Hierarchical Feature Binding in Convolutional Neural Networks: Making Adversarial Attacks Geometrically Challenging

Niels Leadholm, Simon Stringer
Oxford Lab for Theoretical Neuroscience and Artificial Intelligence, University of Oxford

Overview

- Approach robust machine vision with a novel deep-learning architecture using ideas from primate feature binding
  - Feature binding = how separately represented features are encoded in a relationally meaningful way (e.g. a small edge composing part of the larger contour of an object)

- Absence of such representations from current models such as CNNs might partly explain their vulnerability to subtly altered images --> adversarial examples (Szegedy et al., 2014).

- Literature suggests adversarial examples a result of 'off-manifold' perturbations, where decision boundary is poor (Tanay & Griffin, 2016; Khoury & Hadfield-Menell, 2018; Stutz et al., 2019)
  - We aim to capture hierarchical feature binding, providing representations in otherwise vulnerable directions

- CNNs with our modification empirically more robust against broad range of L_0, L_2 and L_∞ attacks in both the black-box and white-box setting on MNIST, FMNIST, and CIFAR-10 (latter results in associated paper under review)
  - However, model appears to still be vulnerable to sufficiently powerful attack (e.g. Brendel et al., 2019)

- Despite persistent vulnerability, evaluations support that modifications do indeed improve decision boundary, and as a result of preserved binding information

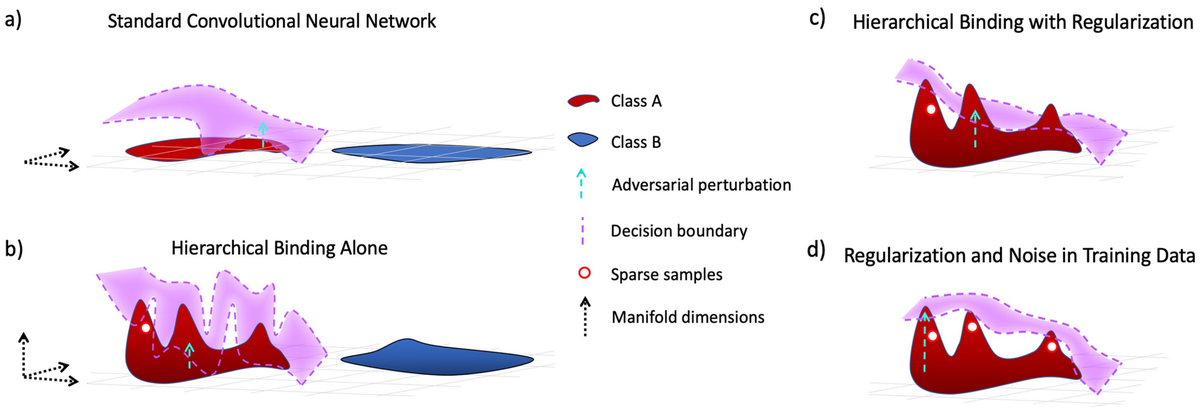

Figure 1: The effect of binding on the decision boundary

Red and blue represent two different object manifolds (e.g. cats and dogs)

a) Object classes often represented in low-dimensions : enables linear separating decision boundaries, but decision boundary unpredictable off the manifold. Potentially many vulnerable directions.

b) Aim is to preserve additional dimensions of variation. Hierarchical binding enables explicit representation of these features alongside more abstract dimensions. Due to the sparsity of samples in high-dimensions, this alone is insufficient.

c) + regularization --> sparse data points can inform a more useful decision boundary, and some enhanced model robustness is seen.

d) + regularization + noise --> further addresses the sampling problem, providing a more robust decision boundary (***Table 1***).

Our model beats adversarial training (Madry et al, 2018) on several black-box attacks, and on the All L_2 and All L_0 metric, but at a smaller cost in clean classification accuracy and with fewer (4x) parameters.

The Neuroscience

Feature binding :
--> the brain's ability to jointly represent features that are encoded separately (e.g. the colour and shape that jointly describe a yellow triangle) (Treisman,1998)
- binding can be ***hierarchical***: for example, when we look at a cat, we see not only that it is a cat, but also the particular spatial features of that feline, from edges to the possible presence of a scarred eye, or the absence of an ear

- Eguchi et al. (2018) proposed how the brain might capture hierarchical binding with spiking activity and temporal coincidence detection
- such binding would be consistent with observed neurons such as ***border-ownership cells*** (Zhou et al. 2000)

Adversarial examples in humans :
- under conditions of time-limited viewing and masking, humans show slight but significant sensitivity to adversarial examples (Elsayed et al, 2018)

- Eguchi et al (2018) proposed top-down and lateral computations needed for biological encoding of binding --> the experimental conditions of Elsayed et al (2018) could have disrupted these and contributed to transfer attack vulnerability

- until now limited literature explaining adversarial robustness to transfer/black-box attacks through biological means


Chihuahua... or ostrich?

Conclusion

- Implement a novel CNN architecture, inspired by recent work in theoretical neuroscience (Eguchi et al., 2018). This architecture seeks to capture hierarchical binding representations that encode ***the causal relations between lower level and higher-level visual features*** of an object.

- Within the framework of adversarial examples as off-manifold perturbations, we provided empirical evidence of enhanced robustness to a broad range of L_0, L_∞ and L_2 norm-measured attacks, all within both black box and white box settings.

Model Description

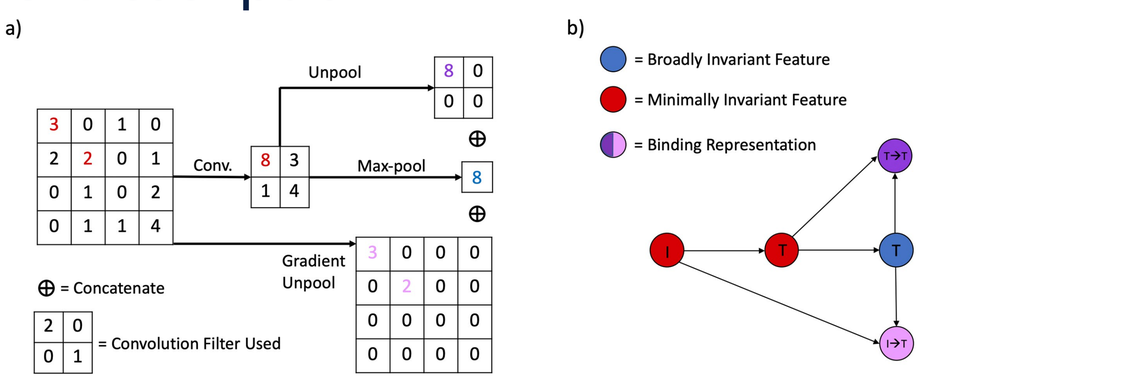

Figure 2: Implementing a Hierarchical Binding Convolutional Neural Netowrk (HBCNN)

- Unpooling captures which low-level features causally drove max-pooled representations
- 'Gradient unpooling' captures which simple features contributed to abstract representations. Broadly: gradient of max-pooled neurons taken w.r.t. neurons in lower layer; this informs a binary mask to preserving the γ most important low-level neurons.
- Unpooling + gradient unpooling representations are concatenated with max-pooling --> all provided to fully-connected layer.
- Figure 2b : how these operations relate to hierarchical feature binding as causal relations among features (Eguchi et al., 2018).

Results

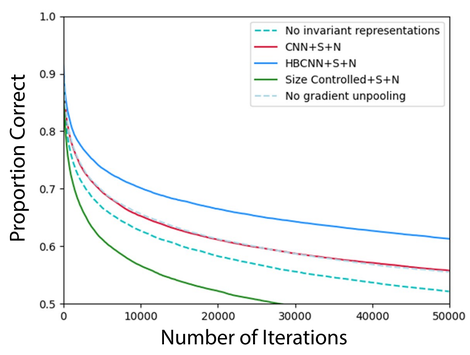

Figure 3: Hierarchical binding reduces the probability of finding vulnerable decision regions

Gaussian ball of constant magnitude requires more iterations to find vulnerable region. Robustness is not just e.g. gradient masking.

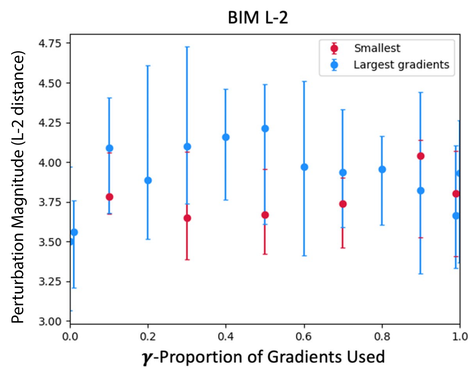

Figure 4: Effect of hierarchical binding dimension and causal role on robustness

Uprojecting most important low-level activations provides benefit, while using smallest gradients to derive binding confers less.

Table 1: MNIST Results

	CNN+S+N	**Ours** HBCNN+S+N	Size-Controlled CNN+S+N	CNN+AT
Clean accuracy	99.10%	99.05%	**99.43%**	98.40%
L_2**-metric** ($\epsilon = 1.5$)				
■ Transfer	3.1 (89%)	**3.2** (89%)	3.0 (92%)	3.7 (95%)
■ Uniform Noise	6.6 (99%)	**11.3** (99%)	6.2 (99%)	8.7 (98%)
■ Gaussian Noise	10.0 (99%)	**10.5** (99%)	9.7 (99%)	5.3 (97%)
■ Boundary	2.5 (88%)	**9.2** (98%)	2.1 (84%)	1.4 (42%)
■ Pointwise	4.4 (96%)	**4.5** (97%)	4.2 (97%)	1.9 (73%)
FGM	8.9 (95%)	9.6 (96%)	**10.0** (97%)	9.0 (98%)
FGM w/GE	8.8 (95%)	9.3 (96%)	**9.8** (97%)	∞ (97%)
DeepFool	7.3 (91%)	**8.1** (93%)	3.6 (89%)	9.4 (94%)
DeepFool w/GE	7.3 (93%)	**7.7** (94%)	4.6 (91%)	9.5 (94%)
BIM	3.6 (84%)	**4.0** (86%)	3.0 (85%)	4.9 (93%)
BIM w/GE	3.6 (84%)	**4.0** (87%)	3.1 (84%)	4.5 (93%)
PGD	2.5 (77%)	**2.8** (79%)	1.9 (71%)	2.8 (86%)
All L_2	2.1 (75%)	**2.5** (78%)	1.8 (69%)	1.4 (39%)
L_∞**-metric** ($\epsilon = 0.3$)				
■ Transfer	0.29 (43%)	**0.33** (60%)	0.26 (33%)	0.39 (94%)
FGSM	0.48 (76%)	**0.62** (82%)	0.50 (76%)	0.44 (95%)
FGSM w/GE	0.50 (78%)	**0.63** (83%)	0.51 (76%)	∞ (95%)
DeepFool	0.83 (81%)	**1.0** (85%)	0.32 (54%)	0.46 (94%)
DeepFool w/GE	**1.0** (86%)	**1.0** (86%)	0.45 (73%)	0.71 (94%)
BIM	0.34 (56%)	**0.44** (66%)	0.25 (31%)	0.36 (93%)
BIM w/GE	0.34 (57%)	**0.45** (67%)	0.25 (31%)	0.63 (93%)
MIM	0.32 (54%)	**0.42** (66%)	0.25 (33%)	0.34 (93%)
MIM w/GE	0.35 (56%)	**0.44** (68%)	0.26 (38%)	0.44 (94%)
PGD	0.22 (28%)	**0.24** (33%)	0.15 (0%)	0.33 (91%)
All L_∞	0.20 (20%)	**0.22** (25%)	0.15 (0%)	0.33 (91%)
L_0**-metric** ($\epsilon = 12$)				
■ Pointwise	22 (75%)	**23** (79%)	21 (76%)	5 (5%)
■ Salt&Pepper Noise	142 (97%)	135 (97%)	**156** (98%)	14 (57%)
All L_0	22 (75%)	**23** (79%)	21 (76%)	5 (5%)

Median L_p distance of a successful adversary; a larger value indicates the model is more difficult to fool. Bold indicates the best performance between the CNN+S+N, Size-Controlled CNN+S+N and HBCNN+S+N; blue indicates the best performance across all networks. HB-CNN = hierarchical binding convolutional neural network (our model); AT = adversarial training; S = label smoothing regularization; N = Gaussian and Salt-and- pepper noise during training; GE=Gradient Estimation. ■ = black-box attack

References

Brendel, W., Rauber, J., K ummerer, M., Ustyuzhaninov, I., Bethge, M., 2019. Accurate, reliable andfast robustness evaluation, in: Advances in Neural Information Processing Systems.
Akihiro Eguchi, James B. Isbister, Nasir Ahmad, and Simon Stringer. The emergence of polychronization and feature binding in a spiking neural network model of the primate ventral visual system. Psychological Review, 2018. ISSN 0033295X. doi: 10.1037/rev0000103.
Elsayed, G.F., Papernot, N., Shankar, S., Kurakin, A., Cheung, B., Goodfellow, I., Sohl-Dickstein, J., 2018. Adversarial examples that fool both computer vision and time-limited humans, in: Advances in Neural Information Processing Systems.
Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings, 2018.
David Stutz, Matthias Hein, and Bernt Schiele. Disentangling adversarial robustness and generalization. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2019. ISBN 9781728132938. doi: 10.1109/CVPR.2019.00714.
Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In 2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings, 2014.
Thomas Tanay and Lewis Griffin. A Boundary Tilting Persepective on the Phenomenon of Adversarial Examples. arXiv preprint arXiv:1608.07690, 8 2016. URL http://arxiv.org/abs/ 1608.07690.
Anne Treisman. Feature binding, attention and object perception. Philosophical Transactions of the Royal Society B: Biological Sciences, 1998. ISSN 09628436. doi: 10.1098/rstb.1998.0284.
H Zhou, H S Friedman, and Rudiger von der Heydt. Coding of border ownership in monkey visual cortex. The Journal of Neuroscience, 20(17):6594–6611, 2000. ISSN 0270-6474. doi: 10.1523/ JNEUROSCI.2797-12.2013.

BBSRC bioscience for the future