A Spectral View of Adversarially Robust Features

Shivam Garg

Vatsal Sharan*
Brian Zhang*
Gregory Valiant

Stanford University
What are adversarial examples?

Adding small amount of well-crafted noise to the test data fools the classifier.
More Questions than Answers

Intense ongoing research efforts, but we still don’t have a good understanding of many basic questions:

• What are the tradeoffs between the amount of data available, accuracy of the trained model, and vulnerability to adversarial examples?

• What properties of the geometry of a dataset make models trained on it vulnerable to adversarial attacks?
More Questions than Answers

Intense ongoing research efforts, but we still don’t have a good understanding of many basic questions:

• What are the tradeoffs between the amount of data available, accuracy of the trained model, and vulnerability to adversarial examples?

• What properties of the geometry of a dataset make models trained on it vulnerable to adversarial attacks?
Simpler Objective: Adversarially Robust Features

- **Robust Classifier**: A function from $\mathbb{R}^d \rightarrow \mathbb{R}$, that doesn’t change much with small perturbations to data, and agrees with true labels.
Simpler Objective: Adversarially Robust Features

- **Robust Classifier**: A function from $\mathbb{R}^d \to \mathbb{R}$, that doesn’t change much with small perturbations to data, and agrees with true labels.

- **Robust Feature**: A function from $\mathbb{R}^d \to \mathbb{R}$, that doesn’t change much with small perturbations to data, and agrees with true labels.

- The function is required to have sufficient variance across data points to preclude the trivial constant function.
Simpler Objective: Adversarially Robust Features

• **Robust Classifier**: A function from \( \mathbb{R}^d \rightarrow \mathbb{R} \), that doesn’t change much with small perturbations to data, and agrees with true labels.

• **Robust Feature**: A function from \( \mathbb{R}^d \rightarrow \mathbb{R} \), that doesn’t change much with small perturbations to data, and agrees with true labels.

• The function is required to have sufficient variance across data points to preclude the trivial constant function.

• Disentangles the challenges of robustness and classification performance

• Train a classifier on top of robust features
Connections to Spectral Graph Theory

- Second eigenvector $\mathbf{v}$ of the Laplacian of a graph is the solution to:

$$\min_{\mathbf{v}} \sum_{(i,j) \in E} (v_i - v_j)^2 \quad \text{s.t.} \quad \sum_i v_i = 0; \quad \sum_i v_i^2 = 1$$

- Assigns values to vertices that change smoothly across neighbors

- Constraints ensure sufficient variance among these values
Connections to Spectral Graph Theory

• Think of input data points as graph vertices with edges denoting some measure of similarity

• Can obtain robust features from the eigenvectors of Laplacian
Connections to Spectral Graph Theory

- Think of input data points as graph vertices with edges denoting some measure of similarity.

- Can obtain robust features from the eigenvectors of Laplacian.

- **Upper bound**: Characterizes the robustness of features in terms of eigenvalues and spectral gap of the Laplacian.

- **Lower bound**: Roughly says that if there exists a robust feature, the spectral approach would find it under certain conditions on the properties of Laplacian.
Create similarity graph according to a given distance metric
[the same metric that we hope to be robust wrt]
Illustration: Extract Feature from 2nd eigenvector

\[ f(x_i) = v_2(x_i) \]
Takeaways

• Disentangling the two goals of robustness and classification performance may help us understand the extent to which a given dataset is vulnerable to adversarial attacks, and ultimately might help us develop better robust classifiers.

• Interesting connections between spectral graph theory and adversarially robust features.
Takeaways

• Disentangling the two goals of robustness and classification performance may help us understand the extent to which a given dataset is vulnerable to adversarial attacks, and ultimately might help us develop better robust classifiers.

• Interesting connections between spectral graph theory and adversarially robust features.

Thank you!