Lecture 8: Visualizing and Understanding

Ali Farhadi, Aditya Kusupati

Lecture 8 - 1 October 31, 2023

Administrative

- Assignment 1 grades will be released this week
- Quiz 1 should follow around the same time
 - Makeup quiz during Matt's office hours
- Project proposals will be graded this week and TA will be assigned
- -
- Assignment 2 due Friday Nov 10th, 11:59pm
 - Start now!
 - Assignment 3 has extra credit opportunities

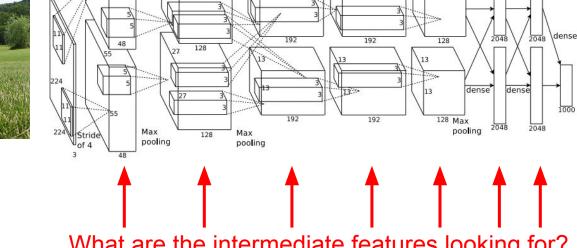
Ali Farhadi, Aditya Kusupati

Lecture 8 - 2 October 31, 2023

Today: What's going on inside ConvNets?

This image is CC0 public domain





Class Scores: 1000 numbers

October 31, 2023

Input Image: 3 x 224 x 224

What are the intermediate features looking for?

Lecture 8 -

3

Krizhevsky et al, "ImageNet Classification with Deep Convolutional Neural Networks", NIPS 2012. Figure reproduced with permission.

Today's agenda

Visualizing what models have learned:

- Visualizing filters
- Visualizing final layer features
- Visualizing activations

Understanding input pixels

- Identifying important pixels
- Saliency via backprop
- Guided backprop to generate images
- Gradient ascent to visualize features

Adversarial perturbations

Ali Farhadi, Aditya Kusupati

Lecture 8 - 4 October 31, 2023

Today's agenda

Visualizing what models have learned:

- Visualizing filters
- Visualizing final layer features
- Visualizing activations

Understanding input pixels

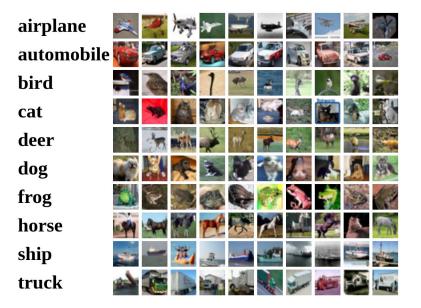
- Identifying important pixels
- Saliency via backprop
- Guided backprop to generate images
- Gradient ascent to visualize features

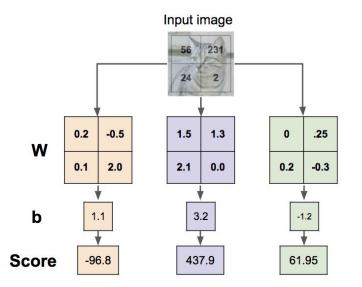
Adversarial perturbations

Ali Farhadi, Aditya Kusupati

Lecture 8 - 5 October 31, 2023

Interpreting a Linear Classifier: Visual Viewpoint



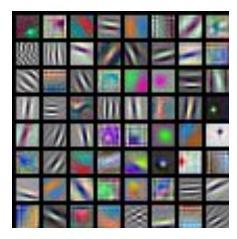




Ali Farhadi, Aditya Kusupati

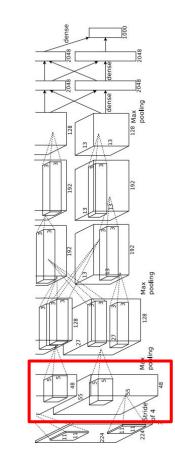
Lecture 8 - 6 October 31, 2023

First Layer: Visualize Filters



AlexNet: 64 x 3 x 11 x 11

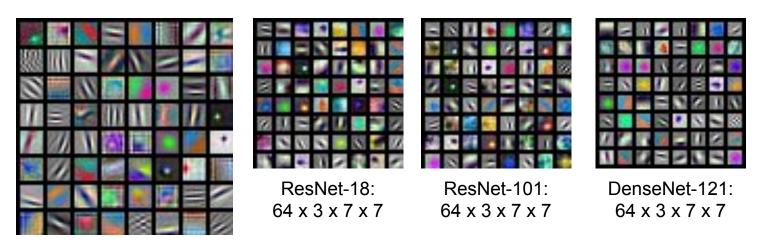
Krizhevsky, "One weird trick for parallelizing convolutional neural networks", arXiv 2014 He et al, "Deep Residual Learning for Image Recognition", CVPR 2016 Huang et al, "Densely Connected Convolutional Networks", CVPR 2017



Ali Farhadi, Aditya Kusupati

Lecture 8 - 7 October 31, 2023

First Layer: Visualize Filters



Max Max pooling Ma

AlexNet: 64 x 3 x 11 x 11

Krizhevsky, "One weird trick for parallelizing convolutional neural networks", arXiv 2014 He et al, "Deep Residual Learning for Image Recognition", CVPR 2016 Huang et al, "Densely Connected Convolutional Networks", CVPR 2017

Ali Farhadi, Aditya Kusupati

Lecture 8 - 8 October <u>31, 2023</u>

Visualize the filters/kernels (raw weights)

We can visualize filters at higher layers, but not that interesting

(these are taken from ConvNetJS CIFAR-10 demo) Weights: 医脊髓膜的 经投资 的复数 医鼻子

Weights:

2011日代1955年間)

layer 1 weights

16 x 3 x 7 x 7 (我们是我们的我们的我们的你的。(你们是你们的你们的你们的你?"我们的你? 如果我想出来的情况。(我们会会会会会会会会会会会会会会)(我们是我们的情况。 · 模試)(並当由目書語並將此約定系書書方的)(物和指面與問時展開目標書種建成種)(書包書記書 而發行的內部保護保護部(保護局局等部所與保護部合管理的)(開設市務運動局部的運動

> layer 2 weights 20 x 16 x 7 x 7

Weights:)(國際軍業局部電気局通知型機械成長務委員會)(局理法院和部長局制務を受消費業の目的な 的)(國產商總理要的認識的原語是認識的的筆意)(與產豐原外國發展的發展的最適是關鍵 · 新聞)(原情察語編集論學的智慧內提的意思和思考)(原本語語語法語書物語語語語語》)(原情 layer 3 weights 主奉祭)(總基總總法法律問題因表記律問題問題》(出於法律法律認識與通知問任論心 医乳肉炎 (再能到我在想能是我好好你会会自己的过去分词) (我没想得到多么没有的爱好。) 20 x 20 x 7 x 7 新新新闻曲台)(李浩章高温等等是否有美国南美国李法高发达)(和美国法国新的科学会社议》 非投行动型法的)(新商業市政技業支援新商業成長要定定局限的)(非常以保以支援支援支援 ※第四半三型にの)(通常会社にお法院を通知をつかまたの所有の)(ごやがほかをお知られる)

調査整備)(開始整約編集整整編装整備の開設)(は非計算は印刷書用計算用的計算)(構成構

医多足骨瘤 医多足骨炎 (口名名法多姓氏美国王的姓氏法名 (加口的名词复数

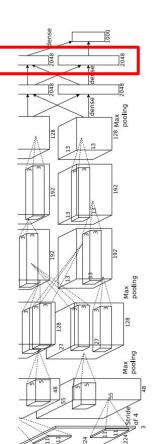
那些這個這些這個是那些的心)(這個這個的關係的這個個個的心心)(美力的水準要將這來非 金剛與內里語)(自然自然的思想是要保護和認識的說)(在亞洲的名言語的思想是可以認識的)(無

Ali Farhadi, Aditya Kusupati

October 31, 2023 Lecture 8 -9

Last Layer

FC7 layer



4096-dimensional feature vector for an image (layer immediately before the classifier)

Run the network on many images, collect the feature vectors

Ali Farhadi, Aditya Kusupati

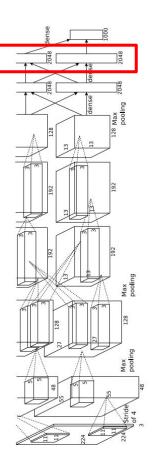
Lecture 8 - 10 October 31, 2023

Last Layer: Nearest Neighbors

4096-dim vector

Test image L2 Nearest neighbors in feature space





Krizhevsky et al, "ImageNet Classification with Deep Convolutional Neural Networks", NIPS 2012. Figures reproduced with permission.

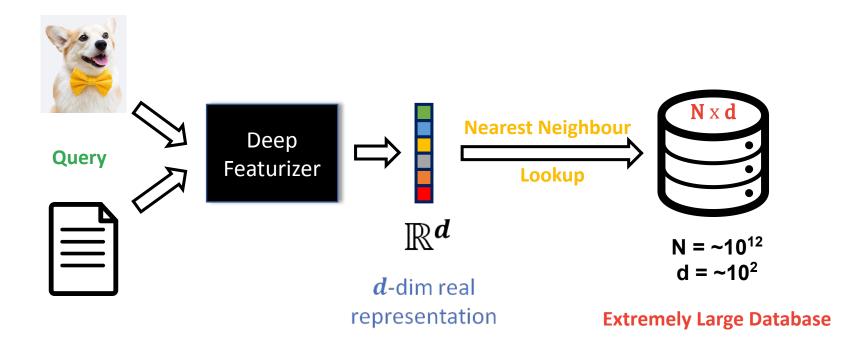
Ali Farhadi, Aditya Kusupati

Recall: Nearest neighbors

in <u>pixel</u> space

Lecture 8 - 11 October 31, 2023

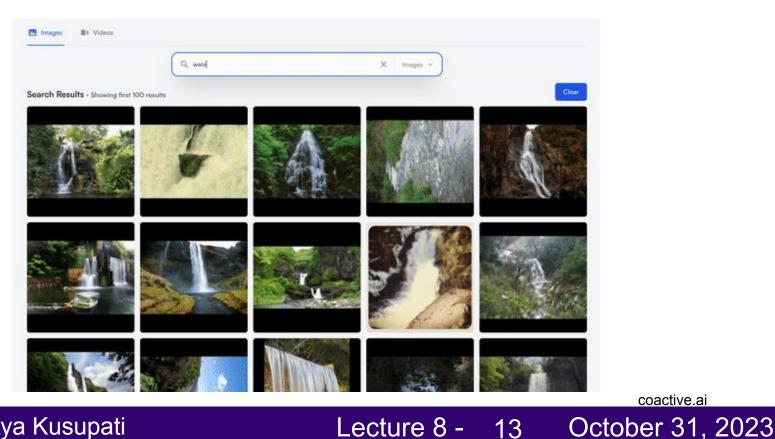
Last Layer: Learned Metric for "Semantic" Search



Ali Farhadi, Aditya Kusupati

Lecture 8 - 12 October 31, 2023

Last Layer: Modern Day Search



coactive.ai

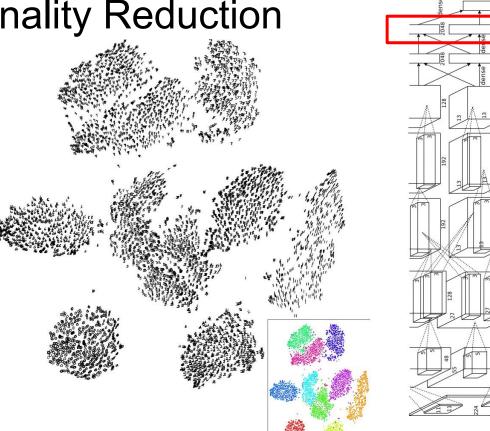
13

Last Layer: Dimensionality Reduction

Visualize the "space" of FC7 feature vectors by reducing dimensionality of vectors from 4096 to 2 dimensions

Simple algorithm: Principal Component Analysis (PCA)

More complex: t-SNE



Van der Maaten and Hinton, "Visualizing Data using t-SNE", JMLR 2008 Figure copyright Laurens van der Maaten and Geoff Hinton, 2008. Reproduced with permission.

Ali Farhadi, Aditya Kusupati

Lecture 8 - 14 October 31, 2023

Vax

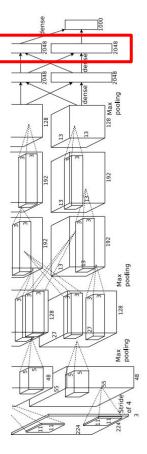
Last Layer: Dimensionality Reduction



Van der Maaten and Hinton, "Visualizing Data using t-SNE", JMLR 2008 Krizhevsky et al, "ImageNet Classification with Deep Convolutional Neural Networks", NIPS 2012. Figure reproduced with permission.





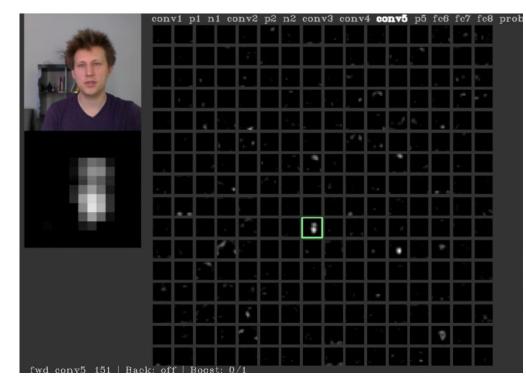


Ali Farhadi, Aditya Kusupati

Lecture 8 - 15 October 31, 2023

Visualizing Activations

conv5 feature map is 128x13x13; visualize as 128 13x13 grayscale images



Lecture 8 -

16

October 31, 2023

Yosinski et al, "Understanding Neural Networks Through Deep Visualization", ICML DL Workshop 2014. Figure copyright Jason Yosinski, 2014. Reproduced with permission.

Today's agenda

Visualizing what models have learned:

- Visualizing filters
- Visualizing final layer features
- Visualizing activations

Understanding input pixels

- Identifying important pixels
- Saliency via backprop
- Guided backprop to generate images
- Gradient ascent to visualize features

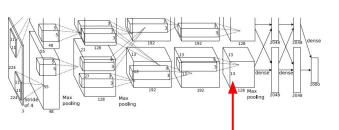
Adversarial perturbations

Ali Farhadi, Aditya Kusupati

Lecture 8 - 17 October 31, 2023

Maximally Activating Patches

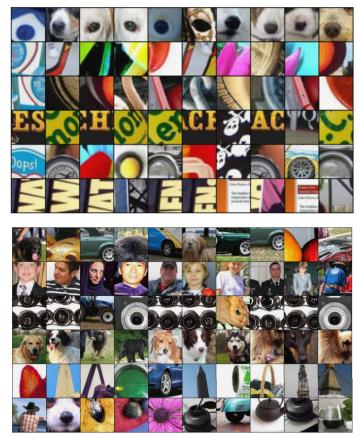




Pick a layer and a channel; e.g. conv5 is 128 x 13 x 13, pick channel 17/128

Run many images through the network, record values of chosen channel

Visualize image patches that correspond to maximal activations



Springenberg et al, "Striving for Simplicity: The All Convolutional Net", ICLR Workshop 2015 Figure copyright Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, Martin Riedmiller, 2015; reproduced with permission.

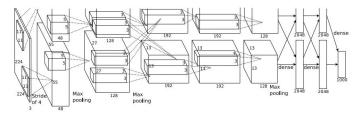
Ali Farhadi, Aditya Kusupati

Lecture 8 - 18 October 31, 2023

Which pixels matter: Saliency via Occlusion

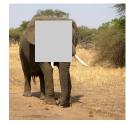
Mask part of the image before feeding to CNN, check how much predicted probabilities change

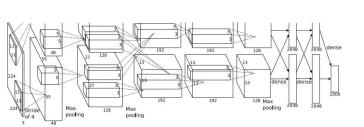




P(elephant) = 0.95

P(elephant) = 0.75





Zeiler and Fergus, "Visualizing and Understanding Convolutional Networks", ECCV 2014

Boat image is CC0 public domain Elephant image is CC0 public domain Go-Karts image is CC0 public domain

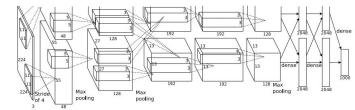
Ali Farhadi, Aditya Kusupati

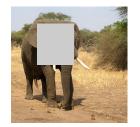
Lecture 8 - 19 October 31, 2023

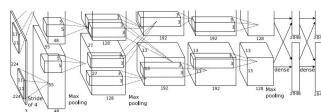
Which pixels matter: Saliency via Occlusion

Mask part of the image before feeding to CNN, check how much predicted probabilities change



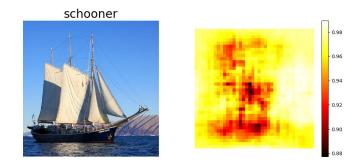




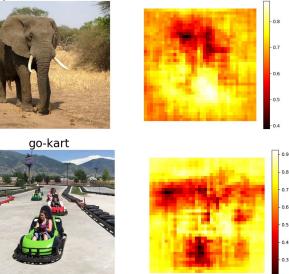


Zeiler and Fergus, "Visualizing and Understanding Convolutional Networks", ECCV 2014

Boat image is <u>CC0 public domain</u> Elephant image is <u>CC0 public domain</u> <u>Go-Karts image</u> is <u>CC0 public domain</u>



African elephant, Loxodonta africana



Ali Farhadi, Aditya Kusupati

Lecture 8 - 20 October 31, 2023

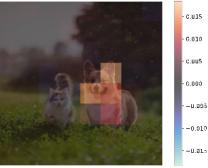
Saliency via Occlusion: Shapley Values







$$P(corgi) = 0.8$$



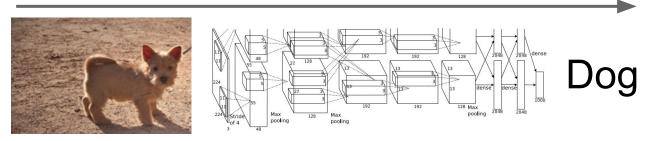
Credit: Ian Covert; Lundberg & Lee 2017

Ali Farhadi, Aditya Kusupati

Lecture 8 - 21 October 31, 2023

Which pixels matter: Saliency via Backprop

Forward pass: Compute probabilities



Lecture 8 -

22

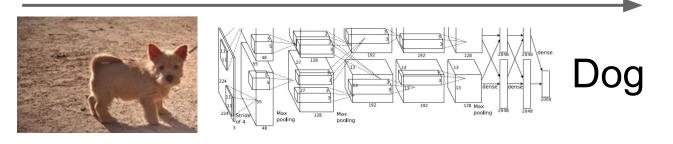
October 31, 2023

Simonyan, Vedaldi, and Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps", ICLR Workshop 2014.

Figures copyright Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, 2014; reproduced with permission.

Which pixels matter: Saliency via Backprop

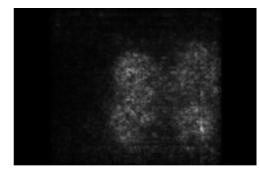
Forward pass: Compute probabilities



Compute gradient of (unnormalized) class score with respect to image pixels, take absolute value and max over RGB channels

Simonyan, Vedaldi, and Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps", ICLR Workshop 2014.

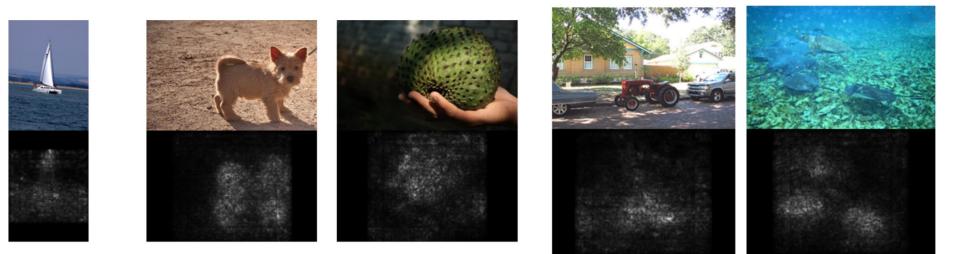
Figures copyright Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, 2014; reproduced with permission.



Ali Farhadi, Aditya Kusupati

Lecture 8 - 23 October 31, 2023

Saliency Maps



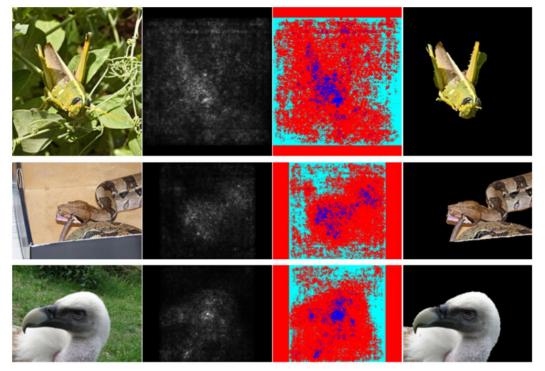
Lecture 8 - 24

October 31, 2023

Simonyan, Vedaldi, and Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps", ICLR Workshop 2014.

Figures copyright Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, 2014; reproduced with permission.

Saliency Maps: Segmentation without supervision



Simonyan, Vedaldi, and Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps", ICLR Workshop 2014.

Figures copyright Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, 2014; reproduced with permission. Rother et al, "Grabcut: Interactive foreground extraction using iterated graph cuts", ACM TOG 2004

Ali Farhadi, Aditya Kusupati

Use GrabCut on

saliency map

Lecture 8 - 25 October 31, 2023

Saliency maps: Uncovers biases

Such methods also find biases

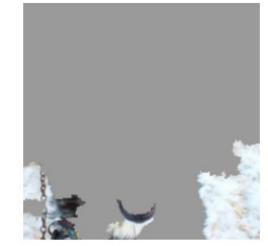
wolf vs dog classifier looks is actually a snow vs no-snow classifier



(a) Husky classified as wolf

Lecture 8 -

26

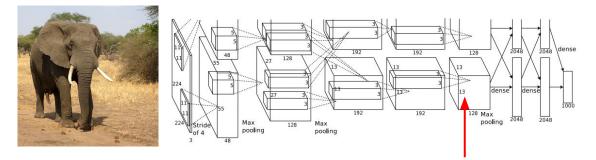


(b) Explanation

October 31, 2023

Figures copyright Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, 2016; reproduced with permission. Ribeiro et al, ""Why Should I Trust You?" Explaining the Predictions of Any Classifier", ACM KDD 2016

Intermediate Features via (guided) backprop



Pick a single intermediate channel, e.g. one value in 128 x 13 x 13 conv5 feature map

Compute gradient of activation value with respect to image pixels

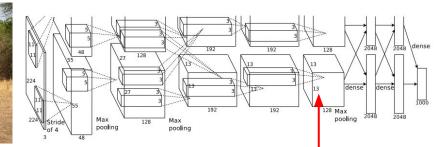
Zeiler and Fergus, "Visualizing and Understanding Convolutional Networks", ECCV 2014 Springenberg et al, "Striving for Simplicity: The All Convolutional Net", ICLR Workshop 2015

Ali Farhadi, Aditya Kusupati

Lecture 8 - 27 October 31, 2023

Intermediate Features via (guided) backprop





b) Forward pass

	-1	5		1	0	5
	-5	-7	\rightarrow	2	0	0
5	2	4		0	2	4

ReLU

Backward pass: backpropagation

	0	2	4
	-2	3	-1
←	6	-3	1
	2	-1	3

Pick a single intermediate neuron, e.g. one value in 128 x 13 x 13 conv5 feature map

Compute gradient of neuron value with respect to image pixels

Zeiler and Fergus, "Visualizing and Understanding Convolutional Networks", ECCV 2014 Springenberg et al, "Striving for Simplicity: The All Convolutional Net", ICLR Workshop 2015 Backward pass: guided backpropagation

0	0	0	Ť	-2	3	-1
6	0	0		6	-3	1
0	0	3		2	-1	3

Images come out nicer if you only backprop positive gradients through each ReLU (guided backprop)

Figure copyright Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, Martin Riedmiller, 2015; reproduced with permission.

Ali Farhadi, Aditya Kusupati

Lecture 8 - 28 October 31, 2023

6

Intermediate features via (guided) backprop





Guided Backprop

29

Lecture 8 -

October 31, 2023

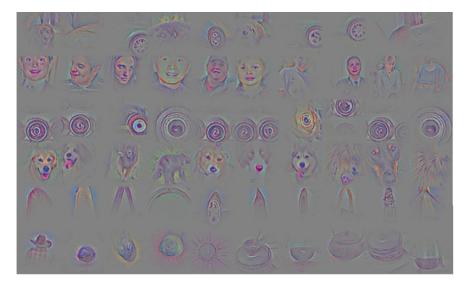
Maximally activating patches (Each row is a different neuron)

Zeiler and Fergus, "Visualizing and Understanding Convolutional Networks", ECCV 2014 Springenberg et al, "Striving for Simplicity: The All Convolutional Net", ICLR Workshop 2015 Figure copyright Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, Martin Riedmiller, 2015; reproduced with permission.

Intermediate features via (guided) backprop



Maximally activating patches (Each row is a different neuron)



Guided Backprop

30

Lecture 8 -

October 31, 2023

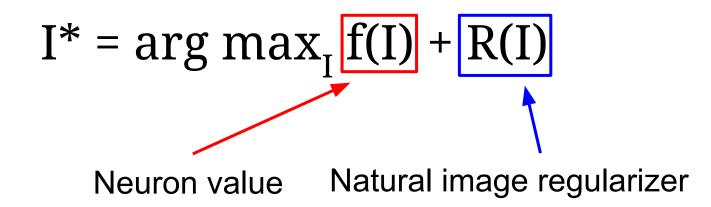
Zeiler and Fergus, "Visualizing and Understanding Convolutional Networks", ECCV 2014 Springenberg et al, "Striving for Simplicity: The All Convolutional Net", ICLR Workshop 2015 Figure copyright Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, Martin Riedmiller, 2015; reproduced with permission.

(Guided) backprop:

Find the part of an image that a neuron responds to

Gradient ascent:

Generate a synthetic image that maximally activates a neuron



Ali Farhadi, Aditya Kusupati

Lecture 8 - 31 October 31, 2023

1. Initialize image to zeros

$$\arg\max_{I} S_c(I) - \lambda \|I\|_2^2$$

score for class c (before Softmax)

October 31, 2023

zero image

Lecture 8 -

32

Repeat:

- 2. Forward image to compute current scores
- 3. Backprop to get gradient of neuron value with respect to image pixels
- 4. Make a small update to the image

Lecture 8 -

33

October 31, 2023

$$\arg\max_{I} S_c(I) - \lambda \|I\|_2^2$$

Simple regularizer: Penalize L2 norm of generated image

Simonyan, Vedaldi, and Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps", ICLR Workshop 2014. Figures copyright Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, 2014; reproduced with permission

$$\arg\max_{I} S_c(I) - \lambda \|I\|_2^2$$

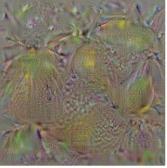
Simple regularizer: Penalize L2 norm of generated image



dumbbell

cup







bell pepper

lemon

husky

Simonyan, Vedaldi, and Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps", ICLR Workshop 2014.

Figures copyright Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, 2014; reproduced with permission.

Ali Farhadi, Aditya Kusupati

Lecture 8 - 34 October 31, 2023

$$\arg\max_{I} S_c(I) - \lambda \|I\|_2^2$$

Simple regularizer: Penalize L2 norm of generated image



Yosinski et al, "Understanding Neural Networks Through Deep Visualization", ICML DL Workshop 2014. Figure copyright Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson, 2014. Reproduced with permission. goose

ostrich

limousine

Ali Farhadi, Aditya Kusupati

Lecture 8 - 35 October 31, 2023

October 31, 2023

Lecture 8 -

36

$$\arg\max_{I} S_c(I) - \lambda \|I\|_2^2$$

Better regularizer: Penalize L2 norm of image; also during optimization periodically

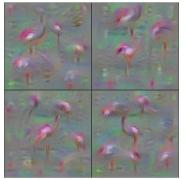
- (1) Gaussian blur image
- (2) Clip pixels with small values to 0
- (3) Clip pixels with small gradients to 0

Yosinski et al, "Understanding Neural Networks Through Deep Visualization", ICML DL Workshop 2014.

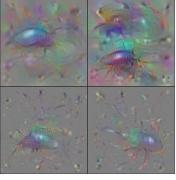
$$\arg\max_{I} S_c(I) - \lambda \|I\|_2^2$$

Better regularizer: Penalize L2 norm of image; also during optimization periodically

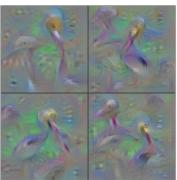
- (1) Gaussian blur image
- (2) Clip pixels with small values to 0
- (3) Clip pixels with small gradients to 0



Flamingo



Ground Beetle



Pelican



Indian Cobra

Yosinski et al, "Understanding Neural Networks Through Deep Visualization", ICML DL Workshop 2014. Figure copyright Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson, 2014. Reproduced with permission.

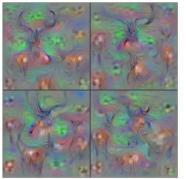
Ali Farhadi, Aditya Kusupati

Lecture 8 - 37 October 31, 2023

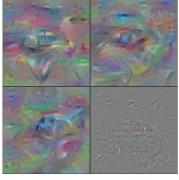
$$\arg\max_{I} S_c(I) - \lambda \|I\|_2^2$$

Better regularizer: Penalize L2 norm of image; also during optimization periodically

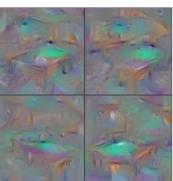
- (1) Gaussian blur image
- (2) Clip pixels with small values to 0
- (3) Clip pixels with small gradients to 0



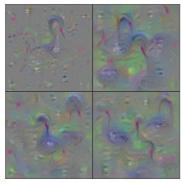
Hartebeest



Station Wagon



Billiard Table



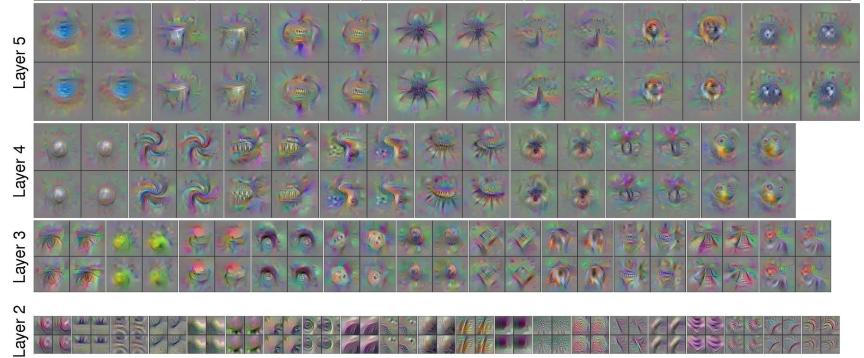
Black Swan

Yosinski et al, "Understanding Neural Networks Through Deep Visualization", ICML DL Workshop 2014. Figure copyright Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson, 2014. Reproduced with permission.

Ali Farhadi, Aditya Kusupati

Lecture 8 - 38 October 31, 2023

Use the same approach to visualize intermediate features



Lecture 8 -

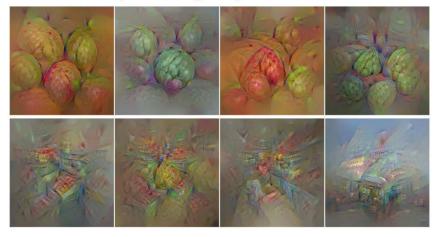
39

October 31, 2023

Yosinski et al, "Understanding Neural Networks Through Deep Visualization", ICML DL Workshop 2014. Figure copyright Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson, 2014. Reproduced with permission.

Adding "multi-faceted" visualization gives even nicer results: (Plus more careful regularization, center-bias)

Reconstructions of multiple feature types (facets) recognized by the same "grocery store" neuron



Corresponding example training set images recognized by the same neuron as in the "grocery store" class



40

Lecture 8 -

October 31, 2023

Nguyen et al, "Multifaceted Feature Visualization: Uncovering the Different Types of Features Learned By Each Neuron in Deep Neural Networks", ICML Visualization for Deep Learning Workshop 2016. Figures copyright Anh Nguyen, Jason Yosinski, and Jeff Clune, 2016; reproduced with permission.



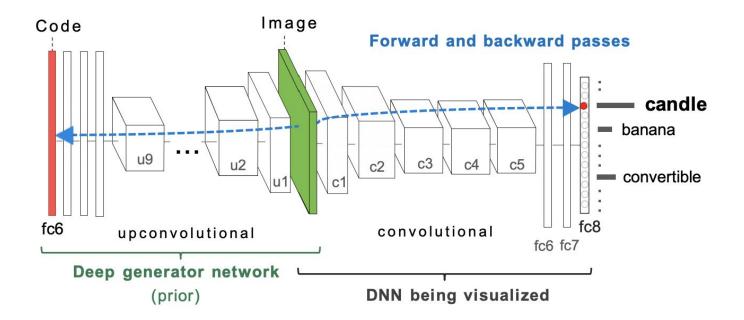
Lecture 8 - 41

Nguyen et al, "Multifaceted Feature Visualization: Uncovering the Different Types of Features Learned By Each Neuron in Deep Neural Networks", ICML Visualization for Deep Learning Workshop 2016. Figures copyright Anh Nguyen, Jason Yosinski, and Jeff Clune, 2016; reproduced with permission.

Ali Farhadi, Aditya Kusupati

October 31, 2023

Optimize in FC6 latent space instead of pixel space:



Lecture 8 -

42

October 31, 2023

Nguyen et al, "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks," NIPS 2016 Figure copyright Nguyen et al, 2016; reproduced with permission.

Optimize in FC6 latent space instead of pixel space:



Lecture 8 - 43

October 31, 2023

Nguyen et al, "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks," NIPS 2016 Figure copyright Nguyen et al, 2016; reproduced with permission.

Today's agenda

Visualizing what models have learned:

- Visualizing filters
- Visualizing final layer features
- Visualizing activations

Understanding input pixels

- Identifying important pixels
- Saliency via backprop
- Guided backprop to generate images
- Gradient ascent to visualize features

Adversarial perturbations

Ali Farhadi, Aditya Kusupati

Lecture 8 - 44 October 31, 2023

Fooling Images / Adversarial Examples

- (1) Start from an arbitrary image
- (2) Pick an arbitrary class
- (3) Modify the image to maximize the class

Lecture 8 -

45

October 31, 2023

(4) Repeat until network is fooled

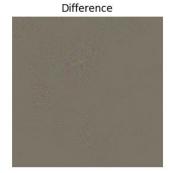
Fooling Images / Adversarial Examples

African elephant

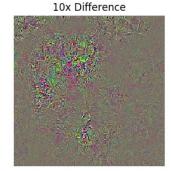




iPod



Difference



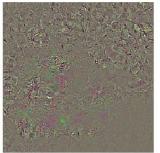
schooner







10x Difference



Boat image is CC0 public domain Elephant image is CC0 public domain

Ali Farhadi, Aditya Kusupati

Lecture 8 - 46 October 31, 2023

Fooling Images / Adversarial Examples

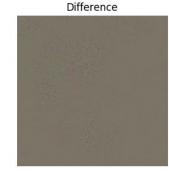
African elephant



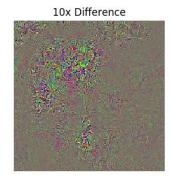
schooner



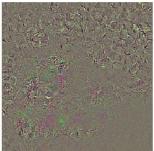
iPod



Difference



10x Difference









Boat image is CC0 public domain Elephant image is CC0 public domain

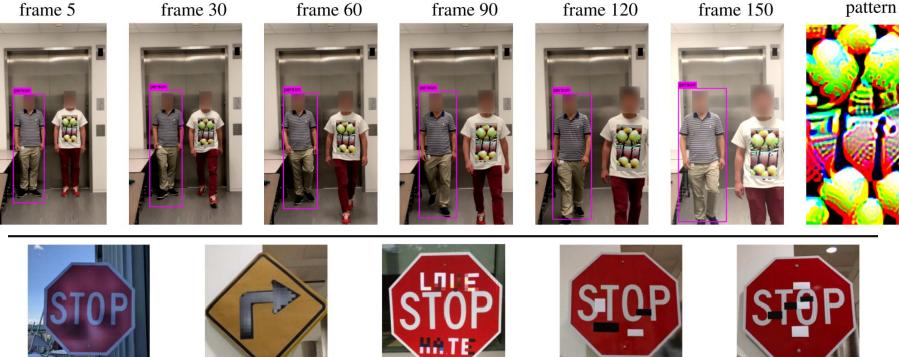
Check out lan Goodfellow's lecture from 2017

Ali Farhadi, Aditya Kusupati

Lecture 8 - 47 October 31, 2023

Fooling Person Detectors and Self-driving Cars





Xu et al., 2019; Eykholt et al., 2018

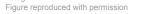
Ali Farhadi, Aditya Kusupati

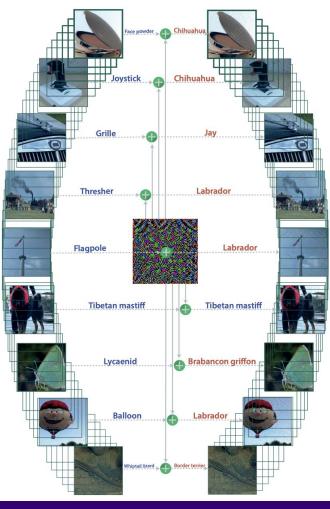
Lecture 8 -October 31, 2023 48

Fooling Images / Adversarial Exa

Universal perturbations

Moosavi-Dezfooli, Seyed-Mohsen, et al. "Universal adversarial perturbations." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.





Ali Farhadi, Aditya Kusupati

Lecture 8 - 49

October 31, 2023



Many methods for understanding CNN representations

Activations: Nearest neighbors, Dimensionality reduction, maximal patches, occlusion Gradients: Saliency maps, class visualization, fooling images, feature inversion

Lecture 8 -

_50

October 31, 2023

Adversarial Examples: To confuse the models

Next time: Self-supervised Learning

Ali Farhadi, Aditya Kusupati

Lecture 8 - 51 October 31, 2023