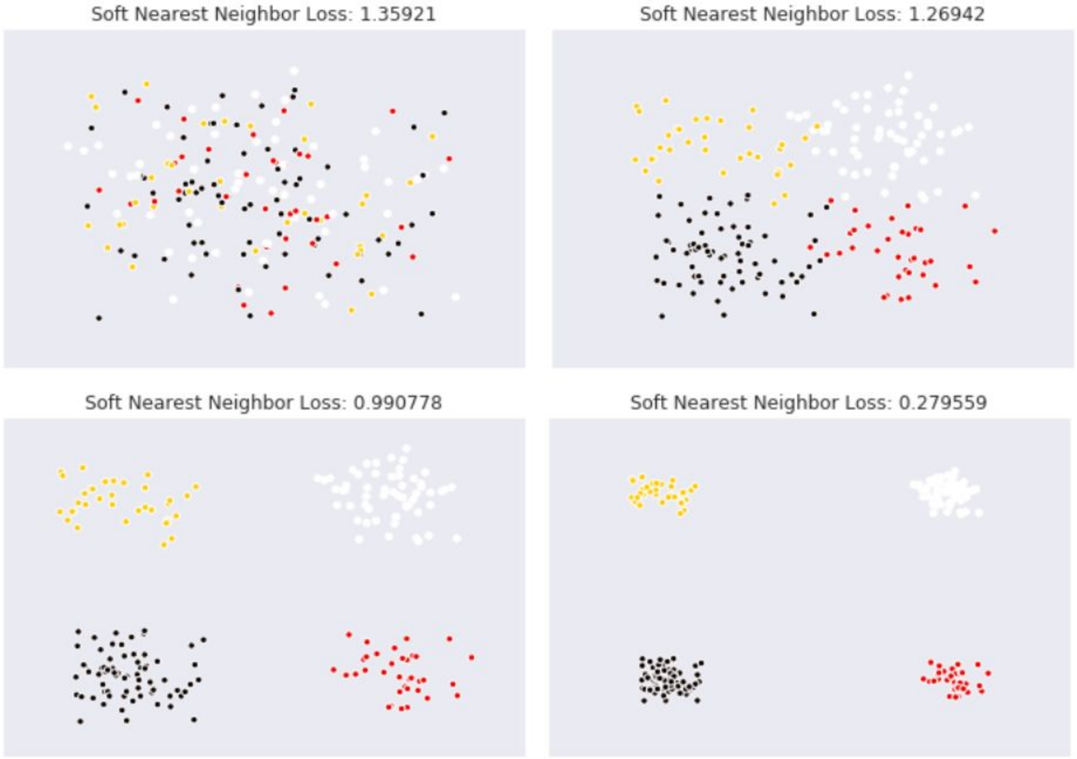
Analyzing and Improving Representations with the Soft Nearest Neighbor Loss

Nicholas Frosst, Nicolas Papernot, Geoffrey Hinton
{frosst,papernot,geoffhinton}@google.com



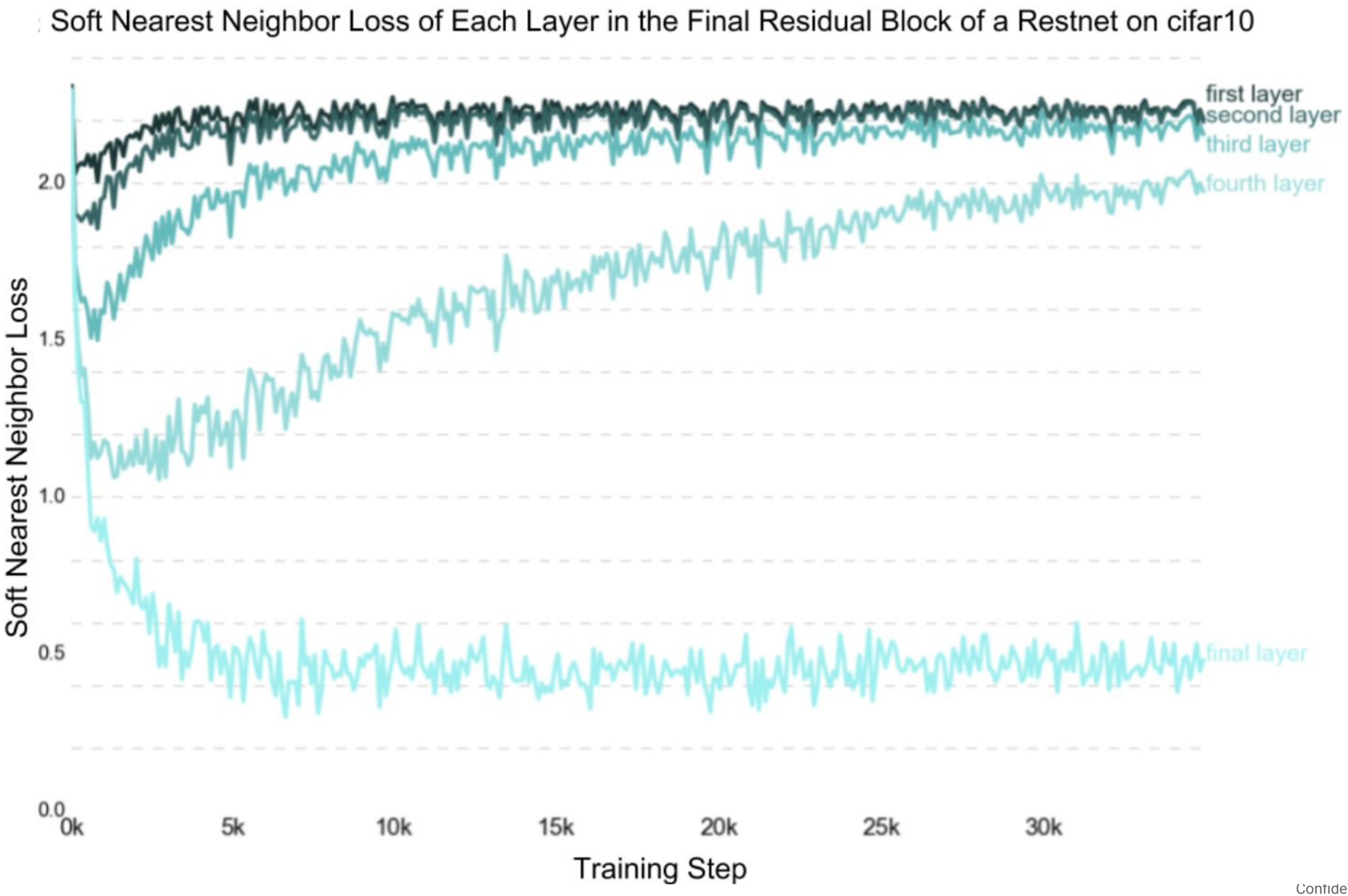
Soft Nearest Neighbor Loss: 1.35921

Soft Nearest Neighbor Loss: 1.26942

Soft Nearest Neighbor Loss: 0.990778

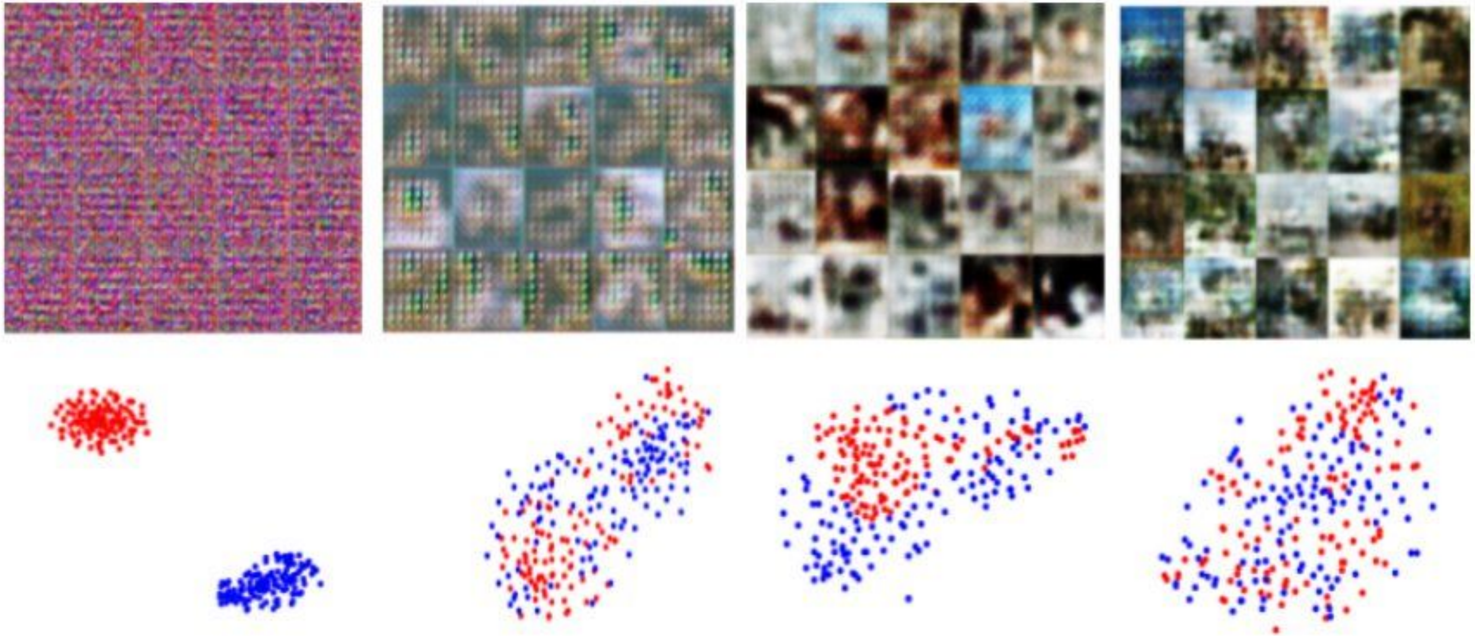
Soft Nearest Neighbor Loss: 0.279559

$$l_{sn}(x, y, T) = -\frac{1}{b} \sum_{i \in 1..b} \log \left(\frac{\displaystyle\sum_{\substack{j \in 1..b \\ j \neq i \\ y_i = y_j}} e^{-\frac{||x_i - x_j||^2}{T}}}{\displaystyle\sum_{\substack{k \in 1..b \\ k \neq i}} e^{-\frac{||x_i - x_k||^2}{T}}} \right)$$

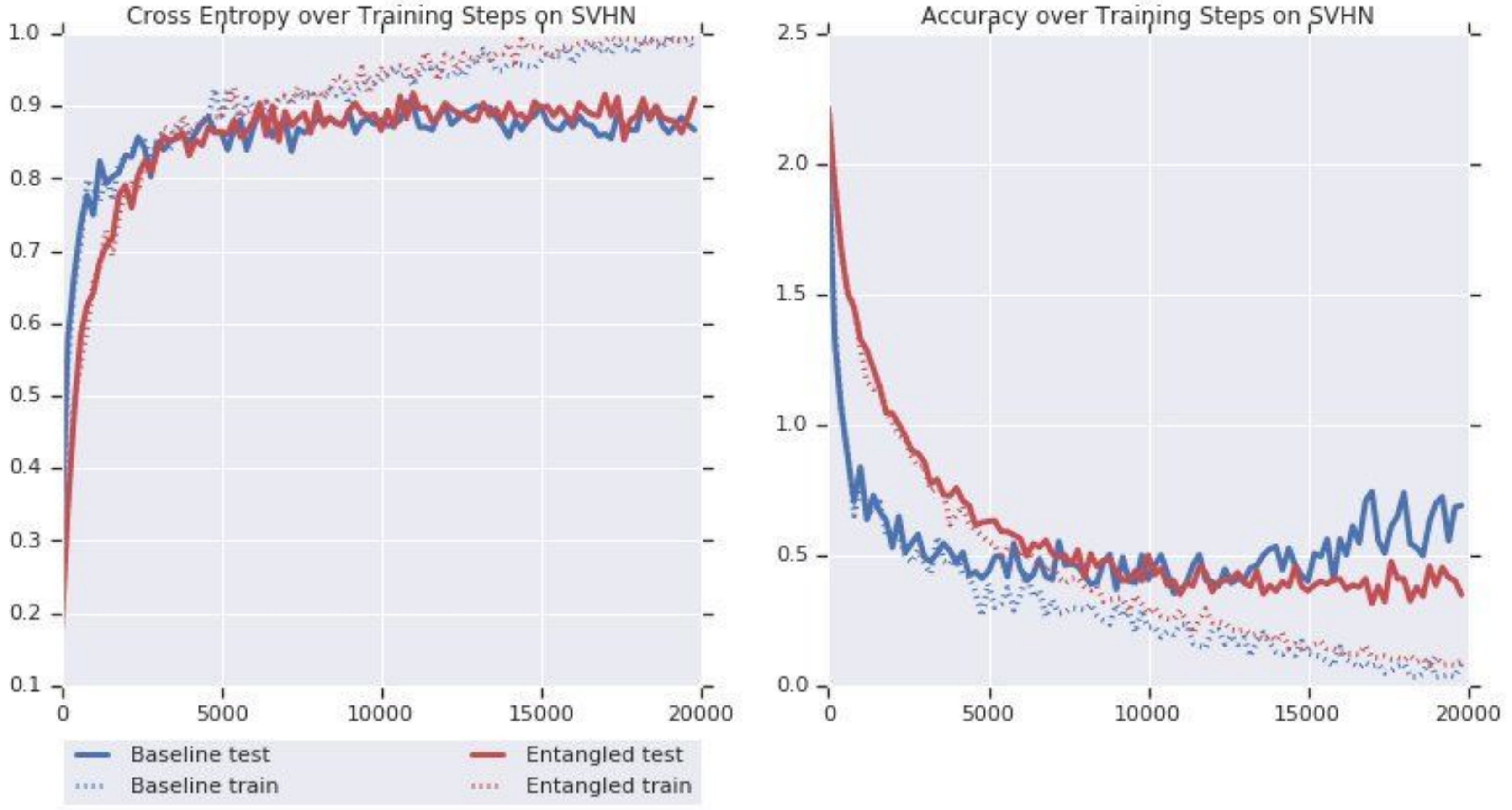
: Soft Nearest Neighbor Loss of Each Layer in the Final Residual Block of a Restnet on cifar10
Soft Nearest Neighbor Loss
2.0
1.5
1.0
0.5
0.0
0k
5k
10k
15k
20k
25k
30k
Training Step
first layer
second layer
third layer
fourth layer
final layer

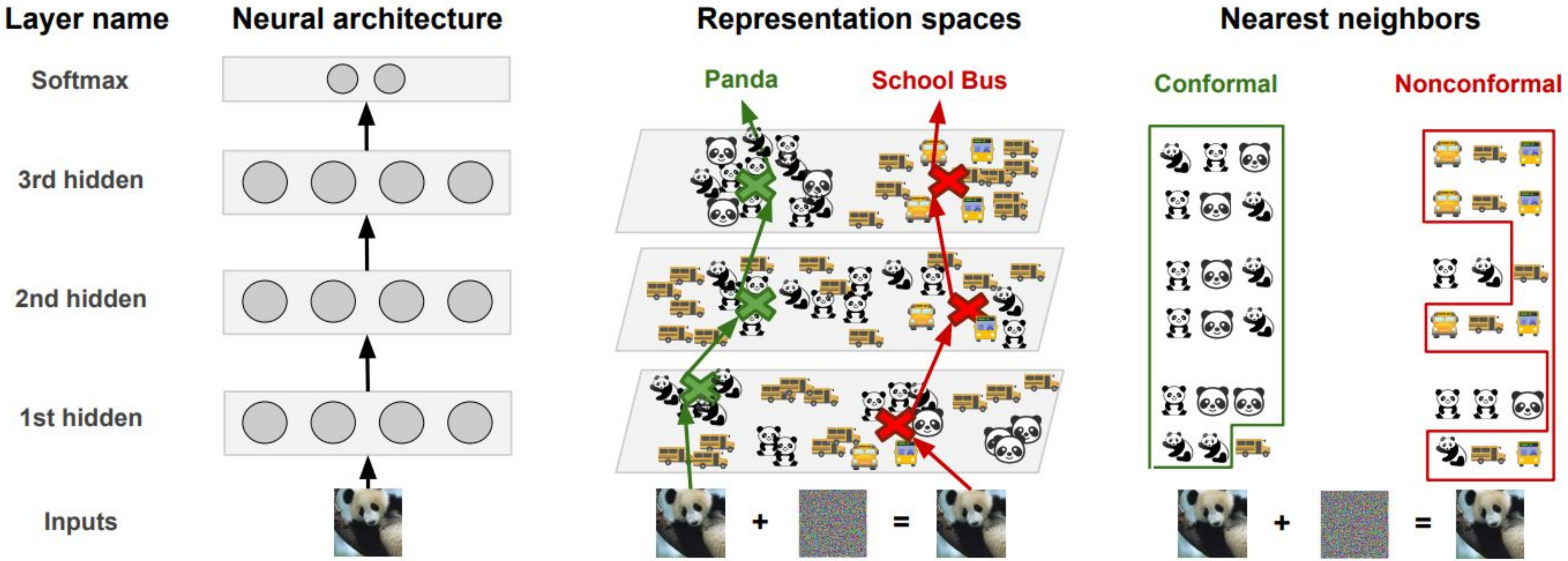
Generated Data

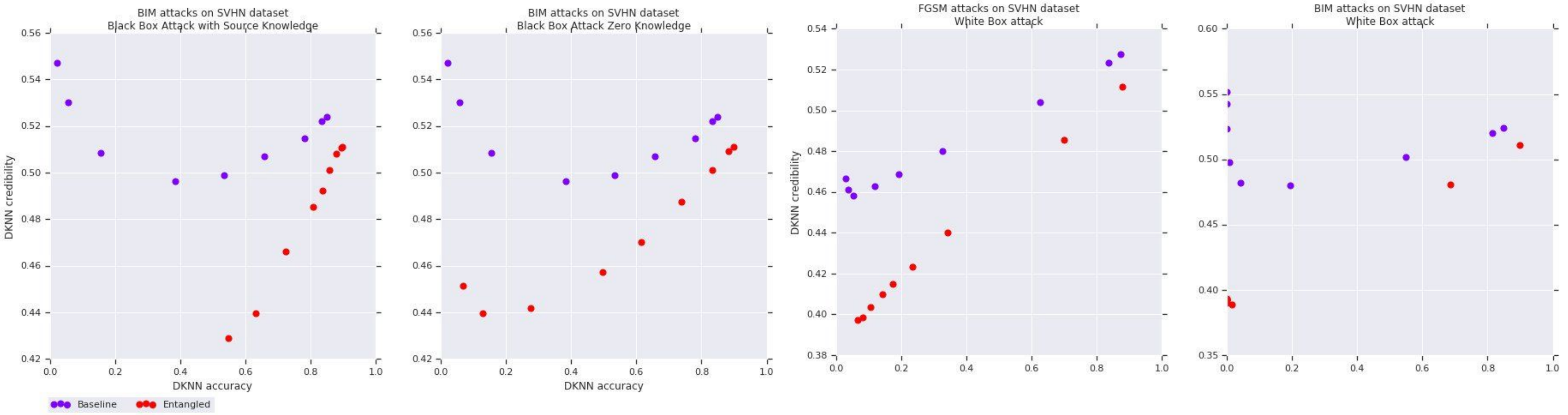
t-SNE of real (blue) and generated (red) data



Training Step	0	400	800	2000
Soft Nearest Neighbor Loss	0.5617	0.6196	0.6208	0.6487


Cross Entropy over Training Steps on SVHN
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
5000
10000
15000
20000
Accuracy over Training Steps on SVHN
2.5
2.0
1.5
1.0
0.5
0.0
0
5000
10000
15000
20000
Baseline test
Baseline train
Entangled test
Entangled train


Layer name
Neural architecture
Representation spaces
Nearest neighbors
Softmax
3rd hidden
2nd hidden
1st hidden
Inputs
Panda
School Bus
Conformal
Nonconformal
+
=
+
=



BIM attacks on SVHN dataset
Black Box Attack with Source Knowledge

BIM attacks on SVHN dataset
Black Box Attack Zero Knowledge

FGSM attacks on SVHN dataset
White Box attack

BIM attacks on SVHN dataset
White Box attack

Thank you!

Come see the full poster Pacific Ballroom #18

Analyzing and Improving Representations with the Soft Nearest Neighbor Loss

Nicholas Frosst, Nicolas Papernot, Geoffrey Hinton
{frosst,papernot,geoffhinton}@google.com