Deep k-Nearest Neighbors

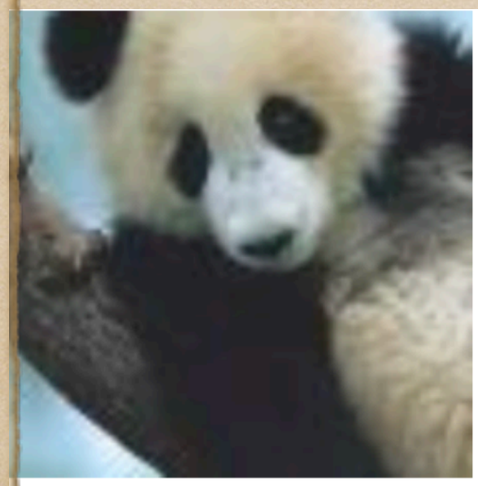
Buhua Liu

DkNN

- Introduction
- Problem Definition
- Methodology
- Experiments
- Conclusion

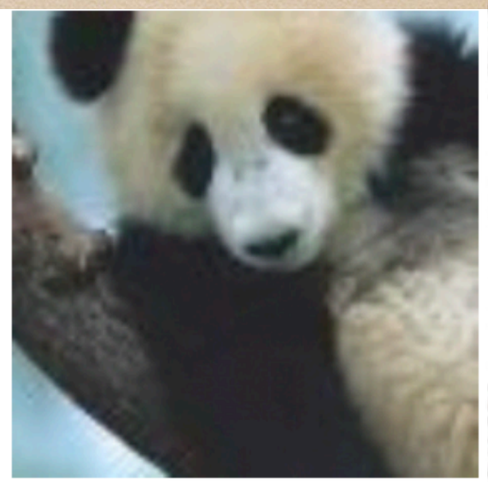
DkNN

- Introduction
- Problem Definition
- Methodology
- Experiments
- Conclusion



$+ .007 \times$

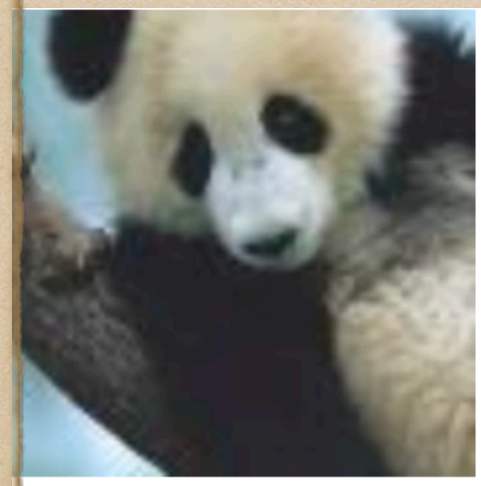
$=$



$\boldsymbol{x}$	$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$	$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
“panda”	“nematode”	“gibbon”
57.7% confidence	8.2% confidence	99.3 % confidence

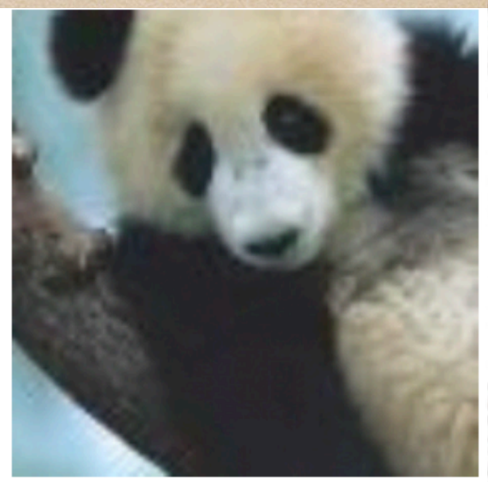
Adversarial Examples

http://arxiv.org/abs/1412.6572



$+ .007 \times$

$=$



$\boldsymbol{x}$

“panda”
57.7% confidence

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

“nematode”
8.2% confidence

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

“gibbon”
99.3 % confidence

Adversarial Examples

http://arxiv.org/abs/1412.6572

DkNN

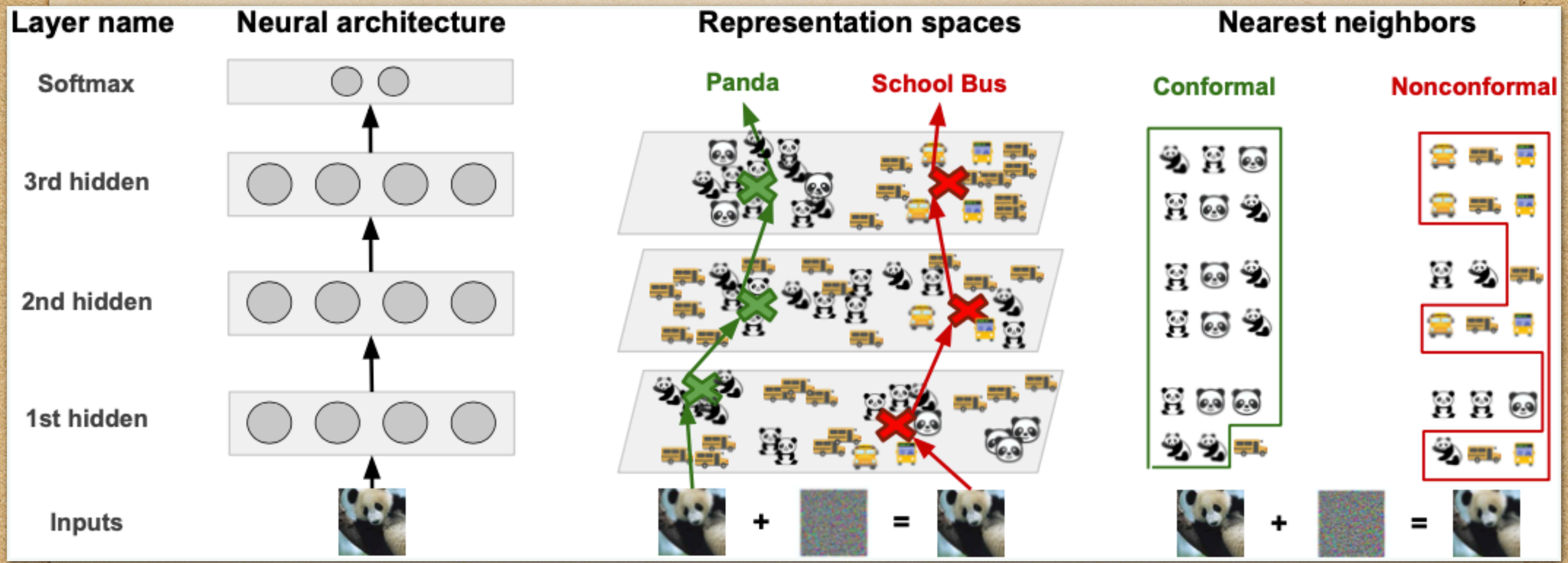
- Introduction
- Problem Definition
- Methodology
- Experiments
- Conclusion

What is wrong?

- Misclassification—Lack of robustness
- Unreliable confidence estimates
- Lack of interpretability—Over-parameterized DNN

DkNN

- Introduction
- Problem Definition
- Methodology
- Experiments
- Conclusion



Intuition behind the DkNN

http://arxiv.org/abs/1803.04765

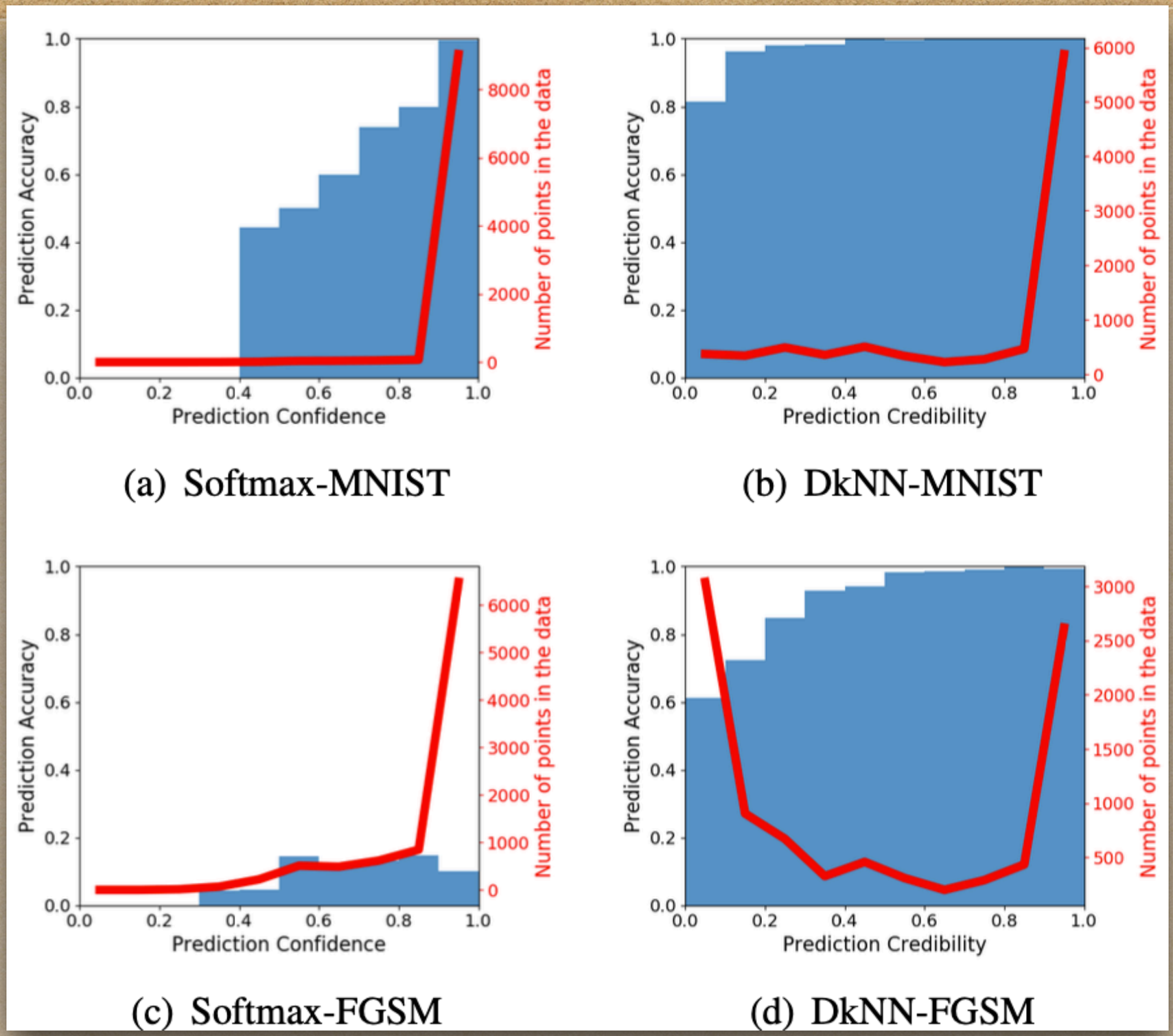
Algorithm 1 – Deep k-Nearest Neighbor.

Input: training data (X, Y), calibration data (X^c, Y^c)
Input: trained neural network f with l layers
Input: number k of neighbors
Input: test input z

// Compute layer-wise k nearest neighbors for test input z
for each layer $\lambda \in 1..l$ **do**
  $\Gamma \leftarrow k$ points in X closest to z found w/ LSH tables
  $\Omega_\lambda \leftarrow \{Y_i : i \in \Gamma\}$ ▷Labels of k inputs found
end for
// Compute prediction, confidence and credibility
$A = \{\alpha(x, y) : (x, y) \in (X^c, Y^c)\}$ ▷Calibration
for each label $j \in 1..n$ **do**
  $\alpha(z, j) \leftarrow \Sigma_{\lambda \in 1..l} |i \in \Omega_\lambda : i \neq j|$ ▷Nonconformity
  $p_j(z) = \frac{|\{\alpha \in A : \alpha \geq \alpha(z,j)\}|}{|A|}$ ▷empirical p-value
end for
prediction $\leftarrow \arg\max_{j \in 1..n} p_j(z)$
confidence $\leftarrow 1 - \max_{j \in 1..n, j \neq prediction} p_j(z)$
credibility $\leftarrow \max_{j \in 1..n} p_j(z)$
return prediction, confidence, credibility

DkNN

- Introduction
- Problem Definition
- Methodology
- Experiments
- Conclusion



Reliability Diagrams

DkNN

- Introduction
- Problem Definition
- Methodology
- Experiments
- Conclusion

Conclusion

- Validate the effectiveness of DkNN
- Hands-on experience on TF2.0

Thank you!