Università degli Studi di Bergamo

Dipartimento di ingegneria Gestionale, dell'informazione della produzione

Evaluating the Robustness of Bayesian Neural Networks By Exploiting Uncertainty Estimation

Relatore:

Prof. Angelo Gargantini

Correlatori:

Dott. Andrea Bombarda

Dott.ssa Silvia Bonfanti

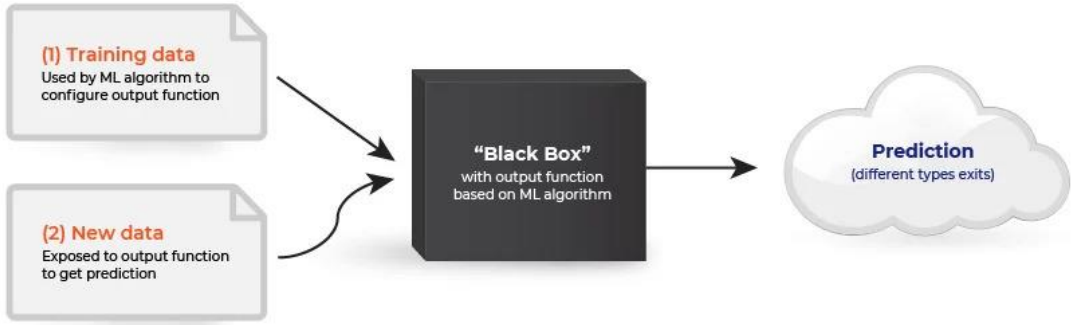
Candidato:

Wasim Essbai

Anno Accademico 2022-2023

Introduzione

- Impiego di modelli di machine learning in diversi settori, spesso eseguendo funzioni **critiche**
- Apprendimento direttamente dai dati
- Utilizzo a **scatola nera**



ma…

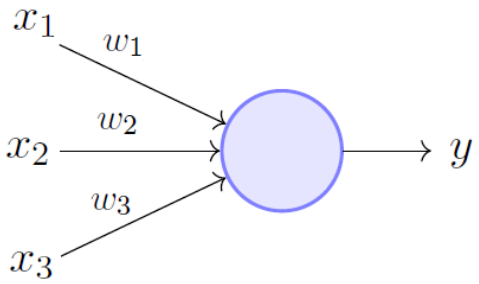
- **Se i dati non sono rappresentativi?**
- **Se ci sono perturbazioni nelle condizioni operative?**

Robustezza

Obiettivi

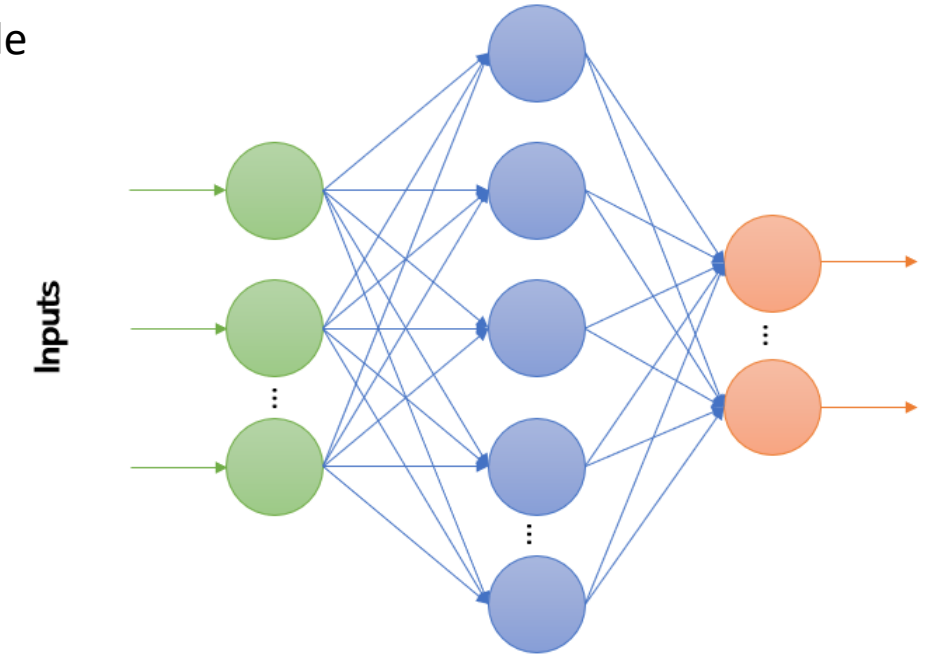
- Come valutare la robustezza delle reti neurali bayesiane (BNN)?
- Le BNN sono più robuste delle reti neurali standard?
- Si può usare la stima di incertezza per avere modelli più robusti?

Reti neurali

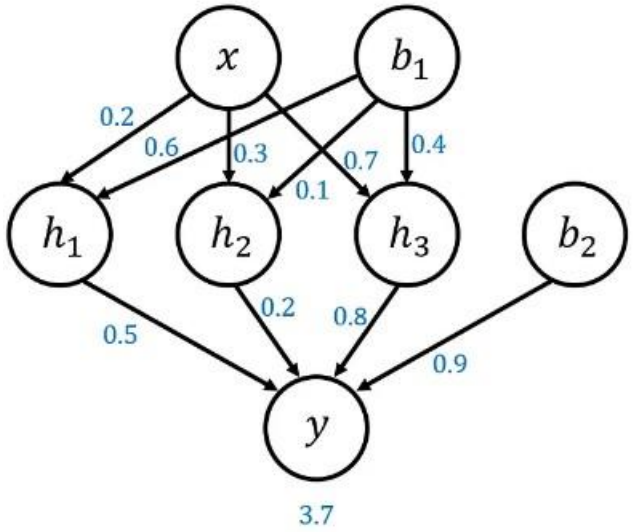


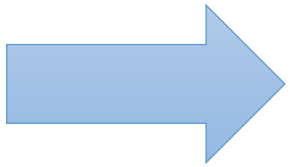
Neuron artificiale

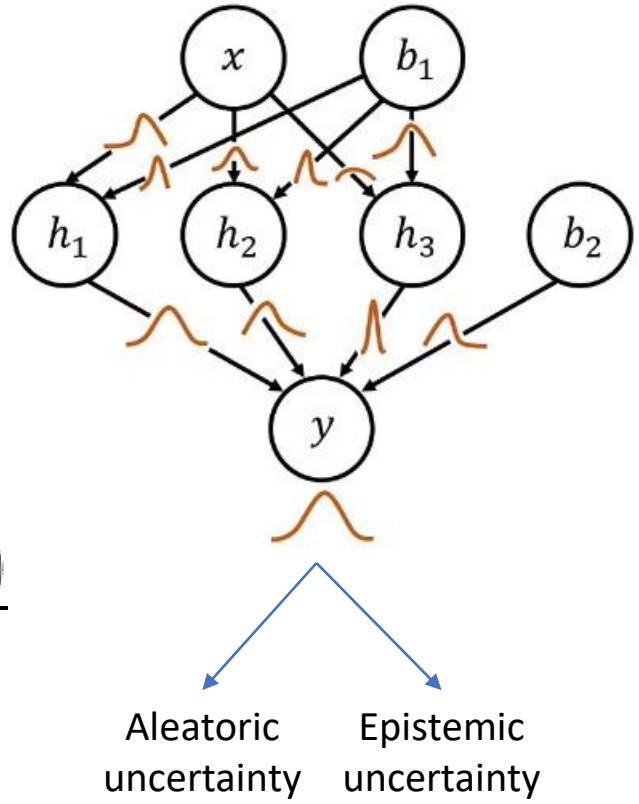
Layer



Reti neurali bayesiane





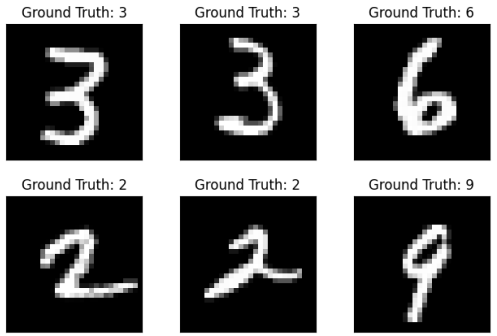


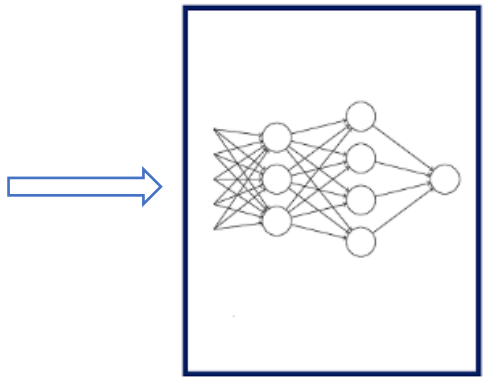
$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Teorema di Bayes

Caso di studio

- Classificazione di immagini
- Dataset: MNIST
- Architettura della rete:
  - Multilayer Perceptron (MLP)
  - Input layer: 784 neuroni
  - Hidden layer: 100 neuroni
  - Output layer: 10 neuroni

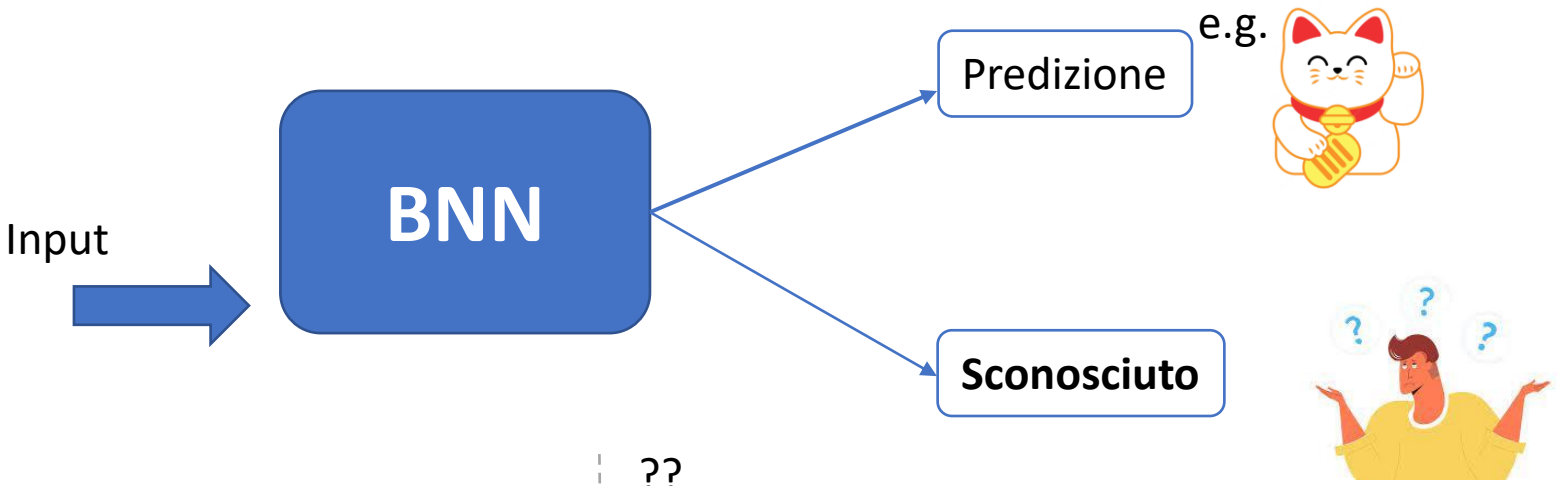


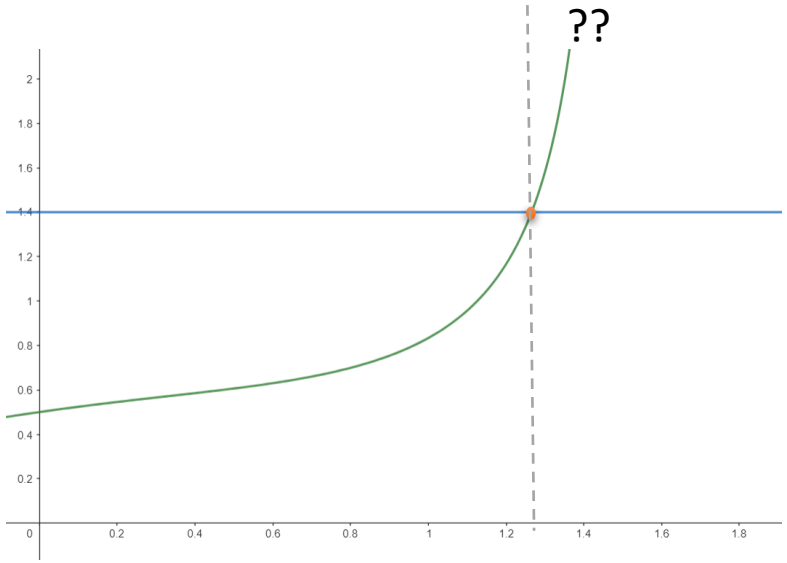


$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \\ y_{10} \end{bmatrix}$$

$Y_i \sim \text{Bernoulli}(y_i)$

Classificazione con incertezza





Idea:
Quando l'incertezza supera una certa soglia il sistema dichiara di non sapere la classe.

<u>Come definire la soglia?</u>

- Una possibile scelta è un'approssimazione lineare

$$threshold = maxUnc(1 - conf\_level)$$

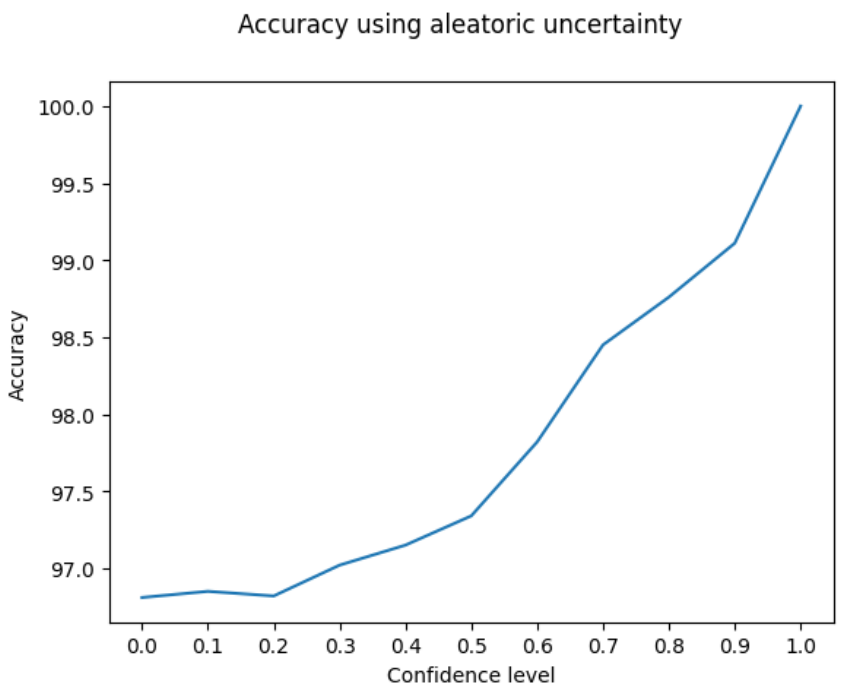
- L'accuracy diventa:

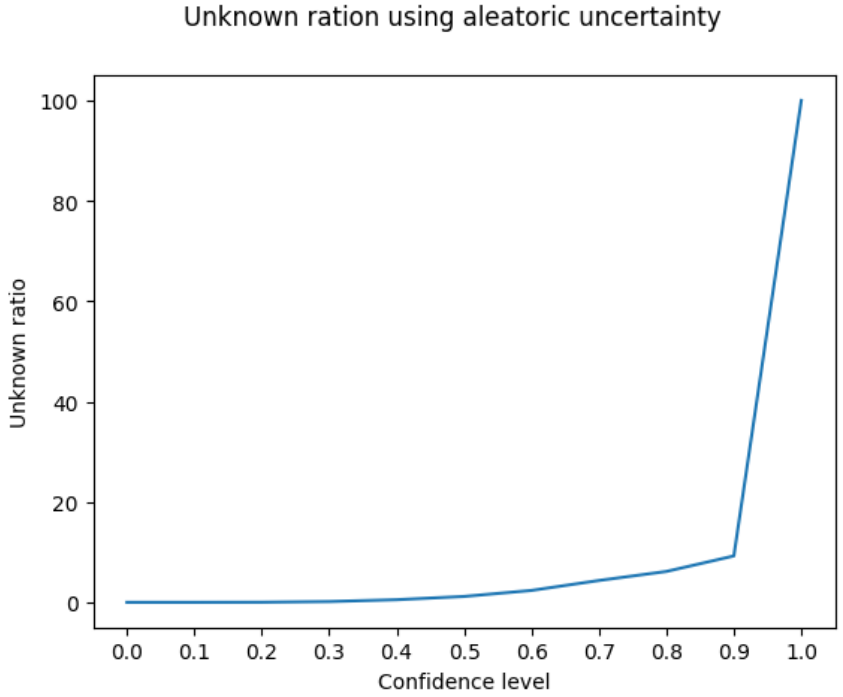
$$acc(C, D) = \frac{\#classificazioni\ corrette}{|D| - \#sconosciuti}$$

Aleatoric uncertainty

$$aleatoric = diag\{\frac{1}{T}\sum_{t=1}^{T}[diag(\hat{p}_t) - \hat{p}_t^{\otimes 2}]\}$$

$$threshold = 0.9(1 - conf\_level)$$

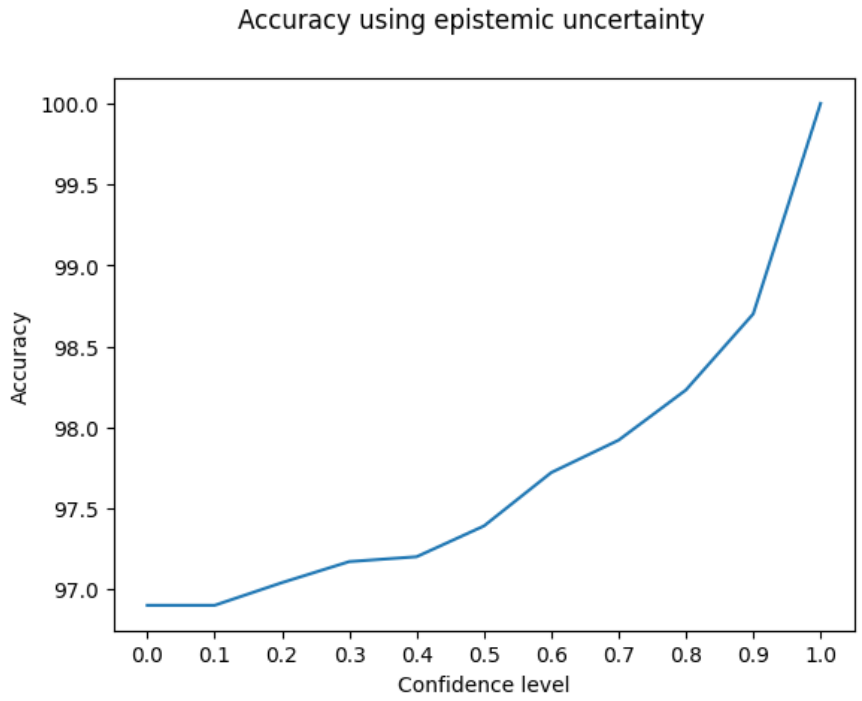


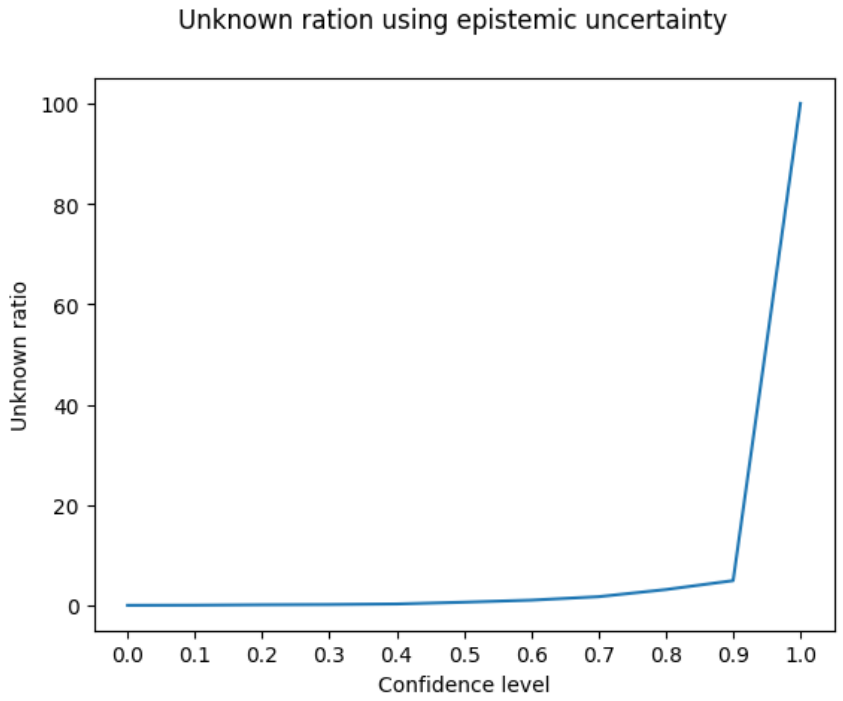


Epistemic uncertainty

$$epistemic = diag\{\frac{1}{T}\sum_{t=1}^{T}[\hat{p}_t - \bar{p}]^{\otimes 2}\}$$

$$threshold = 0.2(1 - conf\_level)$$



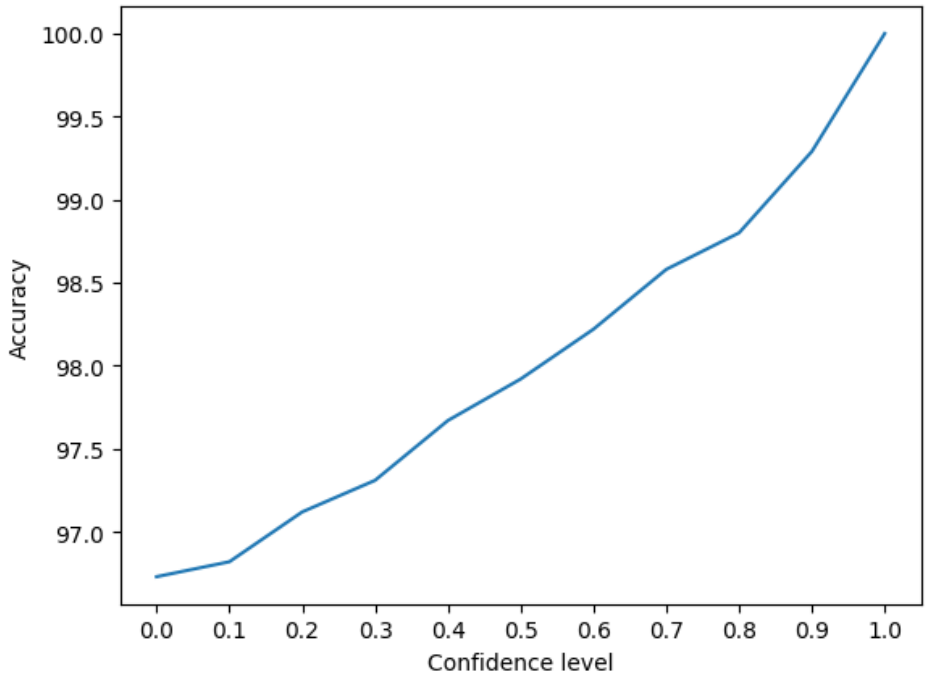


Standard deviation

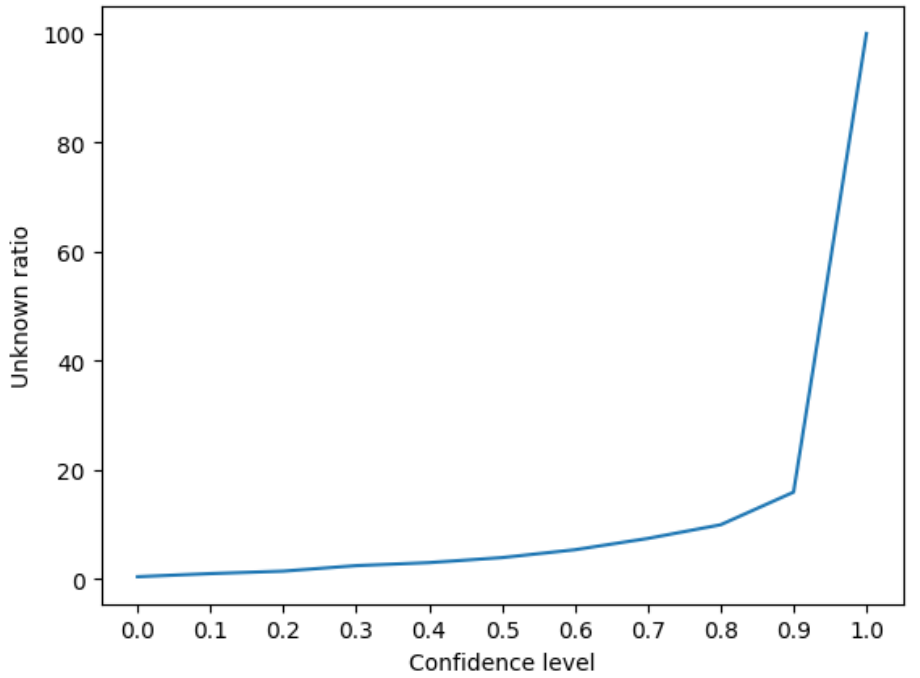
$$Var[Y] = E[(Y - E[Y])^2] = \sum_{i=0}^{9}(i - E[Y])^2 \cdot p_i = \sigma^2$$

$$threshold = 2.87(1 - conf\_level)$$

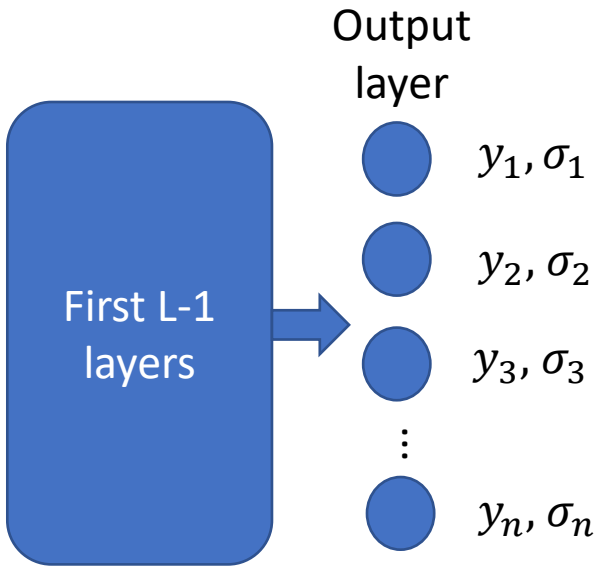
Accuracy using the standard deviation



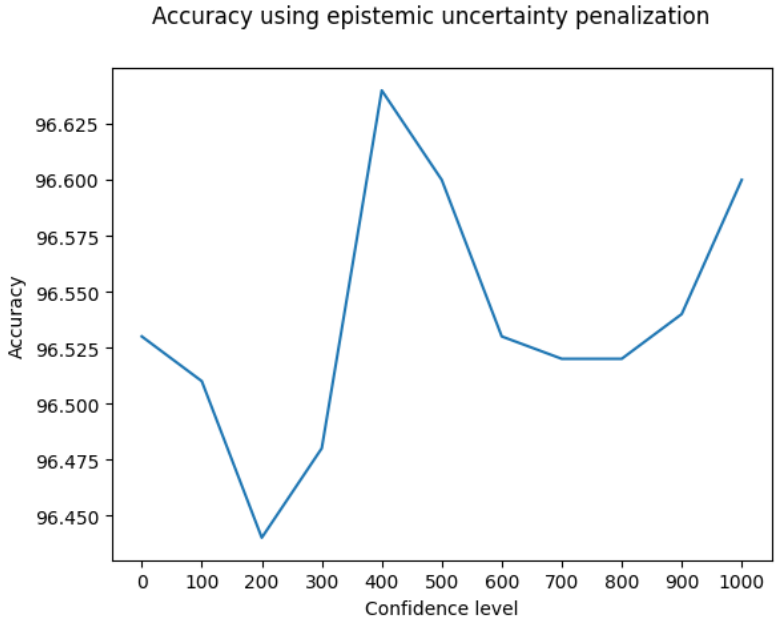
Unknown ration using the standard deviation

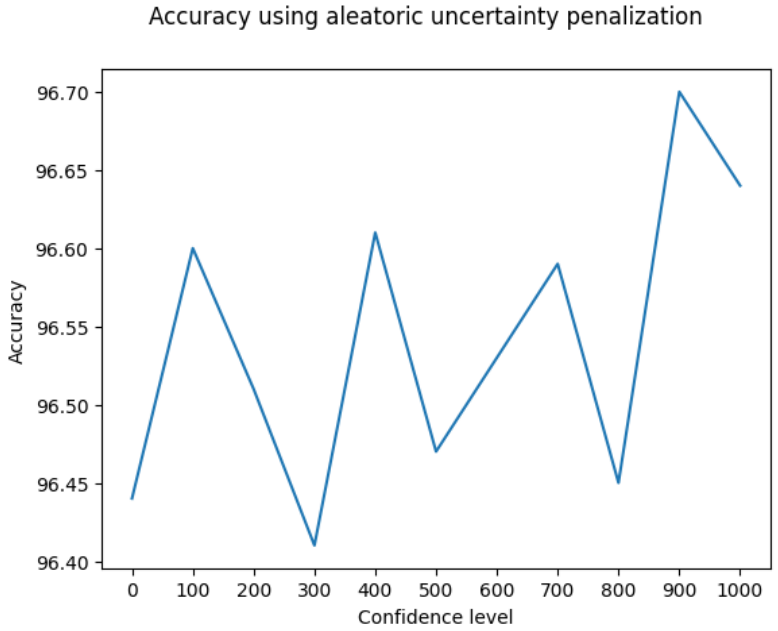


Uncertainty penalization



$$\tilde{y}_i = \frac{y_i}{\sigma_1}$$

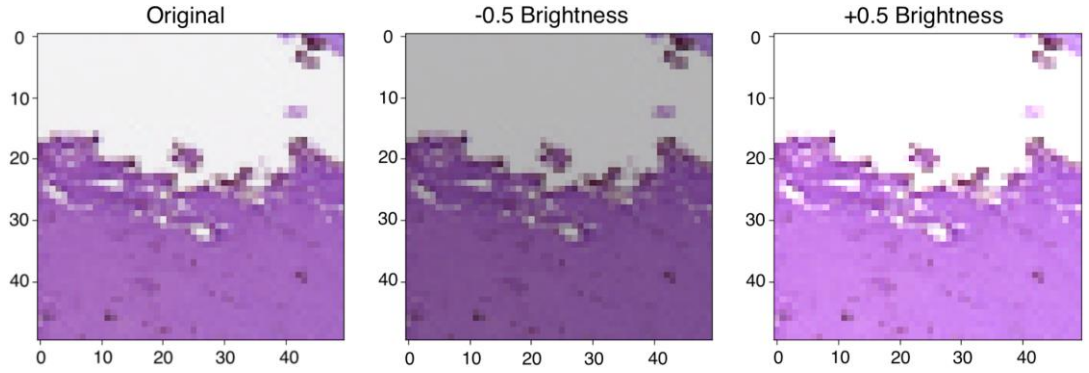




Usando questo approccio l'accuracy non migliora, bensì oscilla attorno al valore nominale.

Alterazioni

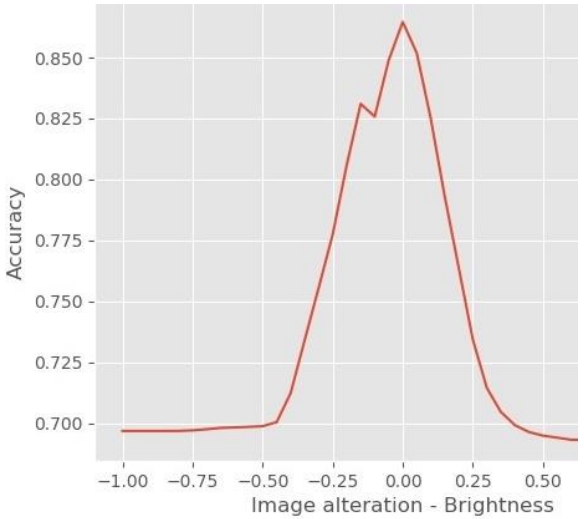
- Trasformazione di un dato in input per simulare possibili problemi nel mondo reale
- Un'alterazione può realizzarsi in un determinato range di valori
  - Esempio: aumento luminosità +- 10%
- Definite da un **esperto di dominio**

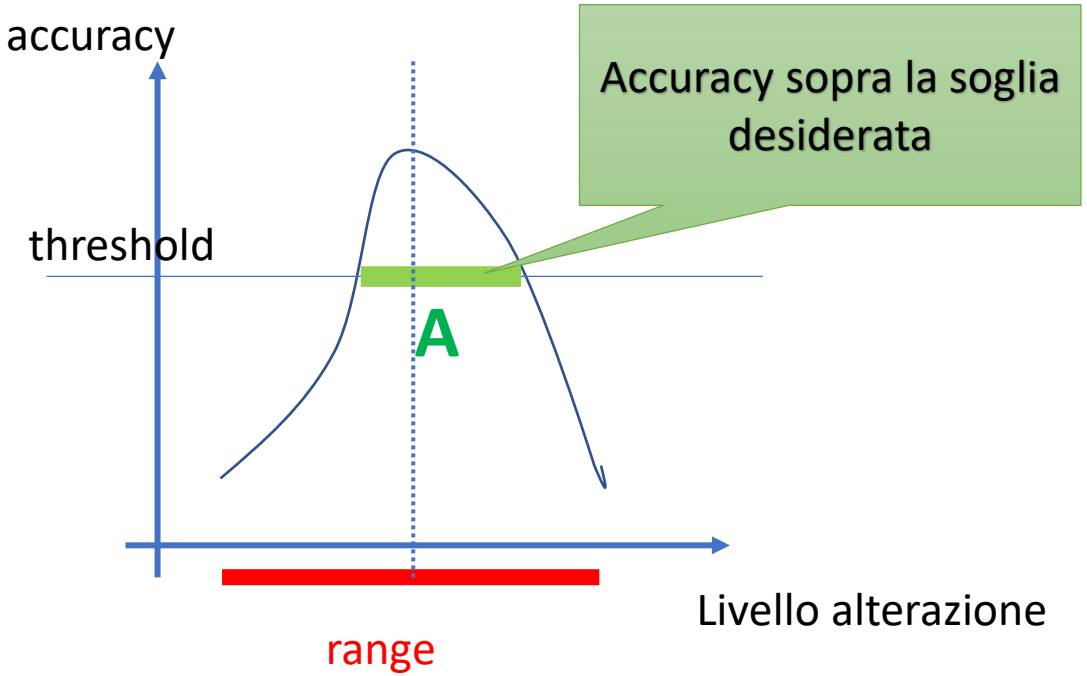


- **Alterazioni considerate:**
  - Rumore gaussiano
  - Sfocatura
  - Luminosità
  - Traslazione
  - Zoom
  - Compressione

Robustezza: intuizione

- Alterando i dati, l'accuracy diminuisce
- L'accuracy non deve scendere sotto una certa soglia



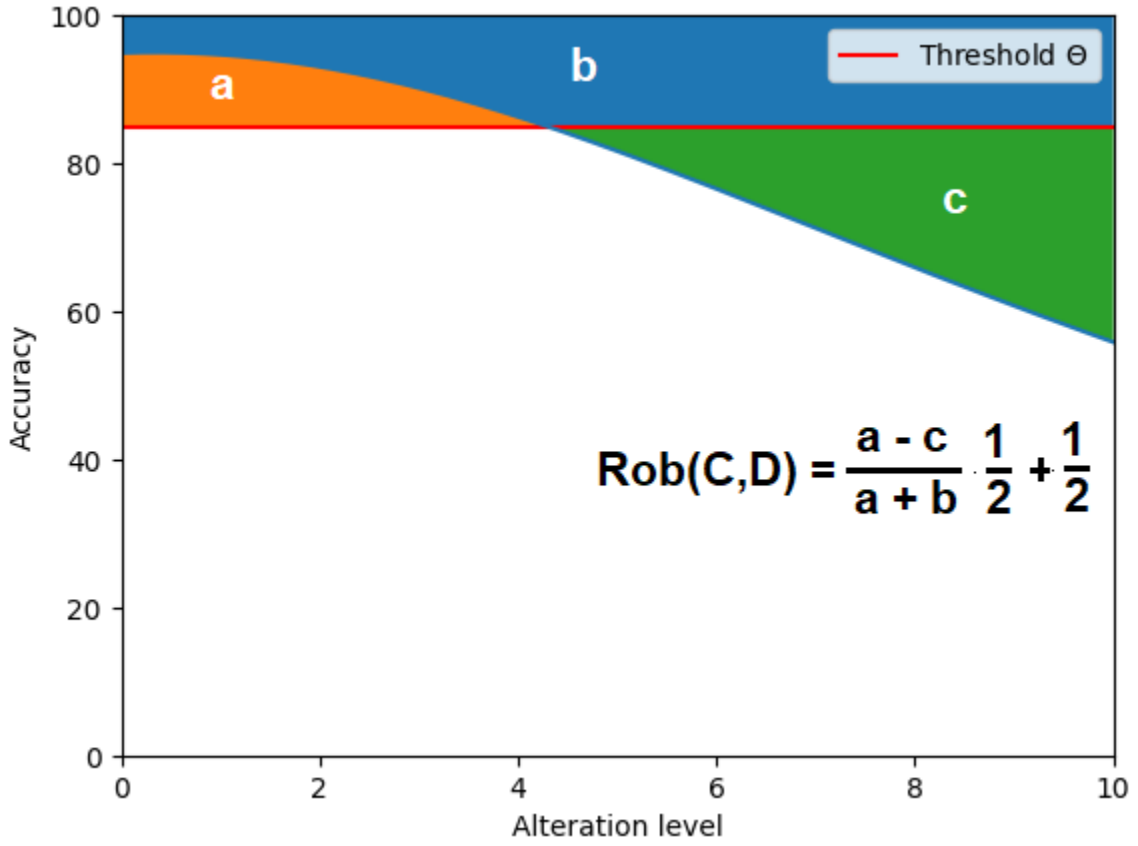


Robustezza: definizione

Definition 4.2.8 *(Robustness extension 3). Let C be a classifier, $tol(x)$ a tolerance function and $dep(x)$ a penalization function. The robustness $rob_A(C, D) \in [0, 1]$ of a classifier C w.r.t. alteration of type A in the range $[L_A, U_A]$ on a dataset D is defined formally as:*

$$rob_A(C, D) = \frac{\int_{L_A}^{U_A} [tol(acc(C, D^{A_i}) - \Theta) - dep(acc(C, D^{A_i}) - \Theta)] \cdot p_A(i)\ di}{2} + \frac{1}{2}$$

where Θ is a threshold referring to minimum accuracy accepted and p_A is probability distribution of the alteration levels.



Robustezza contro l'indecisione

Definition 4.3.1 *(Indecision robustness). Let C be a classifier, $tol(x)$ a tolerance function and $dep(x)$ a penalization function. The robustness against indecision $robInd_A(C, D) \in [0, 1]$ of a classifier C w.r.t. alteration of type A in the range $[L_A, U_A]$ on a dataset D is defined formally as:*

$$robInd_A(C, D) = \frac{\int_{L_A}^{U_A} [tol(\gamma - ind(C, D^{A_i})) - dep(\gamma - ind(C, D^{A_i}))] \cdot p_A(i)\ di}{2} + \frac{1}{2}$$

where γ is a threshold referring to maximum unknown ratio accepted, p_A is probability distribution of the alteration levels and $ind(C, D^{A_i})$ is the unknown ratio of C evaluated on D^{A_i}.

Si possono combinare le due metriche?

- La nuova metrica deve riflettere l'accuracy quando la rete non emette sconosciuti
- Quando ci sono sconosciuti la nuova metrica deve essere minore dell'accuracy

Idea: usare un rapporto

$$eff(C, D) = \frac{acc(C, D)}{ind(C, D) + 1}$$

Prende il nome di effectiveness

Effectiveness

Definition 4.3.2 *(Effectiveness). The effectiveness of a classifier C on dataset D is defined as:*

$$eff(C, D) = \frac{acc(C, D) \cdot (1 - ind(C, D))}{ind(C, D) + 1}$$

Robustezza aumentata

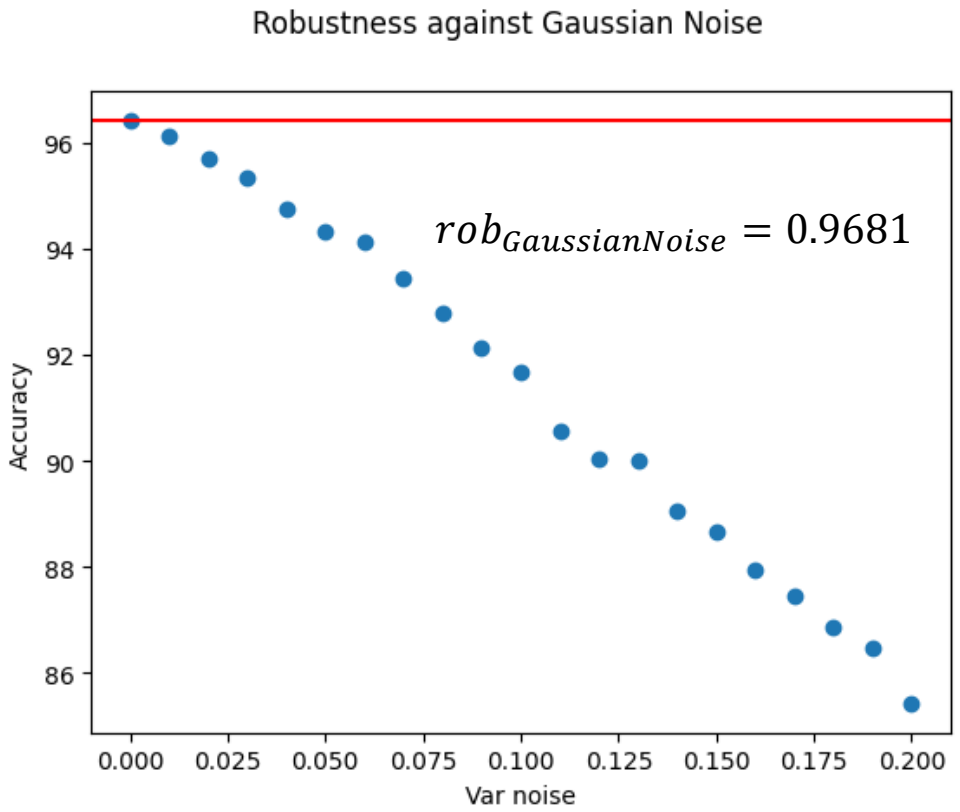
Definition 4.3.3 *(Augmented robustness 3). Let C be a classifier, $tol(x)$ a tolerance function and $dep(x)$ a penalization function. The augmented robustness $robAug_A(C, D) \in [0, 1]$ of a classifier C w.r.t. alteration of type A in the range $[L_A, U_A]$ on a dataset D is defined formally as:*

$$robAug_A(C, D) = \frac{\int_{L_A}^{U_A} [tol(eff(C, D^{A_i}) - \beta) - dep(eff(C, D^{A_i}) - \beta)] \cdot p_A(i)\ di}{2} + \frac{1}{2}$$

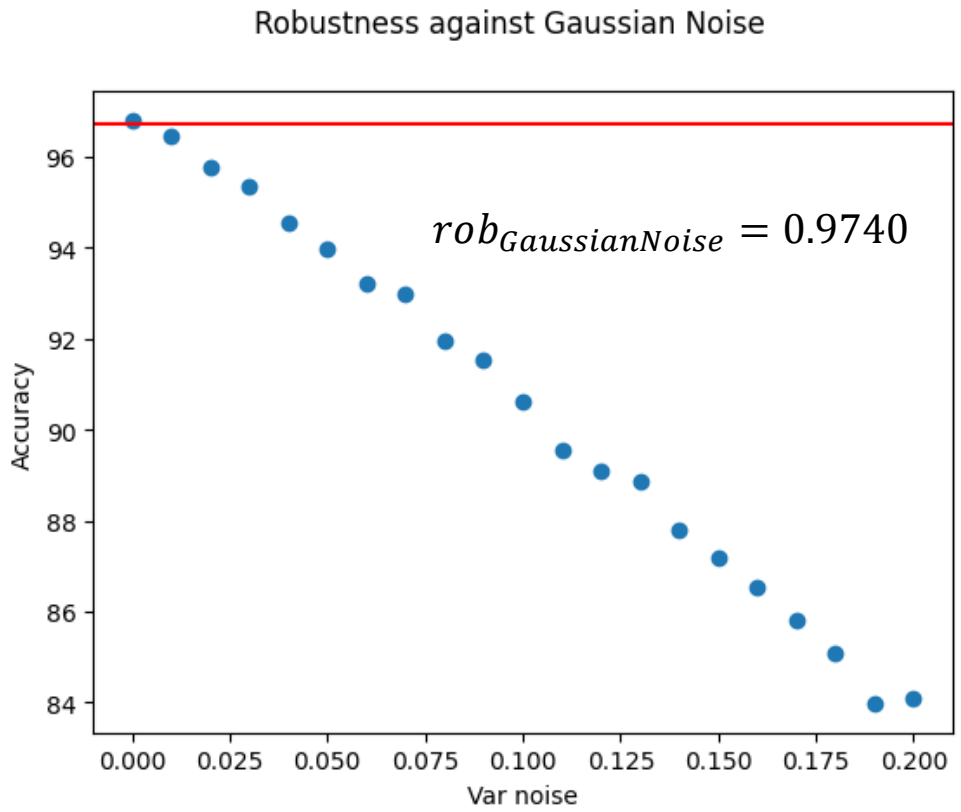
where β is a threshold referring to minimum effectiveness accepted and p_A is probability distribution of the alteration levels.

Robustezza a rumore gaussiano

Risultati senza uso di incertezza



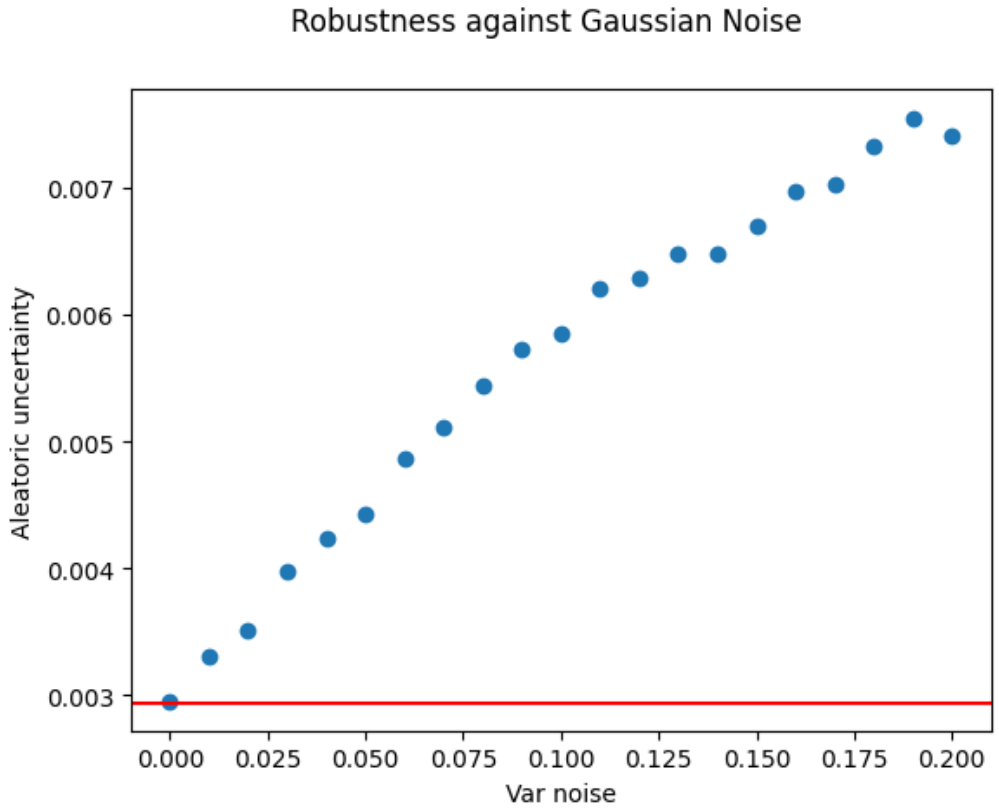
Standard NN

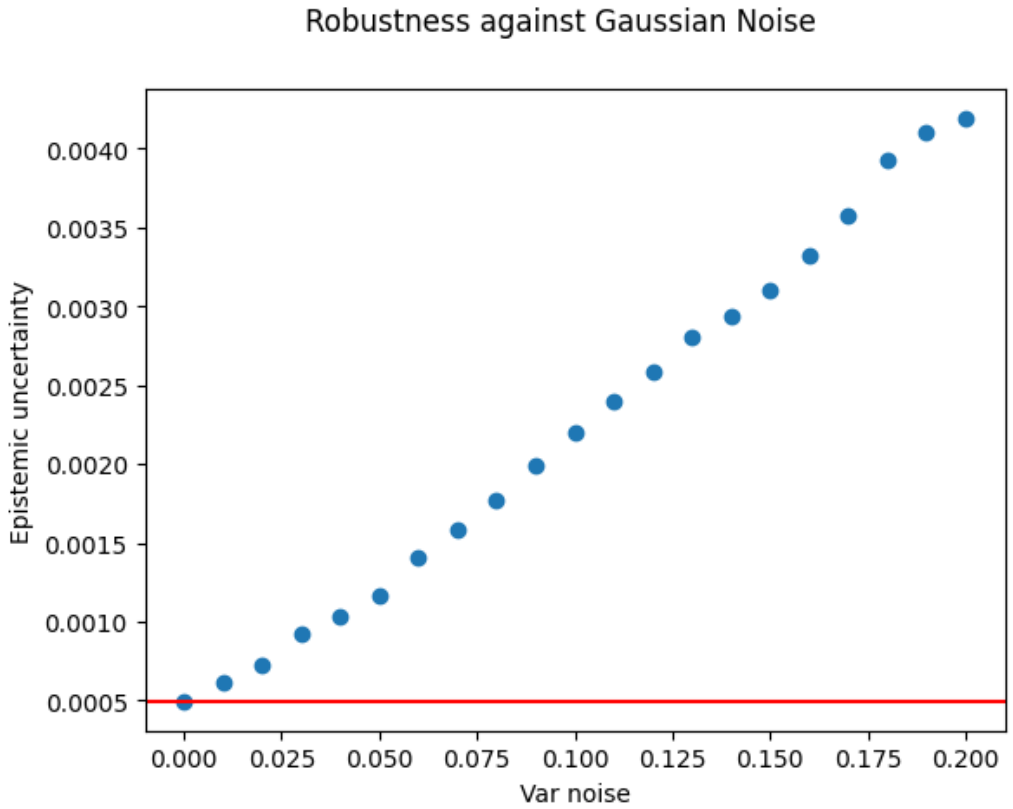


BNN

Robustezza a rumore gaussiano

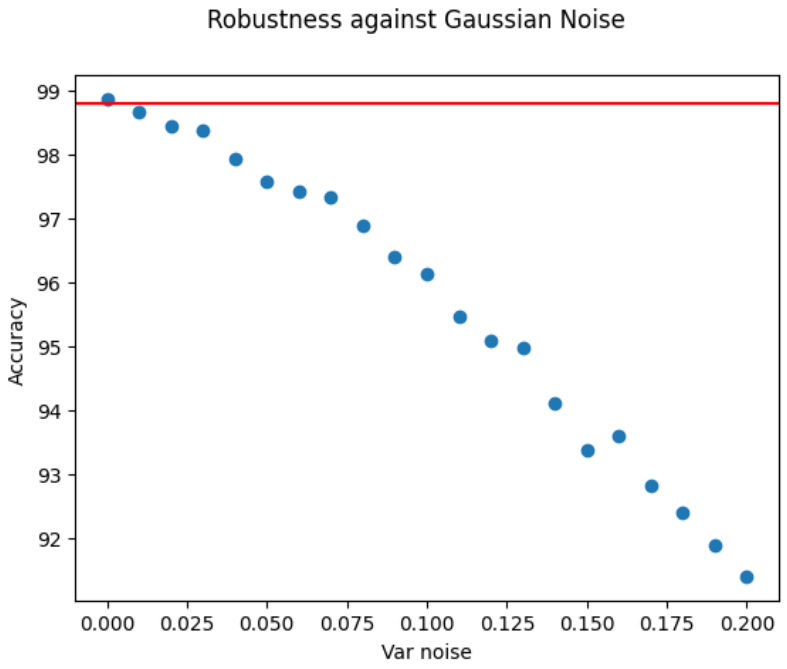
Andamento incertezza



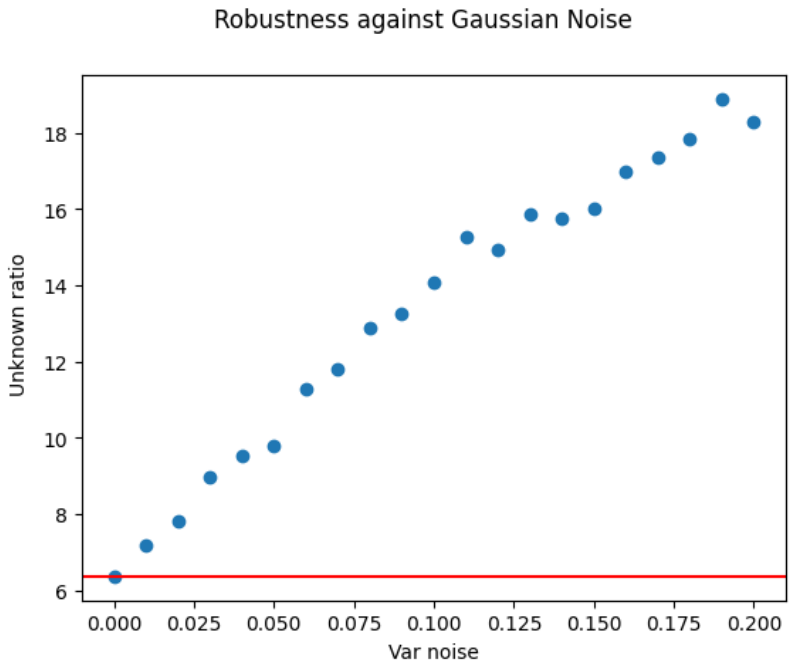


Robustezza a rumore gaussiano

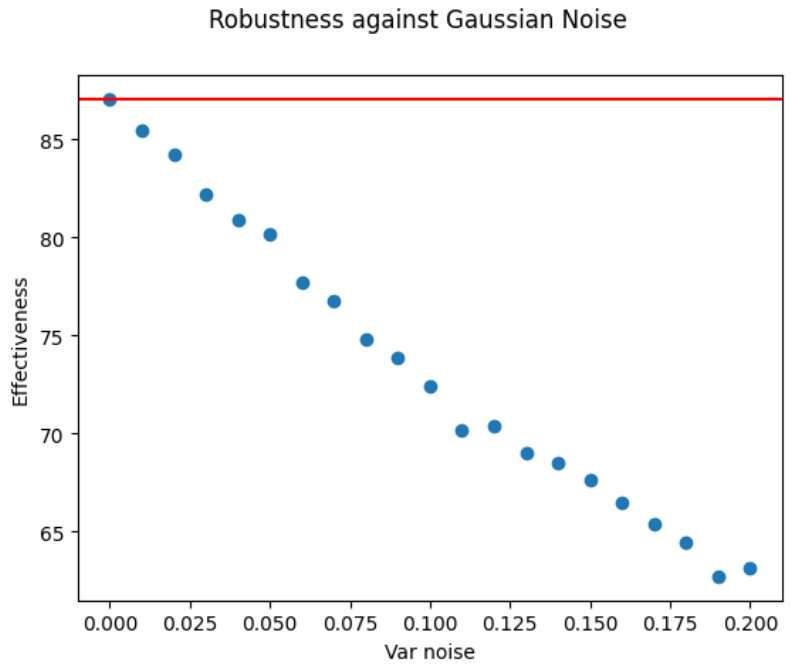
Usando incertezza aleatoria



$rob_{GaussianNoise} = 0.9842$



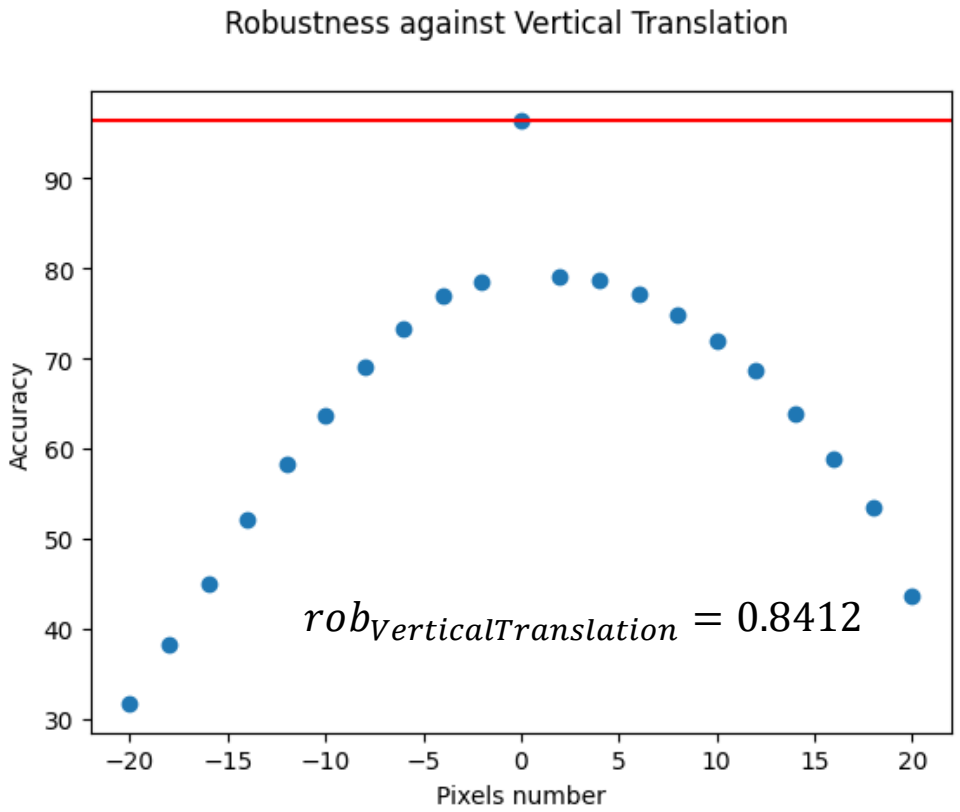
$robInd_{GaussianNoise} = 0.9624$



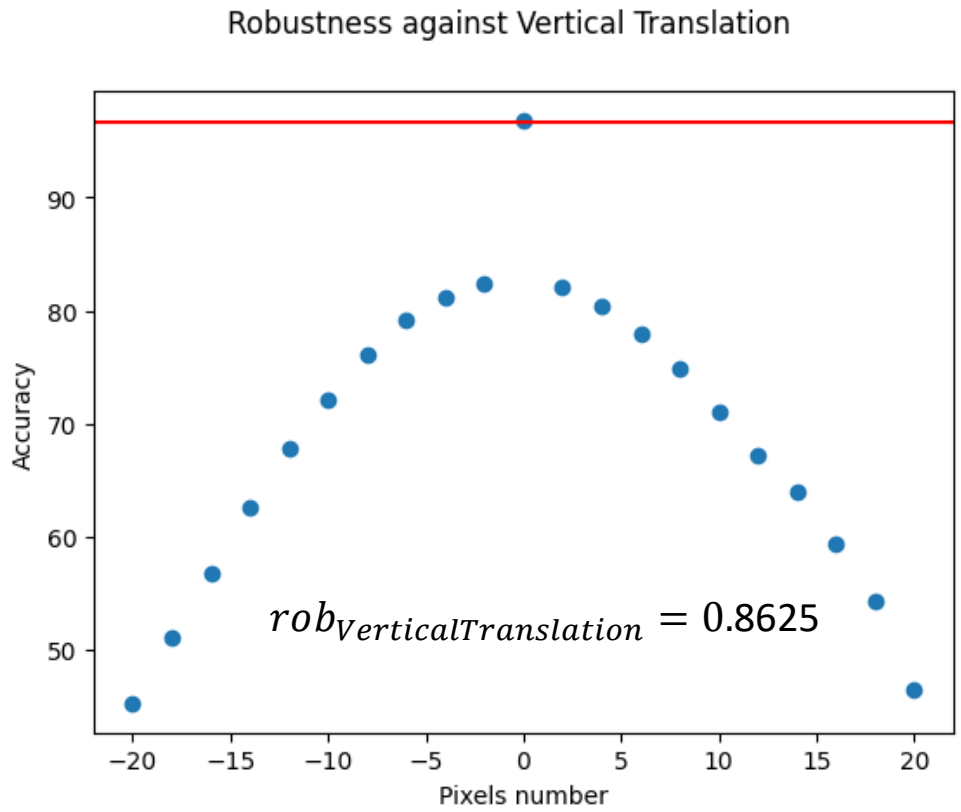
$robAug_{GaussianNoise} = 0.9217$

Robustezza a traslazione verticale

Risultati senza uso di incertezza



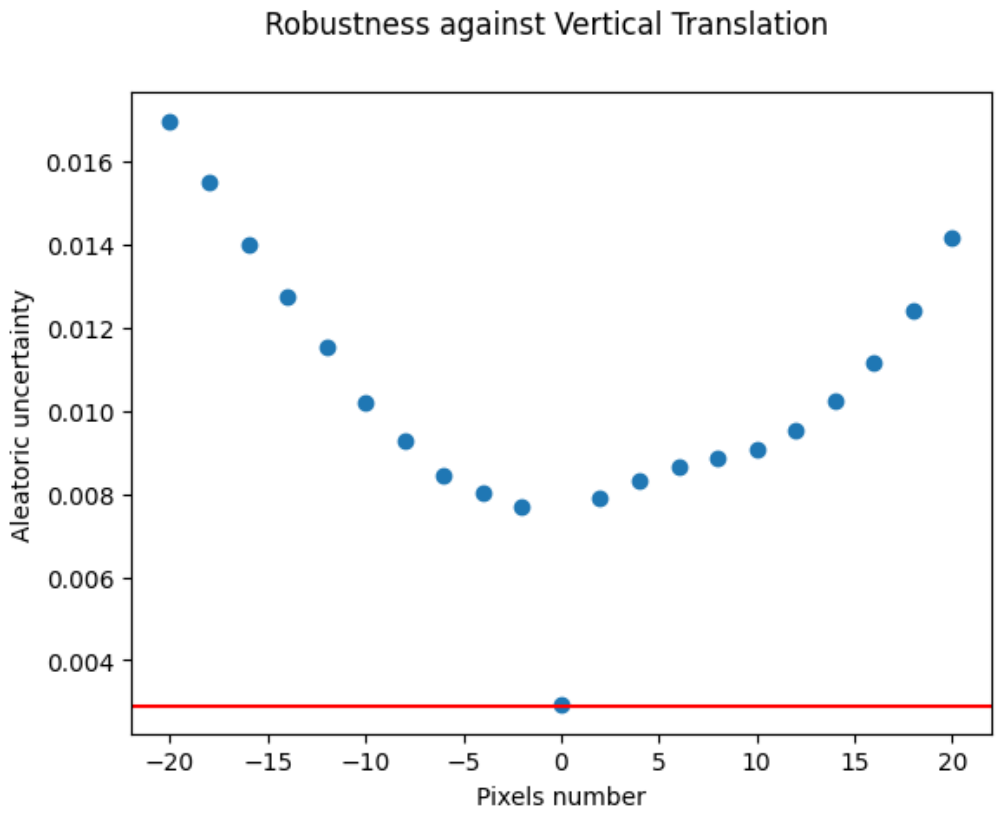
Standard NN

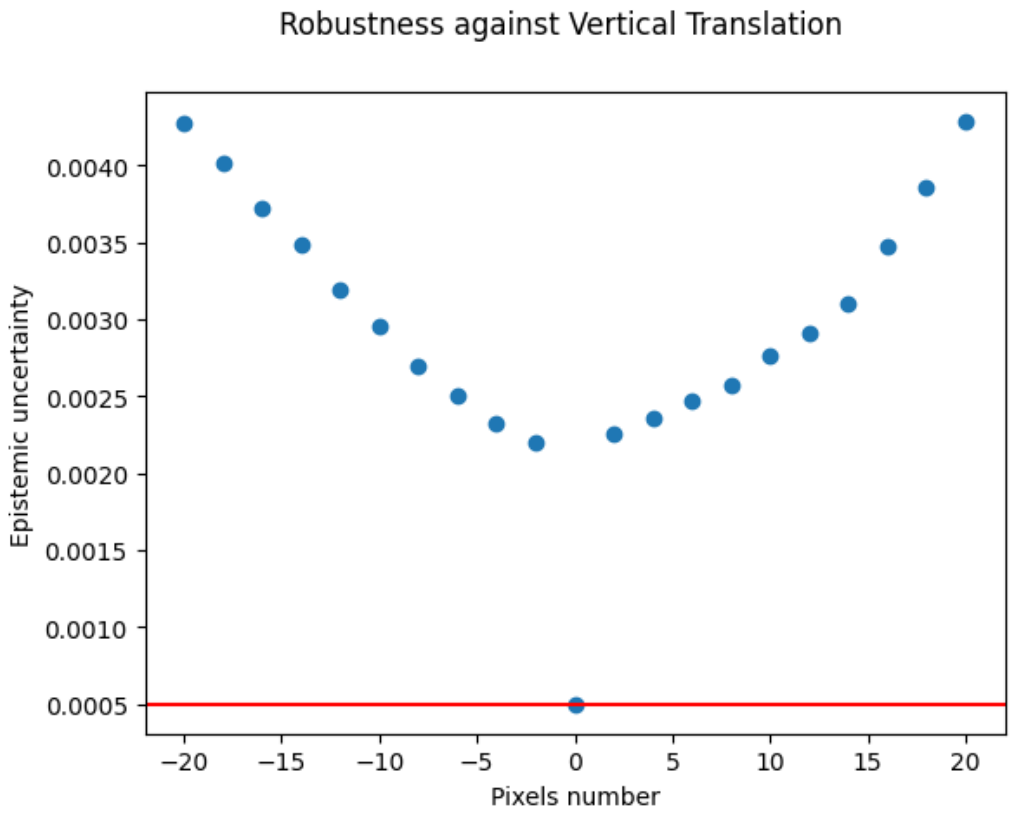


BNN

Robustezza a traslazione verticale

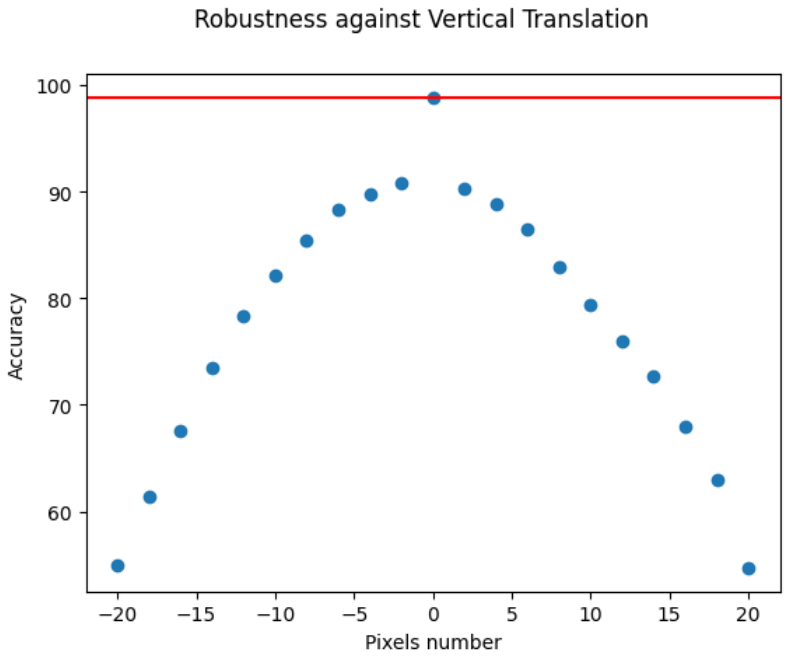
Andamento incertezza



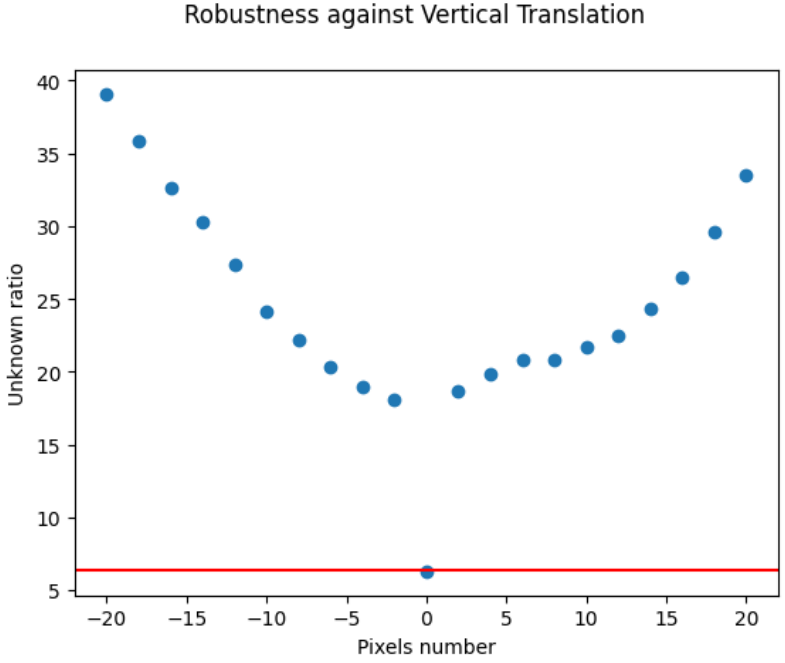


Robustezza a traslazione verticale

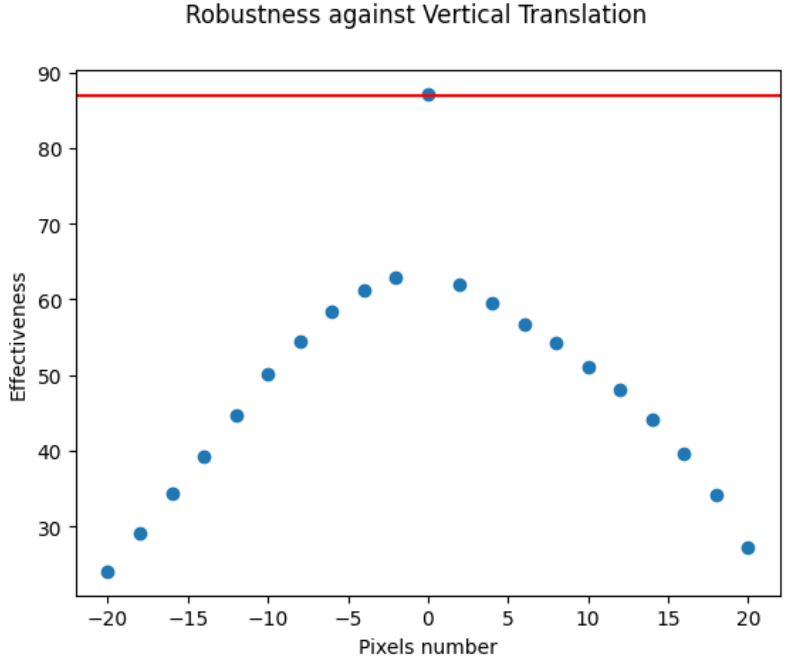
Usando incertezza aleatoria



$rob_{VerticalTranslation} = 0.8994$



$robInd_{VerticalTranslation} = 0.9065$



$robAug_{VerticalTranslation} = 0.7863$

Conclusioni

- Le BNN usate senza incertezza non portano un miglioramento significativo rispetto alle reti neurali standard
- L'incertezza stimata aiuta a diminuire il numero di predizioni errate, aumentando intrinsecamente la robustezza della rete
- Usare solo accuracy non basta
- L'incertezza va inclusa nella valutazione

Università degli Studi di Bergamo

Dipartimento di ingegneria Gestionale, dell'informazione della produzione

Evaluating the Robustness of Bayesian Neural Networks By Exploiting Uncertainty Estimation

Relatore:

Prof. Angelo Gargantini

Correlatori:

Dott. Andrea Bombarda

Dott.ssa Silvia Bonfanti

Candidato:

Wasim Essbai

Anno Accademico 2022-2023