Towards *The Deep Model*: Understanding Visual Recognition Through Computational Models

Panqu Wang
Dissertation Defense
03/23/2017
Summary

Human Visual Recognition (face, object, scene)

Simulate
Explain
Predict

Model
Outline

• **Introduction**

• **Cognitive Modeling: Explaining cognitive phenomena using computational models**
  - Modeling the relationship between face and object recognition
  - Modeling hemispheric asymmetry of face processing
  - Modeling the contribution of central versus peripheral vision on scene processing

• **Engineering: Improving semantic segmentation**
  - Deep learning in pixel-level semantic segmentation
Introduction: Ventral Temporal Cortex (VTC)

Introduction: The Model (TM)

- Focused on modeling and explaining cognitive phenomena.
- Preprocessing with
  - Gabor Filters: V1 complex cells
  - PCA: feature compression and dimensionality reduction
Neural network learns tasks-specific features based on the output. The hidden layer can correspond to:

- Fusiform Face Area (FFA) for face recognition
- Lateral Occipital Complex (LOC) for general object recognition.
Introduction: The Model (TM)

A variant of TM (mixture-of-expert structure)
Introduction: The Deep Model (TDM)

TDM: TM gets deeper
Outline

• Introduction

• Cognitive Modeling: Explaining cognitive phenomena using computational models
  ➢ Modeling the relationship between face and object recognition
  ➢ Modeling hemispheric asymmetry of face processing
  ➢ Modeling the contribution of central versus peripheral vision on scene processing

• Engineering: Improving semantic segmentation
  ➢ Deep learning in pixel-level semantic segmentation
Modeling the relationship between face and object recognition

The Question

Is face recognition *independent* from general object recognition?

**Kanwisher et al. (1997), Tsao et al. (2006):**
Faces are processed in a domain-specific module (FFA), which is not activated for non-face objects.

**Wilmer et al. (2010):**
Cambridge Face Memory Test (CFMT) score shared very little variance (6.7%) with a test of visual memory for abstract art.

**Gauthier et al. (2000), McGugin et al. (2014), Xu (2005):**
FFA is activated by non-face object categories of expertise, such as birds and cars.

**Gauthier et al. (2014):**
Experience ($E$) modulates the domain general ability ($v$) to discriminate face and non-face categories. When $E$ is high for both categories, shared variance is 0.59.

Support for independence

Support for dependence
The Question

Is face recognition *independent* from general object recognition?

**Gauthier et al. (2014):** Experience ($E$) modulates the domain general ability ($v$) that is shared between both face and non-face object recognition.

$$\text{Performance}_{\text{cat}} \propto v \cdot \text{Experience}_{\text{cat}}$$
They showed that, as predicted, as experience ($E$) grows, the shared variance between the face and object recognition performance *increased* monotonically.

Can our model (TM) account for this?
Modeling Procedure

\[ \text{Performance}_{\text{cat}} \propto v \cdot \text{Experience}_{\text{cat}} \]

- We trained different network for each subject, based on that subject's data.
- We obtained \( v \) from the performance on the face recognition score, as human subjects have max \( E \) for face recognition:

\[ \text{Performance}_{\text{face}} \propto v. \quad (\text{Experience}_{\text{face}} = 100\%) \]
Modeling Procedure

1. Train first

2. Train together
Modeling Result (Subordinate Level Categorization)

Object Accuracy (VET)

Face Accuracy (CFMT)

Object Accuracy

Low  | High

Experience with VET categories (O-EXP)
Does Experience Has to be in Subordinate Level?

- How about basic level categorization?

Levels of visual categorization for “chair”
Modeling Result (Basic Level Categorization)

- No obvious experience moderation effect.
- Because the task is easy, not much *experience* is needed to get good performance.
- The type of experience matters - fine level discrimination of homogeneous categories.
Analysis: Role of Experience

Network: Low $E$

- Faces: 87.5%
- Non-faces: 16.67%

Accuracy

Network: High $E$

- Faces: 83.33%
- Non-faces: 87.5%

Accuracy

- More experience separates the objects to a greater extent, and helps the network achieve better performance.
Analysis: Role of Experience

Entropy of hidden units as a function of training epochs.

(Less experience means less diversity, not less training)
Conclusions

• Our experimental results support Gauthier et al.’s hypothesis that face and object recognition are not independent, given sufficient experience.

• Our experiment predicts that the source of the experience moderation effect is not in regions of the brain that are sensitive only to category level, as opposed to regions that are associated with better performance in individuation for objects and faces, such as the FFA.

• The reason for this mediation is that the FFA embodies a transform that amplifies the differences between homogeneous objects.
Outline

• Introduction

• **Cognitive Modeling: Explaining cognitive phenomena using computational models**
  - Modeling the relationship between face and object recognition
  - **Modeling hemispheric asymmetry of face processing**
  - Modeling the contribution of central versus peripheral vision on scene processing

• **Engineering: Improving semantic segmentation**
  - Deep learning in pixel-level semantic segmentation
Modeling the development of hemispheric asymmetry of face processing

Introduction

• Face processing is right hemisphere (RH) biased.
  ➢ Neural activation to faces is stronger in RH. [Kanwisher et al., 1997]
  ➢ Lesion in right FFA results in prosopagnosia. [Bouvier & Engel, 2006]

• Can TM account for this phenomenon?
Modeling Procedure

Two constraints:

- RH develops its function earlier than LH and dominates in infancy. [Chiron et al., 1997]
- Visual acuity changes radically over time. [Peterzell et al., 1995] [Atkinson et al., 1997]
Modeling Procedure

Learning rate vs. iteration graph for Module 1 (RH) and Module 2 (LH).

- RH predominance
- Acuity change
Modeling Procedure

- Four object categories: faces, books, cups, and digits.
- For each category, perform subordinate level classification task on 10 different individuals, with basic level classification on other 3 categories.
- Record the weight of the gating layer for each model “hemisphere” after training converges.
Result

Face expert (control)

Face expert with different learning rates ("waves of plasticity")
Result

Module 1: RH
Module 2: LH

Cup expert (control)

Digit expert (control)

Digit expert (split learning rate)
Result: Gating Node Value vs. Training Time

(a) Face Expert

(b) Book Expert

(c) Cup Expert

(d) Number Expert
Discussion

- TM can account for the development of the RH lateralization of face processing.

- This lateralization rises naturally as a consequence of two realistic cognitive constraints: early development of the RH, and the change of acuity across time.

- We hypothesize that low spatial frequency (LSF) information is the key that contributes to this result. Results show that LSF generates higher accuracy than HSF in face processing.

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Full freq.</th>
<th>Low freq.</th>
<th>High freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>100.00%</td>
<td>100.00%</td>
<td>97.41%</td>
</tr>
<tr>
<td>Testing</td>
<td>98.00%</td>
<td>99.88%</td>
<td>74.90%</td>
</tr>
</tbody>
</table>
Outline

- Introduction

- Cognitive Modeling: Explaining cognitive phenomena using computational models
  - Modeling the relationship between face and object recognition
  - Modeling hemispheric asymmetry of face processing
  - Modeling the contribution of central versus peripheral vision on scene processing

- Engineering: Improving semantic segmentation
  - Deep learning in pixel-level semantic segmentation
Modeling the contribution of central versus peripheral vision on scene processing

Introduction: Central and Peripheral Vision

http://www.codeproject.com/Articles/33648/Polar-View-of-an-Image
Introduction: Log Polar Transformation

http://www.codeproject.com/Articles/33648/Polar-View-of-an-Image
Larson & Loschky (2009)

- Explored the contribution of central versus peripheral vision in scene recognition using the Window and Scotoma paradigm.

![Scene conditions with differing radii in degrees of visual angle](image)
Larson & Loschky (2009)
Modeling Larson & Loschky (2009)
Modeling Cortical Magnification

The dataset
Modeling Larson & Loschky (2009)

Modeling result of visual angle vs. classification accuracy
Analysis

• Why is peripheral vision better?

Hypothesis 1:
Peripheral vision contains better features.

• To test hypothesis 1, we train a network using central or peripheral information only, and test using full image.

• If peripheral vision is better, its corresponding Scotoma condition test accuracy at 5 degrees should be higher than central vision (Window condition).
Analysis

- The network (a) contains only 2.3 million parameters, only 4% of AlexNet.
- The network is trained from scratch to avoid any interference of pre-training.
- Result:

<table>
<thead>
<tr>
<th>Data Used</th>
<th>Central</th>
<th>Peripheral</th>
<th>Full</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.500</td>
<td>0.687</td>
<td>0.715</td>
</tr>
<tr>
<td>SE</td>
<td>0.002</td>
<td>0.006</td>
<td>0.023</td>
</tr>
</tbody>
</table>

Table: Classification accuracy at 5 degrees

- Supports Hypothesis 1: peripheral vision contains better features.
Analysis

• Why is peripheral vision better?

Hypothesis 2:
Peripheral vision creates better internal representations for scene recognition.

• To test hypothesis 2, we train TDM and visualize its internal representations.

• Better features will result in higher gating node values in TDM.
The Deep Model (TDM)

TDM: TM gets deeper
Analysis: Gating Node Value

Comparison of gating node value.
Analysis: Gating Node Value

Gating node value as a function of training iterations. When given a choice, the network uses the peripheral pathway.
Analysis: Hidden Unit Activation

Development of representations of layer \textit{fc6} units
Analysis: Hidden Unit Activation

- Peripheral advantage emerges naturally through network training (or human development).
- It generates a spreading transform that benefits scene recognition, and is stable throughout learning.
- If we force equal gating value throughout training, the central pathway is totally turned off.
Conclusions

- TDM was successfully applied to simulate and explain the data from Larson and Loschky (2009), and support their findings of the peripheral advantage in human scene recognition.

- We proposed and validated two hypotheses that explains the peripheral advantage: 1) peripheral vision contains better features; and 2) peripheral vision generates better internal representations.
Outline

• Introduction

• Cognitive Modeling: Explaining cognitive phenomena using computational models
  - Modeling the relationship between face and object recognition
  - Modeling hemispheric asymmetry of face processing
  - Modeling the contribution of central versus peripheral vision on scene processing

• Engineering: Improving semantic segmentation
  - Deep learning in pixel-level semantic segmentation
Deep learning in pixel-level semantic segmentation

Introduction: Semantic Segmentation

Method

We design better convolutional operations for Semantic Segmentation.

- **Decoding: Dense Upsampling Convolution (DUC)**
  - upsampling through *learning*.
  - achieves pixel-level decoding accuracy.

- **Encoding: Hybrid Dilated Convolution (HDC)**
  - motivated by increasingly large receptive field of the visual processing stream.
  - solving “gridding problem” caused by normal dilated convolution.
Dense Upsampling Convolution (DUC)

- **Previous methods:**
  1. Bilinear Upsampling
     - *not* learnable
     - losing *details*

  2. Deconvolution
     - inserting *extra* values (0s)
     - inserting *extra* layers
Dense Upsampling Convolution (DUC)

- Learning *upsampling* through convolution layer.
- No extra zero is inserted.
- Extremely easy to implement.
- Enables end-to-end training, compatible with all FCN structures.
Dense Upsampling Convolution (DUC)
Dense Upsampling Convolution (DUC)
Dense Upsampling Convolution (DUC)

- Performance on CityScapes Validation set, measured using mean Intersection over Union score:

\[
    mIoU = \frac{TP}{TP + FP + FN}
\]

<table>
<thead>
<tr>
<th>Model</th>
<th>Avg</th>
<th>road</th>
<th>sidewalk</th>
<th>building</th>
<th>wall</th>
<th>fence</th>
<th>pole</th>
<th>traffic light</th>
<th>traffic sign</th>
<th>vegetation</th>
<th>terrain</th>
<th>sky</th>
<th>person</th>
<th>rider</th>
<th>car</th>
<th>truck</th>
<th>bus</th>
<th>train</th>
<th>Motorcycle</th>
<th>bicycle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bilinear Upsample</td>
<td>0.723</td>
<td>0.974</td>
<td>0.802</td>
<td>0.906</td>
<td>0.475</td>
<td>0.519</td>
<td>0.523</td>
<td>0.593</td>
<td>0.7</td>
<td>0.907</td>
<td>0.591</td>
<td>0.923</td>
<td>0.78</td>
<td>0.581</td>
<td>0.933</td>
<td>0.693</td>
<td>0.829</td>
<td>0.66</td>
<td>0.615</td>
<td>0.74</td>
</tr>
<tr>
<td>DUC</td>
<td>0.743</td>
<td>0.981</td>
<td>0.847</td>
<td>0.918</td>
<td>0.509</td>
<td>0.531</td>
<td>0.62</td>
<td>0.653</td>
<td>0.75</td>
<td>0.92</td>
<td>0.626</td>
<td>0.943</td>
<td>0.8</td>
<td>0.585</td>
<td>0.944</td>
<td>0.712</td>
<td>0.822</td>
<td>0.586</td>
<td>0.615</td>
<td>0.75</td>
</tr>
</tbody>
</table>

- Benefit almost all categories.

- Especially helpful for small objects.
Hybrid Dilated Convolution (HDC)

- The “gridding effect” in normal dilated convolution: using the same expansion factor leads to “holes” in the information available.
Hybrid Dilated Convolution (HDC)

- The “gridding effect” in normal dilated convolution:

```
<table>
<thead>
<tr>
<th>Ground truth</th>
</tr>
</thead>
</table>
| ![Ground truth](image1)

<table>
<thead>
<tr>
<th>Regular dilated conv</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image2" alt="Regular dilated conv" /></td>
</tr>
</tbody>
</table>
```
Hybrid Dilated Convolution (HDC)

- Solution: change the dilation rate so that there is no common factor.

Regular dilated conv

\[ r = 2,2,2 \]

HDC

\[ r = 1,2,3 \]
Hybrid Dilated Convolution (HDC)
Hybrid Dilated Convolution (HDC)

- Another benefit of using HDC: enlarge receptive field in-place.
- The network can see more and aggregate global information.

Input | GT | DUC | DUC+HDC
--- | --- | --- | ---
![Image](image0.png) | ![Image](image1.png) | ![Image](image2.png) | ![Image](image3.png)
Experiments: CityScapes Dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>fine</strong></td>
<td></td>
</tr>
<tr>
<td>FCN 8s [22]</td>
<td>65.3%</td>
</tr>
<tr>
<td>Dilation10 [32]</td>
<td>67.1%</td>
</tr>
<tr>
<td>LRR-4x [12]</td>
<td>69.7%</td>
</tr>
<tr>
<td>DeepLabv2-CRF [4]</td>
<td>70.4%</td>
</tr>
<tr>
<td>Adelaide _context [19]</td>
<td>71.6%</td>
</tr>
<tr>
<td>ResNet-DUC-HDC (ours)</td>
<td>76.1%</td>
</tr>
<tr>
<td><strong>ResNet-DUC-HDC-Ensemble</strong></td>
<td>77.6%</td>
</tr>
<tr>
<td><strong>coarse</strong></td>
<td></td>
</tr>
<tr>
<td>LRR-4x [12]</td>
<td>71.8%</td>
</tr>
<tr>
<td>ResNet-DUC-HDC-Coarse</td>
<td>78.5%</td>
</tr>
<tr>
<td><strong>ResNet-DUC-HDC-Coarse (better network)</strong></td>
<td>80.1%</td>
</tr>
</tbody>
</table>

Table 4. Performance on Cityscapes test set.
Experiments: Kitti Road

Dataset
• Contains images of three various categories of road scenes, including 289 training images and 290 test images.

Performance:

<table>
<thead>
<tr>
<th></th>
<th>MaxF</th>
<th>AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>UMROAD</td>
<td>95.64%</td>
<td>93.50%</td>
</tr>
<tr>
<td>UMMROAD</td>
<td>97.62%</td>
<td>95.53%</td>
</tr>
<tr>
<td>UROAD</td>
<td>95.17%</td>
<td>92.73%</td>
</tr>
<tr>
<td>URBANROAD</td>
<td>96.41%</td>
<td>93.88%</td>
</tr>
</tbody>
</table>

Table 5. Performance on different road scenes in KITTI test set. MaxF: Maximum F1-measure, AP: Average precision.
Experiments: Kitti Road

Examples
### Experiments: Pascal VOC 2012

<table>
<thead>
<tr>
<th>Method</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepLabv2-CRF[8]</td>
<td>79.7%</td>
</tr>
<tr>
<td>CentraleSupelec Deep G-CRF[5]</td>
<td>80.2%</td>
</tr>
<tr>
<td>HikSeg_COCO[35]</td>
<td>81.4%</td>
</tr>
<tr>
<td><strong>ResNet-DUC (ours)</strong></td>
<td><strong>83.1%</strong></td>
</tr>
</tbody>
</table>

Table 6. Performance on the Pascal VOC2012 test set.

<table>
<thead>
<tr>
<th>Model</th>
<th>Ave.</th>
<th>plane</th>
<th>bike</th>
<th>bird</th>
<th>boat</th>
<th>bottle</th>
<th>bus</th>
<th>car</th>
<th>cat</th>
<th>chair</th>
<th>cow</th>
<th>table</th>
<th>dog</th>
<th>horse</th>
<th>moto</th>
<th>ppl.</th>
<th>plant</th>
<th>sheep</th>
<th>sofa</th>
<th>train</th>
<th>tv</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepLabv2-CRF</td>
<td>79.7</td>
<td><strong>92.59</strong></td>
<td>59.71</td>
<td>91.27</td>
<td>65.04</td>
<td>76.45</td>
<td>94.78</td>
<td>88.89</td>
<td>92.16</td>
<td>33.97</td>
<td>88.39</td>
<td>67.65</td>
<td>89.20</td>
<td>92.14</td>
<td>86.97</td>
<td>87.07</td>
<td>63.05</td>
<td>88.53</td>
<td>60.17</td>
<td>86.53</td>
<td>74.58</td>
</tr>
<tr>
<td>CentraleSupelec Deep G-CRF</td>
<td>80.2</td>
<td>92.90</td>
<td>61.20</td>
<td>91.00</td>
<td>66.30</td>
<td>77.70</td>
<td><strong>95.30</strong></td>
<td>88.90</td>
<td>92.40</td>
<td>33.80</td>
<td>88.40</td>
<td>69.10</td>
<td>89.80</td>
<td>92.90</td>
<td>87.70</td>
<td>87.50</td>
<td>62.60</td>
<td>89.90</td>
<td>59.20</td>
<td><strong>87.10</strong></td>
<td>74.20</td>
</tr>
<tr>
<td>HikSeg_COCO</td>
<td>81.4</td>
<td>95.00</td>
<td>64.20</td>
<td>91.50</td>
<td><strong>79.00</strong></td>
<td>78.70</td>
<td>93.40</td>
<td>88.40</td>
<td>94.30</td>
<td><strong>45.80</strong></td>
<td>89.60</td>
<td>65.20</td>
<td>90.60</td>
<td>92.80</td>
<td>88.70</td>
<td>87.50</td>
<td>62.40</td>
<td>88.40</td>
<td>56.40</td>
<td>86.20</td>
<td>75.30</td>
</tr>
<tr>
<td><strong>ours</strong></td>
<td><strong>83.1</strong></td>
<td>92.10</td>
<td><strong>64.57</strong></td>
<td><strong>94.73</strong></td>
<td>71.04</td>
<td><strong>80.98</strong></td>
<td><strong>89.73</strong></td>
<td><strong>94.88</strong></td>
<td>45.56</td>
<td><strong>93.72</strong></td>
<td><strong>74.39</strong></td>
<td><strong>91.95</strong></td>
<td><strong>95.07</strong></td>
<td><strong>90.02</strong></td>
<td><strong>88.68</strong></td>
<td><strong>69.12</strong></td>
<td><strong>90.42</strong></td>
<td><strong>62.68</strong></td>
<td><strong>86.38</strong></td>
<td><strong>78.15</strong></td>
<td></td>
</tr>
</tbody>
</table>
Experiments: Pascal VOC 2012

Examples
Conclusions

• We proposed simple yet effective convolutional operations for improving semantic segmentation systems.

• We designed a new dense upsampling convolution (DUC) operation to enable pixel-level prediction on feature maps, and hybrid dilated convolution (HDC) to deal with the gridding problem, effectively enlarging the receptive fields of the network.

• Experimental results demonstrate the effectiveness of our framework by achieving state-of-the-art performance on various semantic segmentation benchmarks.
Summary

Human Visual Recognition (face, object, scene)

Simulate
Explain
Predict

TM, TDM
Publications


Thank you!

- My dissertation committee: Dr. Gary Cottrell, Dr. Nuno Vasconcelos, Dr. Virginia de Sa, Dr. Truong Nguyen, Dr. Bhaskar Rao.

- My collaborators: Isabel Gauthier, Chris Kanan, Ben Cipollini, Vicente Malave, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou

- Gary’s Unbelievable Research Unit (GURU)

- Perceptual Expertise Network (PEN)

- Temporal Dynamics of Learning Center (TDLC)

- My family and friends
Q&A