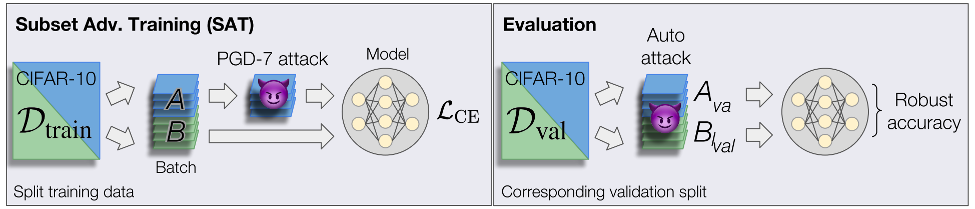
MAX PLANCK INSTITUTE FOR INFORMATICS

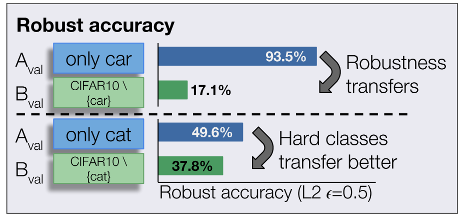
On Adversarial Training without Perturbing All Examples

Max Losch[MPII], Mohamed Omran[MPII], David Stutz[MPII], Mario Fritz[CISPA], Bernt Schiele[MPII]

CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY

In a Nutshell


Subset Adv. Training (SAT)
CIFAR-10
$\mathcal{D}_{\text{train}}$
Split training data
PGD-7 attack
Model
A
B
Batch
$\mathcal{L}_{\text{CE}}$
Evaluation
Auto attack
CIFAR-10
$\mathcal{D}_{\text{val}}$
A_{va}
B_{val}
Robust accuracy
Corresponding validation split



Vanilla adversarial training (AT) and most its variants perturb **every training example**. To what extent is that necessary? We split the training set into subsets A and B, train on $A \cup B$ but construct adv. examples only for examples in A.

Contributions

We propose an analytical tool *Subset Adversarial Training* (SAT) 1 to investigate robustness when only a training subset has been attacked.

2 Adv. robustness transfers to never attacked classes

3 & 4 Harder examples tend to provide best robustness transfer

4 Attacking 50% of training data is sufficient to recover baseline robust accuracy

5 30% reach baseline robust accuracy after transfer to downstream tasks

6 Can be combined with single-step attack training

Paper and Code: `github.com/mlosch/SAT`

[1] Dan Hendrycks et al. "Natural adversarial examples". In: *CVPR* (2021).
[2] Eric Wong, Leslie Rice, and J Zico Kolter. "Fast is better than free: Revisiting adversarial training". In: *ICLR* (2020).

1 Subset Adversarial Training (SAT)

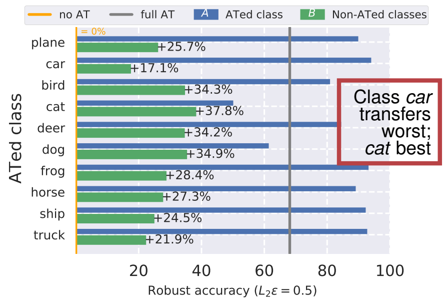
Questions: Does robustness transfer to unseen classes/examples? Does it depend on particular classes/examples?

$$\text{SAT:}\ \min_\theta \mathbb{E}_{(x,y)\in\mathcal{D}_{\text{train}}}\left[\underbrace{w_A \mathbb{1}_{(x,y)\in A} \max_{||\delta||_2\le\epsilon} \mathcal{L}(x+\delta,y;\theta)}_{\text{Adv. Training loss } \forall x\in A} + \underbrace{w_B \mathbb{1}_{(x,y)\in B}\mathcal{L}(x,y;\theta)}_{\text{Vanilla cross-entropy } \forall x\in B}\right]$$

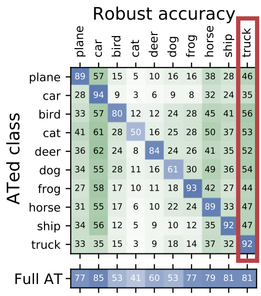
Note: A and B are *fixed* pre training

Examples in B are never attacked

2 Class-subset Splits (CSAT)



Class *car* transfers worst; *cat* best


Robust accuracy

ATed class	plane	car	bird	cat	deer	dog	frog	horse	ship	truck
plane	89	57	15	5	10	16	16	38	28	46
car	28	94	9	3	6	9	8	32	24	35
bird	33	57	80	12	12	24	28	45	41	56
cat	41	61	28	50	16	25	28	50	37	53
deer	36	62	24	8	84	24	26	41	35	52
dog	34	55	28	11	16	61	30	49	36	54
frog	27	58	17	10	11	18	93	42	27	44
horse	31	55	17	6	10	22	24	89	33	47
ship	34	56	12	5	9	10	12	35	92	47
truck	33	35	15	3	9	18	14	37	32	92
Full AT	77	85	53	41	60	53	77	79	81	81

Surprising effect:

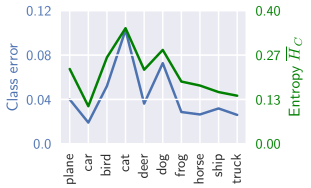
Class *car* transfers worse to *truck* than class *bird*!

- Non-trivial robust accuracy on classes in B
- Characteristics correlate strongly with class *difficulty*

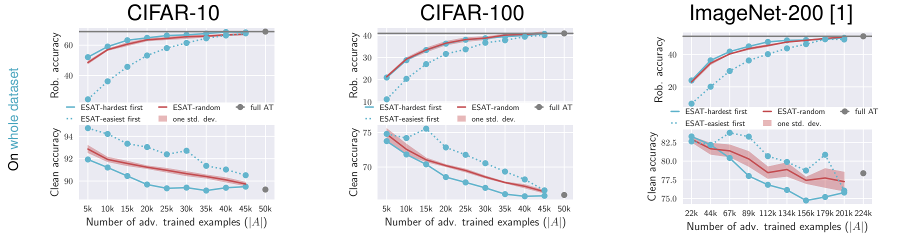
3 Measuring Class Difficulty

As class difficulty metric, we utilize entropy $\mathcal{H}$ over softmax σ. We rank examples once before training.

$$\mathcal{H}(f(x)) = -\sum_{i=1}^{N}\sigma_i(f(x))\cdot\log\sigma_i(f(x))$$

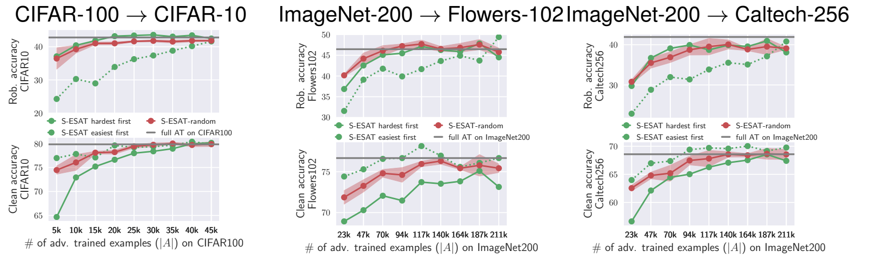


4 Example-subset Splits (ESAT)



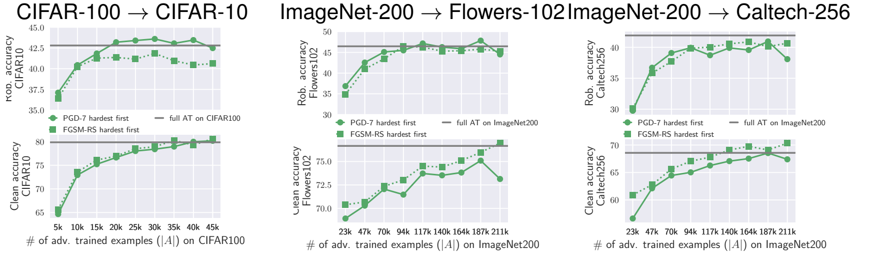
- Hard examples (solid lines) provide best robustness transfer
- 50% of hardest examples provide near baseline rob. accuracy

5 Transfer to Downstream Tasks (S-ESAT)



- Only 30% of examples provide near baseline rob. accuracy

6 Single-step S-ESAT



- Results generalize when combined with single-step attacks (FGSM[2])