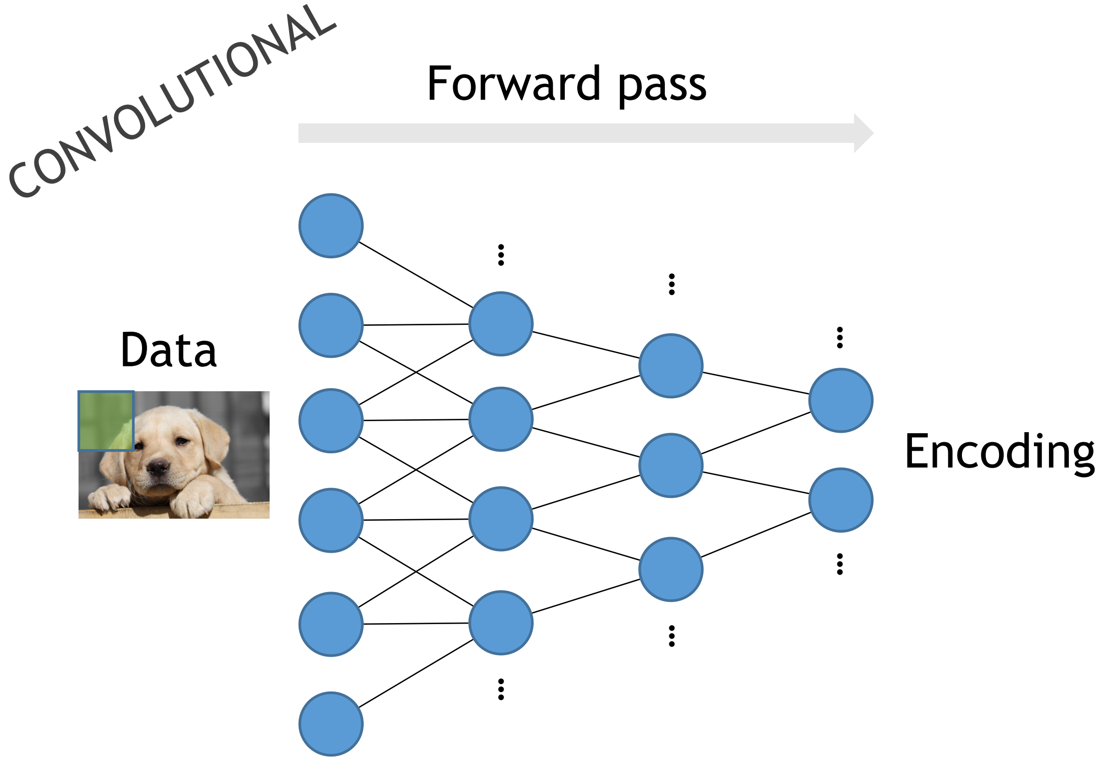
Information Highways and Navigating the Network

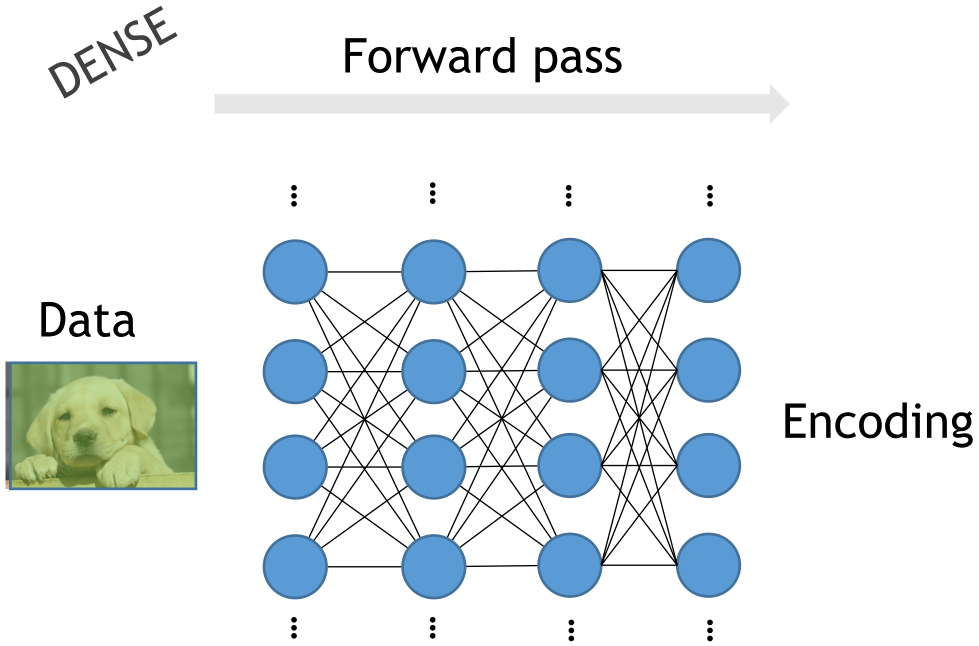
Recovery of activation pathways under the DCSC framework

Overview

- Review of CNNs and their successes
- Robustness, adversarial examples and barriers to further adoption
- Importance and role of activation pathways
- Deep Convolutional Sparse Coding framework (DCSC)
- A probabilistic bound on the recovery of activation pathways by the forward pass algorithm
- Limitations of result and future outlook – from activation pathways to information highways

A Convolutional Neural Network (CNN) is a Neural Network with an enforced local connectivity pattern between layers



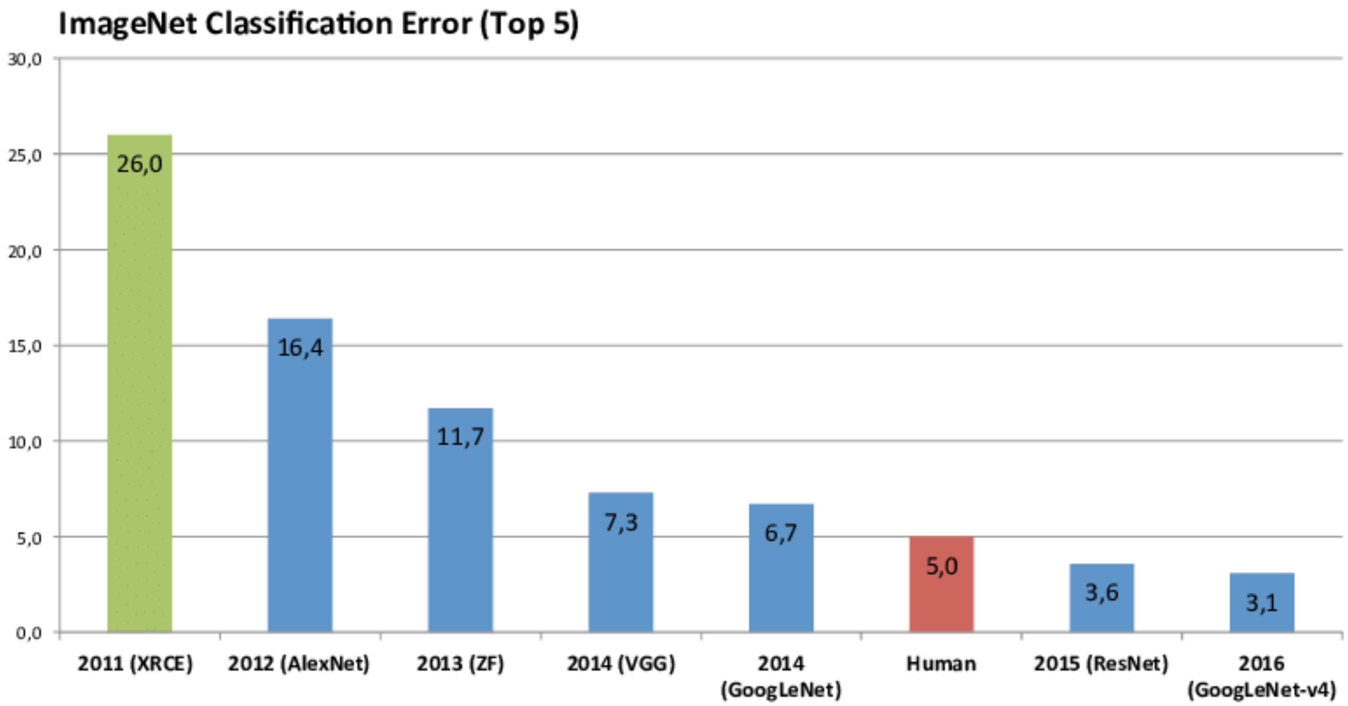


- Can identify local features and provides degree of translation invariance
- Sparsely connected plus typically use weight sharing, therefore far fewer parameters

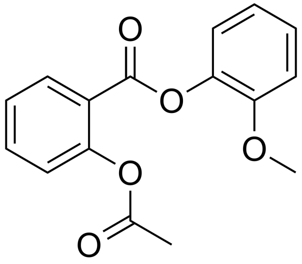
- Identifies global features
- Densely connected with many parameters, therefore harder to train and greater risk of overfitting

Since the arrival of AlexNet CNNs have been the state of the art technique for many tasks in computer vision

<u>Performance:</u>



<u>Wider potential applications:</u>



Drug discovery

Self driving cars

Robotics

Natural language processing

....

Uncertainty around their robustness however poses a barrier to the adoption of CNNs in higher risk applications

<u>Adversarial Examples:</u>

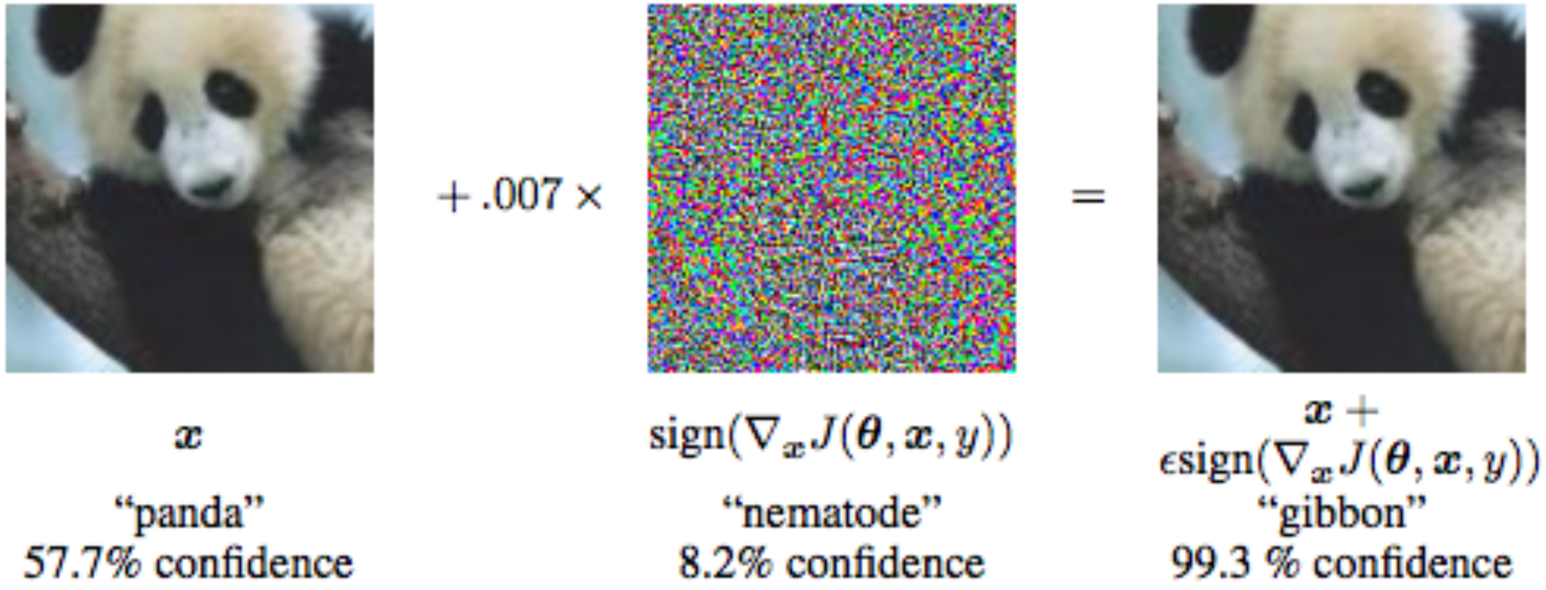


Image from Explaining and Harnessing Adversarial Examples, Ian Goodfellow

<u>Challenges</u>

- Its very difficult to perform formal verification on CNN systems to test their 'correctness'
- We struggle to characterize their success rate theoretically
- We do not fully understand their failure modes

To understand the failure modes of neural networks we need a better understanding of how they work

Approximation
Which functions can the forward pass algorithm approximate? How does this impact network topology and architecture?

Optimization
How do we train, i.e., configure the parameters, of a CNN so that the forward pass approximates a given function?

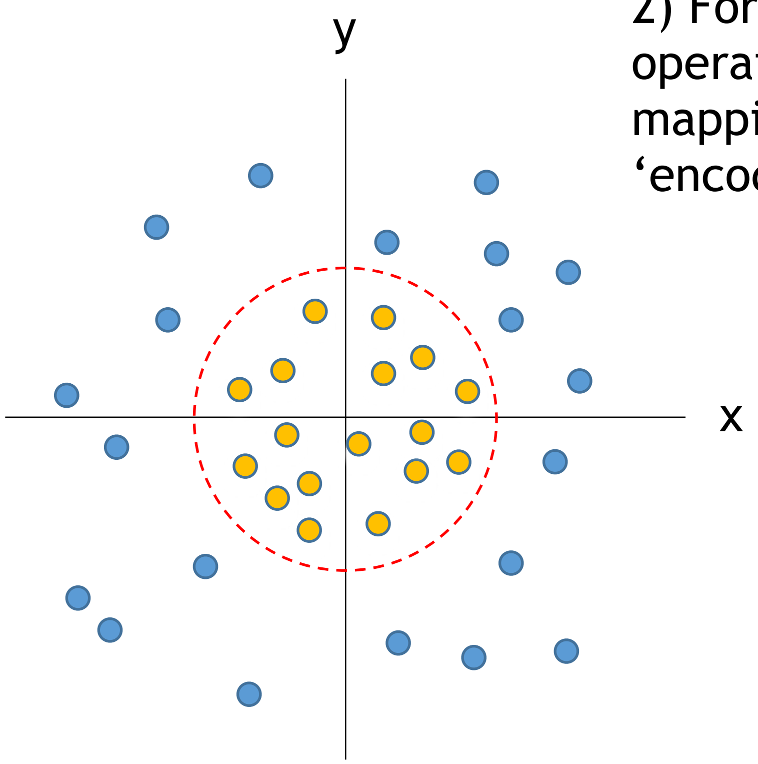
Generalization
How can we be sure that the forward pass algorithm will process new data points to the same accuracy as those in the training set?

Stability
How can we guarantee or determine the robustness of the forward pass to perturbations?

Goal: understand the forward pass algorithm and the way in which ‘knowledge’ is managed in a CNN

The role of the forward pass algorithm is to map the input data into a new space which is well suited for a given task

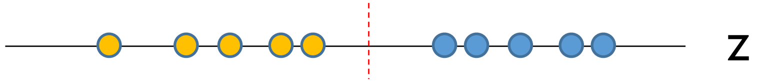
Toy example:



1) The input data points consist of pairs of numbers and belong to one of two classes

2) Forward pass performs operation on input data, mapping it into a new ‘encoded’ space

$z = (x^2 + y^2)^{1/2}$



The radius is a feature we can use to determine the class of a data point

3) Data points can easily be classified by a linear classifier in the encoded data space

During training CNNs learn a hierarchy of features which it uses, for example, to classify the training data

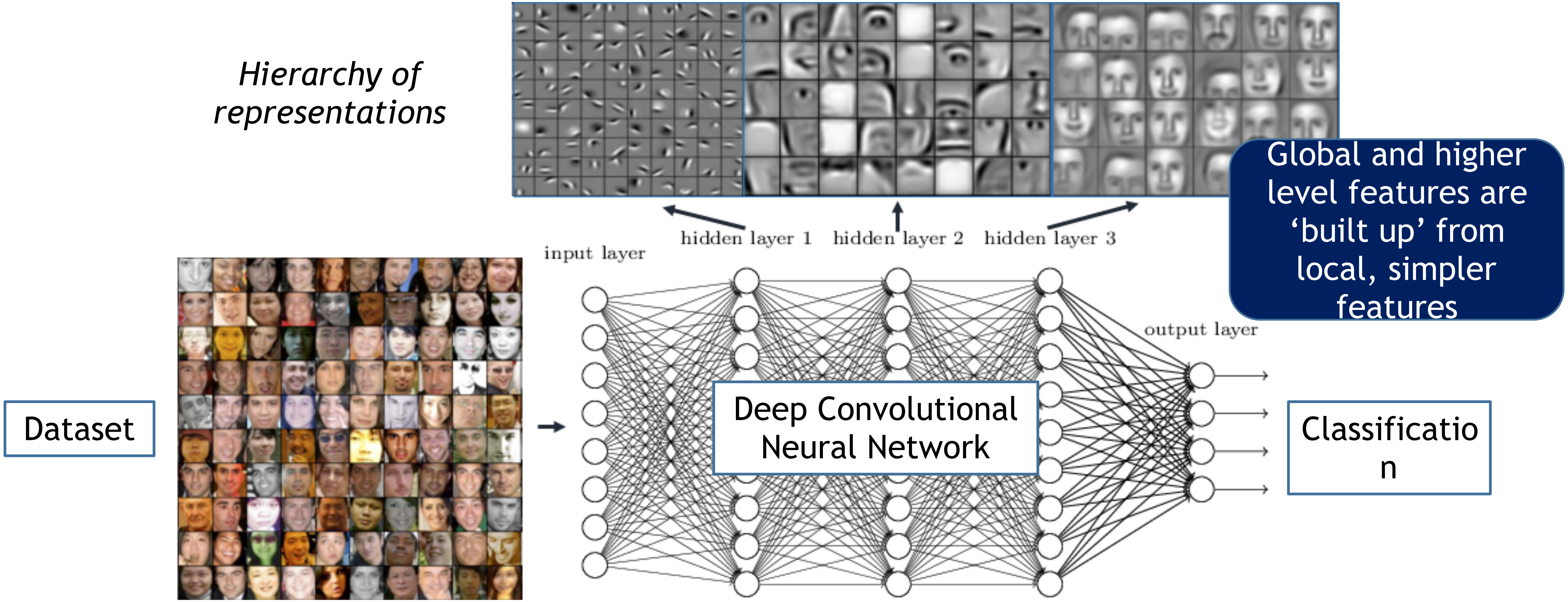
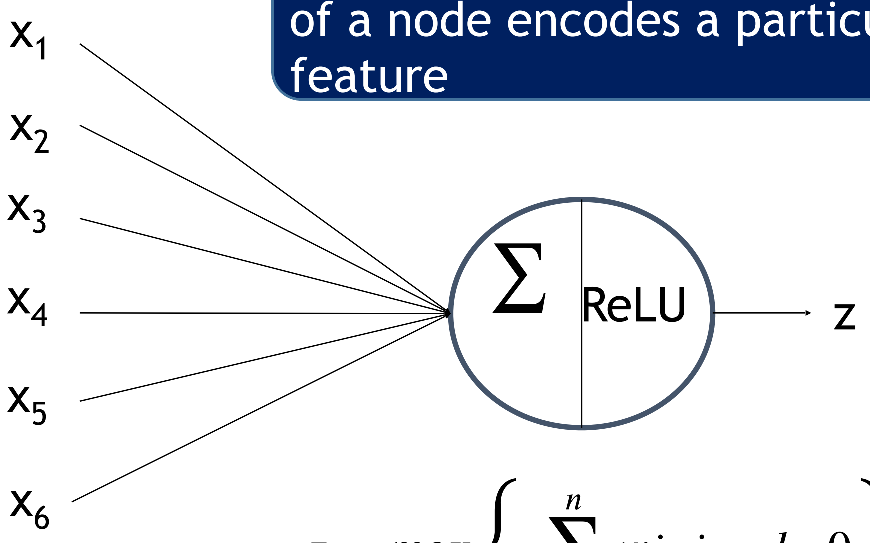


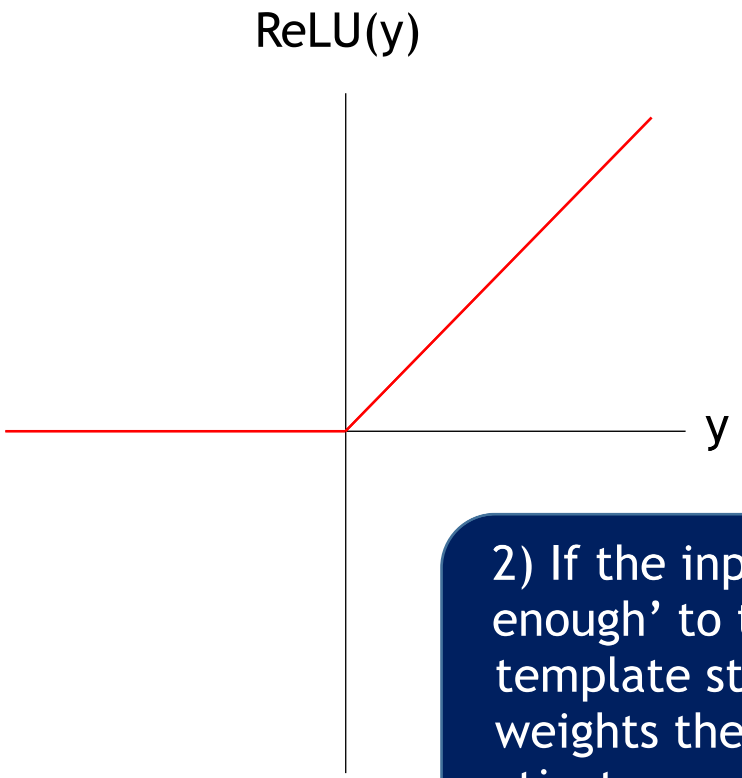
Image from Andrew Ng

The activation of a node in the CNN indicates the presence of a particular feature

1) The set of input weights $\{w_i\}$ of a node encodes a particular feature



$$z = \max\left\{ \sum_{i=1}^{n} wixi \; - b, \; 0 \right\}$$

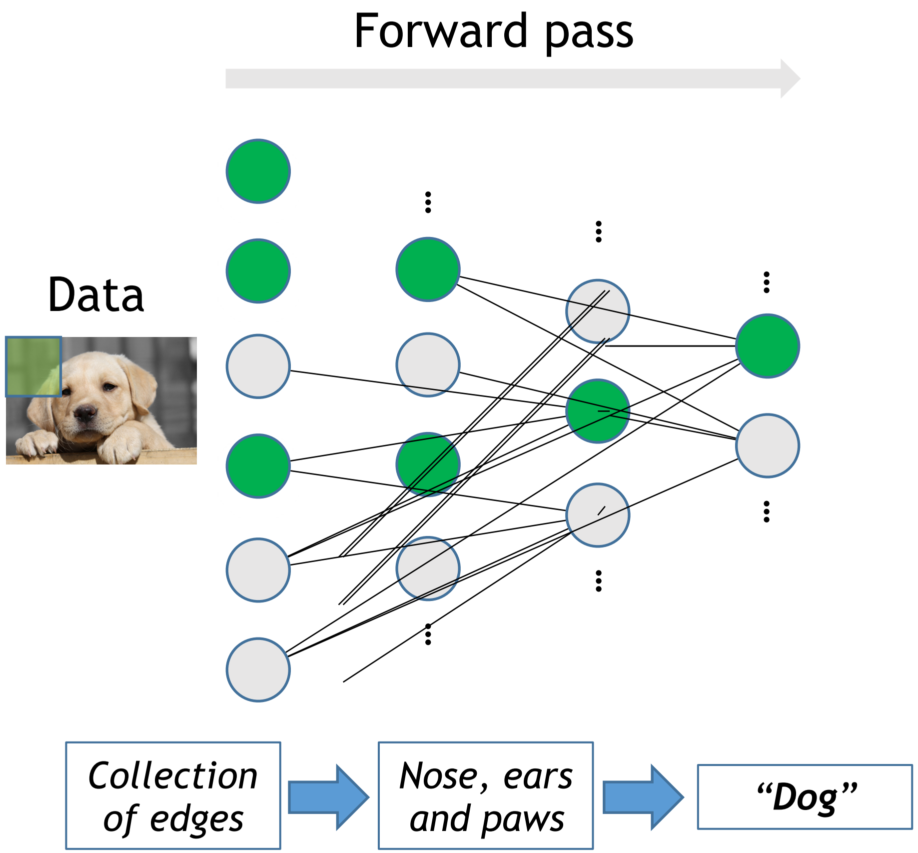


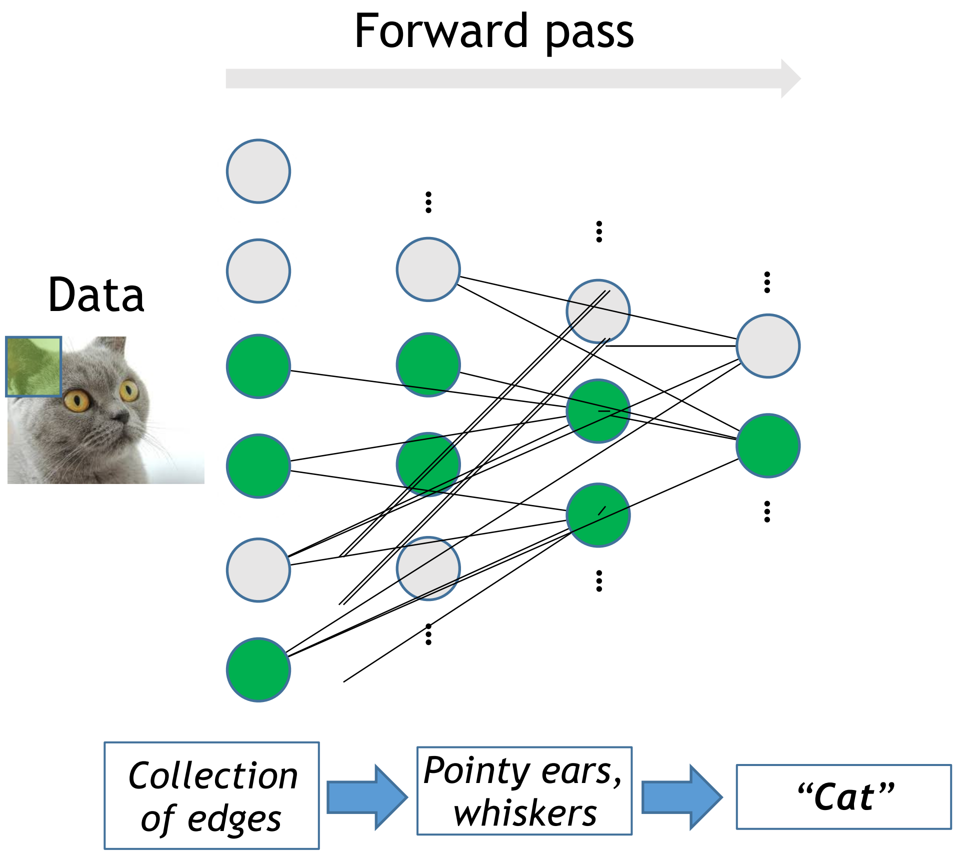
2) If the input is 'close enough' to the feature template stored in the weights then the node activates and is non-zero

3) The pattern of node activations across the network is called the activation pathway

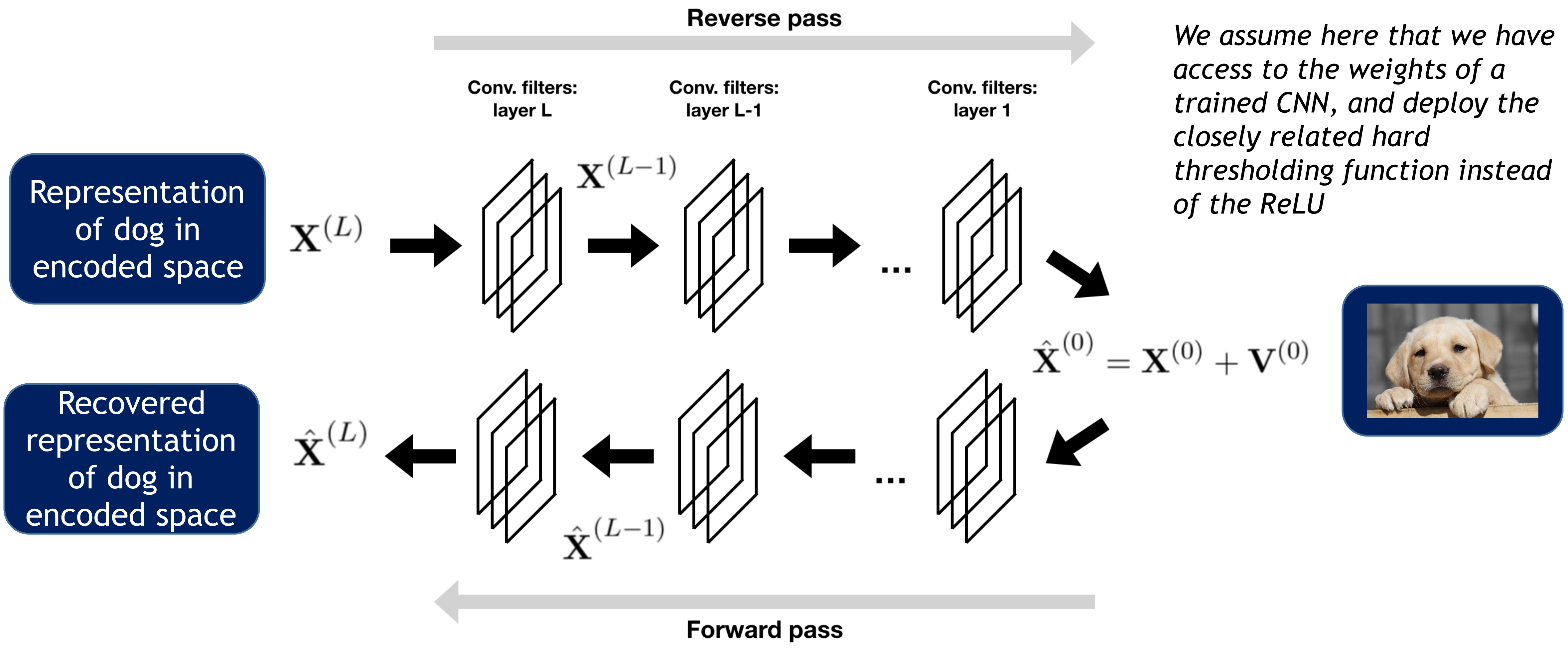
The activation pathway indicates how the CNN interprets new input data in terms of its own hierarchy of representations

The cat and dog images navigate the network via different activation paths





We can analyze the ability of the forward pass to recover the <u>reverse</u> activation pathway using the DCSC framework



We assume here that we have access to the weights of a trained CNN, and deploy the closely related hard thresholding function instead of the ReLU

Building on prior work, we can bound the probability that the forward pass fails to recover a given reverse activation pathway

Theorem 1. *Let $\hat{\boldsymbol{X}}^{(0)}$ be a date matrix consistent with Model (2.2) with $\| \boldsymbol{V}^{(l)} \|_{2,\infty}^{P^{(l)}} \leq \zeta_l$, $\| \boldsymbol{X}^{(l)} \|_{0,\infty}^{Q^{(l)}} \leq S_l$ for all $l = 0, \ldots, L-1$. Further assume that $\boldsymbol{D}^{(l)}$ is a random diagonal matrix with independent Rademacher random variables on the diagonal entries drawn independent of the dictionaries $\boldsymbol{A}^{(l)}$. Then let $\hat{\boldsymbol{X}}^{(l)}$ be computed as in (2.4) and denote as Z_L the event that the location of the non-zeros in $\boldsymbol{X}^{(l)}$ and $\hat{\boldsymbol{X}}^{(l)}$ exactly coincide for $l = 0, 1, \ldots, L$; then the probability this event doesn't hold, $\bar{Z}_L$, is at most*

$$P(\bar{Z}_L) \leq 2dM \sum_{l=1}^{L} n_l \exp\left(- \frac{|X_{min}^{(l)}|^2}{8\left(|X_{max}^{(l)}|^2 \mu_l^2 S_l + \zeta_{l-1}^2\right)} \right) \tag{3.1}$$

Furthermore when Z_L does occur then for all j,

$$\| \hat{\boldsymbol{x}}_j^{(l)} - \boldsymbol{x}_j^{(l)} \|_{2,\infty}^{P^{(l)}} \leq \zeta_l \tag{3.2}$$

where

$$\zeta_l = \sqrt{\| \hat{\boldsymbol{X}}^{(l)} \|_{0,\infty}^{P^{(l)}}} \left(\mu_l (S_l - 1) |X_{max}^{(l)}| + \zeta_{l-1} \right). \tag{3.3}$$

From activation pathways to information highways

Limitations of current results:

1. Is a sufficient but not necessary condition on the probability of recovery
2. For a non-trivial probability bound we require certain sparsity and noise constraints to be satisfied
3. **Exact recovery is a good starting point to understand the forward pass but CNNs do not need to be invertible to perform well**

Future goals: understand how data points of a given class traverse a trained network, i.e., understand the information highways through the network rather than just the activation path of a single data point

Summary

- A deeper understanding of CNNs is required to better facilitate their deployment in higher risk applications
- The activation pathway is indicative of how a CNN processes data, hence a better understanding of it may inform us as to the CNNs' failure modes
- We introduce the DCSC model and the reverse activation pathway as a method of analyzing the efficacy of the forward pass
- Using this approach we are able to bound the probability that the forward pass fails to recover the reverse activation pathway
- In the future we aim to analyze how classes of inputs are routed through the network and hence how CNNs' organize information they gain during training

Questions?