels sufficiently close ( $l_{j+1} - l_j < minStep$ ). Therefore, assuming a correct value for  $\hat{a}$ , the error that could be done comes only from those intervals in IP. Formally

**Theorem 1.** Let C be a CNN and A a given alteration type defined in the range  $[L_A, U_A]$ . Let  $rob_A$  be the robustness computed for C and A by ASAP using a given  $\hat{a}$ . Let  $rob_A^*$  be the real robustness value. Under the assumption that  $\hat{a}$  is a suitable parameter, i.e., the real accuracy curve is included in the areas of two parabolas with concavity depth  $\hat{a}$ , the maximum error of the computed robustness has a guaranteed upper bound defined as follows:

$$|rob_A - rob_A^*| \leq \epsilon_A \quad \text{with} \quad \epsilon_A = \frac{\sum_{(l_j, l_{j+1}) \in IP} (l_{j+1} - l_j)}{|U_A - L_A|}$$