JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

# Trade-off between Robustness and Accuracy of Vision Transformers

**Yanxi Li, Chang Xu**
*School of Computer Science*
*Faculty of Engineering*
*University of Sydney*

THE UNIVERSITY OF
SYDNEY

# Background & Motivation

Deep neural networks (DNNs) excel in computer vision tasks but are susceptible to input perturbations. The **trade-off** between natural accuracy and robustness remains a challenge, even for Vision Transformers (ViTs), which inherently exhibit robustness.

To address this, we propose **TORA-ViTs**, leveraging pretrained ViT models for both accuracy and robustness. TORA-ViTs comprise accuracy and robustness adapters, alongside a gated fusion module that balances the trade-off. Experimental results on ImageNet demonstrate that TORA-ViTs significantly enhance robustness while maintaining competitive accuracy.

# Preliminary

The common **supervised training** objective of vision transformers can be written as:

$$\mathcal{L}_{\mathrm{ACC}}(f;\mathcal{D}) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\ell_{\mathrm{CE}}(f(\boldsymbol{x}),y)].$$

**Adversarial training** is a common method to improve adversarial robustness, which can be formulated as a min-max problem:

$$\mathcal{L}_{\mathrm{ROB}}(f;\mathcal{D}) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\left[\max_{\boldsymbol{x}'\in\mathcal{B}_p(\boldsymbol{x},\varepsilon)} \ell_{\mathrm{CE}}(f(\boldsymbol{x}'),y)\right],$$

where $\mathcal{B}_p(\boldsymbol{x},\varepsilon) = \left\{\boldsymbol{x}' \colon \|\boldsymbol{x}-\boldsymbol{x}'\|_p \le \varepsilon\right\}$ is a $l_p$ ball.

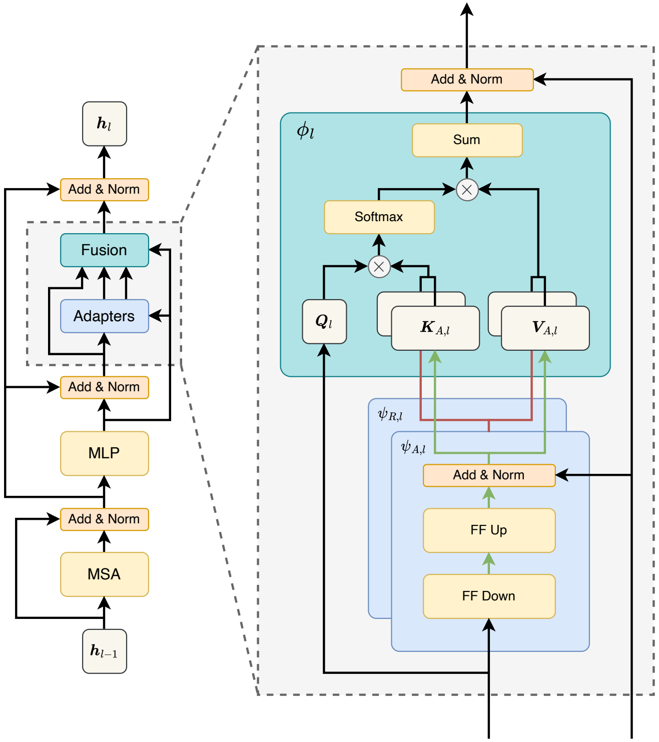
# The Architecture of TORA-ViTs

The architecture of **TORA-ViTs** consists of two major components:

1)a pair of adapters, including an **accuracy adapter** $\psi_{A,l}(\cdot)$ for extracting predictive features and a **robust adapter** $\psi_{R,l}(\cdot)$ for extracting robust features, and

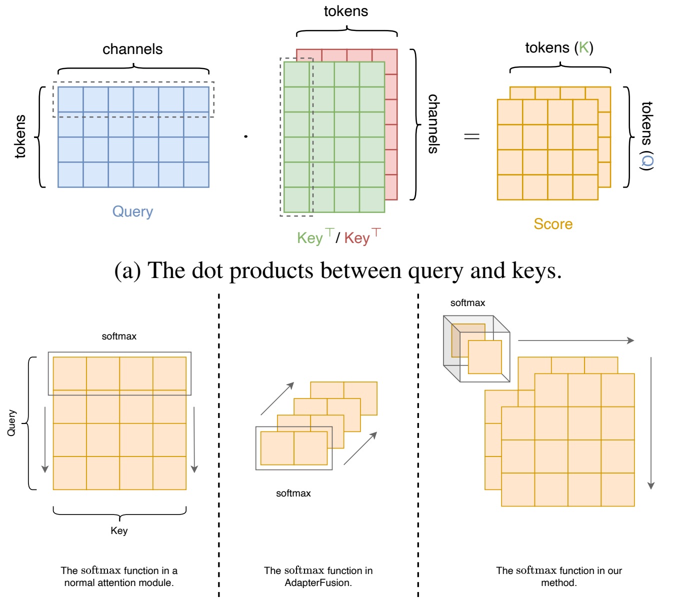
2)a **gated fusion module** $\phi_l(\cdot,\cdot)$ for combining those features as inputs for the next ViT block.

These components are inserted after the MLP layer in each ViT block.



# Attention-based Gated Fusion

To combine the predictive and robust features extracted by the accuracy and robustness adapters in a trade-off-aware manner, we propose an attention-based gated fusion module. We first calculate the dot-product attention score matrices between the features from the ViT blocks and adapters. Then, a softmax function is applied adapter-wise to the score matrices. The softmax results are used as a weighted gate to fuse the predictive and robust features.


channels
tokens
Query
·
tokens
channels
Key$^\top$ / Key$^\top$
=
tokens (K)
tokens (Q)
Score
softmax
Query
Key
The softmax function in a normal attention module.
softmax
The softmax function in AdapterFusion.
softmax
The softmax function in our method.

(a) The dot products between query and keys.

(b) Comparison of various methods to apply the softmax function in the attention mechanism.

# Trade-off Training

**Add** [ACC]/[ROB] **tokens:**

$$\boldsymbol{a}_l = \psi_{A,l}\left(\mathrm{Concat}\left([\mathrm{ACC}]_{l-1}, \boldsymbol{z}_{l,2:,:}\right)\right)$$

$$\boldsymbol{r}_l = \psi_{R,l}\left(\mathrm{Concat}\left([\mathrm{ROB}]_{l-1}, \boldsymbol{z}_{l,2:,:}\right)\right)$$

**Make prediction:**

$$\hat{y} = \frac{1}{2} f_{\mathrm{ACC}}([\mathrm{ACC}]_L) + \frac{1}{2} f_{\mathrm{ROB}}([\mathrm{ROB}]_L)$$

# Trade-off Training

**Two-Phase Trade-off Training**

**Phase 1.** Independent training:

$$\min_{\Psi_R,\Phi} \mathcal{L}_{\mathrm{ROB}}(F;\mathcal{D})$$

$$\min_{\Psi_A,\Phi} \mathcal{L}_{\mathrm{ACC}}(F;\mathcal{D})$$

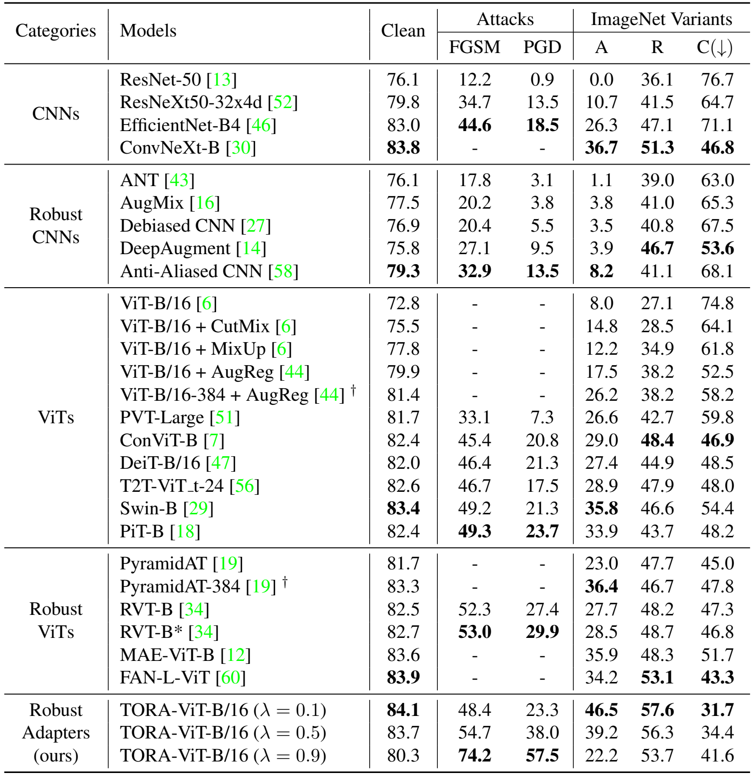
where $F = \{f, \Psi_R, \Psi_A, \Phi\}$, with $\Psi_R = \{\psi_{R,l} \middle| 1 \le l \le L\}$, $\Psi_A = \{\psi_{A,l} \middle| 1 \le l \le L\}$, $\Phi = \{\phi_l | 1 \le l \le L\}$.

**Phase 2.** Joint training:

$$\min_{\Phi} \lambda\, \mathcal{L}_{\mathrm{ROB}}(F;\mathcal{D}) + (1-\lambda)\mathcal{L}_{\mathrm{ACC}}(F;\mathcal{D})$$

# Experiments

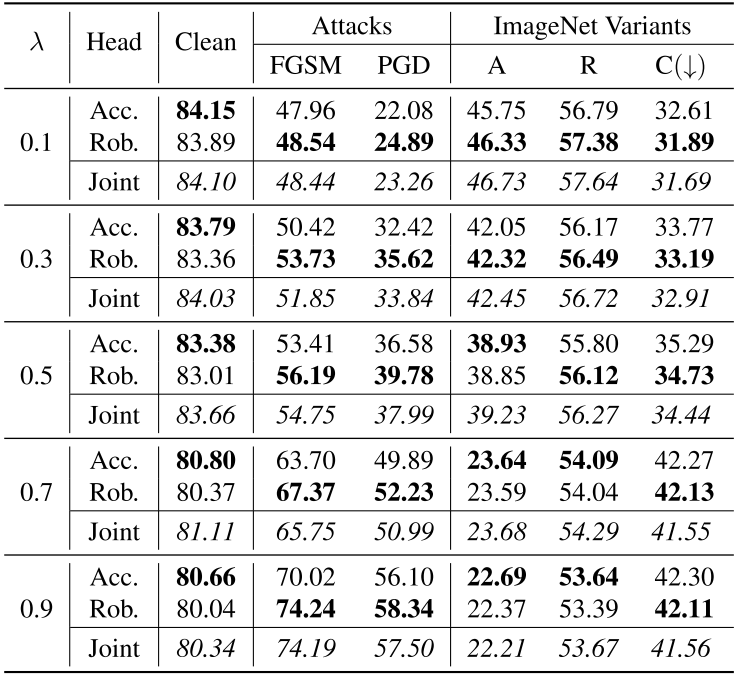
1. Performance on ImageNet-1K and variants (ImageNet-A/R/C)



| Categories | Models | Clean | Attacks | | ImageNet Variants | | |
|---|---|---|---|---|---|---|---|
| | | | FGSM | PGD | A | R | C(↓) |
| CNNs | ResNet-50 [13] | 76.1 | 12.2 | 0.9 | 0.0 | 36.1 | 76.7 |
| | ResNeXt50-32x4d [52] | 79.8 | 34.7 | 13.5 | 10.7 | 41.5 | 64.7 |
| | EfficientNet-B4 [46] | 83.0 | **44.6** | **18.5** | 26.3 | 47.1 | 71.1 |
| | ConvNeXt-B [30] | **83.8** | - | - | **36.7** | **51.3** | **46.8** |
| Robust CNNs | ANT [43] | 76.1 | 17.8 | 3.1 | 1.1 | 39.0 | 63.0 |
| | AugMix [16] | 77.5 | 20.2 | 3.8 | 3.8 | 41.0 | 65.3 |
| | Debiased CNN [27] | 76.9 | 20.4 | 5.5 | 3.5 | 40.8 | 67.5 |
| | DeepAugment [14] | 75.8 | 27.1 | 9.5 | 3.9 | **46.7** | **53.6** |
| | Anti-Aliased CNN [58] | **79.3** | **32.9** | **13.5** | **8.2** | 41.1 | 68.1 |
| ViTs | ViT-B/16 [6] | 72.8 | - | - | 8.0 | 27.1 | 74.8 |
| | ViT-B/16 + CutMix [6] | 75.5 | - | - | 14.8 | 28.5 | 64.1 |
| | ViT-B/16 + MixUp [6] | 77.8 | - | - | 12.2 | 34.9 | 61.8 |
| | ViT-B/16 + AugReg [44] | 79.9 | - | - | 17.5 | 38.2 | 52.5 |
| | ViT-B/16-384 + AugReg [44] † | 81.4 | - | - | 26.2 | 38.2 | 58.2 |
| | PVT-Large [51] | 81.7 | 33.1 | 7.3 | 26.6 | 42.7 | 59.8 |
| | ConViT-B [7] | 82.4 | 45.4 | 20.8 | 29.0 | **48.4** | **46.9** |
| | DeiT-B/16 [47] | 82.0 | 46.4 | 21.3 | 27.4 | 44.9 | 48.5 |
| | T2T-ViT$_t$-24 [56] | 82.6 | 46.7 | 17.5 | 28.9 | 47.9 | 48.0 |
| | Swin-B [29] | **83.4** | 49.2 | 21.3 | **35.8** | 46.6 | 54.4 |
| | PiT-B [18] | 82.4 | **49.3** | **23.7** | 33.9 | 43.7 | 48.2 |
| Robust ViTs | PyramidAT [19] | 81.7 | - | - | 23.0 | 47.7 | 45.0 |
| | PyramidAT-384 [19] † | 83.3 | - | - | **36.4** | 46.7 | 47.8 |
| | RVT-B [34] | 82.5 | 52.3 | 27.4 | 27.7 | 48.2 | 47.3 |
| | RVT-B* [34] | 82.7 | **53.0** | **29.9** | 28.5 | 48.7 | 46.8 |
| | MAE-ViT-B [12] | 83.6 | - | - | 35.9 | 48.3 | 51.7 |
| | FAN-L-ViT [60] | **83.9** | - | - | 34.2 | **53.1** | **43.3** |
| Robust Adapters (ours) | TORA-ViT-B/16 ($\lambda = 0.1$) | **84.1** | 48.4 | 23.3 | **46.5** | **57.6** | **31.7** |
| | TORA-ViT-B/16 ($\lambda = 0.5$) | 83.7 | 54.7 | 38.0 | 39.2 | 56.3 | 34.4 |
| | TORA-ViT-B/16 ($\lambda = 0.9$) | 80.3 | **74.2** | **57.5** | 22.2 | 53.7 | 41.6 |

# Experiments

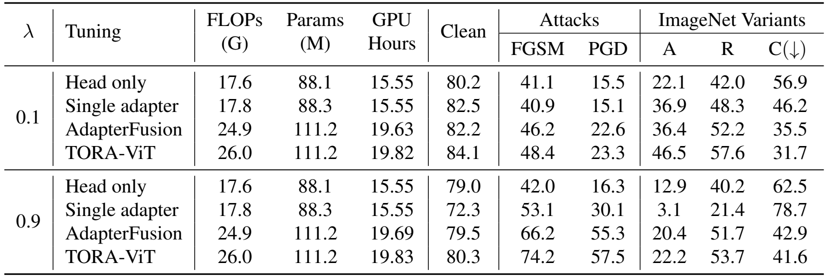
2. Performance of different heads and their joint prediction with different λ.



| $\lambda$ | Head | Clean | Attacks | | ImageNet Variants | | |
|---|---|---|---|---|---|---|---|
| | | | FGSM | PGD | A | R | C(↓) |
| 0.1 | Acc. | **84.15** | 47.96 | 22.08 | 45.75 | 56.79 | 32.61 |
| | Rob. | 83.89 | **48.54** | **24.89** | **46.33** | **57.38** | **31.89** |
| | Joint | *84.10* | *48.44* | *23.26* | *46.73* | *57.64* | *31.69* |
| 0.3 | Acc. | **83.79** | 50.42 | 32.42 | 42.05 | 56.17 | 33.77 |
| | Rob. | 83.36 | **53.73** | **35.62** | **42.32** | **56.49** | **33.19** |
| | Joint | *84.03* | *51.85* | *33.84* | *42.45* | *56.72* | *32.91* |
| 0.5 | Acc. | **83.38** | 53.41 | 36.58 | **38.93** | 55.80 | 35.29 |
| | Rob. | 83.01 | **56.19** | **39.78** | 38.85 | **56.12** | **34.73** |
| | Joint | *83.66* | *54.75* | *37.99* | *39.23* | *56.27* | *34.44* |
| 0.7 | Acc. | **80.80** | 63.70 | 49.89 | **23.64** | **54.09** | 42.27 |
| | Rob. | 80.37 | **67.37** | **52.23** | 23.59 | 54.04 | **42.13** |
| | Joint | *81.11* | *65.75* | *50.99* | *23.68* | *54.29* | *41.55* |
| 0.9 | Acc. | **80.66** | 70.02 | 56.10 | **22.69** | **53.64** | 42.30 |
| | Rob. | 80.04 | **74.24** | **58.34** | 22.37 | 53.39 | **42.11** |
| | Joint | *80.34* | *74.19* | *57.50* | *22.21* | *53.67* | *41.56* |

# Experiments

3. Comparison of different tuning methods



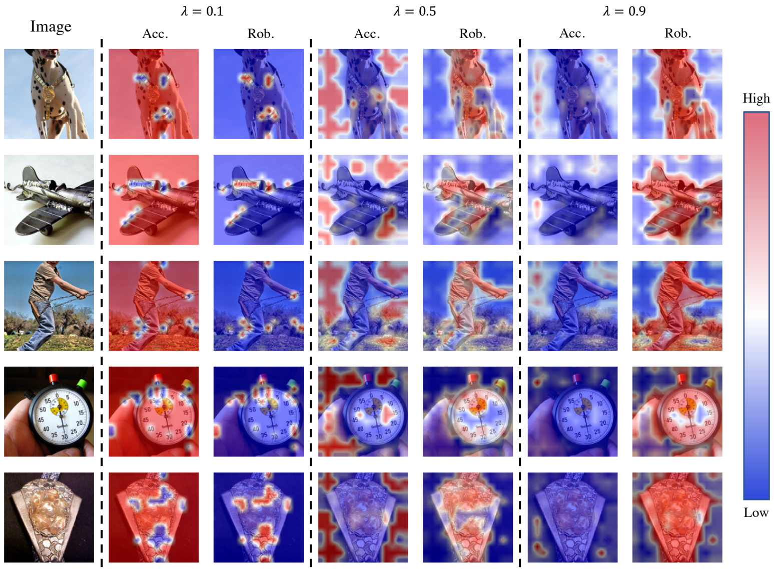
| $\lambda$ | Tuning | FLOPs (G) | Params (M) | GPU Hours | Clean | Attacks | | ImageNet Variants | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | FGSM | PGD | A | R | C(↓) |
| 0.1 | Head only | 17.6 | 88.1 | 15.55 | 80.2 | 41.1 | 15.5 | 22.1 | 42.0 | 56.9 |
| | Single adapter | 17.8 | 88.3 | 15.55 | 82.5 | 40.9 | 15.1 | 36.9 | 48.3 | 46.2 |
| | AdapterFusion | 24.9 | 111.2 | 19.63 | 82.2 | 46.2 | 22.6 | 36.4 | 52.2 | 35.5 |
| | TORA-ViT | 26.0 | 111.2 | 19.82 | 84.1 | 48.4 | 23.3 | 46.5 | 57.6 | 31.7 |
| 0.9 | Head only | 17.6 | 88.1 | 15.55 | 79.0 | 42.0 | 16.3 | 12.9 | 40.2 | 62.5 |
| | Single adapter | 17.8 | 88.3 | 15.55 | 72.3 | 53.1 | 30.1 | 3.1 | 21.4 | 78.7 |
| | AdapterFusion | 24.9 | 111.2 | 19.69 | 79.5 | 66.2 | 55.3 | 20.4 | 51.7 | 42.9 |
| | TORA-ViT | 26.0 | 111.2 | 19.83 | 80.3 | 74.2 | 57.5 | 22.2 | 53.7 | 41.6 |

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

# Visualization of Attention Maps

The visualization shows the attentions for different adapters in the gated fusion module with various ratios ($\lambda$). A color map ranging from *blue* to *white* to *red* is used, where red indicates high attention and blue indicates low attention.

It is evident that the features generated by the accuracy adapter prioritize context, while the features produced by the robustness adapter concentrate on the main object to be classified. This observation aligns with the theory of robust non-predictive and predictive non-robust features.



JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

# Thank you!

THE UNIVERSITY OF SYDNEY