Adversarial Robustness without Adversarial Training: A Teacher-Guided Curriculum Learning Approach

Anindya Sarkar* **Anirban Sarkar*** **Sowrya Gali*** **Vineeth N Balasubramanian**

Indian Institute of Technology Hyderabad

Outline

- Adversarial Training (AT) based Methods for Adversarial Robustness
  - Drawbacks
- Curriculum Learning Based Adversarial Training
- Non-iterative Adversarial Robustness Training
- Robustness and Alignment of Explanation Maps
- Teacher-guided Saliency Based Robust Training
  - First Phase - Enforcing Alignment
  - Second Phase - Model Refinement
- Further Clarifications on Two Phase Training
- Experiments and Results
- Conclusion

Outline

- Adversarial Training (AT) based Methods for Adversarial Robustness
  - Drawbacks
- Curriculum Learning Based Adversarial Training
- Non-iterative Adversarial Robustness Training
- Robustness and Alignment of Explanation Maps
- Teacher-guided Saliency Based Robust Training
  - First Phase - Enforcing Alignment
  - Second Phase - Model Refinement
- Further Clarifications on Two Phase Training
- Experiments and Results
- Conclusion

Adversarial Training (AT)

Adversarial Training (AT) is the standard training method to achieve adversarial robustness based on min-max optimization where inner maximization generates perturbed images within an Є-ball and the outer minimization tunes the model parameters according to the perturbed images.

$$\min_{\theta} \rho(\theta), where$$

$$\rho(\theta) = \mathbb{E}_{(x,y)\ D}[\max_{\delta \in B(\epsilon)} L(\theta, x+\delta, y)]$$

Drawbacks

- Costly due to iterative inner maximization
- Perform poorly on clean data

Outline

- Adversarial Training (AT) based Methods for Adversarial Robustness
  - Drawbacks
- Curriculum Learning Based Adversarial Training
- Non-iterative Adversarial Robustness Training
- Robustness and Alignment of Explanation Maps
- Teacher-guided Saliency Based Robust Training
  - First Phase - Enforcing Alignment
  - Second Phase - Model Refinement
- Further Clarifications on Two Phase Training
- Experiments and Results
- Conclusion

Curriculum Learning Based Adversarial Training

Different Variations

- Progressively increasing the number of PGD steps [Cai et al. 2018]
- Gradually increase the convergence quality of the generated adversarial examples [Wang et al. 2019]
- Learning initially from least adversarial data and progressively utilizes increasingly more adversarial data [Zhang et al. 2020]
- Curriculum loss as inner maximization step which depends on a difficulty parameters that gradually increased as the training progresses [Sitawarin et al. 2020]

Outline

- Adversarial Training (AT) based Methods for Adversarial Robustness
  - Drawbacks
- Curriculum Learning Based Adversarial Training
- Non-iterative Adversarial Robustness Training
- Robustness and Alignment of Explanation Maps
- Teacher-guided Saliency Based Robust Training
  - First Phase - Enforcing Alignment
  - Second Phase - Model Refinement
- Further Clarifications on Two Phase Training
- Experiments and Results
- Conclusion

Non-iterative Adversarial Robustness Training

Different Variations

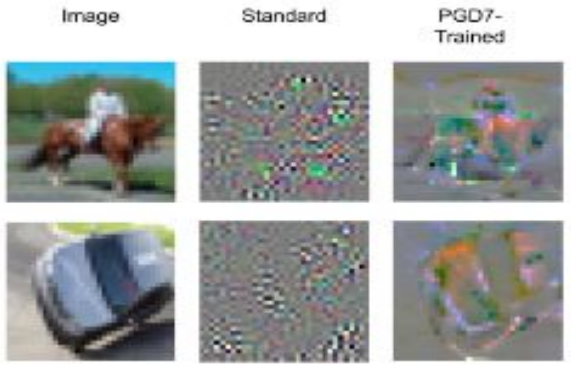
- Single-step adversarial training method using dropout scheduling [Vivek et al. 2020]
- JARN [Chan et al. 2020] improves model robustness by matching the gradient of loss w.r.t. The image to the actual image
- Employs a discriminator to compare between the jacobian and the image saliency [Chan et al. 2020]

Outline

- Adversarial Training (AT) based Methods for Adversarial Robustness
  - Drawbacks
- Curriculum Learning Based Adversarial Training
- Non-iterative Adversarial Robustness Training
- Robustness and Alignment of Explanation Maps
- Teacher-guided Saliency Based Robust Training
  - First Phase - Enforcing Alignment
  - Second Phase - Model Refinement
- Further Clarifications on Two Phase Training
- Experiments and Results
- Conclusion

Robustness and Alignment of Explanation Maps

Attribution maps for adversarially trained models tend to align more to actual image compared to naturally trained models. This connection was studied in [Etmann et al. 2019]



[Chan et al. 2020]

Outline

- Adversarial Training (AT) based Methods for Adversarial Robustness
  - Drawbacks
- Curriculum Learning Based Adversarial Training
- Non-iterative Adversarial Robustness Training
- Robustness and Alignment of Explanation Maps
- Teacher-guided Saliency Based Robust Training
  - First Phase - Enforcing Alignment
  - Second Phase - Model Refinement
- Further Clarifications on Two Phase Training
- Experiments and Results
- Conclusion

Teacher-guided Saliency Based Robust Training

Main Ideas

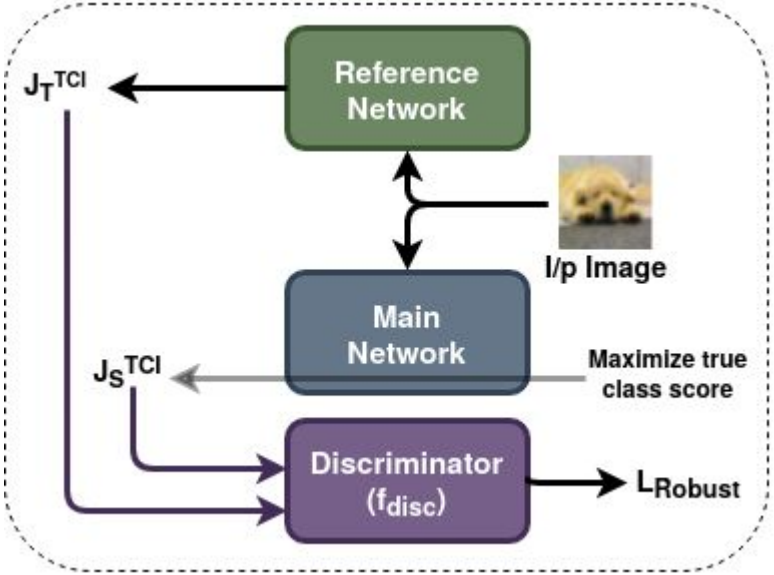
- Enforce alignment between saliency and object features through training
  - Accomplished by forcing saliency of the main model to follow the object features
  - The object features are provided by saliency of a pre-trained reference model
- Model decision of a truly adversarially robust model can be changed only by perturbing the pixels of the object and not any pixel outside of the object in an image.

Teacher-guided Saliency Based Robust Training

First Phase - Enforcing Alignment

- Pre-trained teacher network f_T , student network f_S parametrized by θ and a discriminator network f_{disc} parametrized by Ø. Saliency map from the teacher is $J^{TCI}{}_T$ and the student is $J^{TCI}{}_S$
- Here the following objective function is:

$$\theta_{optimum} = \underset{\theta}{\operatorname{argmin}}[\mathcal{L}_{CE} + \underbrace{(\beta\mathcal{L}_{Robust} + \gamma\mathcal{L}_{diff})}_{\textit{alignment loss}}];$$



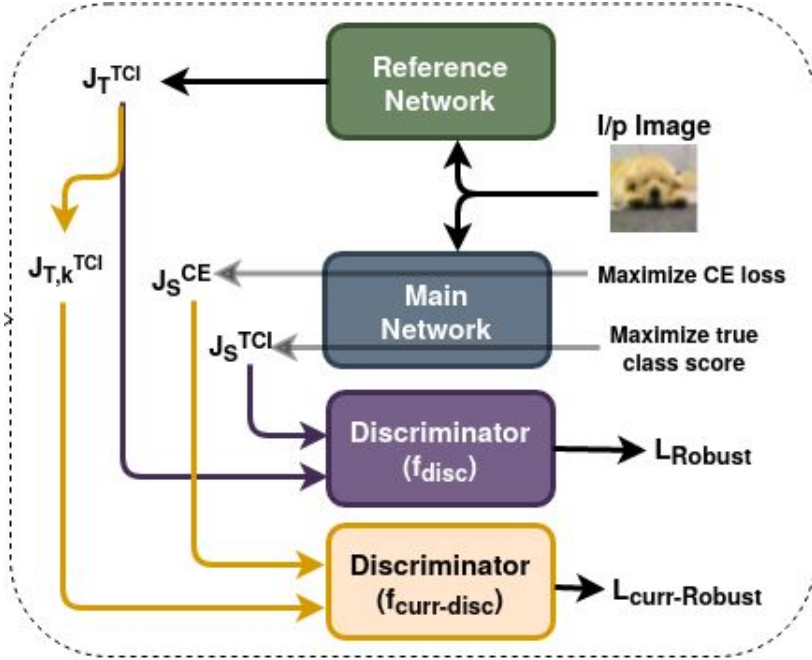
$$\mathcal{L}_{Robust} = \mathbb{E}_{J_T}[\log f_{disc}(J_T^{TCI})] + \mathbb{E}_{J_S^{TCI}}[\log(1 - f_{disc}(J_S^{TCI}))]$$

$$\mathcal{L}_{diff} = ||J_S^{TCI} - J_T^{TCI}||_2^2$$

Teacher-guided Saliency Based Robust Training

Second Phase - Model Refinement

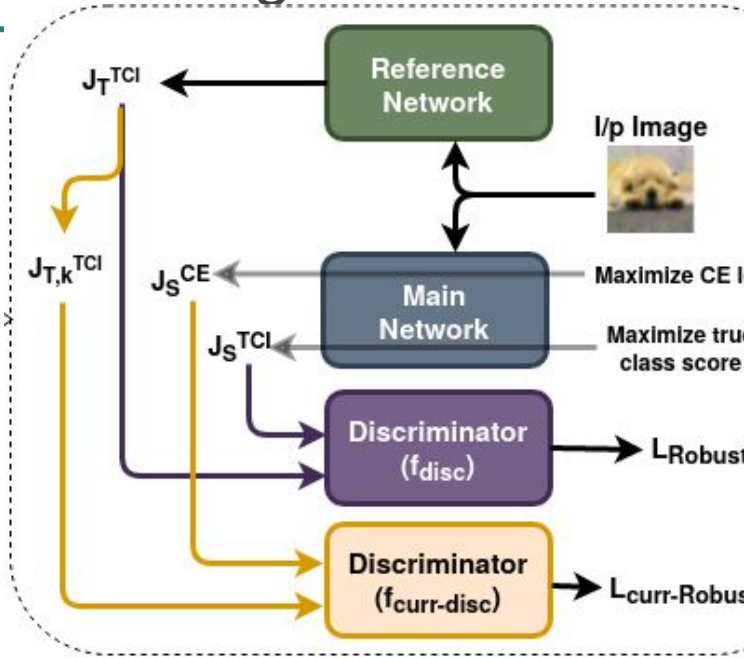
- Ensure that the decision of the model can be changed only by perturbing the object pixels.
- Use curriculum style learning by gradually shortening the set of pixels which are allowed to perturb in order to reduce true class prediction score.
- Restricting the attacker with very few options to perturb object pixels reduces the adversarial attack effect on input image.



Teacher-guided Saliency Based Robust Training

Second Phase - Model Refinement

- Obtain J^{CE}_S by maximizing the CE loss of student w.r.t. input pixels. Top k% of the saliency map from the teacher is $J^{TCI}_{T,k}$. Consider a discriminator network $f_{curr\text{-}disc}$ parametrized by ξ.
- Here the following objective function is:



$$\theta_{optimum} = \underset{\theta}{\operatorname{argmin}}[\mathcal{L}_{CE} + (\underbrace{\beta\mathcal{L}_{Robust} + \gamma\mathcal{L}_{diff}}_{\text{alignment loss}} + \underbrace{\beta\mathcal{L}_{curr-Robust} + \gamma\mathcal{L}_{curr-diff}}_{\text{curriculum loss}})];$$

$$\mathcal{L}_{curr-Robust} = \mathbb{E}_{J^{TCI}_{T,k}}[\log f_{curr-disc}(J^{TCI}_{T,k})] + \mathbb{E}_{J^{CE}_S}[\log(1 - f_{curr-disc}(J^{CE}_S)]$$

$$\mathcal{L}_{curr-diff} = ||J^{CE}_S - J^{TCI}_{T,k}||^2_2$$

Outline

- Adversarial Training (AT) based Methods for Adversarial Robustness
  - Drawbacks
- Curriculum Learning Based Adversarial Training
- Non-iterative Adversarial Robustness Training
- Robustness and Alignment of Explanation Maps
- Teacher-guided Saliency Based Robust Training
  - First Phase - Enforcing Alignment
  - Second Phase - Model Refinement
- Further Clarifications on Two Phase Training
- Experiments and Results
- Conclusion

Further Clarifications on Two Phase Training

Why J^{TCI}_S in first phase and include J^{CE}_S in second phase?

- Our method is motivated by alignment of saliency map with object features
- Including J^{CE}_S forces the model to learn the allowed set of pixels, to be perturbed, to reduce the class score

Why curriculum learning?

- Enforce that most of the pixels, which are allowed to change, should belong to the most discriminative parts of the object
- Fewer pixels should be considered from the lesser discriminative parts of the object

Outline

- Adversarial Training (AT) based Methods for Adversarial Robustness
  - Drawbacks
- Curriculum Learning Based Adversarial Training
- Non-iterative Adversarial Robustness Training
- Robustness and Alignment of Explanation Maps
- Teacher-guided Saliency Based Robust Training
  - First Phase - Enforcing Alignment
  - Second Phase - Model Refinement
- Further Clarifications on Two Phase Training
- Experiments and Results
- Conclusion

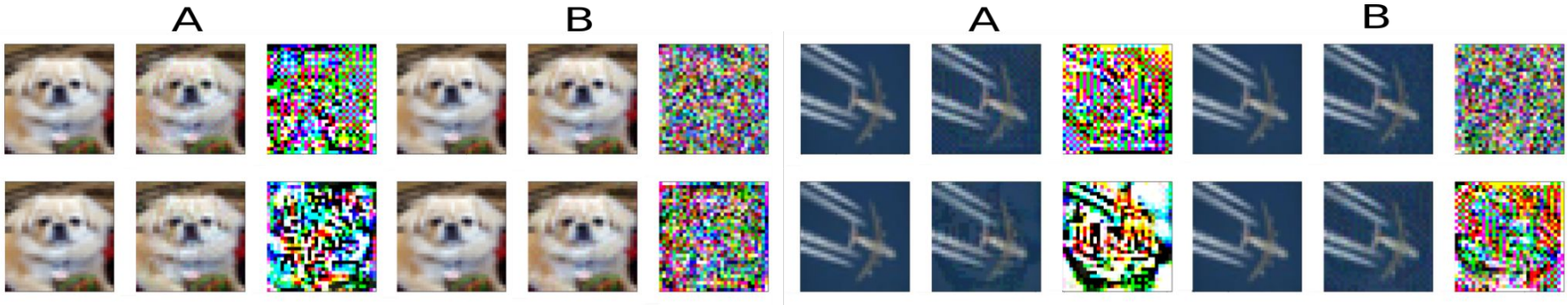
Experiments and Results

Results with CIFAR-10

Type	Curriculum	Methods	Clean	FGSM	PGD-5	PGD-10	PGD-20	C&W	AA
Iterative Methods	*NO*	AT(PGD-7)[Madry et al. 2017]	87.25	56.22	55.50	47.30	45.90	46.80	44.04
		FNT[Xie et al. 2019]	87.31	NA	NA	46.99	46.65	NA	NA
		LAT[Singh et al. 2019]	87.80	NA	NA	53.84	53.71	NA	49.12
		TRADES[Zhang et al. 2019]*	84.92	61.06	NA	NA	56.61	51.98	53.08
		GAIRAT[Zhang et al. 2020]	85.75	NA	NA	NA	57.81	NA	NA
		AWP-AT[Wu et al. 2020]*	85.57	62.90	NA	NA	58.14	55.96	54.04
		MART[Wang et al. 2019]*	84.17	67.51	NA	NA	58.56	54.58	NA
	YES	CAT18[Cai et al. 2018]	77.43	57.17	NA	NA	46.06	42.28	NA
		Dynamic AT[Wang et al. 2019]	85.03	63.53	NA	NA	48.70	47.27	NA
		FAT[Zhang et al. 2020]	87.00	65.94	NA	NA	49.86	48.65	53.51
		ATES[Sitawarin et al. 2020]*	86.84	NA	NA	NA	55.06	NA	50.72
Non-Iterative Methods	*NO*	SADS[Babu et al. 2020]+	82.01	51.99	NA	45.66	NA	NA	NA
		JARN-AT1[Chan et al. 2019]	84.80	67.20	50.00	27.60	15.50	NA	0.26
		IGAM[Chan et al. 2020]	88.70	54.00	52.50	47.60	45.10	NA	NA
		AT-Free[Shafahi et al. 2019]	85.96	NA	NA	NA	46.82	46.60	41.47
	YES	**OURS**	**90.63**	**67.84**	**63.81**	**61.44**	**59.59**	**61.83**	**54.71**

Experiments and Results

Visualizations for effect of curriculum learning



Outline

- Adversarial Training (AT) based Methods for Adversarial Robustness
  - Drawbacks
- Curriculum Learning Based Adversarial Training
- Non-iterative Adversarial Robustness Training
- Robustness and Alignment of Explanation Maps
- Teacher-guided Saliency Based Robust Training
  - First Phase - Enforcing Alignment
  - Second Phase - Model Refinement
- Further Clarifications on Two Phase Training
- Experiments and Results
- Conclusion

Conclusion

- Propose a non-iterative method to achieve adversarial robustness based on standard model training
- Much faster compared to traditional adversarial training (AT) based methods
- Significantly outperforms SOTA methods on adversarial accuracy without affecting natural accuracy
- Greatly applicable for practical applicability

Thank You