Generalist: Decoupling Natural and Robust Generalization

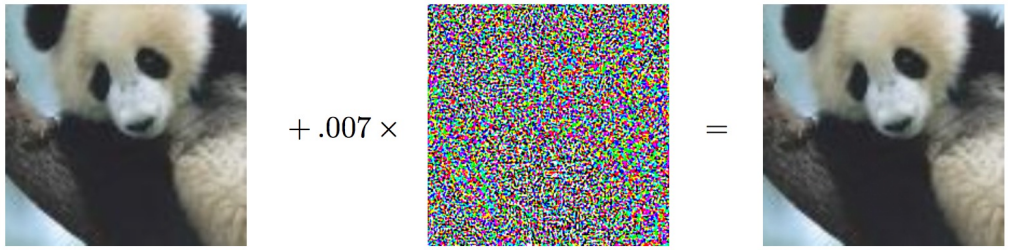
Hongjun Wang Yisen Wang

Peking University

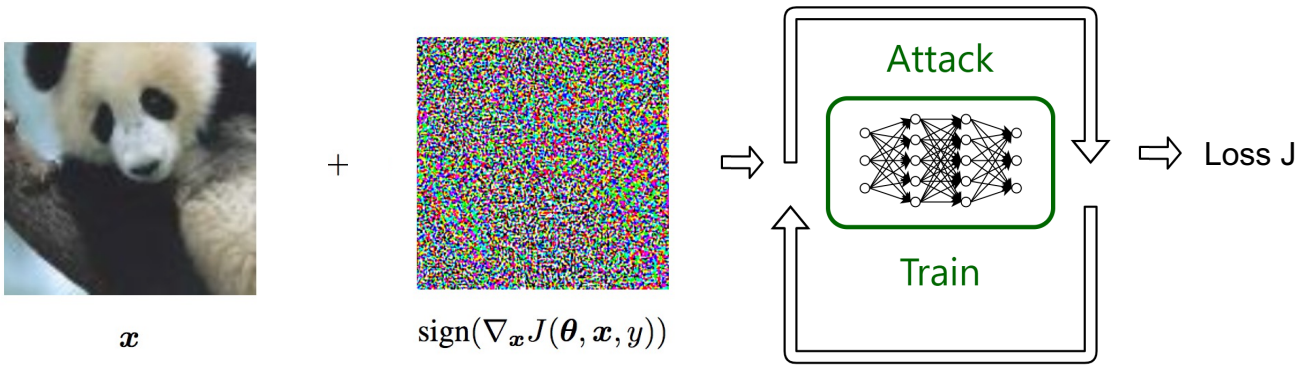
Paper: https://arxiv.org/pdf/2303.13813.pdf
Code: https://github.com/PKU-ML/Generalist

Adversarial Training

Adversarial attack

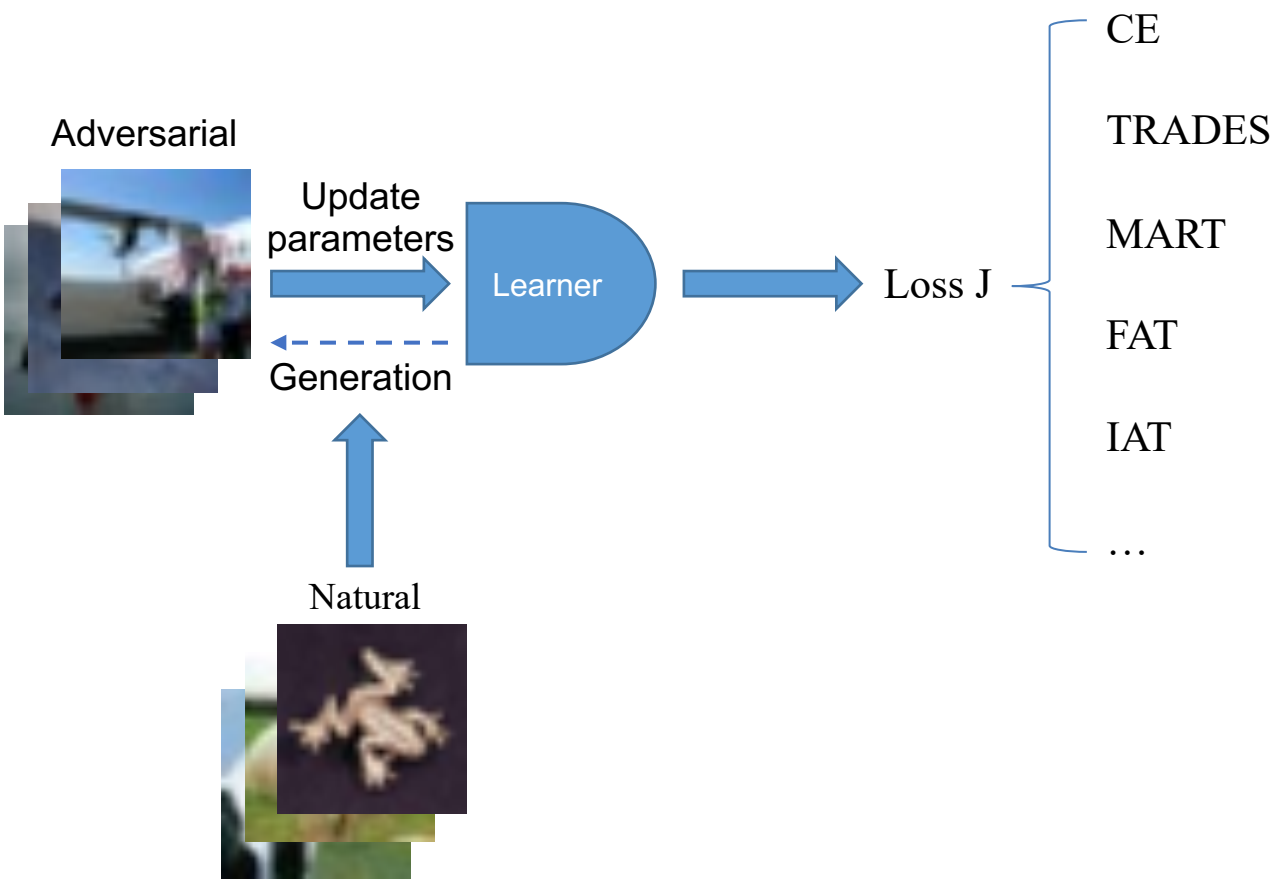


Adversarial training



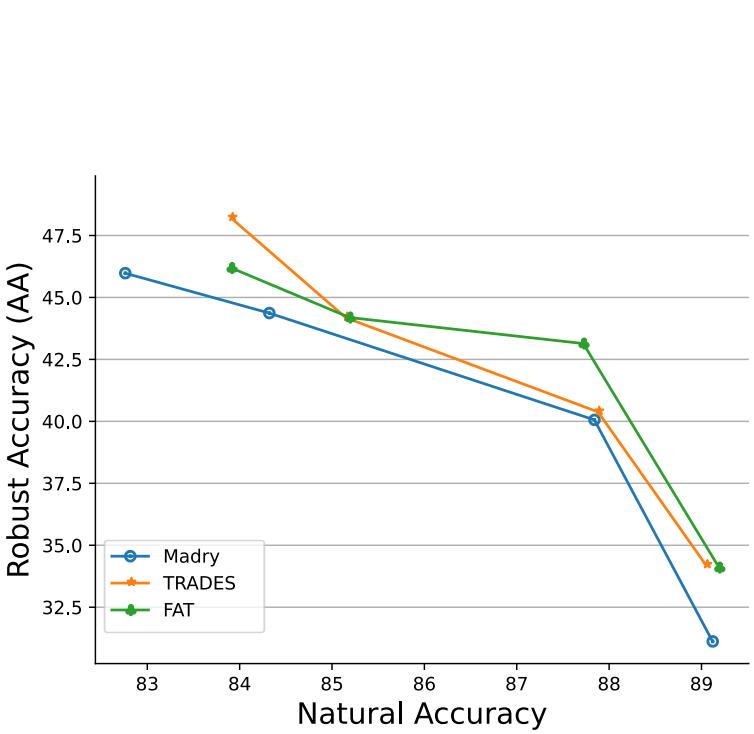
Goodfellow et. al. in Explaining and Harnessing Adversarial Examples

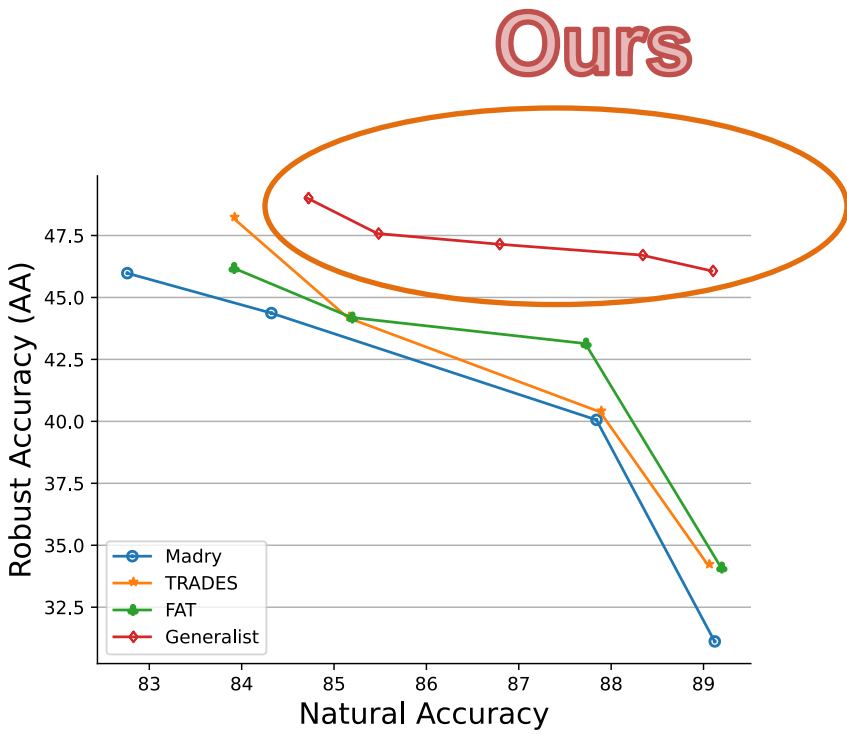
One learner for all?



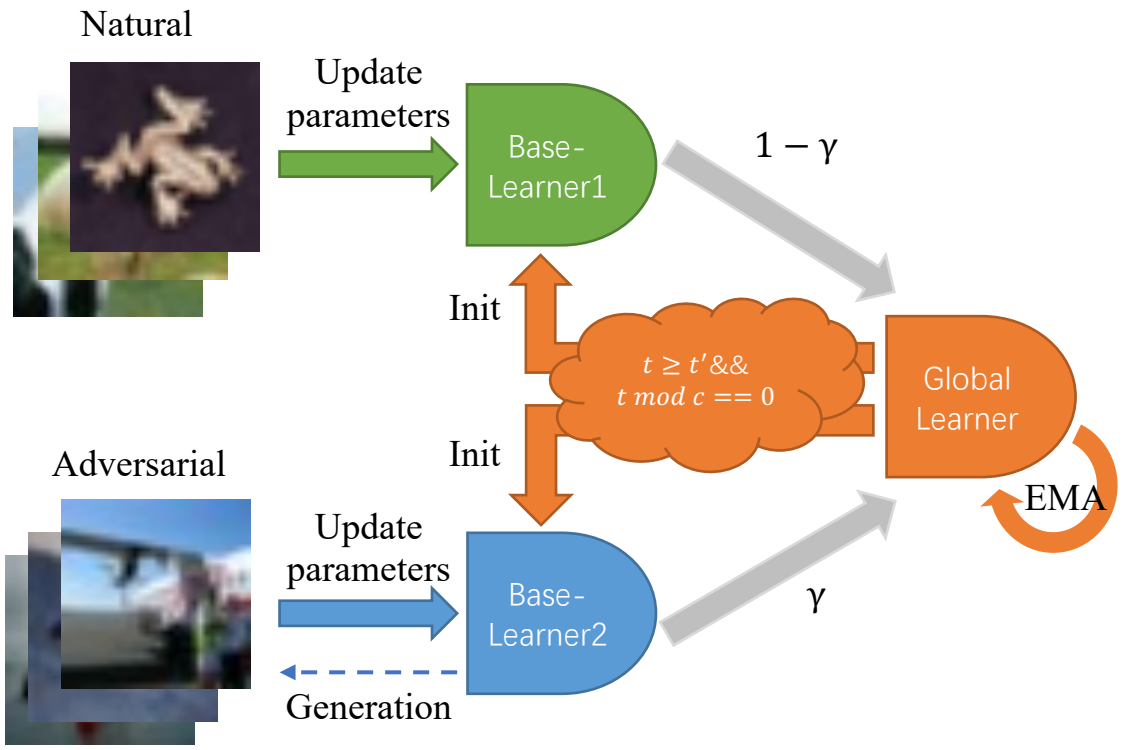
Motivation

- Undesirable increase in the natural error when the adversarial error decreases (e.g. TRADES, FAT)
- Not flexible training configurations in the joint training framework



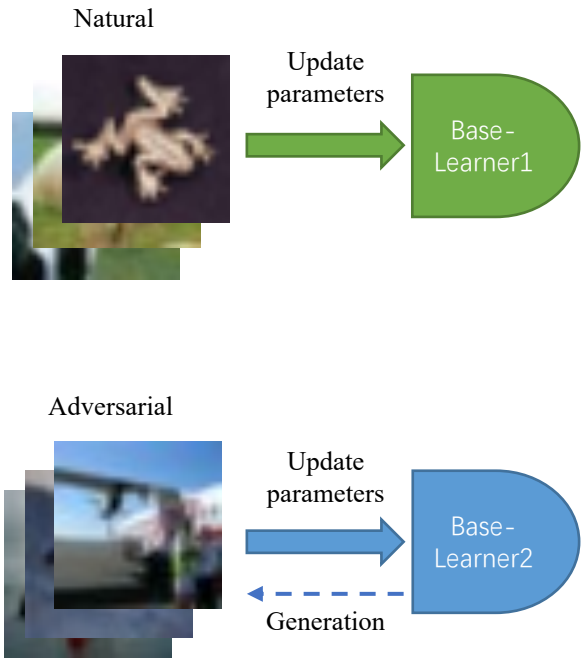


Framework


Natural
Update parameters
Base-Learner1
$1-\gamma$
Init
$t \geq t' \&\&$
$t \bmod c == 0$
Global Learner
Init
EMA
Adversarial
Update parameters
Base-Learner2
γ
Generation

Framework

- Epoch < t'



for t $\leftarrow 1, 2, \cdots, T$ **do**

Sample a minibatch (x, y) from data distribution $\mathcal{D}_1$

/ Parallel-1: Update parameters of base learner-1 over $\mathcal{D}_1$ */*

(Optional) Performing model ensembling, data augmentation or label smoothing, etc.

$\theta_n \leftarrow \mathcal{Z}_n \left[\mathbb{E}_{(x,y)}(\nabla_\theta \ell_1(x, y; \theta_n)), \tau_n\right]$

/ Parallel-2: Update parameters of base learner-2 over $\mathcal{D}_2$ */*

$x'_0 \leftarrow x + \varepsilon, \varepsilon \sim \text{Uniform}(-\varepsilon, \varepsilon)$.

for k $\leftarrow 1, 2, \cdots, K$ **do**

$$x'_k \leftarrow \Pi_{x'_k \in \mathbb{B}_\varepsilon(x)} \left(\kappa \,\text{sign}\left(x'_{k-1} + \nabla_{x'_{k-1}} \ell_2(x'_{k-1}, y; \theta_r)\right)\right)$$

end for

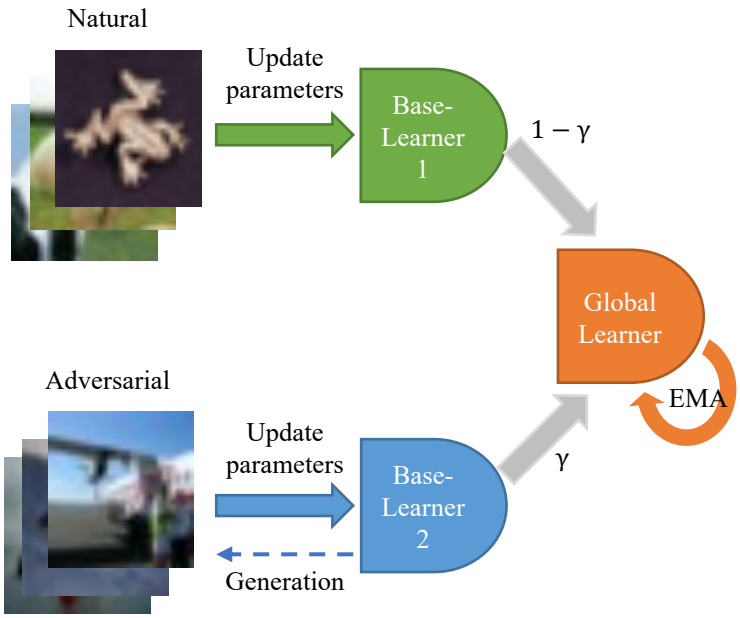
(Optional) Performing model ensembling, data augmentation or label smoothing, etc.

$\theta_r \leftarrow \mathcal{Z}_r \left[\mathbb{E}_{(x',y)}(\nabla_\theta \ell_2(x'_K, y; \theta_r)), \tau_r\right]$

end for

Framework

- Epoch < t'



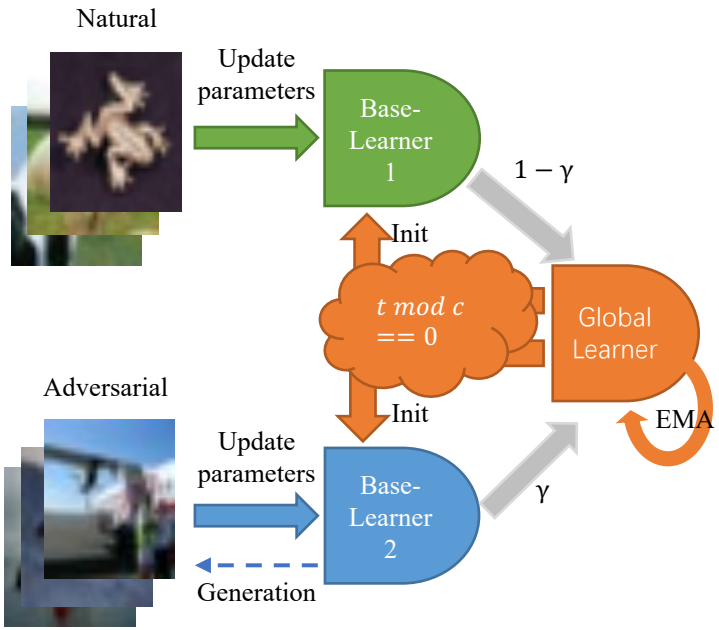
for t $\leftarrow 1, 2, \cdots, T$ **do**
  Sample a minibatch (x, y) from data distribution $\mathcal{D}_1$

/ For the global learner*/*
$\theta_g \leftarrow \alpha'\theta_g + (1-\alpha')(\gamma\theta_r + (1-\gamma)\theta_n)$

end if
end for

Framework

- Epoch $\geq t'$



for t $\leftarrow 1, 2, \cdots, T$ **do**
  Sample a minibatch (x, y) from data distribution $\mathcal{D}_1$

  if $t \geq t'$ and $t \mod c == 0$ **then**
    $\theta_r, \theta_n \leftarrow \theta_g$
  end if
end for

Advantages

- Decouple task-aware assignments from joint training
  - Each base learner can wield customized strategies (e.g., EMA, augmentations) for better performance
  - Lower error in sub-tasks results in a lower error bound for the global learner (Theorem 1)

- Initialize base learners from the global learner
  - Enable fast learning within a given assignment and improve generalization (Claim in Section 3.3)

Experiments

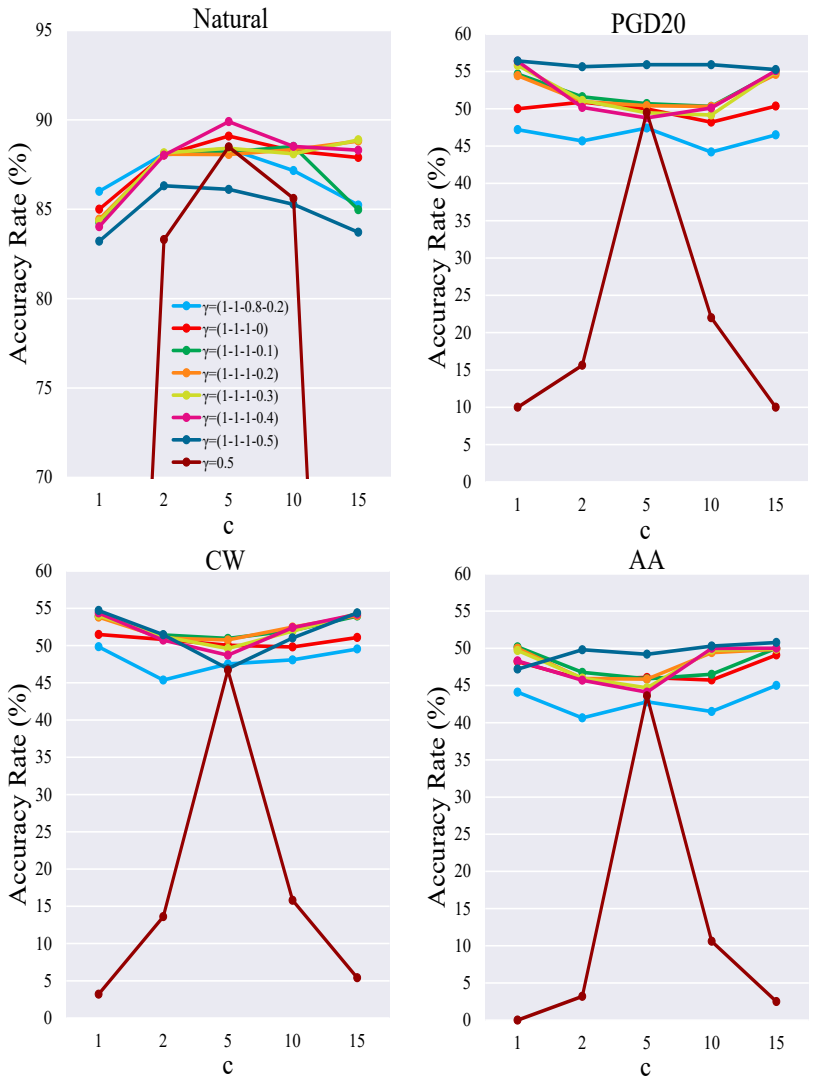
- ResNet-18 on CIFAR-10

Method	NAT	PGD20	PGD100	MIM	CW	APGD_{ce}	APGD_{dlr}	APGD_{t}	FAT_{t}	Square	AA
NT	**93.04**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AT ($\beta = 1$)	84.32	48.29	48.12	47.95	49.57	**47.47**	48.57	45.14	46.17	54.21	44.37
AT ($\beta = 1/2$)	87.84	44.51	44.53	47.30	44.93	40.58	42.55	40.20	44.56	50.76	40.06
TRADES ($\lambda = 6$)	83.91	**54.25**	**52.21**	**55.65**	**52.22**	**53.47**	**50.89**	**48.23**	**48.53**	**55.75**	**48.20**
TRADES ($\lambda = 1$)	87.88	45.58	45.60	47.91	45.05	42.95	42.49	40.38	43.89	53.49	40.32
FAT	87.72	46.69	46.81	47.03	49.66	46.20	47.51	44.88	45.76	52.98	43.14
IAT	84.60	40.83	40.87	43.07	39.57	37.56	37.95	35.13	36.06	49.30	35.13
RST	84.71	44.23	44.31	45.33	42.82	41.25	42.01	40.41	46.54	50.49	37.68
Generalist	**89.09**	**50.01**	**50.00**	**52.19**	**50.04**	46.53	**48.70**	**46.37**	**47.32**	**56.68**	**46.07**

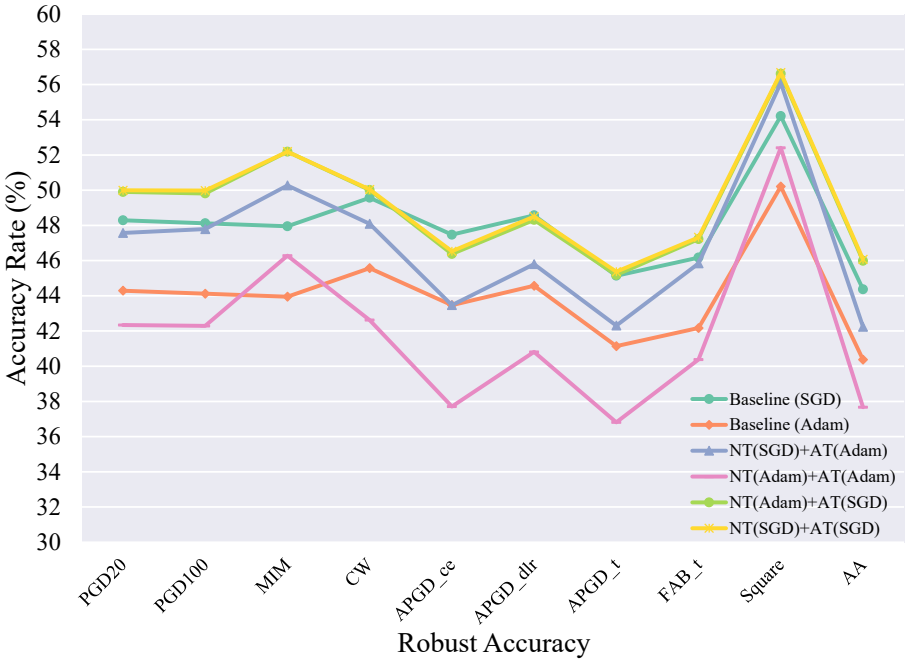
- WRN-32-10 on CIFAR-10

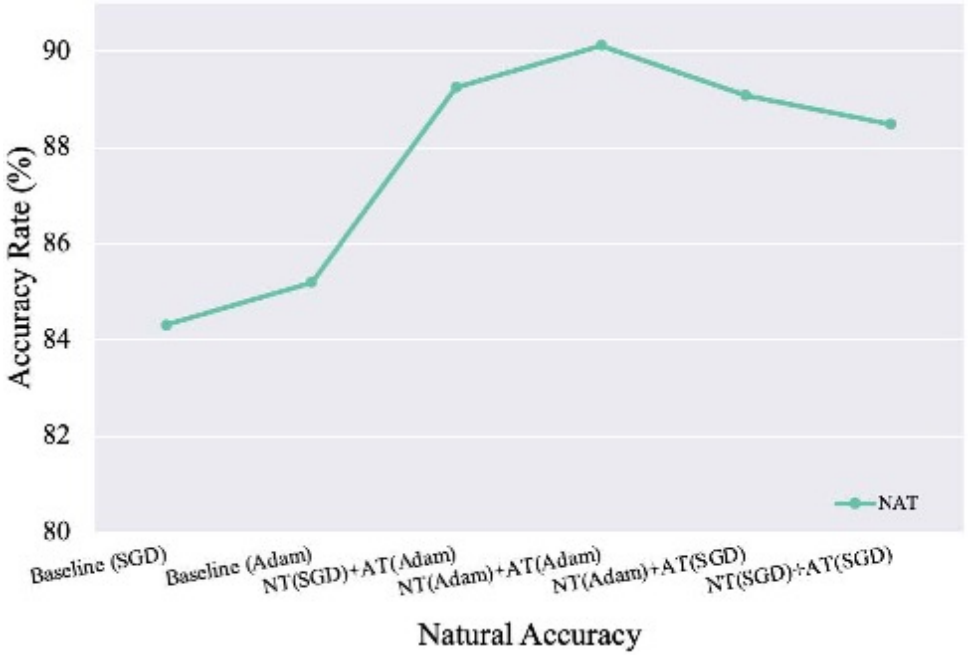
Method	NAT	PGD20	PGD100	MIM	CW	APGD_{ce}	APGD_{dlr}	APGD_{t}	FAT_{t}	Square	AA
NT	**93.30**	0.01	0.02	0.05	0.00	0.00	0.00	0.00	0.87	0.28	0.00
AT ($\beta = 1$)	87.32	49.01	48.83	48.25	52.80	48.83	49.00	46.34	48.17	54.26	46.11
AT ($\beta = 1/2$)	89.27	48.95	48.86	51.35	49.56	45.98	47.66	44.89	46.42	56.83	44.81
TRADES ($\lambda = 6$)	85.11	**54.58**	**54.82**	**55.67**	**54.91**	**54.89**	**55.50**	**52.71**	**52.61**	**57.62**	**52.19**
TRADES ($\lambda = 1$)	87.20	51.33	51.65	52.47	53.19	51.60	51.88	49.97	50.01	54.83	49.81
FAT	89.65	48.74	48.69	48.24	52.11	48.50	48.81	46.70	46.17	51.51	44.73
IAT	87.93	50.55	50.72	52.37	48.71	47.71	46.55	43.84	45.78	56.52	43.80
RST	87.27	46.55	46.76	47.02	45.99	45.73	46.58	45.78	43.18	52.44	41.52
Generalist	**91.03**	**56.88**	**56.92**	**58.87**	**57.23**	**53.94**	**55.80**	**53.00**	**53.65**	**63.10**	**52.91**

Communication frequency and mixing ratio

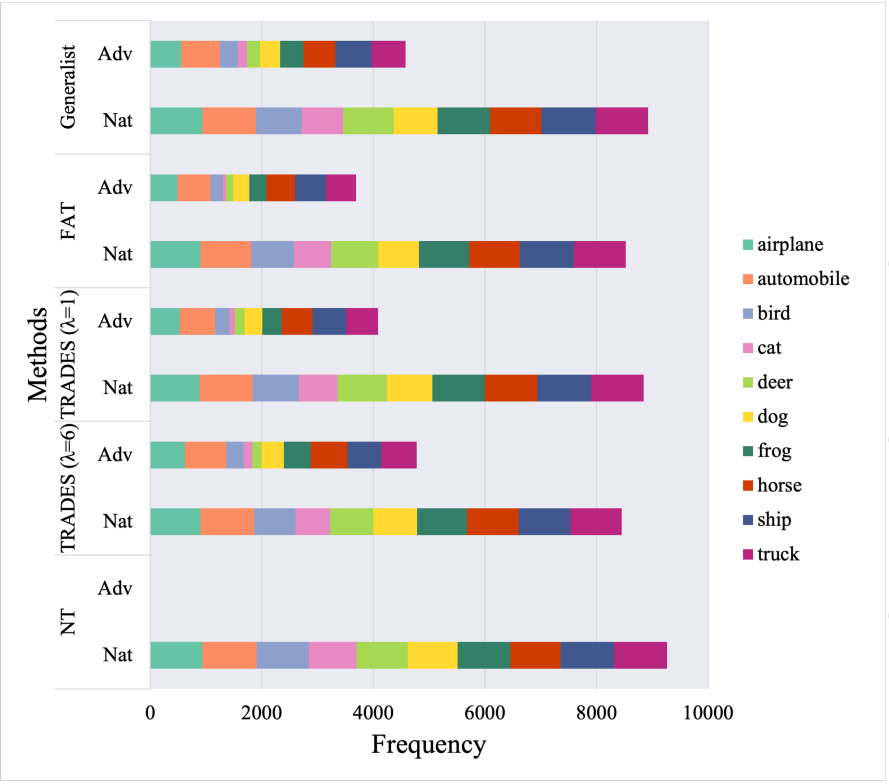


Different Optimizers





Visualization



(1) Easy


TRADES: dog
Generalist: horse
Label: horse


TRADES: horse
Generalist: bird
Label: bird


TRADES: dog
Generalist: frog
Label: frog


TRADES: airplane
Generalist: bird
Label: bird


TRADES: airplane
Generalist: cat
Label: cat

(2) Hard


TRADES: bird
Generalist: cat
Label: cat


TRADES: horse
Generalist: dog
Label: dog


TRADES: truck
Generalist: bird
Label: bird


TRADES: cat
Generalist: deer
Label: deer


TRADES: truck
Generalist: cat
Label: cat

(3) Adv (TRADES)


TRADES: airplane
Generalist: ship
Label: ship


TRADES: dog
Generalist: cat
Label: cat


TRADES: truck
Generalist: airplane
Label: airplane


TRADES: dog
Generalist: bird
Label: bird


TRADES: horse
Generalist: deer
Label: derr

(4) Adv (FAT)


FAT: automobile
Generalist: truck
Label: truck


FAT: airplane
Generalist: bird
Label: bird


FAT: airplane
Generalist: ship
Label: ship


FAT: ship
Generalist: horse
Label: horse


FAT: frog
Generalist: deer
Label: deer

Conclusion

- Propose a bi-expert framework named Generalist for mitigating the tradeoff between natural and robust generalization
- By decoupling from the joint training paradigm, each base learner can wield customized strategies based on data distribution
- Theoretically and empirically justify the effectiveness of Generalist

Poster Session

Fri 23 Jun
1:30 a.m. CST — 3 a.m. CST

West Building Exhibit Halls ABC 388