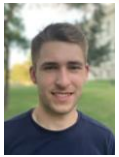


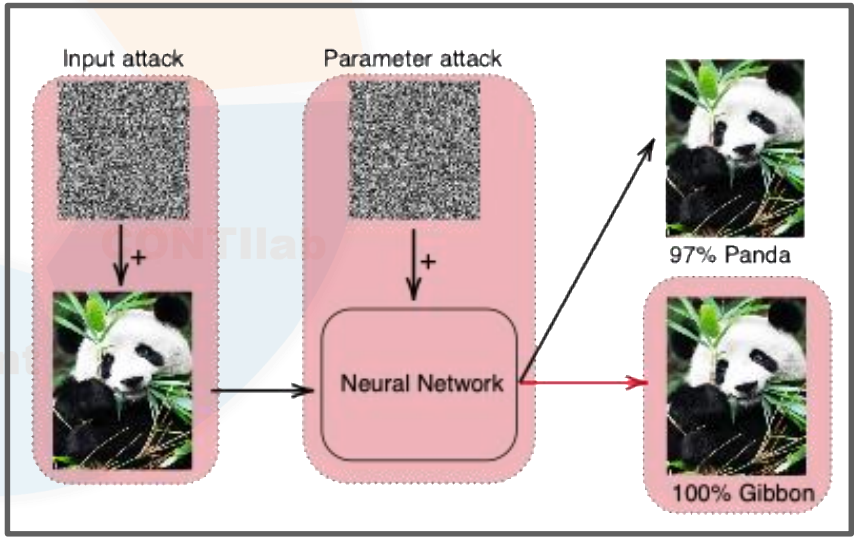
Name: Attila Szász, MSc student, second year

Project type: thesis project

Topic: Parameter Robustness of Neural Networks

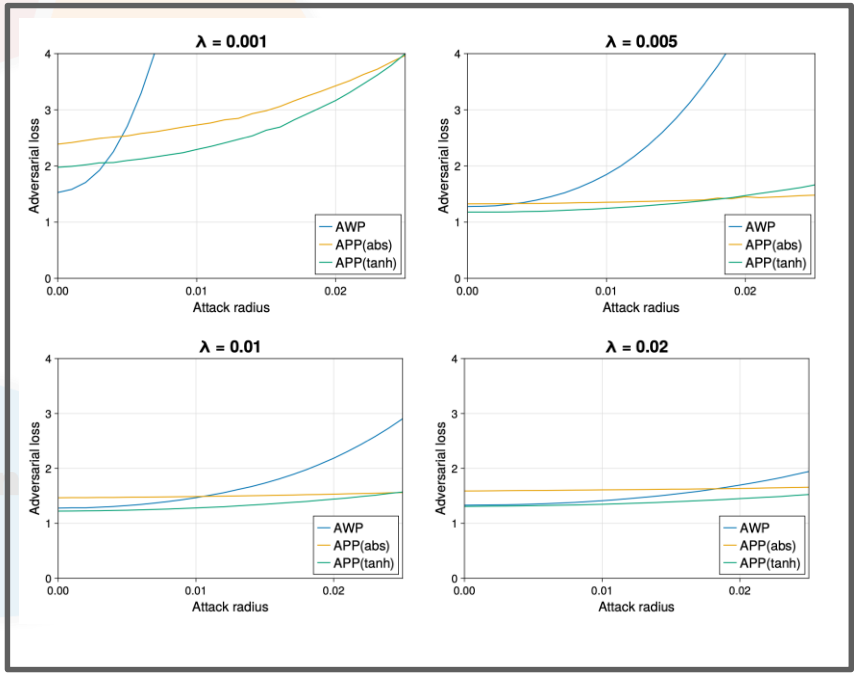
Supervisors: Dr. Balázs Bánhelyi

- Neural networks can be highly sensitive to small perturbations in both input and parameter space.
- During our research, we focused on parameter perturbations.
- The **Adversarial Parameter Propagation (APP)** algorithm was defined, which was able to improve both input and parameter robustness.
- $\theta = \arg\min_{\theta} \left(\max_{\check{\theta} \epsilon B_{p1}(\theta,\lambda)} \left(\max_{\check{x} \epsilon B_{p2}(x,\varepsilon)} L_{\check{\theta}}(\check{x}, y) \right) \right)$
- The APP computes precise bounds for the output of the neural networks and then optimizes the parameters of the model based on the worst-case scenario determined from these bounds.
- Why was only naive interval arithmetic used?



Experiments

- Techniques for reducing the widening of the intervals was defined.
- $\text{Radius}(x) = \frac{\tanh(sx^2)}{m}\lambda$
- The restriction on the parameter value range.
- Neural networks were trained for classification with APP and AWP (Adversarial Weight Perturbation).
- CIFAR-10 dataset
- CNN4 architecture
- APP with and without the regularization techniques
- We examined the following:
    - The normal accuracy and adversarial accuracy of the trained networks.
    - The input robustness of the alternative networks within the defined neighborhood of the midpoint network.
    - Resistance to Adversarial Parameter Attack (APP)



Results & future work

- The APP induces stronger flattening effect on the weight loss landscape than the AWP.
- The APP with our radius narrowing technique improves both normal accuracy and input adversarial accuracy.
- The APP without the radius narrowing technique improves the resistance to APP attack.
- The APP always optimizes the midpoint network.
- Other result: The implementation of the whole training system in julia.
- Future works:
  - Wider networks
  - New bound propagation methods (statistical approaches)

ϵ	λ	Algorithm	Accuracy	AutoAttack
$\frac{2}{255}$	0.001	APP (tanh)	68.15 %	42.75%
		APP (abs)	59.48%	36.08%
		AWP	59.88%	39.01%
	0.005	APP (tanh)	67.17%	49.92%
		APP (abs)	60.15%	44.92%
		AWP	61.9%	43.33%
	0.01	APP (tanh)	63.8%	49.63%
		APP (abs)	52.71%	42.85%
		AWP	60.68%	44.16%
	0.02	APP (tanh)	60.77%	48.18%
		APP (abs)	48.81%	41.94%
		AWP	58.1%	43.70%

ϵ	λ	Algorithm	$\hat{\lambda}$ 0.02	0.04	0.06	0.08	0.1
$\frac{2}{255}$	0.001	APP (tanh)	27%	31%	36%	36%	41%
		APP (abs)	14%	18%	**19%**	**24%**	**27%**
		AWP	14%	28%	30%	37%	45%
	0.005	APP (tanh)	19%	24%	28%	30%	28%
		APP (abs)	13%	24%	24%	33%	29%
		AWP	8%	18%	34%	44%	49%
	0.01	APP (tanh)	12%	25%	22%	27%	**27%**
		APP (abs)	9%	18%	20%	**24%**	**27%**
		AWP	**6%**	17%	26%	38%	38%
	0.02	APP (tanh)	13%	19%	29%	31%	31%
		APP (abs)	9%	23%	27%	28%	28%
		AWP	21%	**13%**	23%	37%	32%