A geometric perspective on the robustness of deep networks

Seyed-Mohsen Moosavi-Dezfooli

Amirkabir Artificial Intelligence Summer Summit
July 2019
Are we ready?
Adversarial perturbations

Universal adversarial perturbations, Moosavi et al., CVPR 2017.
“Geometry is not true, it is advantageous.”

Henri Poincaré
Adversarial perturbations
How large is the “space” of adversarial examples?

Universal perturbations
What causes the vulnerability of deep networks to universal perturbations?

Adversarial training
What geometric features contribute to a better robustness properties?
Geometry of adversarial perturbations
\[ r^* = \arg\min_r \|r\|_2 \quad \text{s.t.} \quad \hat{k}(x + r) \neq \hat{k}(x) \]

**Geometric interpretation of adversarial perturbations**

\[ x \in \mathbb{R}^d \]
Adversarial examples are “blind spots”.


Deep classifiers are “too linear”.

- Explaining and harnessing adversarial examples, Goodfellow et al., ICLR 2015.
Robustness of classifiers: from adversarial to random noise, Fawzi, Moosavi, Frossard, *NIPS 2016.*
Decision boundary of CNNs is almost flat along *random* directions.

- Robustness of classifiers: from adversarial to random noise, Fawzi, Moosavi, Frossard, *NIPS 2016.*
Adversarial perturbations constrained to a random subspace of dimension \( m \). 

\[
\begin{align*}
\mathbf{r}_S(x) &= \arg \min_{\mathbf{r} \in \mathcal{S}} \| \mathbf{r} \| \quad \text{s.t.} \quad \hat{k}(x + \mathbf{r}) \neq \hat{k}(x)
\end{align*}
\]

For low curvature classifiers, w.h.p., we have

\[
\mathbf{r}_S(x) = \mathbf{r}^*_{S(x)}
\]

Space of adversarial perturbations
Structured additive perturbations

Robustness of classifiers: from adversarial to random noise, Fawzi, Moosavi, Frossard, *NIPS 2016.*

The “space” of adversarial examples is quite vast.
Geometry of adversarial examples

Decision boundary is “locally” almost flat. Datapoints lie close to the decision boundary.

Flatness can be used to construct diverse set of perturbations. design efficient attacks.
Geometry of universal perturbations
Universal adversarial perturbations (UAP)

Universal adversarial perturbations, Moosavi et al., CVPR 2017.
Diversity of UAPs

VGG-19  VGG-16  VGG-F  CaffeNet
ResNet-152  GoogLeNet

Diversity of perturbations
Why do universal perturbations exist?
Robustness of classifiers to universal perturbations, Moosavi et al., ICLR 2018.
Normals to the decision boundary are “globally” correlated.

Robustness of classifiers to universal perturbations, Moosavi et al., *ICLR 2018.*
The flat model only partially explains the universality.

- Robustness of classifiers to universal perturbations, Moosavi et al., ICLR 2018.
The principal curvatures of the decision boundary:

- Robustness of classifiers to universal perturbations, Moosavi et al., *ICLR 2018*. 

![Graph showing the principal curvatures of the decision boundary.](image-url)
The principal curvatures of the decision boundary:

- Robustness of classifiers to universal perturbations, Moosavi et al., *ICLR 2018.*
The principal curvatures of the decision boundary:

- Robustness of classifiers to universal perturbations, Moosavi et al., *ICLR 2018.*
The principal curvatures of the decision boundary:

- Robustness of classifiers to universal perturbations, Moosavi et al., *ICLR 2018.*
Normal sections of the decision boundary (for different datapoints) along a single direction:

- **UAP direction**
- **Random direction**

Robustness of classifiers to universal perturbations, Moosavi et al., *ICLR 2018.*
The curved model better explains the existence of universal perturbations.

Robustness of classifiers to universal perturbations, Moosavi et al., *ICLR 2018.*
Universality of perturbations
Shared curved directions explain this vulnerability.

A possible solution
Regularizing the geometry to combat against universal perturbations.

Why are deep nets curved?
- With friends like these, who needs adversaries?, Jetley et al., NeurIPS 2018.
Geometry of adversarial training
In a nutshell

Adversarial perturbations

Image batch

$x$

$\mathbf{x} + \mathbf{r}$

Training

Adversarial training

Curvature regularization
One of the most effective methods to improve adversarial robustness...

- Obfuscated gradients give a false sense of security, Athalye et al., *ICML 2018*. (Best paper)
Curvature profiles of normally and adversarially trained networks:

Robustness via curvature regularisation, and vice versa, Moosavi et al., CVPR 2019.
Curvature Regularization (CURE)

Robustness via curvature regularisation, and vice versa, Moosavi et al., CVPR 2019.
<table>
<thead>
<tr>
<th>AT</th>
<th>CURE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Implicit regularization</td>
<td>Explicit regularization</td>
</tr>
<tr>
<td>Time consuming</td>
<td>3x to 5x faster</td>
</tr>
<tr>
<td>SOTA robustness</td>
<td>On par with SOTA</td>
</tr>
</tbody>
</table>

- Robustness via curvature regularisation, and vice versa, Moosavi et al., *CVPR 2019*. 
Inherently more robust classifiers
Curvature regularization can significantly improve the robustness properties.

Counter-intuitive observation
Due to a more linear nature, an adversarially trained net is “easier” to fool.

A better trade-off?
- Adversarial Robustness through Local Linearization, Qin et al., arXiv.
Future challenges
Architectures
Batch-norm, dropout, depth, width, etc.

Data
# of modes, convexity, distinguishability, etc.

Training
Batch size, solver, learning rate, etc.

- Spatially transformed adversarial examples, Xiao et al., *ICLR 2018*.
Robustness may be at odds with accuracy, Tsipras et al., NeurIPS 2018.
ETHZ
Zürich, Switzerland

Google
Zürich
Interested in my research?

smoosavi.me

moosavi.sm@gmail.com