The disadvantage of CNN versus DBN image classification under adversarial conditions

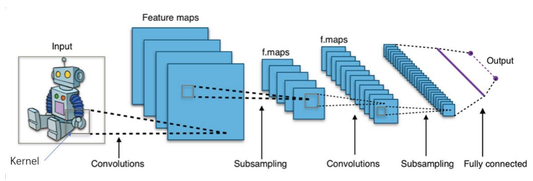
Tao (Andy) Yang†, Daniel L. Silver †,*
† Acadia University

Summary

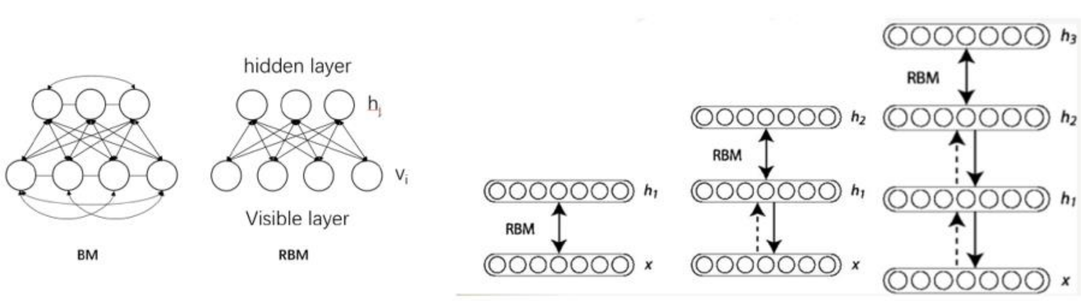
- We compare Convolutional Neural Networks (CNN) and Deep Belief Network's (DBN) ability to withstand common image classification attacks.
- The results show that the DBN models generally perform better under attack than the CNN models.

Introduction

- Focus is on images created to attack the performance of CNN and DBN classification models.We show that the CNN inductive bias, which assumes that important features can be extracted from adjacent pixels, fails under certain adversarial attacks.
- In contrast, DBNs, which are pre-trained using unsupervised examples, learn features with no assumption about the proximity of pixels, and do better under the same attacks.

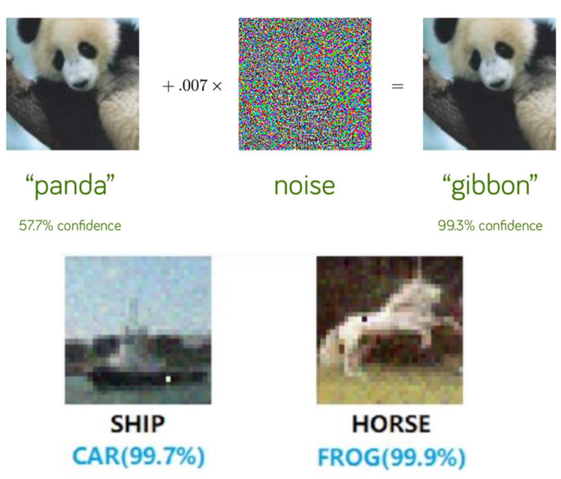


Source: https://en.wikipedia.org/wiki/Convolutional_neural_network



Background

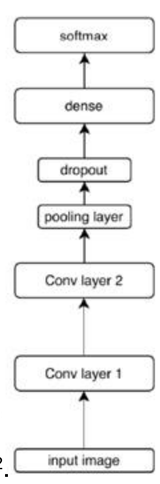
- It is possible to trick trained neural networks by carefully modifying the pixels of the images.
- This is called an adversarial attack, and if done well, it will make the network think an image of a panda is an image of a gibbon.
- In 2018, it was shown that by only changing one pixel, a trained CNN network will incorrectly think a horse is a frog[1].

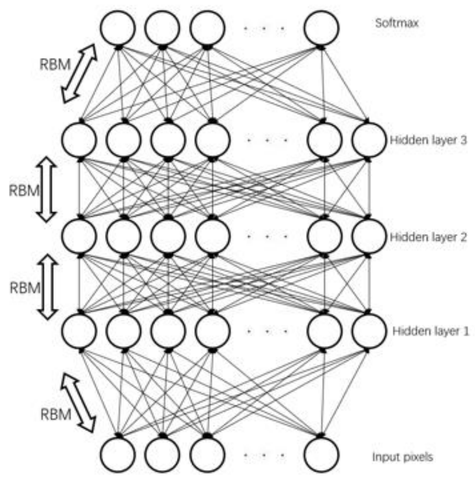


- Most images in the CIFAR-10 dataset can be modified similarly to fool CNN models

Theory and approach

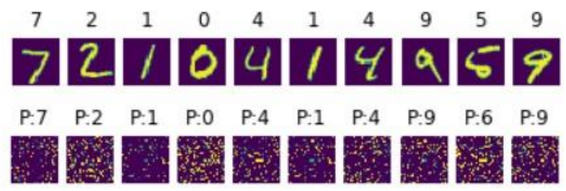
- Our theory is that CNN makes a strong inductive bias assumption about the relationship between pixels that are proximal to each other, and this will be a disadvantage for CNN when people intentionally modify the pixels.
- We wish to show that the CNN inductive bias, according to proximity of pixels, which assumes pixels adjacent to each other will tend to have the same colour and brightness, fails under certain adversarial attacks.
- DBNs, which learn features with a less specific inductive bias with no assumptions about the proximity of pixels, will not fail in the same way[2].





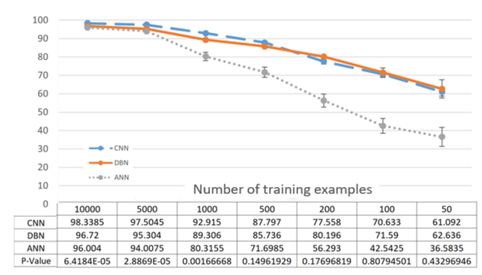
Empirical studies

- This section aims to demonstrate empirically that a semi-supervised approach using DBNs is superior to the strictly supervised CNN approach when under attack.
- The first experiment uses the MNIST data set to show the weakness of the CNN algorithm, as compared to the DBN algorithm, when the kernel assumption no longer holds for the input data.



Examples of images used in the first experiment: top row shows original images, bottom row shows mixed-pixel images.

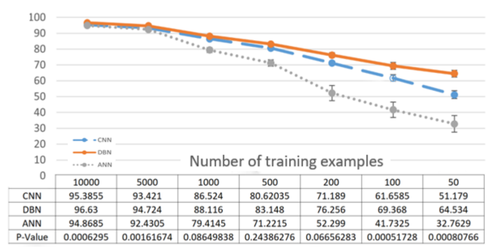
- CNN has an advantage over DBN when the training samples are high, but then loses the advantage as the numbers of labelled training examples decreases.
- This is expected since CNN relies on a large number of human-labeled training samples to learn an accurate model.



	10000	5000	1000	500	200	100	50
CNN	98.3385	97.5045	92.915	87.797	77.558	70.633	61.092
DBN	96.72	95.304	89.306	85.736	80.196	71.59	62.636
ANN	96.004	94.0075	80.3155	71.6985	56.293	42.5425	36.5835
P-Value	6.4184E-05	2.8869E-05	0.00166668	0.14961929	0.17696819	0.80794501	0.43296946

CNN and DBN accuracy on original test images by number of training examples.

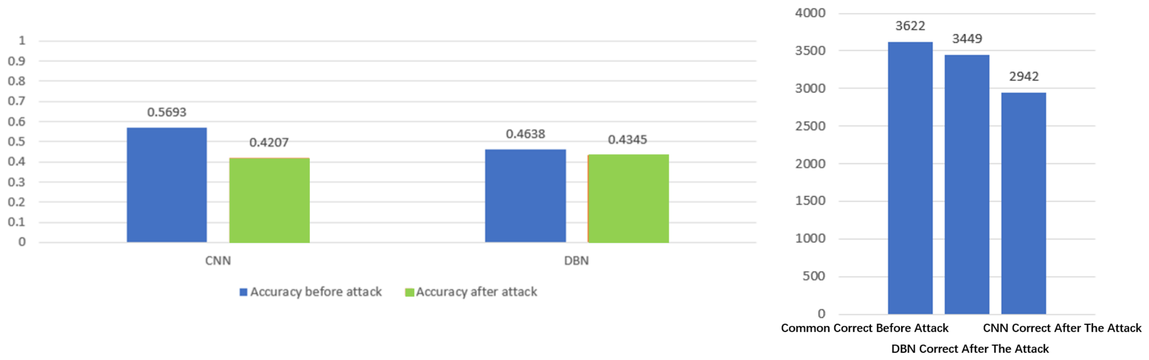
- Mixing up the pixels of the images adversely affects the performance of the CNN models
- But DBN models are not affected – their level of accuracy is as before the attack



	10000	5000	1000	500	200	100	50
CNN	95.3855	93.421	86.524	80.62035	71.189	61.6585	51.179
DBN	96.63	94.724	88.116	83.148	76.256	69.368	64.534
ANN	94.8685	92.4305	79.4145	71.2215	52.299	41.7325	32.7629
P-Value	0.0006295	0.00161674	0.08649838	0.24386276	0.06656283	0.00051728	0.00080766

CNN and DBN accuracy on mixed-pixel test images by number of training examples.

- The second experiment tests the performance of CNN and DBN models against noisy one-pixel attack images and shows the significant loss in model accuracy that CNN suffers versus DBN.



Left: Performance on test examples before and after the one-pixel attack.
Right: Performance on test images before and after the one-pixel attack.

Conclusion

- CNNs inductive bias becomes a disadvantage when the assumption of the relationship between proximal pixels no longer holds.
- In comparison, the DBN algorithm assumes that the best internal representation can be developed by pre-training the network using large sets of unlabelled examples from the same input space without other a priori assumptions.
- This suggests that in the context of lifelong machine learning (or continual learning) systems that the DBN approach of learning the underlying internal representation is superior to the highly engineered CNN approach[3].

References

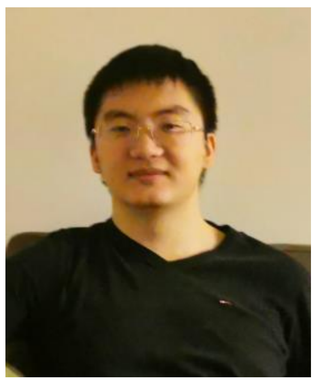
- J. Su, D. V. Vargas, and K. Sakurai. "One Pixel Attack for Fooling Deep Neural Networks". In: IEEE Trans. on Evol. Comp. 23.5 (2019), pp. 828–841. doi: 10.1109/tevc.2019.2890858.
- G. E. Hinton. "A Practical Guide to Training Restricted Boltzmann Machines". In: Neural Networks: Tricks of the Trade, 2nd Edition, LNCS 7700 (2012), pp. 599–619.
- A. Anand. Contrastive Self-Supervised Learning. https://ankeshanand.com/blog/2020/01/26/contrative-self-supervised-learning.html. 2020.

Acknowledgements

First and foremost, I would like to show my deepest gratitude to my supervisor, Dr. Daniel Silver. Without his help, the goal of this work would not have been realized.
I shall extend my thanks to all professors in the computer science department at Acadia University for their help. I would also like to give a special thank to Dr. Jim Diamond, Dr. Greg Lee, Dr. Zhang Haiyi and Professor Duane Currie, as they gave me accurate and efficient help on many deep questions. They also provide extra convenience and help which is much more than I should get.
Last but not least, I would like to thank all my friends and family for their financial and moral support.

Canadian AI 2021
(Vancouver, May 25-28, 2021)

ACADIA UNIVERSITY



Tao Yang
137660y@acadiau.ca