

1 Finding the minimal change that will get rid of the network's WaterMark

Given a watermarked trained neural network as described [here](#). We tested what is the minimal change to the network last layer in order to "remove" some watermarks from the network.

1.1 Defining the problem

Given a neural network N with an output size m the network decision for an input x is defined as the coordinate with the maximal value, if the network output is the vector y the decision is $\operatorname{argmax}_{i \in [m]} \{y_i\}$

Given a watermarked network N with a set of K watermarks (A set of inputs to the network) $\{x_1, \dots, x_K\}$ we'll mark the network last layer L such that L is a $m \times n$ matrix where n is the layer's number of neurons and m is the network output size. The change to the last layer will be a matrix with the same dimension as L we'll mark as ε , such that $\varepsilon_{i,j}$ is the change to the last layer matrix entry $L_{i,j}$. We'll measure the overall change to the layer as $\|\varepsilon\|_\infty = \max_{i,j} \{|\varepsilon_{i,j}|\}$.

For a certain input x we're only interested in the input to the last layer we'll mark the input to the last layer v . v is a $n \times 1$ vector. So the original network output $y = Lv$ and the changed network output is $y' = (L + \varepsilon)v$. For a single input x we need to find the minimal ε so that the $\operatorname{argmax}_{i \in [m]} \{y_i\} \neq \operatorname{argmax}_{i \in [m]} \{y'_i\}$

Denote $d := \operatorname{argmax}_{i \in [m]} \{y_i\}$

For some $d' \in [m]$, $d' \neq d$ finding ε with minimal $\|\varepsilon\|_\infty$ such that $y' = (L + \varepsilon)v$ and $d' = \operatorname{argmax}_{i \in [m]} \{y'_i\}$ can be described in a Linear Programming form like so:

$$\begin{aligned} \text{Minimize : } & c \\ \text{Subject to : } & \forall i, j \quad -c \leq \varepsilon_{i,j} \leq c \\ & y' = (L + \varepsilon)v \\ & y'_d \leq y'_{d'} \end{aligned}$$

Using the same method we can find how to change the network to more than one input.

Given inputs x_1, \dots, x_k and their respective inputs to the last layer v_1, \dots, v_k and their respected outputs and decisions $\{y_1, \dots, y_k\} \{d_1, \dots, d_k\}$ we want to find ε such that

$$\forall 1 \leq j \leq k \quad d_j \neq \operatorname{argmax}_{i \in [m]} \{((L + \varepsilon)v_j)_i\}$$

Assuming we choose our new desired output $\{d'_1, \dots, d'_k\}$ And now our LP looks like this:

$$\begin{aligned} \text{Minimize : } & c \\ \text{Subject to : } & \forall i, j \quad -c \leq \varepsilon_{i,j} \leq c \\ & \forall j \quad y'_j = (L + \varepsilon)v_j \\ & \forall j \quad (y'_j)_{d_j} \leq (y'_j)_{d'_j} \end{aligned}$$