Restricting In-variance and Co-variance of Representations for Adversarial Defense in Vision Transformers

Jeevithan Alagurajah and Henry Chu*

University of Louisiana at Lafayette

Lafayette, Louisiana, U.S.A.

* Email: chu@louisiana.edu

Henry Chu is the Lockheed Martin Professor at the University of Louisiana at Lafayette, where he teaches in the School of Computing and Informatics and leads the Informatics Research Institute. His technical interests are in machine vision and machine learning. His recent work in deep learning is in understanding the representation and flow of information within the architecture, with applications in adversarial defense and explainable AI in image classification. His recent research is sponsored by the U.S. National Science Foundation, U.S. Department of Energy, Louisiana Board of Regents, and the Louisiana Department of Health.

Chu received his B.S.E. *summa cum laude* and M.S.E., both in computer engineering, from the University of Michigan, U.S.A., and a Ph.D. in electrical and computer engineering from Purdue University, U.S.A. He is a senior member of the Institute for Electrical and Electronics Engineers, and a professional member of the Association for Computing Machinery.

Paramount Pictures; Skydance Media; TC Productions

Paramount Pictures; Skydance Media; TC Productions

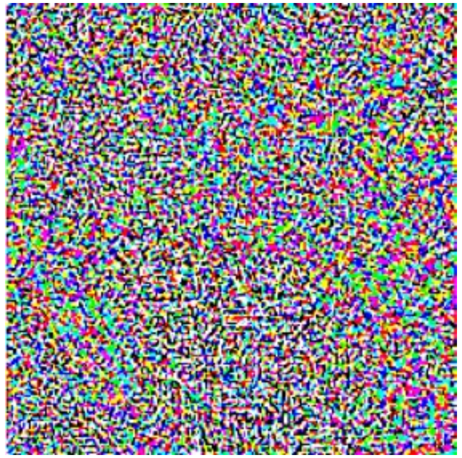
In MI7, US intelligence tries to locate Ethan Hunt (Tom Cruise) at Abu Dhabi Airport using facial-recognition software, but every time they think that they have found him, it turns out to be someone else — a handy trick pulled off by Hunt's pals Benji Dunn (Simon Pegg) and Luther Stickell (Ving Rhames).

The Evening Standard

How can facial-recognition algorithms be fooled?

Adversarial Attacks on Image Classification

$+ .007 \times$



$=$

$\boldsymbol{x}$

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"panda"
57.7% confidence

"nematode"
8.2% confidence

"gibbon"
99.3 % confidence

GoogLeNet prediction

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).

Adversarial In-variance and Co-variance (AICR) Loss

An objective function that creates maximum separation between classes and minimum variance between same class adversarial image and clean images

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{x}', \boldsymbol{y}) = \sum_{i=1}^{N} (\mathcal{L}_{CE}(\boldsymbol{x}_i, \boldsymbol{x}'_i, \boldsymbol{y}_i) + \mathcal{L}'(\boldsymbol{x}_i, \boldsymbol{x}'_i, \boldsymbol{y}_i))$$

Cross-entropy for classification accuracy

1. Attract-Repulse: To create maximum separation between different classes and make same class samples to pull closer
2. Variance: To make clean and adversarial samples to become closer

where

$\mathcal{L}'(\boldsymbol{x}_i, \boldsymbol{x}'_i, \boldsymbol{y}_i) = \sum_{l=1}^{n} (\mathcal{L}_{AR}(\boldsymbol{h}_i^{(l)}, \boldsymbol{h}_i'^{(l)}, \boldsymbol{y}_i) + \alpha \mathcal{L}_{var}(\boldsymbol{h}_i^{(l)}, \boldsymbol{h}_i'^{(l)}, \boldsymbol{y}_i))$

$\boldsymbol{h}^{(l)} = \mathcal{G}_{\boldsymbol{\phi}}^{(l)}(\mathcal{F}_{\boldsymbol{\theta}}^{(l)}(\boldsymbol{x}))$ and $\boldsymbol{h}'^{(l)} = \mathcal{G}_{\boldsymbol{\phi}}^{(l)}(\mathcal{F}_{\boldsymbol{\theta}}^{(l)}(\boldsymbol{x}'))$

$\mathcal{L}_{CE}$: Cross entropy loss

$\mathcal{F}_{\boldsymbol{\theta}}^{(l)}$: CNN representation extractor

$\mathcal{G}_{\boldsymbol{\phi}}^{(l)}$: Auxiliary mapping function

N: Number of instances

n: Number of layers that the loss function is being used

AICR Performance in CNN Adversarial Training

No defense

Trained using AICR loss function

Trained using AICR loss function and adversarial training with samples generated using FGSM

Objective	clean	White-Box Attacks FGSM	BIM	CW	MIM	PGD	Black-Box Attacks FGSM	BIM	CW	MIM	PGD
Mnist ($\epsilon = 0.3, c = 10$)											
$\mathcal{L}_{CE}$	**99.21**	7.1	0.8	4.3	.1	0.0	53.7	37.5	34.6	33.1	36.3
$\mathcal{L}_{AICR}$	99.17	94.8	**90.6**	**98.8**	**90.7**	**90.8**	95.0	95.5	**99.0**	94.5	96.8
$\mathcal{L}_{AICR} + AT_{FGSM}$	98.99	**98.4**	84.4	98.6	87.4	70.3	**97.4**	**97.0**	98.6	**97.1**	**97.8**
FashionMnist ($\epsilon = 0.3, c = 10$)											
$\mathcal{L}_{CE}$	**91.51**	7.9	0.1	0.2	0.01	0.0	42.6	21.3	29.6	32.1	27.7
$\mathcal{L}_{AICR}$	90.86	**67.2**	**56.9**	**57.8**	**55.8**	**46.6**	**82.6**	**84.2**	**88.6**	**81.8**	**85.8**
$\mathcal{L}_{AICR} + AT_{FGSM}$	91.43	59.6	48.7	23.9	49.0	29.9	74.3	71.3	87.1	68.5	74.7
CIFAR10 ($\epsilon = 0.03, c = 0.1$)											
$\mathcal{L}_{CE}$	90.70	20.4	0.0	0.6	0.0	0.0	38.4	29.6	30.3	28.5	27.6
$\mathcal{L}_{AICR}$	92.42	82.4	78.4	81.2	**79.8**	**78.6**	85.4	84.2	86.3	85.4	82.8
$\mathcal{L}_{AICR} + AT_{FGSM}$	**92.99**	**87.0**	**78.6**	**83.4**	79.0	72.3	**88.0**	**86.4**	**87.2**	**85.7**	**83.6**
CIFAR100 ($\epsilon = 0.03, c = 0.1$)											
$\mathcal{L}_{CE}$	**72.53**	19.5	4.1	1.6	3.4	0.17	39.5	32.8	37.2	34.6	28.9
$\mathcal{L}_{AICR}$	69.9	**40.2**	**26.8**	**31.2**	26.3	**24.2**	**57.6**	36.4	**41.7**	44.9	**47.2**
$\mathcal{L}_{AICR} + AT_{FGSM}$	70.2	43.2	23.4	26.4	**27.4**	23.1	53.5	**37.8**	38.9	**46.7**	42.5
SVHN ($\epsilon = 0.03, c = 0.1$)											
$\mathcal{L}_{CE}$	93.75	29.9	5.7	7.1	8.3	9.4	54.3	39.3	33.4	31.4	29.4
$\mathcal{L}_{AICR}$	**94.46**	78.9	47.4	51.7	**53.4**	42.1	83.2	78.9	**87.7**	**76.5**	**86.4**
$\mathcal{L}_{AICR} + AT_{FGSM}$	92.32	**82.1**	**51.1**	**57.8**	52.0	**56.7**	**83.4**	**79.8**	82.3	73.2	82.6

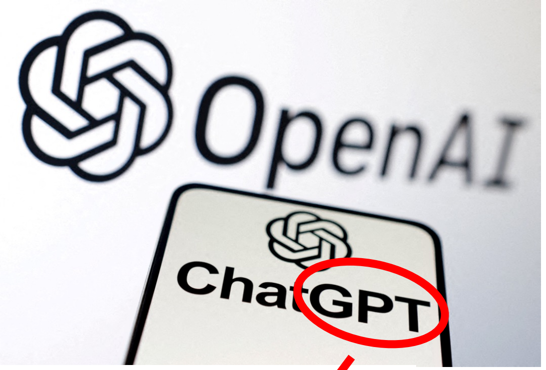
No attacks

More effective attacks

More realistic scenario

The AICR loss function is effective in training CNN to defend against adversarial attacks

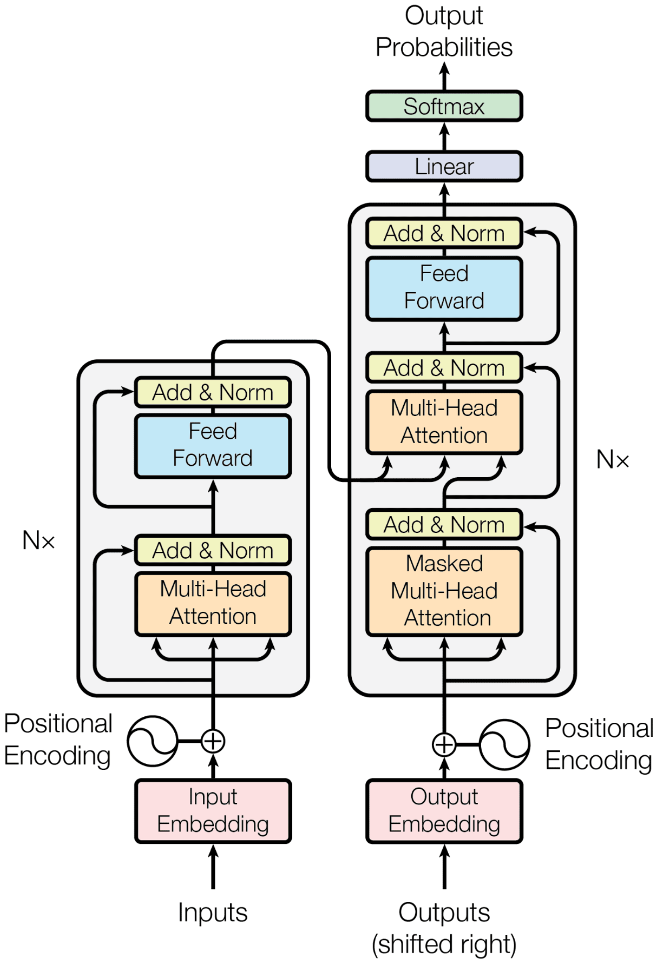
Deep Learning and Vision Transformers



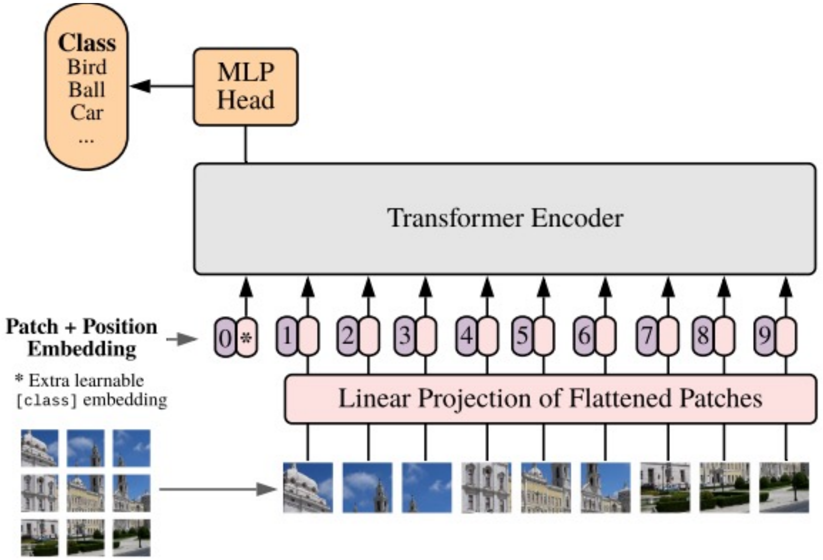
Reuters

Generative
Pre-trained
Transformer

Basic Transformer Architecture



Vision Transformer (ViT) Architecture



ViTs tend to outperform CNNs by a larger margin on large, complex datasets (e.g., ImageNet-21k) due to their superior ability to model long-range dependencies

AICR in Vision Transformers

How do we adopt the AICR loss function to the Vision Transformer architecture?

Component in the AICR loss function for adversarial defense

$$\mathcal{L}'(\boldsymbol{x}_i, \boldsymbol{x}'_i, \boldsymbol{y}_i) = \sum_{l=1}^{n} (\mathcal{L}_{AR}(\boldsymbol{h}_i^{(l)}, \boldsymbol{h}_i'^{(l)}, \boldsymbol{y}_i) + \alpha \mathcal{L}_{var}(\boldsymbol{h}_i^{(l)}, \boldsymbol{h}_i'^{(l)}, \boldsymbol{y}_i))$$

Attract-Repulse: To create maximum separation between different classes and make same class samples to pull closer

Variance: To make clean and adversarial samples to become closer

The attract-repulse loss depends on the average of the representations of each class. This is not possible to determine in ViT

AICR in ViT depends only on the variance loss function

Experiments and Results

	Attacks	ϵ	ViT	*ViT-C*	*ViT-All*
	No-attack	-	**80.1**	78.9	79.6
Fast Gradient Sign Method	FGSM	0.1	15.2	15.8	**16.2**
		0.2	2.7	1.8	**3.6**
Projected Gradient Descent	PGD	0.1	8.5	**9.9**	9.2
		0.2	0.15	**0.33**	0.16
Basic Iterative Method	BIM	0.1	8.4	**9.9**	9.1
		0.2	0.15	**0.33**	0.16
Momentum Iterative Method	MIM	0.1	8.8	**10.3**	9.6
		0.2	0.17	**0.37**	0.22

No defense

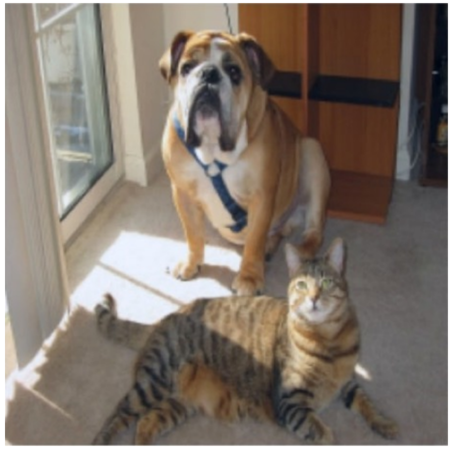
At the classification head

At the head and patches

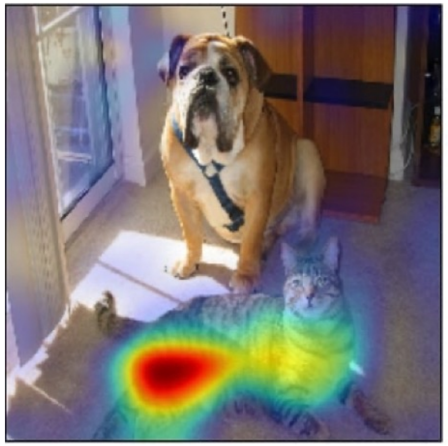
Trained using AICR loss function

Attacks Lead to Attention Shift

- Adversarial attacks succeed when they shift the attention of a classification network when presented with a perturbed copy of an image
- Gradient-weighted Class Activation Mapping (Grad-CAM), is a visualization technique of which parts of an image are most important to the model for classifying a particular object or scene



Original



Grad-Cam ‘Cat’

Grad-Cam ‘Dog’

Selvaraju, R. R., *et al.*, “Grad-CAM: Visual explanations from deep networks via gradient-based localization”, *arXiv e-prints*, 2016. doi:10.48550/arXiv.1610.02391

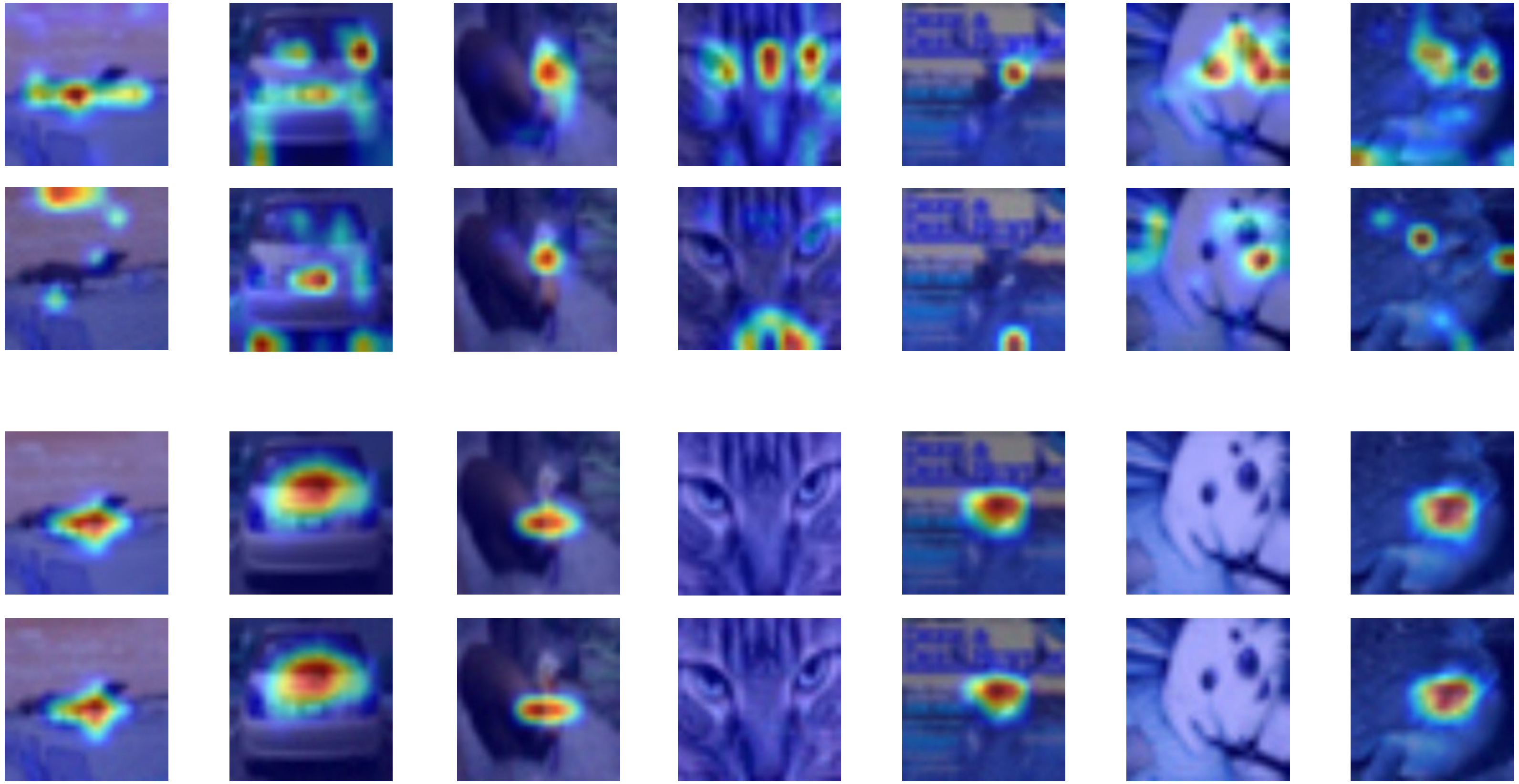
Undefended Attacks Lead to Attention Shifts in ViT

Clean image with no defense

Adversarial image with no defense

Clean image with AICR defense

Adversarial image with AICR defense



Significant attention shift

Minimal attention shift

Conclusion

- Image classification is the key component of many computer vision methods
- Adversarial attacks against image classification can lead to poor performance of computer vision tasks
- AICR loss was shown to be effective against adversarial attacks against CNN classification networks
- Vision transformers (ViTs) often have better image classification performance than CNNs
- We showed the efficacy of adopting the AICR loss to the ViTs

For more information

Jeevithan Alagurajah and Henry Chu*

University of Louisiana at Lafayette

Lafayette, Louisiana, U.S.A.

* Email: chu@louisiana.edu