COD492:- Repair of Deep Neural Networks

Yash Bansal

2022CS51133

Faculty Supervisors:-
Prof. Priyanka Golia
Prof. Kumar Madhukar

Motivation:-

- DNNs are deployed in safety-critical applications (healthcare, autonomous driving)
- High-performing models can still fail on specific inputs
- Retraining is expensive, offers no guarantees, and may introduce new failures

Goal:-

- Develop a method to repair an already trained DNN instead of retraining.
- Identify problematic neurons/layers causing erroneous behavior.
- Make minimal changes to weights to correct issues.
- Preserve accuracy and other properties for non-problematic inputs.

Problem Definition:-

- **DNN Verification Problem**
  - Given trained DNN N and property Q, check if N satisfies Q
  - Check satisfiability of $N \wedge (\neg Q)$
  - Output: UNSAT (property holds) or SAT with counterexamples $x_1, x_2, \ldots, x_n$

- **DNN Repair Problem**
  - Given: trained DNN N, property Q, and counterexamples X
  - Find new network N' where Q holds for all $x \in X$

$$\text{Distance}_L(N, N') = \left(\sum_{i=1}^{m} |w_i - w_i'|^L \right)^{1/L}$$

$$\begin{aligned} \min_{W'} \quad & \text{Distance}_L(W, W') \\ \text{subject to} \quad & N'(x_i) \models Q, \quad \forall x_i \in X \end{aligned}$$

Work Done Till Mid-Term:-

- Understood Veristable solver framework, DPLL(T) verification, and neuron stability concepts
- Debugged solver and analyzed counterexample traces to identify problematic weights
- Explored DPLL tree structure but found limited actionable patterns for repair
- Transitioned to optimization-based approach for more direct repair insights

Optimization Based Approach:-

- Created simple hand-tuned neural networks with well-defined properties
- Generated counterexamples for these controlled test cases
- Used Gurobi optimizer to compute exact optimal weight updates
- Established baseline for minimal modifications that fix property violations

Gradient-Based Repair:-

- Computed nearest correct output and loss for each counterexample

$$\min_{y^*} \quad \|y^* - \hat{y}_i\|^2$$
$$\text{subject to} \quad y^* \models Q$$

$$\mathcal{L}_i = \|N(x_i) - y_i^*\|^2$$

- Calculated gradients of weights and hidden layer activations

$$\Delta z^{(l)} = -\alpha \cdot \nabla_{z^{(l)}} \mathcal{L}_i$$

Gradient-Based Repair:-

- Optimal hidden layer output changes are highly proportional to gradients
- This proportionality provides tractable repair direction without full optimization

$$\begin{aligned} \min_{\Delta W^{(l)}} \quad & \|\Delta W^{(l)}\|_\infty \\ \text{subject to} \quad & (W^{(l)} + \Delta W^{(l)}) \cdot a^{(l-1)} = z^{(l)} + \Delta z^{(l)} \end{aligned}$$

Layer-wise Repair Implementation:-

- Adopted approach from "Minimal Modifications" [1] paper and "Minimal Multi-Layer Modifications" [2]
- Formulate single layer modification as DNN verification problem for scalability
- Perturb one hidden layer at a time proportionally to computed gradients
- Apply modifications recursively in layer-wise fashion to propagate repairs efficiently

References: [1] Minimal Modifications of Deep Neural Networks -https://easychair.org/publications/open/CWhF
[2] Minimal Multi-Layer Modifications of DNNs - https://arxiv.org/abs/2110.09929

Results:-

- Successfully repaired small networks using gradient-proportional perturbations
- Weight updates comparable to Gurobi's optimal solutions
- Scalability to larger networks yet to be tested
- Approach shows promise: Achieves optimization-quality repairs with gradient-method efficiency

Thank You