Lecture 11: Visualizing and Understanding
Last time: Lots of Computer Vision Tasks

Classification

Semantic Segmentation

Object Detection

Instance Segmentation

CAT

GRASS, CAT, TREE, SKY

DOG, DOG, CAT

DOG, DOG, CAT

No spatial extent

No objects, just pixels

Multiple Object

This image is CC0 public domain
Visualizing and Understanding

Source: Zoiner Tejada
Visualizing and Understanding

Source: Zoiner Tejada

DiT, Peebles & Xie, 2023
Visualizing and Understanding: Challenges

- Lots of parameters
- Complex transformations
- Non-intuitive latent feature spaces
- …
Computer Vision is everywhere!
Visualizing and Understanding: Challenges

Artificial neural networks

Human brain: ~100B neurons (?)
Image source: MIT News
Attention as a tool for understanding

Yarbus, Eye Movements and Vision, 1967
Attention as a tool for understanding

Fei-Fei Li, Ehsan Adeli, Ruohan Zhang, Zane Durante, Chen Wang
Attention of ANNs

Das et al., Computer Vision and Image Understanding, 2017
Attention of ANNs

Das et al., Computer Vision and Image Understanding, 2017
Today: What’s going on inside ConvNets?

What are the intermediate features looking for?

Class Scores: 1000 numbers

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

Style transfer
- Features inversion
- Deep dream
- Texture synthesis
- Neural style transfer
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

Style transfer
- Features inversion
- Deep dream
- Texture synthesis
- Neural style transfer
Interpreting a Linear Classifier: Visual Viewpoint

<table>
<thead>
<tr>
<th>airplane</th>
<th>automobile</th>
<th>bird</th>
<th>cat</th>
<th>deer</th>
<th>dog</th>
<th>frog</th>
<th>horse</th>
<th>ship</th>
<th>truck</th>
</tr>
</thead>
</table>

Input image

\[ W = \begin{bmatrix} 0.2 & -0.5 \\ 0.1 \quad 2.0 \end{bmatrix}, \quad b = \begin{bmatrix} 1.1 \\ 3.2 \end{bmatrix}, \quad \text{Score} = -96.8 \]

\[ W = \begin{bmatrix} 1.5 & 1.3 \\ 2.1 \quad 0.0 \end{bmatrix}, \quad b = \begin{bmatrix} 0 \quad 0.2 \end{bmatrix}, \quad \text{Score} = 437.9 \]

\[ W = \begin{bmatrix} 0 \quad 0.2 \end{bmatrix}, \quad b = \begin{bmatrix} -1.2 \end{bmatrix}, \quad \text{Score} = 61.95 \]
First Layer: Visualize Filters

AlexNet:
64 x 3 x 11 x 11

Huang et al, “Densely Connected Convolutional Networks”, CVPR 2017
First Layer: Visualize Filters

AlexNet:  
64 x 3 x 11 x 11

ResNet-18:  
64 x 3 x 7 x 7

ResNet-101:  
64 x 3 x 7 x 7

DenseNet-121:  
64 x 3 x 7 x 7

Huang et al, “Densely Connected Convolutional Networks”, CVPR 2017
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)
4096-dimensional feature vector for an image (layer immediately before the classifier)

Run the network on many images, collect the feature vectors
Last Layer: Nearest Neighbors

**Recall**: Nearest neighbors in pixel space

Figures reproduced with permission.
Last Layer: Nearest Neighbors

Recall: Nearest neighbors in pixel space

Figures reproduced with permission.
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.
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.

See high-resolution versions at http://cs.stanford.edu/people/karpathy/cnnembed/
Visualizing Activations

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

David Bau*, Bolei Zhou*, Aditya Khosla, Aude Oliva, Antonio Torralba


<table>
<thead>
<tr>
<th>House</th>
<th>Dog</th>
<th>Train</th>
<th>Plant</th>
<th>Airplane</th>
</tr>
</thead>
<tbody>
<tr>
<td>res5c unit 1410</td>
<td>res5c unit 1573</td>
<td>res5c unit 924</td>
<td>res5c unit 264</td>
<td>res5c unit 1243</td>
</tr>
<tr>
<td>IoU=0.142</td>
<td>IoU=0.216</td>
<td>IoU=0.293</td>
<td>IoU=0.126</td>
<td>IoU=0.172</td>
</tr>
<tr>
<td>res5c unit 301</td>
<td>res5c unit 1718</td>
<td>res5c unit 2001</td>
<td>res5c unit 766</td>
<td>res5c unit 1379</td>
</tr>
<tr>
<td>IoU=0.087</td>
<td>IoU=0.193</td>
<td>IoU=0.255</td>
<td>IoU=0.092</td>
<td>IoU=0.156</td>
</tr>
<tr>
<td>inception_4e unit 789</td>
<td>inception_4e unit 750</td>
<td>inception_5b unit 626</td>
<td>inception_4e unit 56</td>
<td>inception_4e unit 92</td>
</tr>
<tr>
<td>IoU=0.137</td>
<td>IoU=0.203</td>
<td>IoU=0.145</td>
<td>IoU=0.139</td>
<td>IoU=0.164</td>
</tr>
<tr>
<td>inception_4e unit 175</td>
<td>inception_5b unit 437</td>
<td>inception_5b unit 415</td>
<td>inception_4e unit 714</td>
<td>inception_4e unit 759</td>
</tr>
<tr>
<td>IoU=0.115</td>
<td>IoU=0.108</td>
<td>IoU=0.143</td>
<td>IoU=0.105</td>
<td>IoU=0.144</td>
</tr>
<tr>
<td>conv5_3 unit 243</td>
<td>conv5_3 unit 142</td>
<td>conv5_3 unit 463</td>
<td>conv5_3 unit 85</td>
<td>conv5_3 unit 151</td>
</tr>
<tr>
<td>IoU=0.070</td>
<td>IoU=0.205</td>
<td>IoU=0.126</td>
<td>IoU=0.086</td>
<td>IoU=0.150</td>
</tr>
<tr>
<td>conv5_3 unit 102</td>
<td>conv5_3 unit 491</td>
<td>conv5_3 unit 402</td>
<td>conv4_3 unit 336</td>
<td>conv5_3 unit 204</td>
</tr>
<tr>
<td>IoU=0.070</td>
<td>IoU=0.112</td>
<td>IoU=0.058</td>
<td>IoU=0.068</td>
<td>IoU=0.077</td>
</tr>
</tbody>
</table>
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

Style transfer
- Deep dream
- Features inversion
- Texture synthesis
- Neural style transfer
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

Figure copyright Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, Martin Riedmiller, 2015; reproduced with permission.
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
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
Which pixels matter: Saliency via Backprop

Forward pass: Compute probabilities

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

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

Saliency Maps: Segmentation without supervision

Use GrabCut on saliency map
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  (b) Explanation

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
Intermediate Features via (guided) backprop

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

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

Zeiler and Fergus, “Visualizing and Understanding Convolutional Networks”, ECCV 2014

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

Zeiler and Fergus, “Visualizing and Understanding Convolutional Networks”, ECCV 2014
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

Zeiler and Fergus, “Visualizing and Understanding Convolutional Networks”, ECCV 2014
Figure copyright Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, Martin Riedmiller, 2015; reproduced with permission.
Visualizing CNN features: Gradient Ascent

(Guided) backprop:
Find the part of an image that a neuron responds to

Gradient ascent:
Generate a synthetic image that maximally activates a neuron

\[ l^* = \arg \max_l [f(l) + R(l)] \]

Neuron value Natural image regularizer
Visualizing CNN features: Gradient Ascent

1. Initialize image to zeros

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
Visualizing CNN features: Gradient Ascent

\[ \arg \max_I S_c(I) - \lambda \| I \|_2^2 \]

Simple regularizer: Penalize L2 norm of generated image

Figures copyright Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, 2014; reproduced with permission.
Visualizing CNN features: Gradient Ascent

\[
\arg \max_I S_c(I) - \lambda \|I\|_2^2
\]

Simple regularizer: Penalize L2 norm of generated image

---

Visualizing CNN features: Gradient Ascent

$$\arg \max_I S_c(I) - \lambda \| I \|^2_2$$

Simple regularizer: Penalize L2 norm of generated image

Visualizing CNN features: Gradient Ascent

\[ \arg \max_I S_c(I) - \lambda \|I\|^2 \]

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

Visualizing CNN features: Gradient Ascent

$$\arg\max_I S_c(I) - \lambda \|I\|_2^2$$

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. Figure copyright Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson, 2014. Reproduced with permission.
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

Visualizing CNN features: Gradient Ascent

Use the same approach to visualize intermediate features

Visualizing CNN features: Gradient Ascent

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

Visualizing CNN features: Gradient Ascent


Figures copyright Anh Nguyen, Jason Yosinski, and Jeff Clune, 2016; reproduced with permission.
Visualizing CNN features: Gradient Ascent

Optimize in FC6 latent space instead of pixel space:

Figure copyright Nguyen et al, 2016; reproduced with permission.
Visualizing CNN features: Gradient Ascent

Optimize in FC6 latent space instead of pixel space:

Figure copyright Nguyen et al, 2016; reproduced with permission.
Visualizing ViT features

Chen et al., When Vision Transformers Outperform Resnets Without Pre-training Or Strong Data Augmentations, ICLR 2022; Paul and Chen, Vision Transformers are Robust Learners, AAAI 2022. Reproduced for educational purposes.
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

Style transfer
- Deep dream
- Features inversion
- Texture synthesis
- Neural style transfer
Fooling Images / Adversarial Examples

(1) Start from an arbitrary image
(2) Pick an arbitrary class
(3) Modify the image to maximize the class
(4) Repeat until network is fooled
Fooling Images / Adversarial Examples

African elephant  koala  Difference  10x Difference

schooner  iPod  Difference  10x Difference

Boat image is CC0 public domain
Elephant image is CC0 public domain
Fooling Images / Adversarial Examples

Universal perturbations

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

Style transfer
- Features inversion
- Deep dream
- Texture synthesis
- Neural style transfer
Feature Inversion

Given a CNN feature vector for an image, find a new image that:
- Matches the given feature vector
- “looks natural” (image prior regularization)

\[ x^* = \underset{x \in \mathbb{R}^{H \times W \times C}}{\text{argmin}} \ l(\Phi(x), \Phi_0) + \lambda \mathcal{R}(x) \]

\[ l(\Phi(x), \Phi_0) = \| \Phi(x) - \Phi_0 \|^2 \]

\[ \mathcal{R}_{V^\beta}(x) = \sum_{i,j} \left( (x_{i,j+1} - x_{i,j})^2 + (x_{i+1,j} - x_{ij})^2 \right)^{\beta/2} \]

Feature Inversion

Reconstructing from different layers of VGG-16

Figure from Johnson, Alahi, and Fei-Fei, “Perceptual Losses for Real-Time Style Transfer and Super-Resolution”, ECCV 2016. Copyright Springer, 2016.
Reproduced for educational purposes.
DeepDream: Amplify existing features

Rather than synthesizing an image to maximize a specific neuron, instead try to **amplify** the neuron activations at some layer in the network.

Choose an image and a layer in a CNN; repeat:

1. Forward: compute activations at chosen layer
2. Set gradient of chosen layer *equal to its activation*
3. Backward: Compute gradient on image
4. Update image

*Mordvintsev, Olah, and Tyka, “Inceptionism: Going Deeper into Neural Networks”, Google Research Blog. Images are licensed under CC-BY 4.0*
DeepDream: Amplify existing features

Rather than synthesizing an image to maximize a specific neuron, instead try to amplify the neuron activations at some layer in the network.

Choose an image and a layer in a CNN; repeat:
1. Forward: compute activations at chosen layer
2. Set gradient of chosen layer equal to its activation
3. Backward: Compute gradient on image
4. Update image

Equivalent to:
\[ I^* = \arg \max_i \sum_i f_i(I)^2 \]

Mordvintsev, Olah, and Tyka, “Inceptionism: Going Deeper into Neural Networks”, Google Research Blog. Images are licensed under CC-BY 4.0.
DeepDream: Amplify existing features

```python
def objective_L2(dst):
    dst.diff[:] = dst.data

def make_step(net, step_size=1.5, end='inception_4c/output',
             jitter=32, clip=True, objective=objective_L2):
    '''Basic gradient ascent step.'''

    src = net.blobs['data']  # input image is stored in Net's 'data' blob
dst = net.blobs[end]

    ox, oy = np.random.randint(-jitter, jitter+1, 2)
    src.data[0] = np.roll(np.roll(src.data[0], ox, -1), oy, -2)  # apply jitter shift

    net.forward(end=end)
    objective(dst)  # specify the optimization objective
    net.backward(start=end)
    g = src.diff[0]
    # apply normalized ascent step to the input image
    src.data[:] += step_size / np.abs(g).mean() * g

    ox, oy = np.random.randint(-jitter, jitter+1, 2)
    src.data[0] = np.roll(np.roll(src.data[0], -ox, -1), -oy, -2)  # unshift image

    if clip:
        bias = net.transformer.mean['data']
        src.data[:] = np.clip(src.data, -bias, 255-bias)
```

**Code** is very simple but it uses a couple tricks:

(Code is licensed under Apache 2.0)
DeepDream: Amplify existing features

Code is very simple but it uses a couple tricks:

(Code is licensed under Apache 2.0)

Jitter image

L1 Normalize gradients

```python
def objective_L2(dst):
    dst.diff[:] = dst.data

def make_step(net, step_size=1.5, end='inception_4c/output',
             jitter=32, clip=True, objective=objective_L2):
    '''Basic gradient ascent step.'''

    src = net.blobs['data']  # input image is stored in Net's 'data' blob
    dst = net.blobs[end]

    ox, oy = np.random.randint(-jitter, jitter+1, 2)
    src.data[0] = np.roll(np.roll(src.data[0], ox, -1), oy, -2)  # apply jitter shift

    net.forward(end=end)
    objective(dst)  # specify the optimization objective
    net.backward(start=end)
    g = src.diff[0]

    # apply normalized ascent step to the input image
    src.data[:] += step_size / np.abs(g).mean() * g

    src.data[0] = np.roll(np.roll(src.data[0], -ox, -1), -oy, -2)  # unshift image

    if clip:
        bias = net.transformer.mean['data']
        src.data[:] = np.clip(src.data, -bias, 255-bias)
```
DeepDream: Amplify existing features

```python
def objective_L2(dst):
    dst.diff[:] = dst.data

def make_step(net, step_size=1.5, end='inception_4c/output',
                jitter=32, clip=True, objective=objective_L2):
    """Basic gradient ascent step."

    src = net.blobs['data']  # input image is stored in Net's 'data' blob
    dst = net.blobs[end]

    ox, oy = np.random.randint(-jitter, jitter+1, 2)
    src.data[0] = np.roll(np.roll(src.data[0], ox, -1), oy, -2)  # apply jitter shift

    net.forward(end=end)
    objective(dst)  # specify the optimization objective
    net.backward(start=end)
    g = src.diff[0]

    # apply normalized ascent step to the input image
    src.data[:] += step_size / np.abs(g).mean() * g

    src.data[0] = np.roll(np.roll(src.data[0], -ox, -1), -oy, -2)  # unshift image

    if clip:
        bias = net.transformer.mean['data']
        src.data[:] = np.clip(src.data, -bias, 255-bias)
```

Code is very simple but it uses a couple tricks:

- Jitter image
- L1 Normalize gradients
- Clip pixel values

Also uses multiscale processing for a fractal effect (not shown)
"Admiral Dog!"
"The Pig-Snail"
"The Camel-Bird"
"The Dog-Fish"
Texture Synthesis

Given a sample patch of some texture, can we generate a bigger image of the same texture?

Input → Output

Output image is licensed under the MIT license
Texture Synthesis: Nearest Neighbor

Generate pixels one at a time in scanline order; form neighborhood of already generated pixels and copy nearest neighbor from input.

Texture Synthesis: Nearest Neighbor

Images licensed under the MIT license
Neural Texture Synthesis: Gram Matrix

Each layer of CNN gives $C \times H \times W$ tensor of features; $H \times W$ grid of $C$-dimensional vectors
Neural Texture Synthesis: Gram Matrix

Each layer of CNN gives $C \times H \times W$ tensor of features; $H \times W$ grid of $C$-dimensional vectors

Outer product of two $C$-dimensional vectors gives $C \times C$ matrix measuring co-occurrence
Neural Texture Synthesis: Gram Matrix

Each layer of CNN gives $C \times H \times W$ tensor of features; $H \times W$ grid of $C$-dimensional vectors

Outer product of two $C$-dimensional vectors gives $C \times C$ matrix measuring co-occurrence

Average over all pairs of vectors, giving Gram matrix of shape $C \times C$
Neural Texture Synthesis: Gram Matrix

Each layer of CNN gives $C \times H \times W$ tensor of features; $H \times W$ grid of $C$-dimensional vectors

Outer product of two $C$-dimensional vectors gives $C \times C$ matrix measuring co-occurrence

Average over all $HW$ pairs of vectors, giving **Gram matrix** of shape $C \times C$

Efficient to compute; reshape features from $C \times H \times W$ to $=C \times HW$

then compute $G = FF^T$
Neural Texture Synthesis

1. Pretrain a CNN on ImageNet (VGG-19)
2. Run input texture forward through CNN, record activations on every layer; layer $i$ gives feature map of shape $C_i \times H_i \times W_i$
3. At each layer compute the Gram matrix giving outer product of features:

$$G_{ij}^l = \sum_k F_{ik}^l F_{jk}^l$$ (shape $C_i \times C_i$)

Figure copyright Leon Gatys, Alexander S. Ecker, and Matthias Bethge, 2015. Reproduced with permission.
Neural Texture Synthesis

1. Pretrain a CNN on ImageNet (VGG-19)
2. Run input texture forward through CNN, record activations on every layer; layer $i$ gives feature map of shape $C_i \times H_i \times W_i$
3. At each layer compute the Gram matrix giving outer product of features:

$$G_{ij}^l = \sum_k F_{ik}^l F_{jk}^l \text{ (shape } C_i \times C_i)$$

4. Initialize generated image from random noise
5. Pass generated image through CNN, compute Gram matrix on each layer
6. Compute loss: weighted sum of L2 distance between Gram matrices
7. Backprop to get gradient on image
8. Make gradient step on image
9. GOTO 5

Gatys, Ecker, and Bethge, “Texture Synthesis Using Convolutional Neural Networks”, NIPS 2015
Figure copyright Leon Gatys, Alexander S. Ecker, and Matthias Bethge, 2015. Reproduced with permission.
Neural Texture Synthesis

1. Pretrain a CNN on ImageNet (VGG-19)
2. Run input texture forward through CNN, record activations on every layer; layer $i$ gives feature map of shape $C_i \times H_i \times W_i$
3. At each layer compute the Gram matrix giving outer product of features:

$$G^l_{ij} = \sum_k F^l_{ik} F^l_{jk} \text{ (shape $C_i \times C_i$)}$$

4. Initialize generated image from random noise
5. Pass generated image through CNN, compute Gram matrix on each layer
6. Compute loss: weighted sum of L2 distance between Gram matrices

$$L(\tilde{x}, \hat{x}) = \sum_{l=0}^{L} w_l E_l$$

$$E_l = \frac{1}{4N_l^2 M_l^2} \sum_{i,j} \left( G^l_{ij} - \hat{G}^l_{ij} \right)^2$$

---

Gatys, Ecker, and Bethge, “Texture Synthesis Using Convolutional Neural Networks”, NIPS 2015
Figure copyright Leon Gatys, Alexander S. Ecker, and Matthias Bethge, 2015. Reproduced with permission.
Neural Texture Synthesis

1. Pretrain a CNN on ImageNet (VGG-19)
2. Run input texture forward through CNN, record activations on every layer; layer $i$ gives feature map of shape $C_i \times H_i \times W_i$
3. At each layer compute the Gram matrix giving outer product of features:

$$G_{ij}^{l} = \sum_{k} F_{ik}^{l} F_{jk}^{l} \quad \text{(shape } C_i \times C_i)$$

4. Initialize generated image from random noise
5. Pass generated image through CNN, compute Gram matrix on each layer
6. Compute loss: weighted sum of L2 distance between Gram matrices
7. Backprop to get gradient on image
8. Make gradient step on image
9. GOTO 5

Gatys, Ecker, and Bethge, “Texture Synthesis Using Convolutional Neural Networks”, NIPS 2015
Figure copyright Leon Gatys, Alexander S. Ecker, and Matthias Bethge, 2015. Reproduced with permission.
Neural Texture Synthesis

Reconstructing texture from higher layers recovers larger features from the input texture.
Neural Texture Synthesis: Texture = Artwork

Texture synthesis (Gram reconstruction)

Figure from Johnson, Alahi, and Fei-Fei, "Perceptual Losses for Real-Time Style Transfer and Super-Resolution", ECCV 2016. Copyright Springer, 2016. Reproduced for educational purposes.
Neural Style Transfer: Feature + Gram Reconstruction

Texture synthesis (Gram reconstruction)

Feature reconstruction

Figure from Johnson, Alahi, and Fei-Fei, “Perceptual Losses for Real-Time Style Transfer and Super-Resolution”, ECCV 2016. Copyright Springer, 2016. Reproduced for educational purposes.
Neural Style Transfer

Content Image

Style Image

Starry Night by Van Gogh is in the public domain

Gatys, Ecker, and Bethge, “Texture Synthesis Using Convolutional Neural Networks”, NIPS 2015
Neural Style Transfer

Content Image

Style Image

Style Transfer!

This image is licensed under CC-BY 3.0


Starry Night by Van Gogh is in the public domain

This image copyright Justin Johnson, 2015. Reproduced with permission.
Gatys, Ecker, and Bethge, "Image style transfer using convolutional neural networks", CVPR 2016
Figure adapted from Johnson, Alahi, and Fei-Fei, "Perceptual Losses for Real-Time Style Transfer and Super-Resolution", ECCV 2016. Copyright Springer, 2016. Reproduced for educational purposes.
Gatys, Ecker, and Bethge, "Image style transfer using convolutional neural networks", CVPR 2016
Figure adapted from Johnson, Alahi, and Fei-Fei, "Perceptual Losses for Real-Time Style Transfer and Super-Resolution", ECCV 2016. Copyright Springer, 2016. Reproduced for educational purposes.
Neural Style Transfer

Example outputs from Lua torch implementation

Gatys, Ecker, and Bethge, "Image style transfer using convolutional neural networks", CVPR 2016
Figure copyright Justin Johnson, 2015.
Neural Style Transfer

More weight to content loss

More weight to style loss
Neural Style Transfer

Resizing style image before running style transfer algorithm can transfer different types of features

Gatys, Ecker, and Bethge, "Image style transfer using convolutional neural networks", CVPR 2016
Figure copyright Justin Johnson, 2015.
Neural Style Transfer: Multiple Style Images

Mix style from multiple images by taking a weighted average of Gram matrices

Gatys, Ecker, and Bethge, "Image style transfer using convolutional neural networks", CVPR 2016
Figure copyright Justin Johnson, 2015.
Neural Style Transfer

**Problem:** Style transfer requires many forward / backward passes through VGG; very slow!
Neural Style Transfer

**Problem:** Style transfer requires many forward / backward passes through VGG; very slow!

**Solution:** Train another neural network to perform style transfer for us!
Fast Style Transfer

1. Train a feedforward network for each style
2. Use pretrained CNN to compute same losses as before
3. After training, stylize images using a single forward pass


Figure copyright Springer, 2016. Reproduced for educational purposes.
Fast Style Transfer

Figure copyright Springer, 2016. Reproduced for educational purposes.

https://github.com/jcjohnson/fast-neural-style
Remember Normalization Methods?

- **Batch Norm**: Normalizes over the entire batch.
- **Layer Norm**: Normalizes over the entire layer, but not the entire batch.
- **Instance Norm**: Normalizes over each instance, but not the entire batch or layer.
- **Group Norm**: Normalizes over a group of instances, not the entire batch or layer.
Remember Normalization Methods?

Instance Normalization was developed for style transfer!
Fast Style Transfer

Replacing batch normalization with Instance Normalization improves results

One Network, Many Styles

Figure copyright Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur, 2016; reproduced with permission.
One Network, Many Styles

Use the same network for multiple styles using \textit{conditional instance normalization}: learn separate scale and shift parameters per style.

\begin{equation}
    x_{\text{norm}} = (x - \mu) / \sigma \quad z = \gamma_s x_{\text{norm}} + \beta_s
\end{equation}

Single network can blend styles after training.

Figure copyright Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur, 2016; reproduced with permission.
Summary

Many methods for understanding CNN representations

Activations: Nearest neighbors, dimensionality reduction, maximal patches, occlusion

Gradients: Saliency maps, class visualization, fooling images, feature inversion

Fun: DeepDream, style transfer
Next time:

5/9 Midterm

5/14 Self-supervised Learning