Imperial College London

Understanding Deep CNNs via Interpretable Individual Units

Individual Project for MSc degree

Author: Lin Li Supervisor: Prof. Wayne Luk
Acknowledgement: Dr. Ce Guo, Dr. Ruizhe Zhao

Contents

Importance

- know why a CNN works well and why it does not work
- improve the design of CNNs to perform better

Understanding so far

- a unit is activated by a particular pattern of input
- the representations are learnt **hierarchically** in order of **edges, colours textures, part of objects and objects**

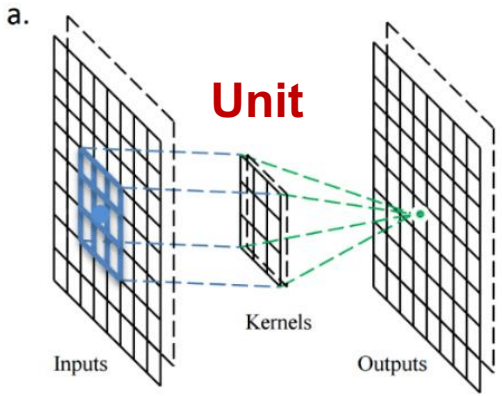
Contributions

- identify the **specific meaning** of units and analyse the distribution
- visualise the **filtering effect** of units and study how it changes with the depth
- diagnose some **classification errors** from viewing through semantic units

Terms

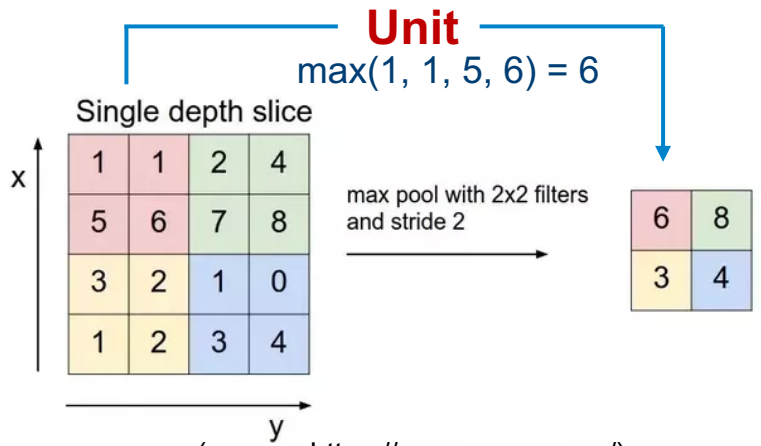
- **unit**: performs specific operations on specific inputs to **produce a feature map**

Convolution layer:



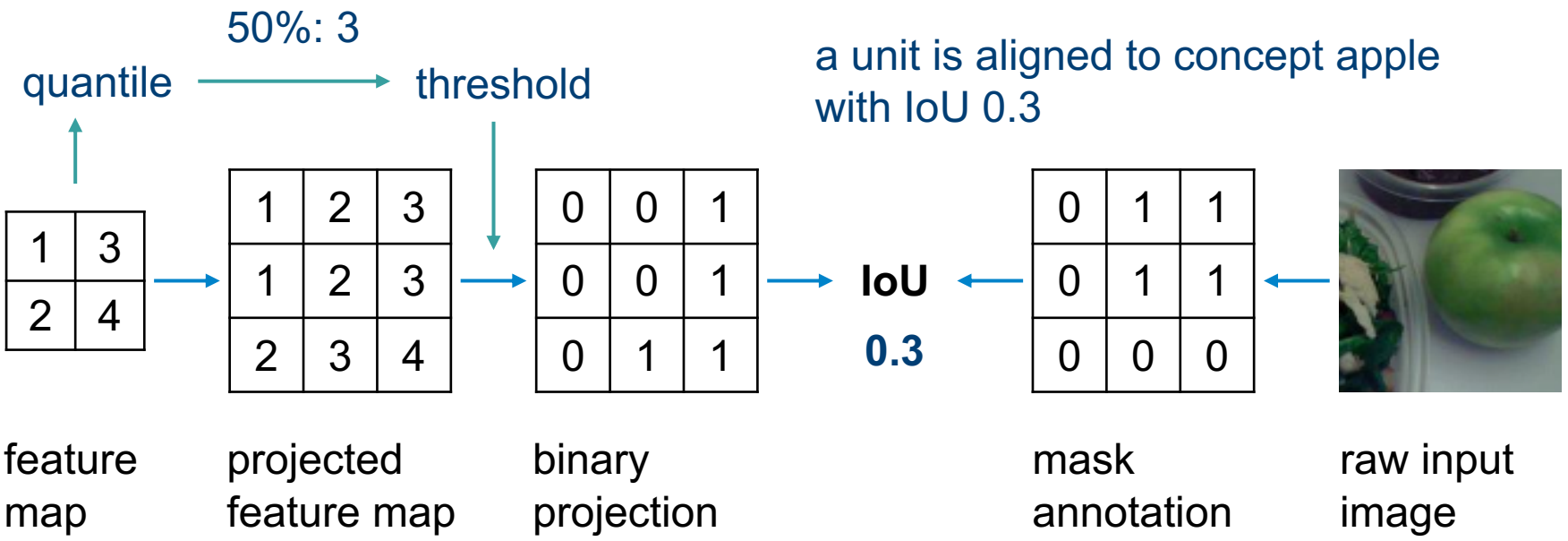
(source: https://www.researchgate.net)

Max pooling layer:



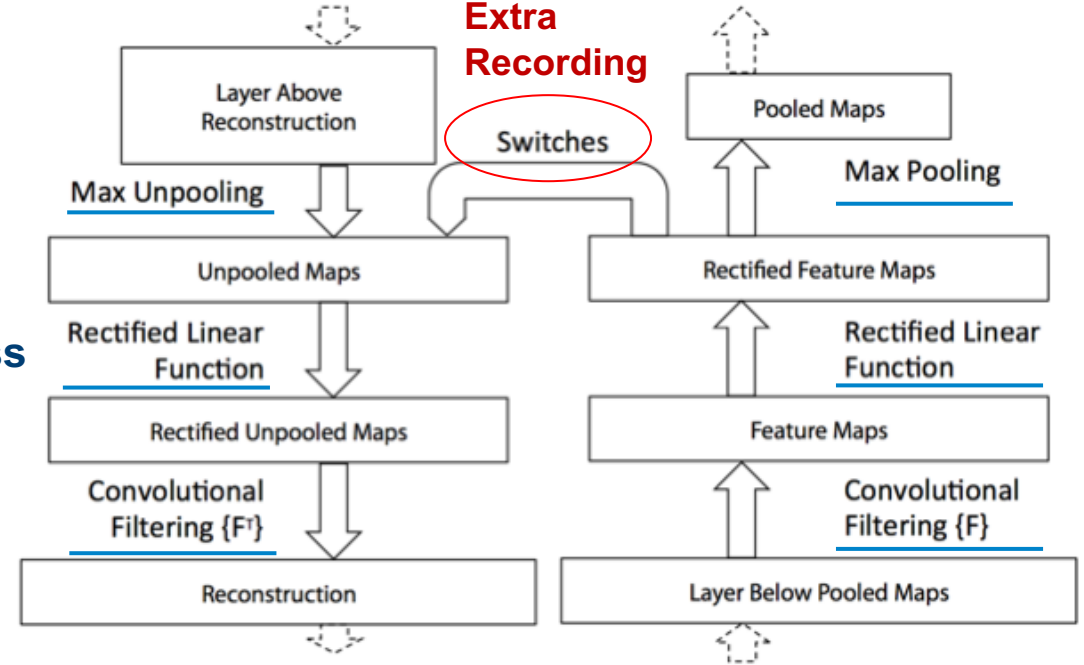
(source: https://www.quora.com/)

Network Dissection



DeConvNet

DeConvNet
Reversed process



ConvNet
Forward pass

(source: Visualizing and Understanding Convolutional Networks)

Identification

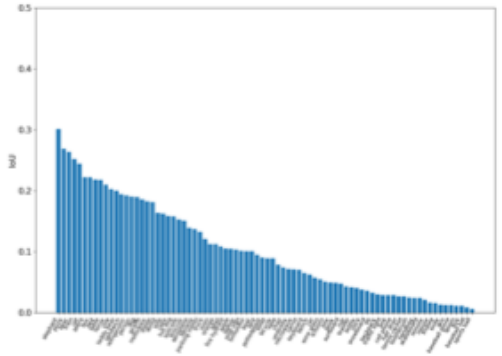
semantic meanings of units							
conv3_2_243		conv4_2_137		conv5_1_407		conv5_3_426	
concept	IoU	concept	IoU	concept	IoU	concept	IoU
bed	0.28	pizza	0.33	zebra	0.46	elephant	0.30
table	0.25	cat	0.29	giraffe	0.32	cat	0.28
pizza	0.25	elephant	0.26	bed	0.30	bear	0.27
bus	0.22	zebra	0.25	aeroplane	0.29	cow	0.27
train	0.22	motorbike	0.25	motorbike	0.26	zebra	0.27
matched units of concepts in different layers							
dog - conv3_2		dog - conv5_3		bear - conv4_1		bear - conv5_2	
unit id	IoU	unit id	IoU	unit id	IoU	unit id	IoU
245	0.21	274	0.32	257	0.39	7	0.43
220	0.21	437	0.30	50	0.35	138	0.41
0	0.21	198	0.27	330	0.34	278	0.41
16	0.20	141	0.26	419	0.33	47	0.41
69	0.20	421	0.26	179	0.33	83	0.39

Table 5.1: Top 5 semantic meanings of units and matched units of concepts

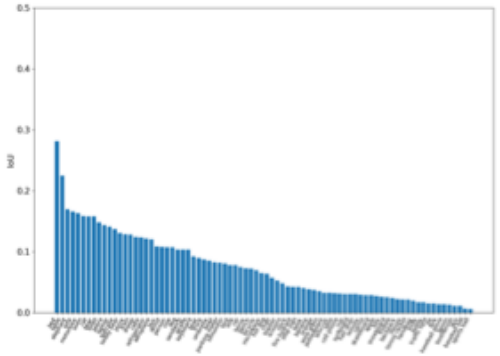
Clusters of Classes

- **Mammal:** Dog, sheep, cat, horse, elephant, cow
- **Vehicle:** bicycle, motorbike, car, truck

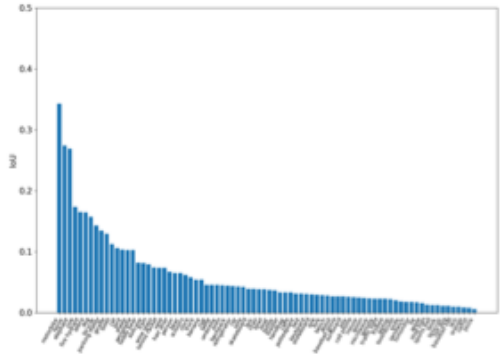
Distribution of concepts for a unit



a unit in **con3**



a unit in **con4**



a unit in **con5**

- **Semantic meaning of units** become more **specific**
- The units in **higher layers** become **more powerful** on **detecting top meanings**, while the ability of detecting others is suppressed

Part detectors

object	part	unit id	IoU	part	unit id	IoU
person	head	conv5_1_104	0.26	hair	conv4_3_80	0.13
	torso	conv5_2_313	0.21			
cat	head	conv5_2_86	0.34	torso	conv5_3_274	0.33
dog	head	conv5_2_248	0.31	muzzle	conv5_2_469	0.18
	torso	conv5_3_274	0.25			
horse	head	conv5_3_314	0.20	muzzle	conv5_3_145	0.14
	torso	conv5_3_325	0.33	neck	conv5_3_478	0.12
sheep	head	conv5_2_248	0.20	torso	conv5_2_143	0.34
cow	head	conv5_3_314	0.24	muzzle	conv5_3_145	0.14
	torso	conv5_3_325	0.32			
bird	head	conv5_2_16	0.11	torso	conv5_2_60	0.24
	wing	conv5_2_60	0.10			
aeroplane	body	conv5_2_469	0.29	stern	conv5_2_78	0.13
	engine	conv4_3_164	0.13			
bicycle	wheel	conv5_2_29	0.18			
motorbike	wheel	conv5_3_422	0.16	saddle	conv5_3_347	0.14
bus	wheel	conv4_3_419	0.13	window	conv5_3_92	0.31
	door	conv4_3_146	0.12			
train	head	conv5_3_443	0.44	coach	pool5_417	0.26
car	wheel	conv4_3_419	0.13	door	conv5_3_484	0.19
	window	conv5_3_190	0.14			
bottle	body	conv5_2_469	0.19			
tvmonitor	screen	conv5_3_435	0.26			
pottedplant	pot	conv5_2_417	0.12	plant	conv4_3_335	0.25

Table 5.3: Identified part detectors for each object-level class

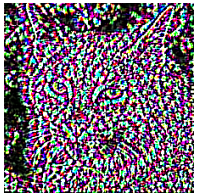
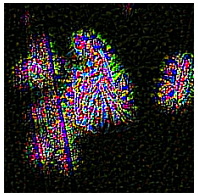
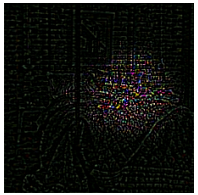
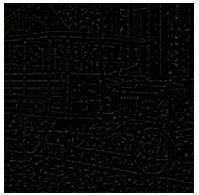
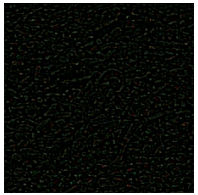
Distribution

- **identified parts** only take a **small proportion** of all parts
- part detectors emerge at **lower layers**

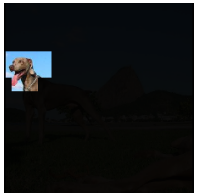
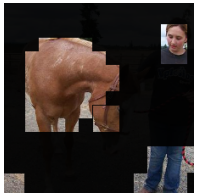
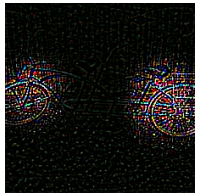
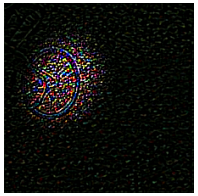
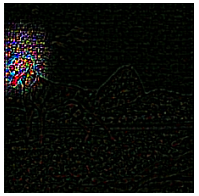
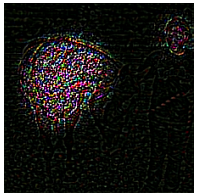
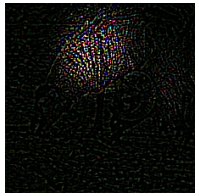
Common Part

- **Mammal** group: **head**, **torso**
- **Vehicle** group: **wheel**

Objects

	cat	person	bed	car	sheep
Original image					
	cat unit	person unit	bed unit	car unit	bed unit
Activation projection					
DeConvNet					

Part of objects

	bicycle	bicycle	dog	horse	sheep
Original image					
	wheel unit	wheel unit	head unit	torso and person	torso unit
Activation projection					
DeConvNet					

Filtering effect

Top sensitive units of **concept Bus** in different layers

- **irrelevant** pixels are **filtered out progressively**
- filtering effect is **enhanced as** the layer **depth increased**

Cross meanings

A **horse** is predicted as a **dog**

unique to dog

dog units

input

top 6 sensitive units of **dog** in the last convolution layer **conv5_3**

horse units

input

top 6 sensitive units of **horse** in the last convolution layer **conv5_3**

- **multiple units** with the **same meanings confuse** the model
- the units with **unique meaning** play a more important role in this case

Wrong activations

A **bed** is predicted as **not a bed**

wrong

activate irrelevant area & deactivate target

right

input

top 6 sensitive units of **bed** in the last convolution layer **conv5_3**

- **more** target concept detectors are **wrongly activated**
- **problems** arise **from** the **units not** the **fully-connected layers**

Network Dissection assessment

the **first quantitative framework** for identifying meanings of units

vs. other visualisation techniques

- identifies meanings by **analyzing all samples**, while others work **case by case**

vs. statistics-based methods

- utlises the information of **objects' locations**, while others do not
- can also **visualise** the procedures and results, while others can not

Network Dissection assessment

Advantages:

- **generalize well** among different datasets
- **easy** to be **scaled** to more categories

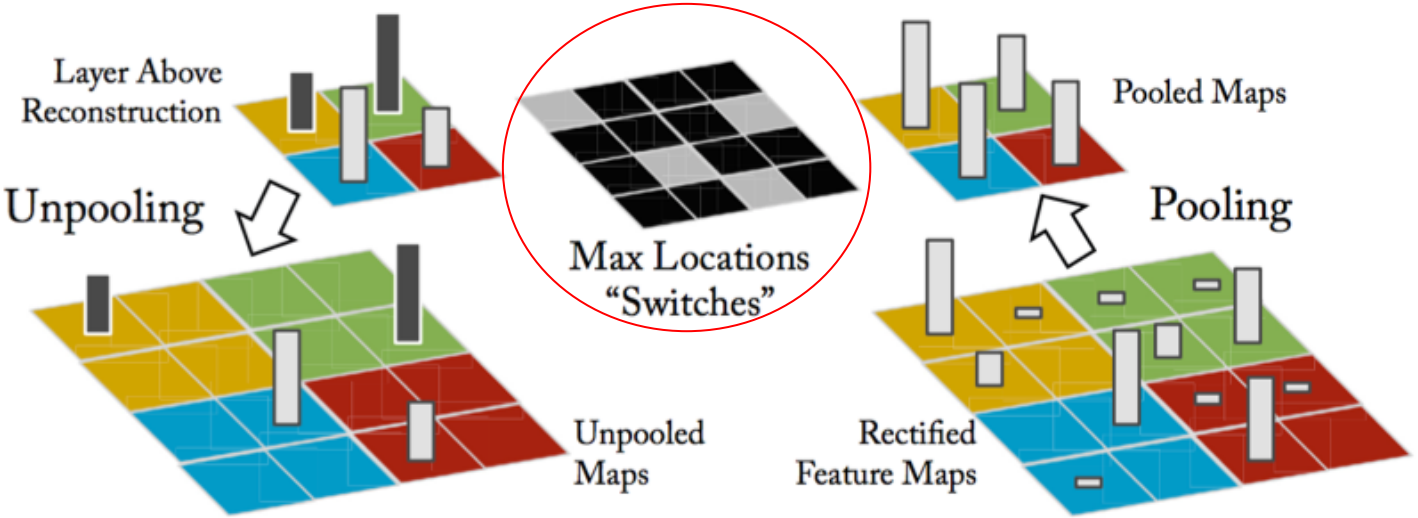
Disadvantages:

- heavily **data-dependent**
- perform **badly** for **small size objects**

Summary

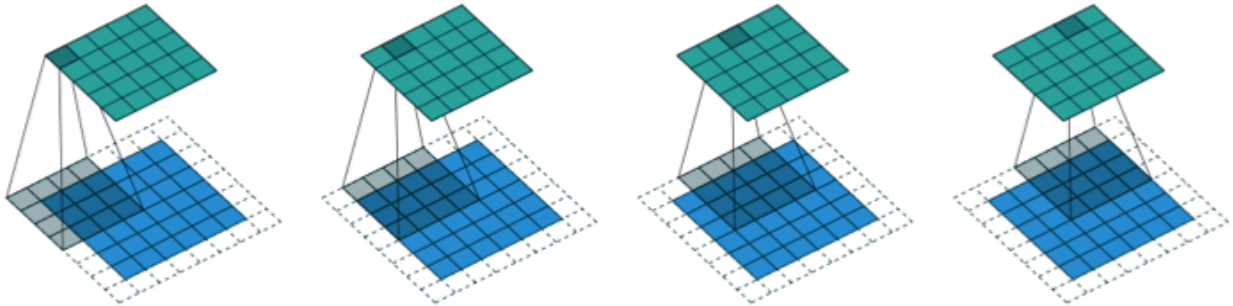
1. Network Dissection works well on **identifying object and part detectors**
2. its results are **consistent to** results of **DeConvNet**
3. some groups of classes **cluster** due to **shared part detectors**
4. the semantic **meaning** of units turns **more specific as** the **depth increased**
5. the **filtering effect** of units is **enhanced as** the **depth increased**
6. study the **roles of units** for some **classification errors**

Unpooling:



(source: Visualizing and Understanding Convolutional Networks)

Deconvolution:



(source: A guide to convolution arithmetic for deep learning)

$F \cdot I = O$ **transposed convolution matrix** $F^T \cdot O = I$

ConvNet **DeConvNet**