On the Vulnerability of Capsule Networks to Adversarial Attacks

Felix Michels*, Tobias Uelwer*, Eric Upschulte*, Stefan Harmeling

Department of Computer Science,
Heinrich-Heine-Universität Düsseldorf, Germany
*Equal contribution

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

Our Contribution

- We extensively evaluate the vulnerability of capsule networks to different adversarial attacks.
- Our experiments show that capsule networks can be fooled by white-box and black-box attacks as easily as convolutional neural networks.
- Adversarial examples can be transferred between capsule networks and convolutional neural networks.

Introduction

Recently capsule networks (CapsNets) [1] have been shown to be a reasonable alternative to convolutional neural networks (ConvNets). For our experiments we focus on CapsNets using the dynamic routing algorithm [1]. Frosst et al. [2] state that CapsNets are more robust against white-box adversarial attacks than other architectures.

Methods

- **Carlini-Wagner attack** (targeted, white-box) [3]: Solves an unconstrained optimization problem to calculate a perturbation for a specified label.
- **Boundary attack** (untargeted, black-box) [4]: Performs a random walk close to the decision boundary while the norm of the perturbation is minimized.
- **DeepFool attack** (untargeted, white-box) [5]: Approximates the network output by a Taylor polynomial and computes perturbations exactly for this approximation.
- **Universal perturbation** (untargeted, white-box) [6]: Calculates and combines multiple adversarial examples using FGSM [7].

Architectures

For each dataset we adapt the model architecture.

Network	MNIST	Fashion-MNIST	SVHN	CIFAR10
ConvNet	99.39%	92.90%	92.57%	88.22%
CapsNet	99.40%	92.65%	92.35%	88.21%

Table 1: Test accuracies achieved by our networks.

The test accuracies of our models are not state-of-the-art. However, we found our models to be suitable for the given task, since the similar performances of both architectures ensure comparability.

Results

Our experiments show that the vulnerability of CapsNets and ConvNets is similar and it is hard to decide which model is more prone to adversarial attacks than the other:

Attack	Network	MNIST	Fashion	SVHN	CIFAR10
CW	ConvNet	1.40	0.51	0.67	0.37
	CapsNet	1.82	0.50	0.60	0.23
Boundary	ConvNet	3.07	1.24	2.42	1.38
	CapsNet	3.26	0.93	1.88	0.72
DeepFool	ConvNet	1.07	0.31	0.41	0.23
	CapsNet	2.02	0.55	0.80	0.16
Universal	ConvNet	6.71	2.61	2.46	2.45
	CapsNet	11.45	5.31	8.59	2.70

Table 2: Average Euclidean norm of the perturbations for each attack and architecture.

The Carlini-Wagner, the boundary and the DeepFool attack calculate adversarial examples that lead to misclassification, whereas the universal perturbations are considered adversarial if the test accuracy on the batch is less than 50%.

Transferability of Adversarial Examples

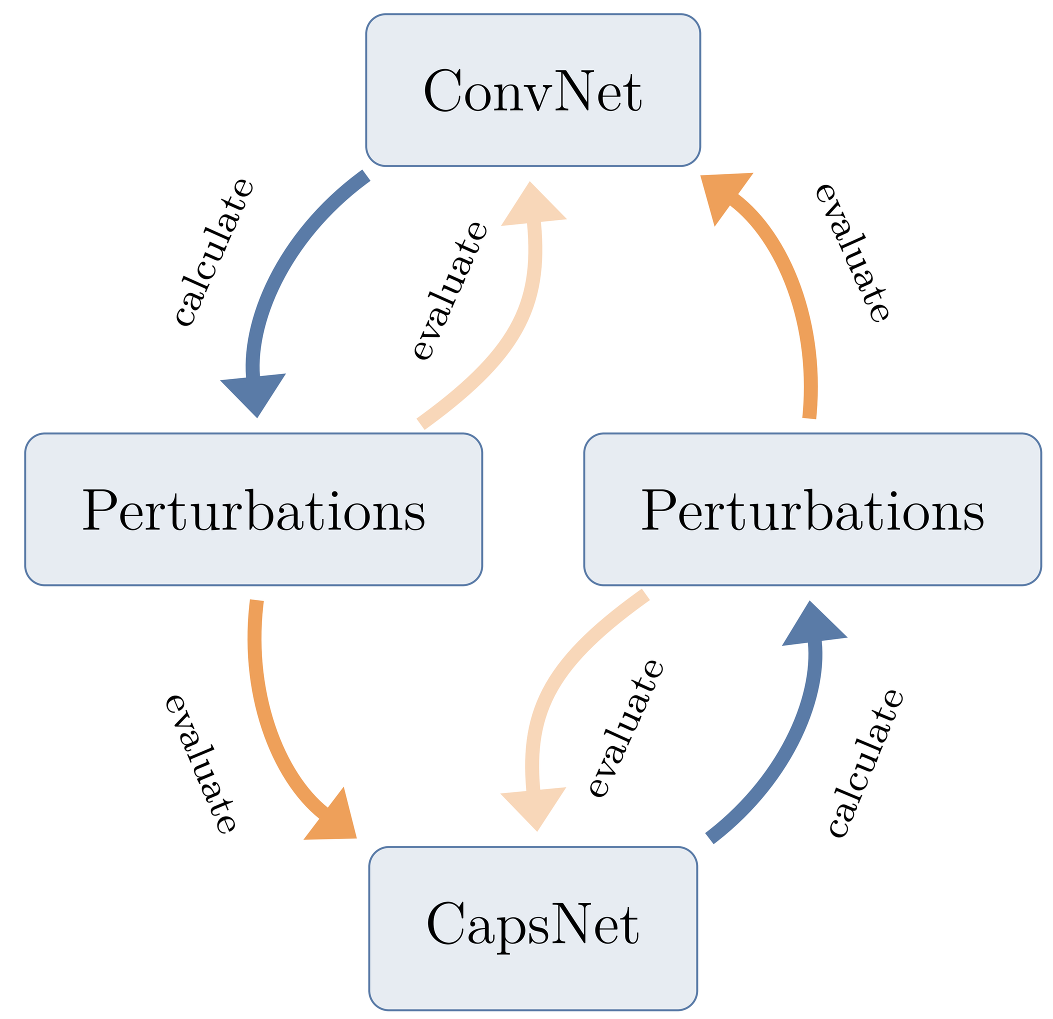

Figure 1: Our evaluation procedure. The light orange arrows show the usual application of adversarial perturbations.

To quantify the transferability we evaluate the perturbations on the other model. The fooling rates corresponding to the (dark) orange arrows are shown in Tab. 3. Especially the smaller universal perturbations calculated on the ConvNet generalize well to the CapsNet (see also Tab. 2).

Attack	Network	MNIST	Fashion	SVHN	CIFAR10
CW	ConvNet	0.8%	1.2%	2.8%	2.4%
	CapsNet	2.0%	2.0%	3.8%	2.0%
Boundary	ConvNet	8.8%	9.5%	10.5%	13.4%
	CapsNet	14.2%	14.6%	12.9%	26.1%
DeepFool	ConvNet	4.3%	8.5%	13.5%	11.8%
	CapsNet	0.9%	10.9%	10.8%	14.1%
Universal	ConvNet	4.9%	20.4%	35.0%	25.9%
	CapsNet	38.2%	25.7%	53.4%	47.2%

Table 3: Fooling rates of adversarial examples calculated for a CapsNet and evaluated on a ConvNet and vice versa.

Conclusion

Our experiments show that CapsNets are not in general more robust to white-box attacks. With sufficiently sophisticated attacks CapsNets can be fooled as easily as ConvNets. Moreover, we show that adversarial examples can be transferred between the two architectures. To fully understand the possibly distinguishable roles of the convolutional and capsule layers with respect to adversarial attacks, we are currently examining the effects of attacks on the activation level of single neurons. However, this analysis is not finished yet and beyond the scope of this work.

References

[1] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton.
Dynamic routing between capsules.
In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3856–3866. Curran Associates, Inc., 2017.

[2] Nicholas Frosst, Sara Sabour, and Geoffrey Hinton.
DARCCC: Detecting adversaries by reconstruction from class conditional capsules.
arXiv preprint arXiv:1811.06969, 2018.

[3] Nicholas Carlini and David Wagner.
Towards evaluating the robustness of neural networks.
In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017.

[4] Wieland Brendel, Jonas Rauber, and Matthias Bethge.
Decision-based adversarial attacks: Reliable attacks against black-box machine learning models.
arXiv preprint arXiv:1712.04248, 2017.

[5] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard.
Deepfool: A simple and accurate method to fool deep neural Networks.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016.

[6] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard.
Universal adversarial perturbations.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017.

[7] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy.
Explaining and harnessing adversarial examples.
arXiv preprint arXiv:1412.6572, 2014.

Contact Information

- Mail: `tobias.uelwer@hhu.de`
- Mail: `felix.michels@hhu.de`