Visualizing and Understanding Convolutional Networks

ZEILER, Matthew D.; FERGUS, Rob. Visualizing and understanding convolutional networks. En European conference on computer vision. Springer, Cham, 2014. p. 818-833.

Master: Intelligent Systems

Subject: SJK003 - Machine Learning

Author: Pablo Muñoz Alcaide

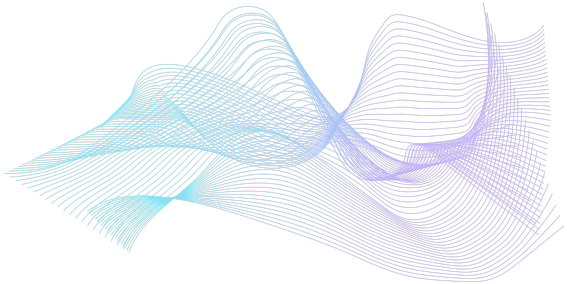
GENERALITAT VALENCIANA

valgrAI Valencian Graduate School and Research Network of Artificial Intelligence

Table of Contents

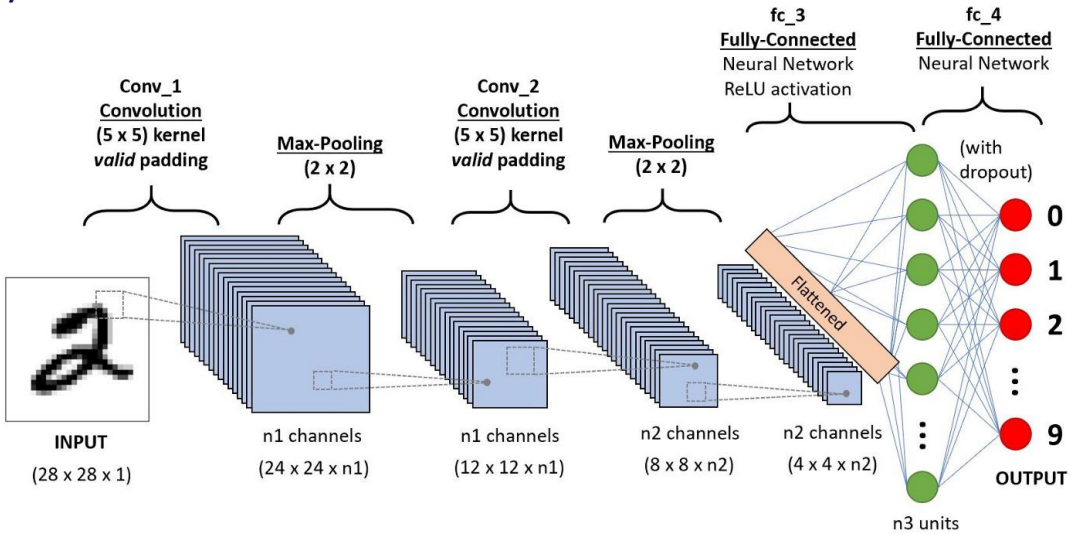
01

Introduction & Motivation

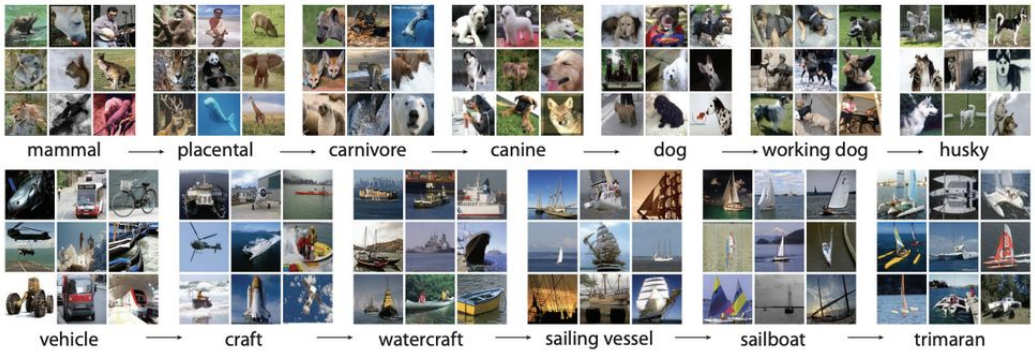


Convolutional Networks (CNN)

Convolutional neural networks are simply neural networks that use convolution in place of general matrix multiplication in at least one of their layers.



Convolutional Networks (CNN)



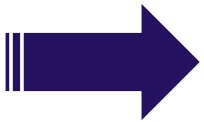
Convnets have demonstrated incredibly performance in different tasks. Such as:

- Image Classification (ImageNet)
- Face detection
- Digit Classification

Why CNNs perform so well?

Despite that encouraging progress, CNNs have some shortcomings.

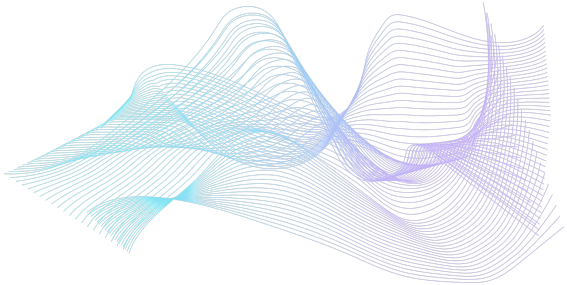
- “Black Box” Nature.
- Low information about the internal operations.
- Limited insight of CNN behaviour.
- Performance is unpredictable.
- Development reduced to trial-and-error.



How could this problem be solved?

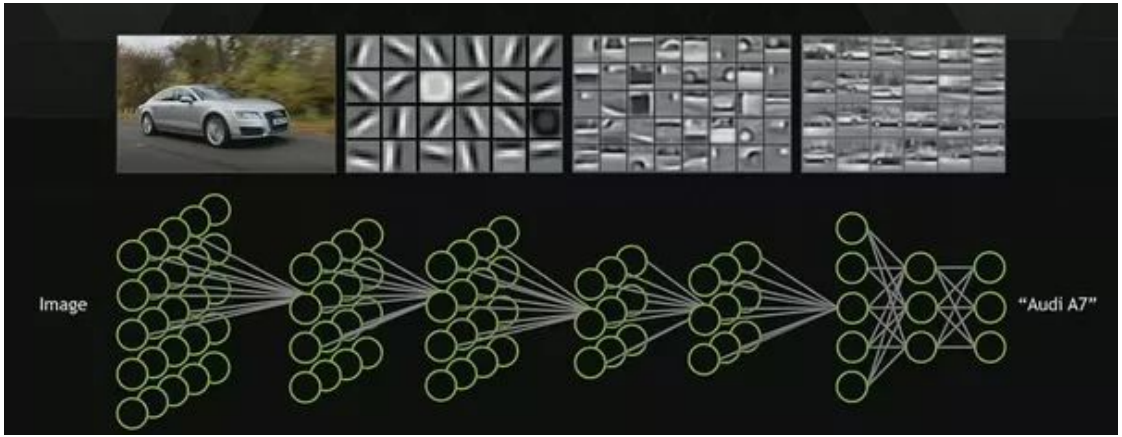
02

Visualisation of CNN



What is Visualisation of CNNs?

Visualization of CNN refers to the creation of graphical representations that offer insights into the network's learned features activities and decision processes.



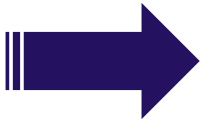
Typical problems related to visualisation

- Visualisation techniques are limited to the 1st layer. (Projections to pixel space are possible)
- In deeper layers there are another techniques such as Gradient descents, hessians…
- That techniques are limited or too complex.

Typical problems related to visualisation

- Visualisation techniques are limited to the 1st layer. (Projections to pixel space are possible)
- In deeper layers there are another techniques such as Gradient descents, hessians…
- That techniques are limited or too complex.

Zeiler et al.



Solution

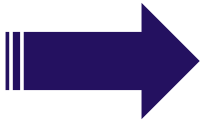
Solution proposed by Zeiler *et al.* provides a visualization technique that:

- Reveals the input pattern that excite individual feature maps at any layer.

Typical problems related to visualisation

- Visualisation techniques are limited to the 1st layer. (Projections to pixel space are possible)
- In deeper layers there are another techniques such as Gradient descents, hessians…
- That techniques are limited or too complex.

Zeiler et al.



Solution

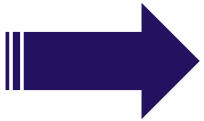
Solution proposed by Zeiler *et al.* provides a visualization technique that:

- Reveals the input pattern that excite individual feature maps at any layer.
- Allows to observe the evolution of features during training.

Typical problems related to visualisation

- Visualisation techniques are limited to the 1st layer. (Projections to pixel space are possible)
- In deeper layers there are another techniques such as Gradient descents, hessians…
- That techniques are limited or too complex.

Zeiler et al.

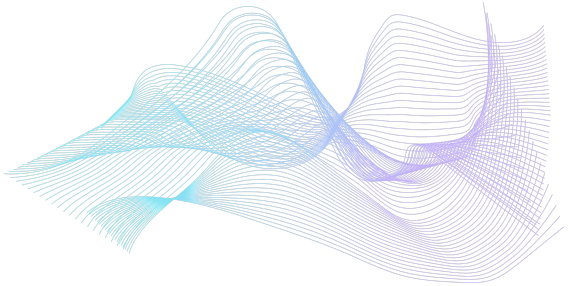


Solution

Solution proposed by Zeiler *et al.* provides a visualization technique that:

- Reveals the input pattern that excite individual feature maps at any layer.
- Allows to observe the evolution of features during training.
- Allows to diagnose potential problems with the model.

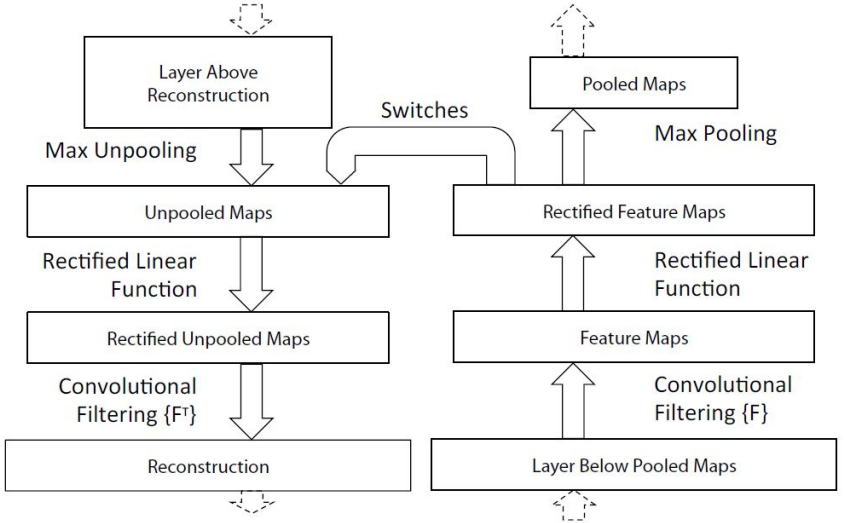
03



Zeiler *et al.* Solution

Decovnet Solution:

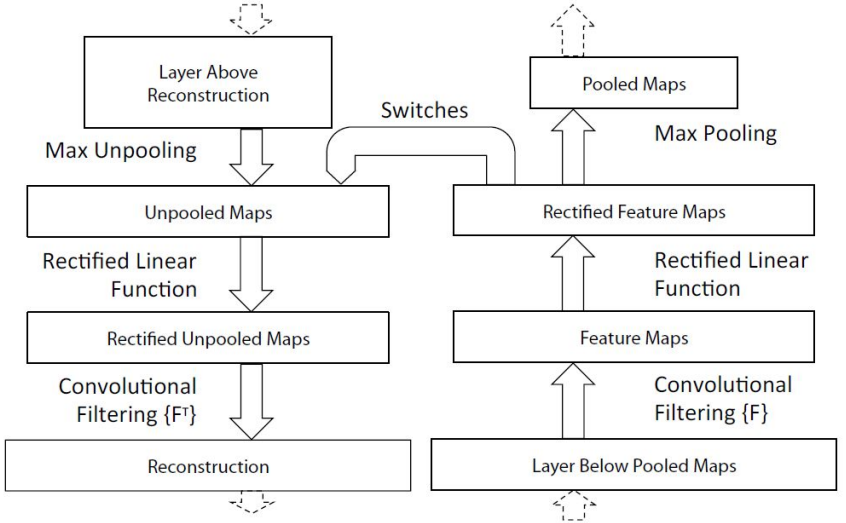
Deconvolutional Network allows to map feature activities back to the input pixel space.



Decovnet Solution:

Deconvolutional Network allows to map feature activities back to the input pixel space.

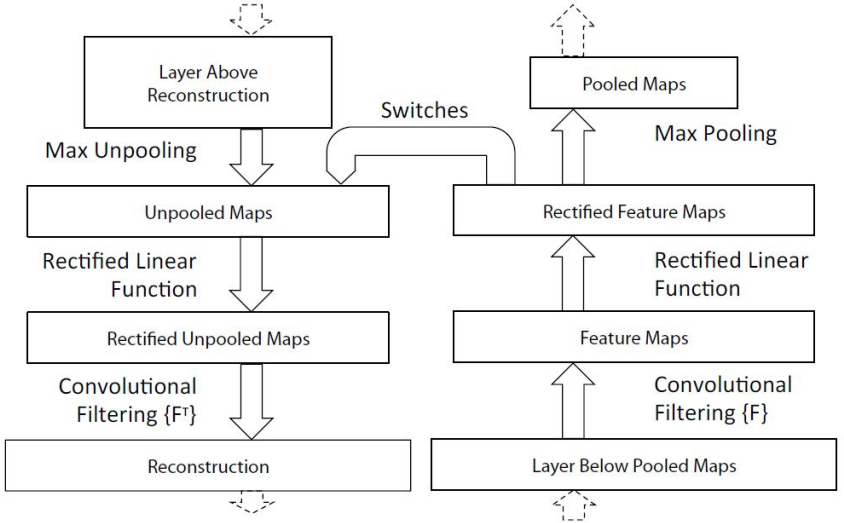
- Decovnet can be thought as a CNN but in reverse. (Features to pixels)



Decovnet Solution:

Deconvolutional Network allows to map feature activities back to the input pixel space.

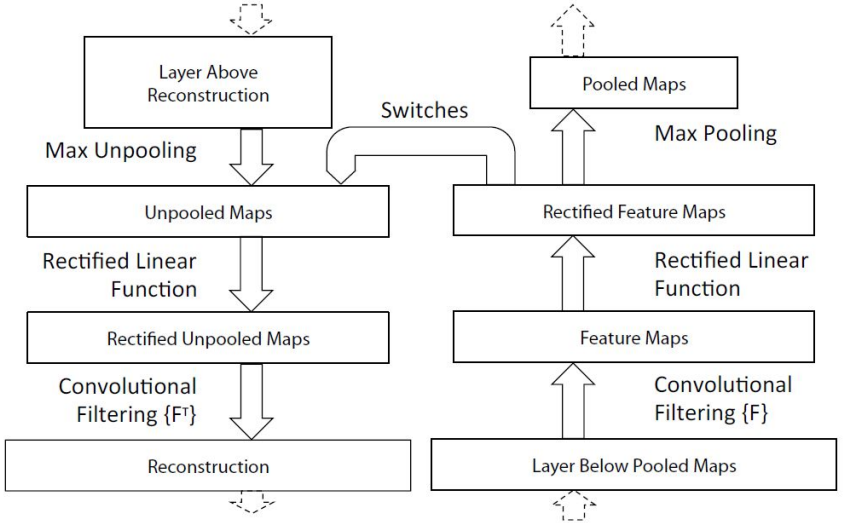
- Decovnet can be thought as a CNN but in reverse. (Features to pixels)
- Decovnet is attached to each CNN layer.



Decovnet Solution:

Deconvolutional Network allows to map feature activities back to the input pixel space.

- Decovnet can be thought as a CNN but in reverse. (Features to pixels)
- Decovnet is attached to each CNN layer.

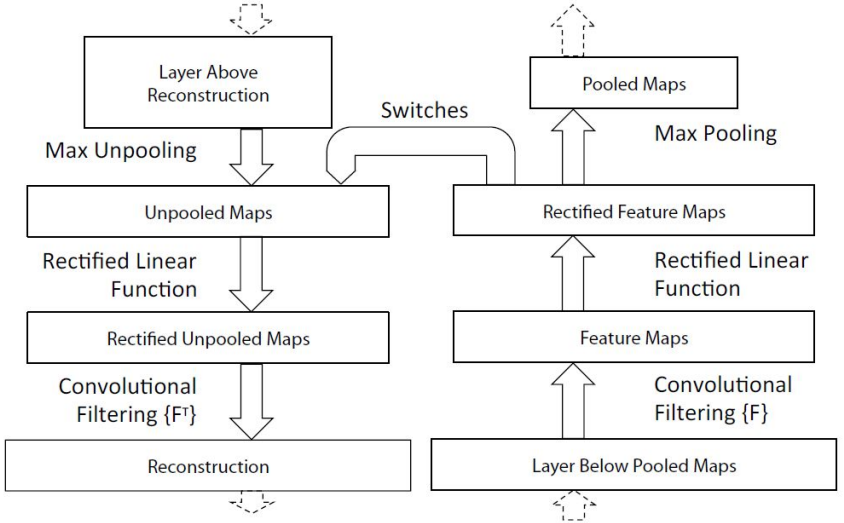


- Image is computed throughout the layers in the CNN.

Decovnet Solution:

Deconvolutional Network allows to map feature activities back to the input pixel space.

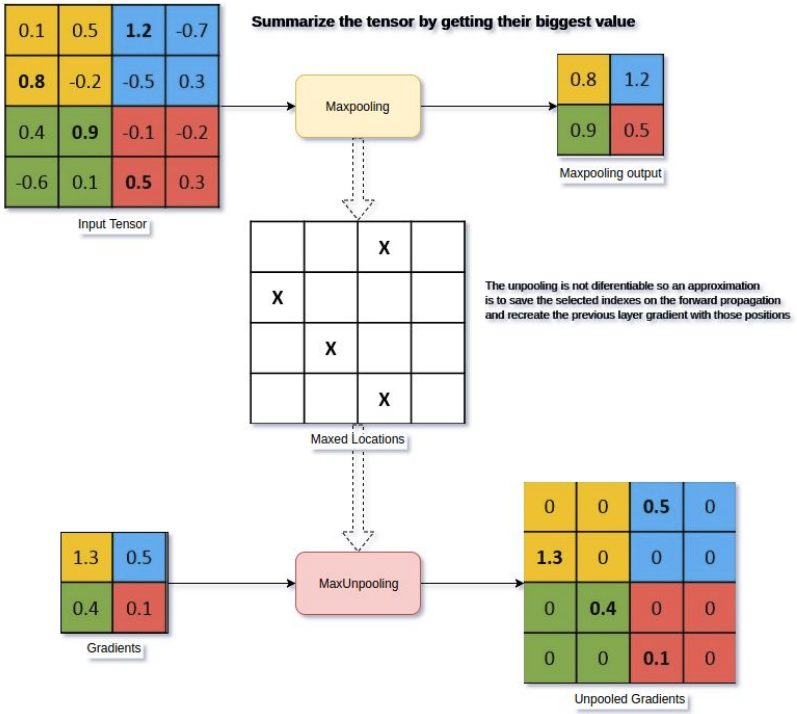
- Decovnet can be thought as a CNN but in reverse. (Features to pixels)
- Decovnet is attached to each CNN layer.



- Image is computed throughout the layers in the CNN.
- To examine a given CNN layer activation:
  - Set all other activations in the layer to zero.
  - Pass the feature map as input to the attached decovnet layer.
  - Reconstruct: Unpool, rectify and filter.
  - Repeat until pixel space is reached.

Unpooling: Understanding Deconvnet

- Max-pooling is non-invertible.

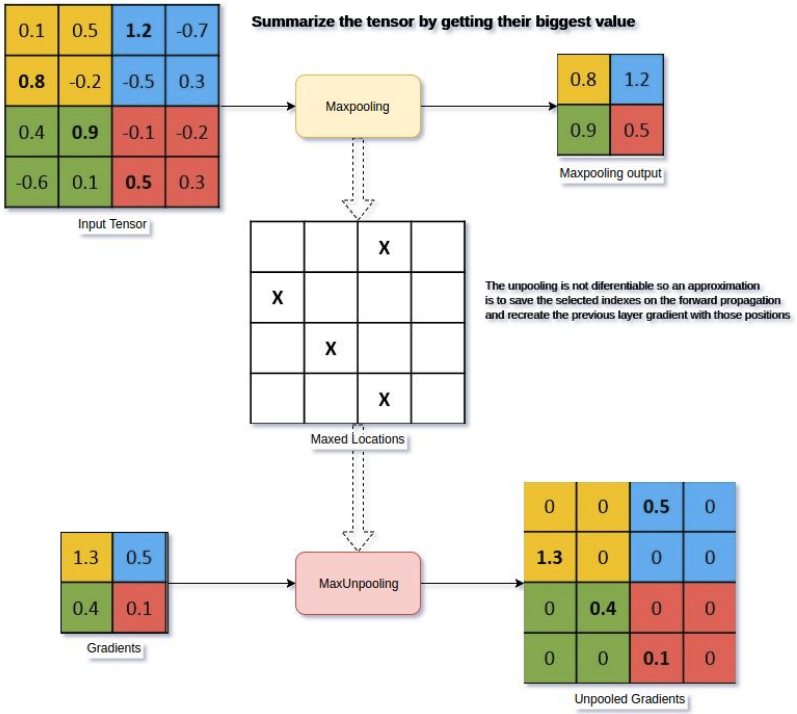


Unpooling: Understanding Deconvnet

- Max-pooling is non-invertible.
- Approximated Inverse:

- Save locations of the maxima within each pooling region. (***Switch*** variables)

- Unpooling operation use this switches to place the reconstructions from the layer above into appropriate locations.



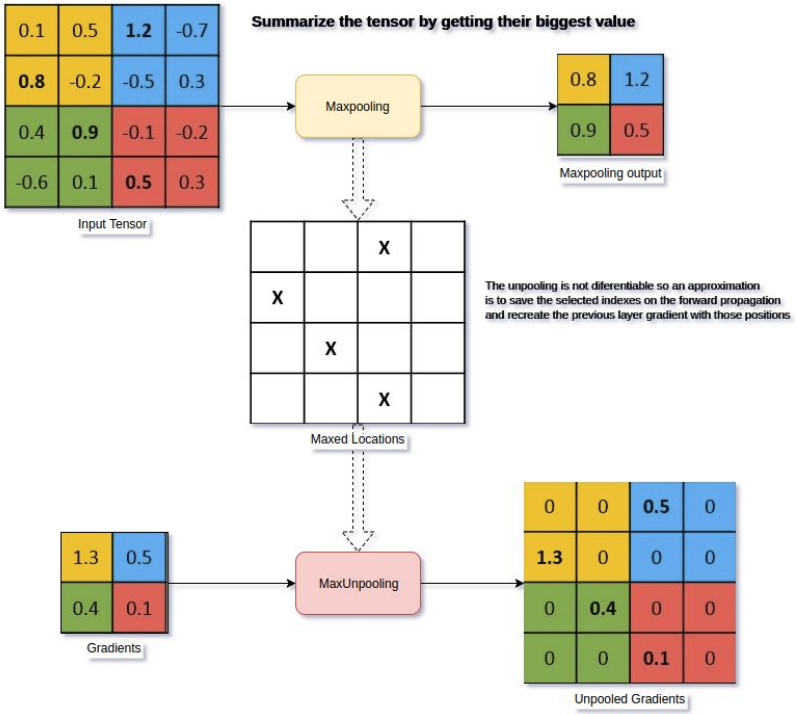
Unpooling: Understanding Deconvnet

- Max-pooling is non-invertible.
- Approximated Inverse:

- Save locations of the maxima within each pooling region. (***Switch*** variables)

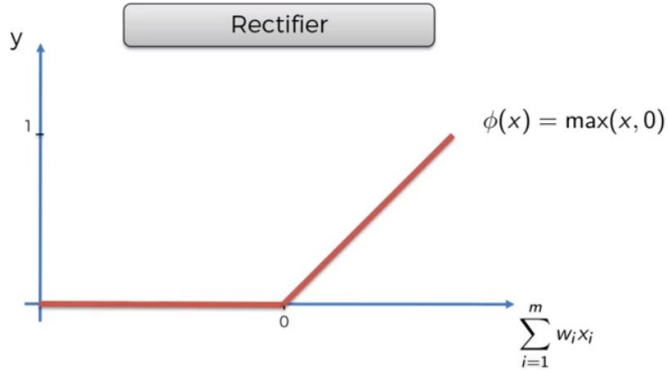
- Unpooling operation use this switches to place the reconstructions from the layer above into appropriate locations.

- Unpooling preserves the structure of the stimulus.



Rectification: Understanding Deconvnet

- CNN use ReLU to rectify feature maps, ensuring feature maps are always positive.



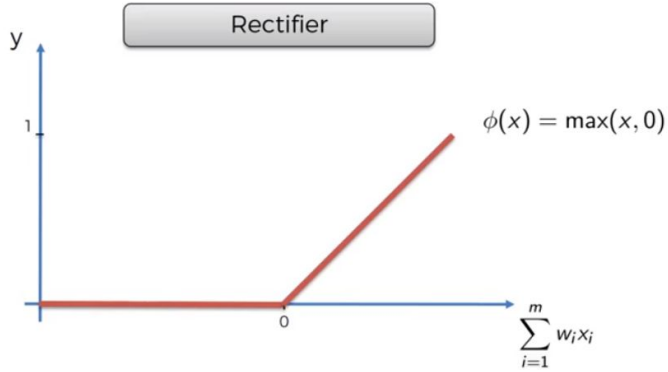
2	1	-3
1	2	1
2	1	-1

Applying ReLU

2	1	0
1	2	1
2	1	0

Rectification: Understanding Deconvnet

- CNN use ReLU to rectify feature maps, ensuring feature maps are always positive.
- ReLU introduces non-linearity.



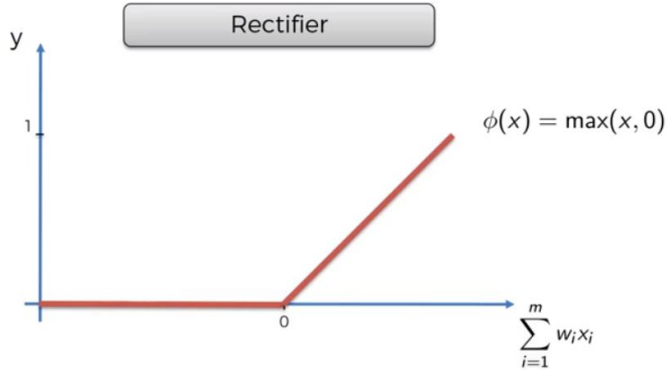
2	1	-3
1	2	1
2	1	-1

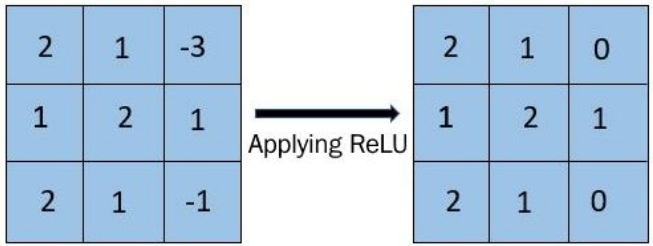
Applying ReLU

2	1	0
1	2	1
2	1	0

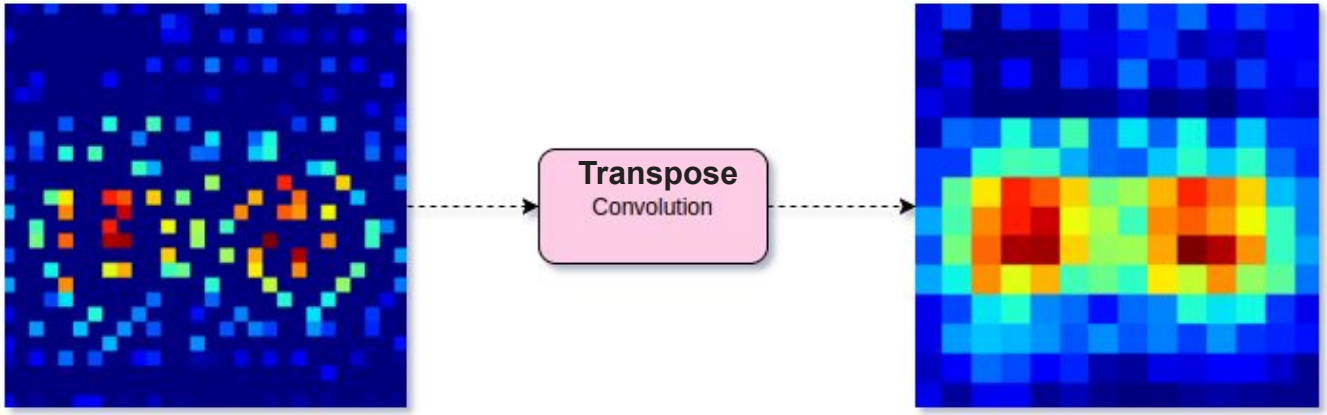
Rectification: Understanding Deconvnet

- CNN use ReLU to rectify feature maps, ensuring feature maps are always positive.
- ReLU introduces non-linearity.
- In the decovnet: We pass the reconstructed signal through ReLU to obtain valid features reconstructions.





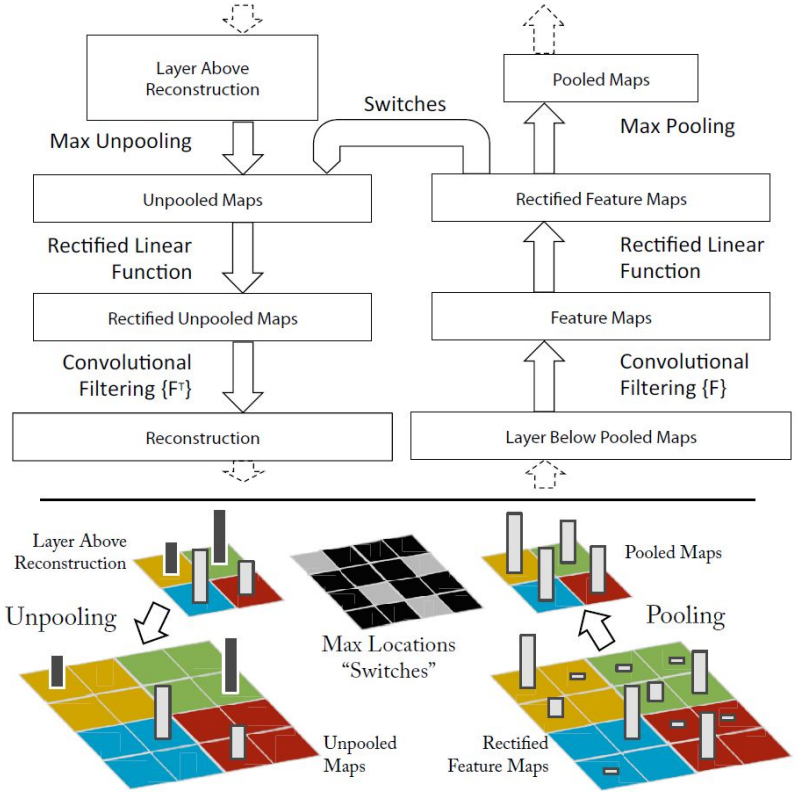
Filtering: Understanding Deconvnet



The unpooling and rectified results are sparse, so we need to apply approximated inverse convolution (Transpose of Convolution applied in the CNN) to the rectified map.

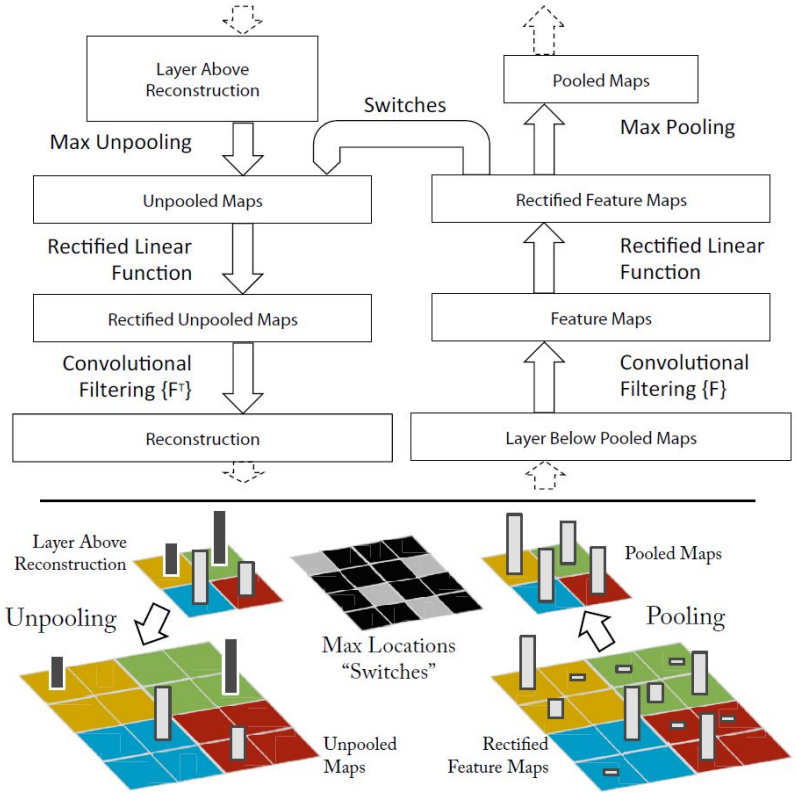
Whole Architecture

- The model is trained discriminatively, so it shows which parts of the input image are discriminative.



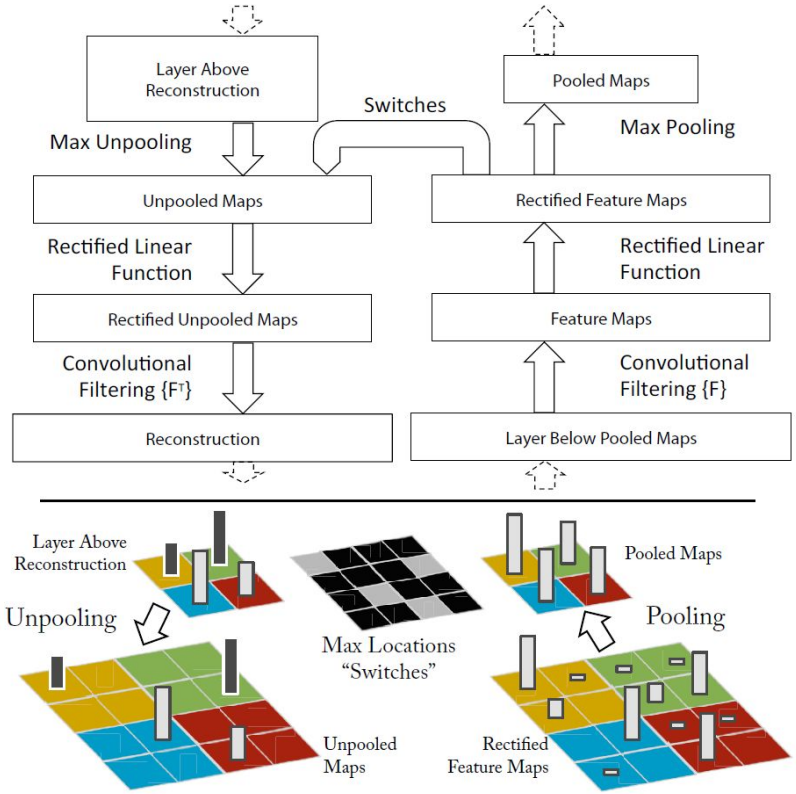
Whole Architecture

- The model is trained discriminatively, so it shows which parts of the input image are discriminative.
- Shortcoming: Only visualizes a single activation, not the joint activity present in a layer.



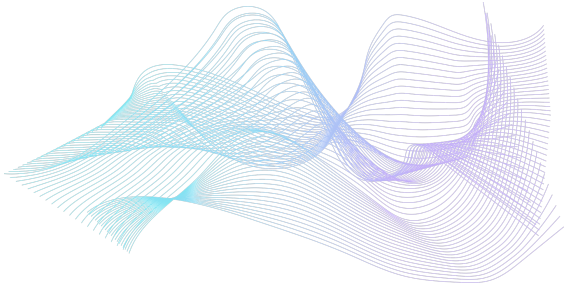
Whole Architecture

- The model is trained discriminatively, so it shows which parts of the input image are discriminative.
- Shortcoming: Only visualizes a single activation, not the joint activity present in a layer.
- Nevertheless, the visualizations are accurate representations of the input stimuli that stimulates the feature map.



04

Understanding the Visualisation

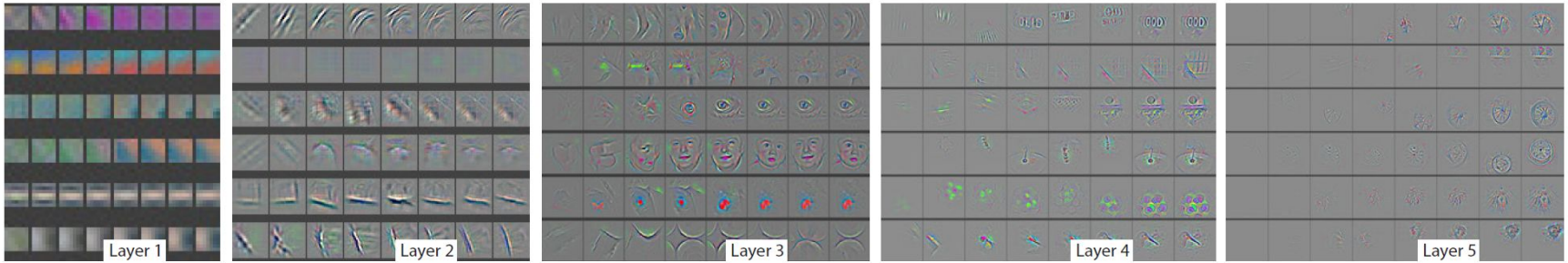


Feature visualisation

At left, the top 9 activations of a selected layer, each projected separately down to pixel space from the given images. Thus, revealing the different structures that excite that map and showing its invariance to input deformations.

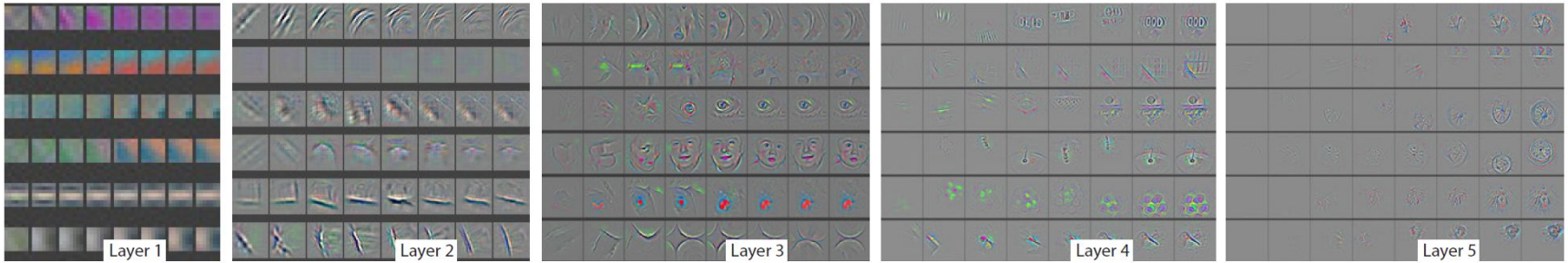
Feature visualisation

Evolution of random subset of features through epochs [1,2,5,10,20,30,40,64]. This visualization shows the strongest activations of the given features, projected down to pixel space.



Feature visualisation

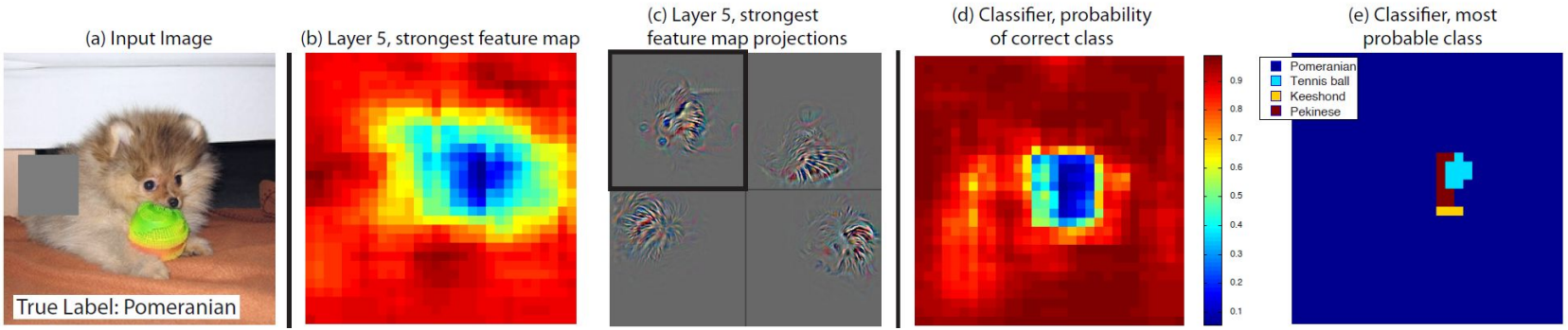
Evolution of random subset of features through epochs [1,2,5,10,20,30,40,64]. This visualization shows the strongest activations of the given features, projected down to pixel space.



- Each layer show the hierarchical nature of the features in the network. (Layer 2: Responds to corners. Layer 3: Similar textures. Layer 4: Significant variation…)
- Lower layers can be seen to converge in a few epochs. Upper Layers, need a considerable number of epochs.

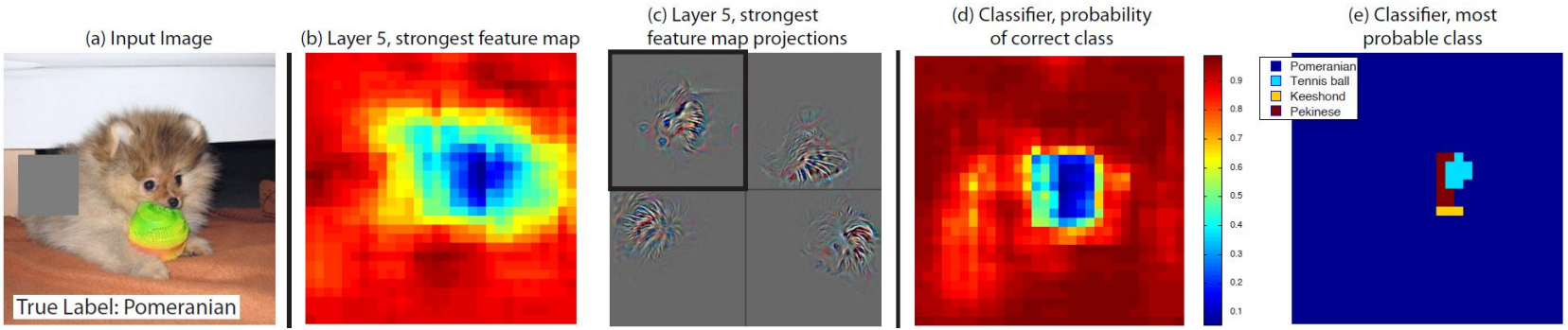
Occlusion Sensitivity

The model is identifying the location of the object, or just using the surrounding context?



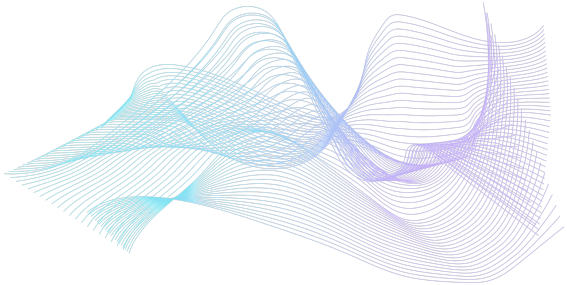
Occlusion Sensitivity

The model is identifying the location of the object, or just using the surrounding context?



Covering-ups of different portions of the image was made. The model localize the objects within the scene, as the probability of the correct class drops significantly when the object is occluded.

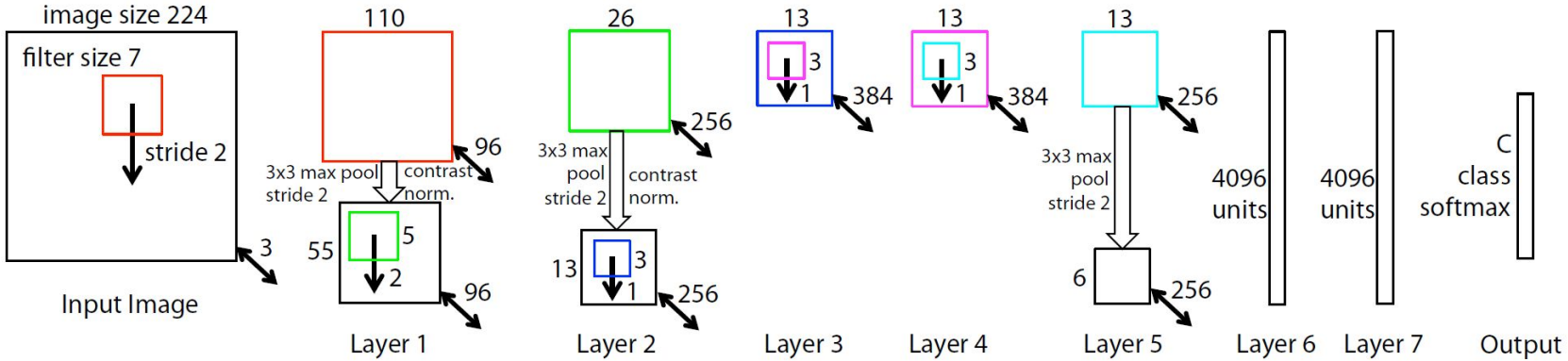
05



Experiments & Results

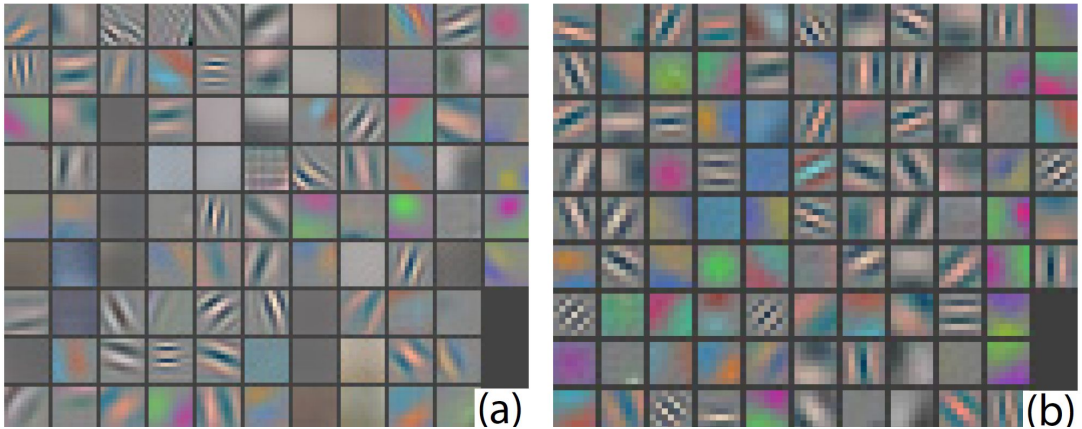
Architecture of the CNN used

The initial choice was Krizhevsky *et al.*'s model from "*ImageNet classification with deep convolutional neural networks*", setting records in the 2012 ImageNet classification. However, the visualisation of the features using the decovnet solution led to the discovery of some improvements.



Why this architecture?

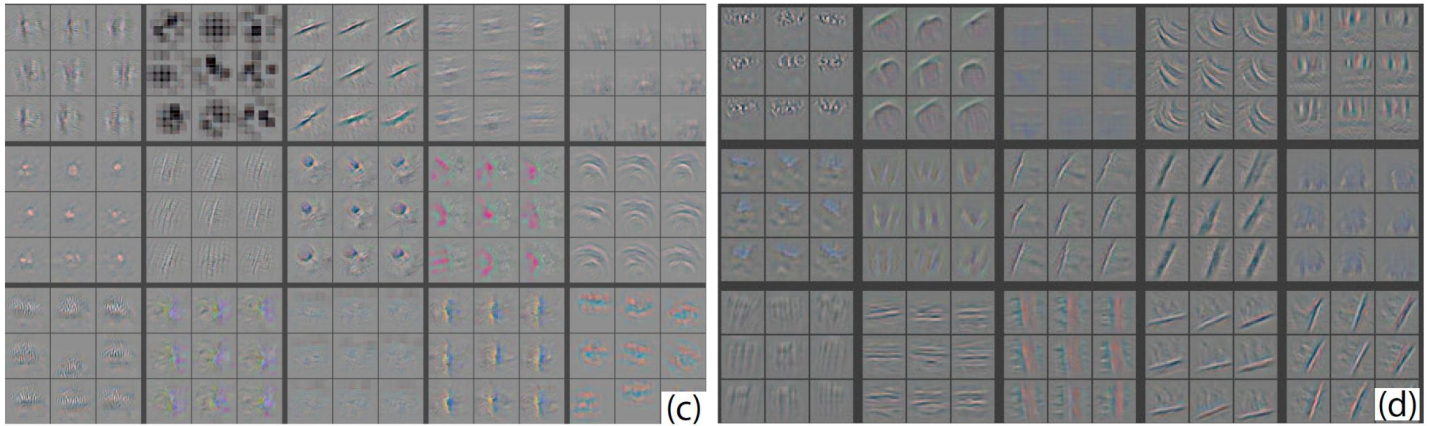
Visualising the first layer of Krizhevsky *et al.* **a)**, it was seen that the filters were a mixture of extremely high and low frequencies.



Zeiler *et al.* therefore reduced the first filter layer from 11x11 to 7x7 to solve that problem **b)**.

Why this architecture?

The 2nd layer visualization **c)** shows aliasing artifacts caused by the large stride of 4 used in the 1st layer convolutions.



Zeiler *et al.* therefore reduce the stride of the 1st layer to 2, retaining much more information **b)**.

ImageNet 2012

Zeiler *et al.* model achieves:

- 1.7% less test error than Krizhevsky *et al.*

1.3M/50k/100k training/validation/test images, spread over 1000 categories.

ImageNet 2012

Zeiler *et al.* model achieves:

- 1.7% less test error than Krizhevsky *et al.*
- Some experiments, removing and modifying layers, show that the overall depth of the model is important for obtaining good performance.

1.3M/50k/100k training/validation/test images, spread over 1000 categories.

ImageNet 2012

Zeiler *et al.* model achieves:

- 1.7% less test error than Krizhevsky *et al.*
- Some experiments, removing and modifying layers, show that the overall depth of the model is important for obtaining good performance.
- That experiments show the importance of the convolutional part in obtaining state-of-the-art performance.

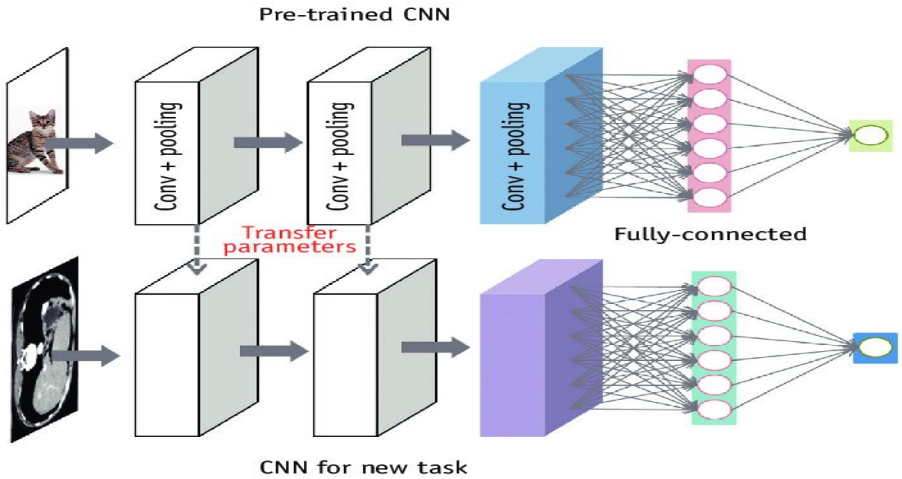
1.3M/50k/100k training/validation/test images, spread over 1000 categories.

Feature Generalization

The ImageNet experiment and feature visualization indicate the convolutional layers learned complex invariances.

Feature Generalization

The ImageNet experiment and feature visualization indicate the convolutional layers learned complex invariances. To validate the generalization, the first seven layers of the ImageNet-trained model remained fixed, and a new softmax classifier was trained using Caltech-101, Caltech-256, and PASCAL VOC 2012 datasets.



Caltech-101 and Caltech-256

The results highlight the efficacy of the ImageNet feature extractor, surpassing all competitors in the field of Caltech-101 and Caltech-256 image classification.

Caltech-101 and Caltech-256

The results highlight the efficacy of the ImageNet feature extractor, surpassing all competitors in the field of Caltech-101 and Caltech-256 image classification.

# Train	Acc % 15/class	Acc % 30/class
Bo *et al.* [3]	–	81.4 ± 0.33
Yang *et al.* [17]	73.2	84.3
Non-pretrained convnet	22.8 ± 1.5	46.5 ± 1.7
ImageNet-pretrained convnet	$\mathbf{83.8 \pm 0.5}$	$\mathbf{86.5 \pm 0.5}$

a) Caltech-101 results.

Caltech-101 and Caltech-256

The results highlight the efficacy of the ImageNet feature extractor, surpassing all competitors in the field of Caltech-101 and Caltech-256 image classification.

# Train	Acc % 15/class	Acc % 30/class
Bo *et al.* [3]	–	81.4 ± 0.33
Yang *et al.* [17]	73.2	84.3
Non-pretrained convnet	22.8 ± 1.5	46.5 ± 1.7
ImageNet-pretrained convnet	$\mathbf{83.8 \pm 0.5}$	$\mathbf{86.5 \pm 0.5}$

a) Caltech-101 results.

# Train	Acc % 15/class	Acc % 30/class	Acc % 45/class	Acc % 60/class
Sohn *et al.* [24]	35.1	42.1	45.7	47.9
Bo *et al.* [3]	40.5 ± 0.4	48.0 ± 0.2	51.9 ± 0.2	55.2 ± 0.3
Non-pretr.	9.0 ± 1.4	22.5 ± 0.7	31.2 ± 0.5	38.8 ± 1.4
ImageNet-pretr.	$\mathbf{65.7 \pm 0.2}$	$\mathbf{70.6 \pm 0.2}$	$\mathbf{72.7 \pm 0.4}$	$\mathbf{74.2 \pm 0.3}$

b) Caltech-256 results.

PASCAL VOC 2012

A 20-way softmax was applied to the ImageNet-pretrained CNN. While this method provides a single exclusive prediction per image, it may not be ideal for PASCAL's scenes with multiple objects.

PASCAL VOC 2012

A 20-way softmax was applied to the ImageNet-pretrained CNN. While this method provides a single exclusive prediction per image, it may not be ideal for PASCAL's scenes with multiple objects.

Acc %	[22]	[27]	[21]	Ours	Acc %	[22]	[27]	[21]	Ours
Airplane	92.0	**97.3**	94.6	96.0	Dining table	63.2	**77.8**	69.0	67.7
Bicycle	74.2	**84.2**	82.9	77.1	Dog	68.9	83.0	**92.1**	87.8
Bird	73.0	80.8	88.2	**88.4**	Horse	78.2	87.5	**93.4**	86.0
Boat	77.5	85.3	60.3	**85.5**	Motorbike	81.0	**90.1**	88.6	85.1
Bottle	54.3	**60.8**	60.3	55.8	Person	91.6	95.0	**96.1**	90.9
Bus	85.2	**89.9**	89.0	85.8	Potted plant	55.9	57.8	**64.3**	52.2
Car	81.9	**86.8**	84.4	78.6	Sheep	69.4	79.2	**86.6**	83.6
Cat	76.4	89.3	90.7	**91.2**	Sofa	65.4	**73.4**	62.3	61.1
Chair	65.2	**75.4**	72.1	65.0	Train	86.7	**94.5**	91.1	91.8
Cow	63.2	77.8	**86.8**	74.4	Tv	77.4	**80.7**	79.8	76.1
Mean	74.3	82.2	**82.8**	79.0	# won	0	**11**	6	3

Results on the test set, compared to leading methods, show a mean performance 3.2% lower than the leading competition result. However, our model outperformed them on 5 classes, sometimes significantly.

Feature Analysis

How discriminative are the features in each layer is analized adjusting the number of layers retained from the ImageNet model and applying either a linear SVM or softmax classifier on top.

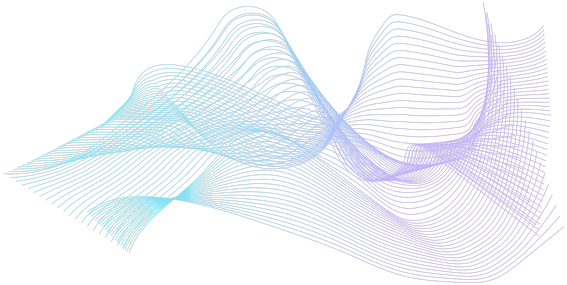
Feature Analysis

How discriminative are the features in each layer is analized adjusting the number of layers retained from the ImageNet model and applying either a linear SVM or softmax classifier on top.

- The deeper the layer, the more powerful features.

	Cal-101 (30/class)	Cal-256 (60/class)
SVM (1)	44.8 ± 0.7	24.6 ± 0.4
SVM (2)	66.2 ± 0.5	39.6 ± 0.3
SVM (3)	72.3 ± 0.4	46.0 ± 0.3
SVM (4)	76.6 ± 0.4	51.3 ± 0.1
SVM (5)	$\mathbf{86.2 \pm 0.8}$	65.6 ± 0.3
SVM (7)	$\mathbf{85.5 \pm 0.4}$	$\mathbf{71.7 \pm 0.2}$
Softmax (5)	82.9 ± 0.4	65.7 ± 0.5
Softmax (7)	$\mathbf{85.4 \pm 0.4}$	$\mathbf{72.6 \pm 0.1}$

06



Conclusions

Conclusions

- Model activity visualization show that features have compositionality, increasing invariance and class discrimination as we ascend the layers.

Conclusions

- Model activity visualization show that features have compositionality, increasing invariance and class discrimination as we ascend the layers.
- Visualisation identifies and addresses model problems, improving results in ImageNet 2012 classification.

Conclusions

- Model activity visualization show that features have compositionality, increasing invariance and class discrimination as we ascend the layers.
- Visualisation identifies and addresses model problems, improving results in ImageNet 2012 classification.
- Occlusion experiments reveal the model's sensitivity to local image structure over broad scene context.

Conclusions

- Model activity visualization show that features have compositionality, increasing invariance and class discrimination as we ascend the layers.
- Visualisation identifies and addresses model problems, improving results in ImageNet 2012 classification.
- Occlusion experiments reveal the model's sensitivity to local image structure over broad scene context.
- The ImageNet-trained model effectively generalizes to similar datasets, surpassing reported results, especially in Caltech-256.

Conclusions

- Model activity visualization show that features have compositionality, increasing invariance and class discrimination as we ascend the layers.
- Visualisation identifies and addresses model problems, improving results in ImageNet 2012 classification.
- Occlusion experiments reveal the model's sensitivity to local image structure over broad scene context.
- The ImageNet-trained model effectively generalizes to similar datasets, surpassing reported results, especially in Caltech-256.
- Generalization to PASCAL data, while less robust, remains within 3.2% of the best result without task-specific tuning.

Conclusions

- Model activity visualization show that features have compositionality, increasing invariance and class discrimination as we ascend the layers.
- Visualisation identifies and addresses model problems, improving results in ImageNet 2012 classification.
- Occlusion experiments reveal the model's sensitivity to local image structure over broad scene context.
- The ImageNet-trained model effectively generalizes to similar datasets, surpassing reported results, especially in Caltech-256.
- Generalization to PASCAL data, while less robust, remains within 3.2% of the best result without task-specific tuning.
- The use of a different loss function, could enhance performance, especially in object detection.

Thanks!

Do you have any questions?

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**