Our old friend

The Felleman-vanEssen Diagram
Feedbacks everywhere

Gilbert & Li (2013)
Of course we soft-pedaled them earlier ....
Neural data has dynamics
Neural data has dynamics

Hierarchical structure can be seen in the dynamics

---

**Internal Consistency of Recorded V4 and IT Neurons**

- **V4** internal consistency
- **IT** internal consistency

Median Spearman-Corrected Split-Half Internal Consistency

<table>
<thead>
<tr>
<th>time (ms)</th>
<th>Median Spearman-Corrected Split-Half Internal Consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>0.2</td>
</tr>
<tr>
<td>80</td>
<td>0.4</td>
</tr>
<tr>
<td>100</td>
<td>0.6</td>
</tr>
<tr>
<td>120</td>
<td>0.7</td>
</tr>
<tr>
<td>140</td>
<td>0.6</td>
</tr>
<tr>
<td>160</td>
<td>0.5</td>
</tr>
<tr>
<td>180</td>
<td>0.4</td>
</tr>
<tr>
<td>200</td>
<td>0.3</td>
</tr>
</tbody>
</table>
Limitations of Feedforward Structures
Limitations of Feedforward Structures
Limitations of Feedforward Structures
Limitations of Feedforward Structures
Limitations of Feedforward Structures
Limitations of Feedforward Structures
Limitations of Feedforward Structures

Diagram showing a sequence of operations including fully connected (fc) layers and convolution (conv) layers, with a bypass and concatenation (concat) indicated.
Limitations of Feedforward Structures
Limitations of Feedforward Structures
Limitations of Feedforward Structures
Limitations of Feedforward Structures
Limitations of Feedforward Structures
Limitations of Feedforward Structures
Limitations of Feedforward Structures

\[
\text{state}_{t+1} = \text{conv}_{t+1} + \lambda \cdot \text{conv}_t
\]
Limitations of Feedforward Structures
Limitations of Feedforward Structures

\[ \text{duration} \]
- 1
- 2
- 3
- 4
- 5

\[ \text{impulse transfer function} \]

\[ \text{bypass} \]
\[ \text{concat} \]

\[ \text{memory} \]
\[ \text{state}_{t+1} = \text{conv}_{t+1} + \lambda \cdot \text{conv}_t \]
Limitations of Feedforward Structures
What task(s)?

a) vanilla categorization
What task(s)?

a) vanilla categorization

b) time-discounting

\[ L = \sum_{t} \gamma^{t} \cdot L_t \]

be accurate but also fast
What task(s)?

a) vanilla categorization

b) time-discounting

\[ L = \sum_{t} \gamma^t \cdot L_t \]

be accurate but also fast

c) heavy occlusion &c
Neural data has dynamics

IT trajectories fit with feedforward models

![Graph showing neural fit over time with different mapping types: fixed mapping, time-varying mapping, and internal consistency.](image-url)
“Vision is an active process, where higher order cognitive influences affect the operations performed by cortical neurons.”
“Vision is an active process, where higher order cognitive influences affect the operations performed by cortical neurons.”

“Top-down influences include various forms of attention, including spatial, object oriented and feature oriented attention.”
“Vision is an active process, where higher order cognitive influences affect the operations performed by cortical neurons.”

“Top-down influences include various forms of attention, including spatial, object oriented and feature oriented attention.”

“Top-down influences …. [also include] perceptual task, object expectation, scene segmentation, efference copy, working memory, and the encoding and recall of learned information.”
Biological views on function

Task-dependent changes in neural tuning and information content in V1

T1: “Which green line is closer to the red?”
Biological views on function

Task-dependent changes in neural tuning and information content in V1

T1: “Which green line is closer to the red?”

T2: “Which black line is closer to the red?”
Task-dependent changes in neural tuning and information content in V1

T1: “Which green line is closer to the red?”

T2: “Which black line is closer to the red?”
Biological views on function

Task-dependent changes in neural tuning and information content in V1

T1: “Which green line is closer to the red?”

T2: “Which black line is closer to the red?”
Biological views on function

Task-dependent changes in neural tuning and information content in V1

T1: “Which green line is closer to the red?”

T2: “Which black line is closer to the red?”
Biological views on function

Task-dependent changes in neural tuning and information content in IT

Comparison of Primate Prefrontal and Premotor Cortex Neuronal Activity during Visual Categorization

Jason A. Cromer, Jefferson E. Roy, Timothy J. Buschman, and Earl K. Miller
Biological views on function

Task-dependent changes in neural tuning and information content in IT

Cromer & Miller 2010
Adaptive shape processing in primary visual cortex

Justin N. J. McManus, Wu Li, and Charles D. Gilbert

Contributed by Charles D. Gilbert, April 18, 2011 (sent for review March 4, 2011)
Adaptive shape processing in primary visual cortex

Justin N. J. McManus\textsuperscript{a}, Wu Li\textsuperscript{b}, and Charles D. Gilbert\textsuperscript{a,1}

Contributed by Charles D. Gilbert, April 16, 2011 (sent for review March 4, 2011)

an entirely different mode of selectivity, for circular shapes. The difference between the mean tuning surfaces under the line and circle/wave tasks was statistically significant (monkey A, total number of surfaces, $n = 53$, $P = 4 \times 10^{-5}$; monkey B, $n = 63$, $P = 0.007$; and monkey C, $n = 62$, $P = 0.003$).
This process suggests that expectation of an object creates a set of filters that are selective for the object’s components and thus, a role of top-down processes in object recognition. The idea is further supported by the transfer of perceptual learning between objects with shared components.
Biological views on function

Efference copy = copy of motor instructions, for (e.g.) stability
Biological views on function

Efference copy = copy of motor instructions, for (e.g.) stability
Biological views on function

Efference copy = copy of motor instructions, for (e.g.) stability

Efference copies are created with our own movement but not those of other people. This is why other people can tickle us (no efference copies of the movements that touch us) but we cannot tickle ourselves (efference copies tell us that we are stimulating ourselves).
Biological views on function

Efference copy = copy of motor instructions, for (e.g.) stability

Summer & Wurtz 2006
Biological views on function

Efference copy = copy of motor instructions, for (e.g.) stability

Superior Colliculus (SC) — “issues motor commands”

Medial Dorsal (MD) of thalamus — “routing”

Frontal Eye Field (FEF) — moves the eyes

Summer & Wurtz 2006
Biological views on function

Efference copy = copy of motor instructions, for (e.g.) stability

Superior Colliculus (SC) — “issues motor commands”

Medial Dorsal (MD) of thalamus — “routing”

Frontal Eye Field (FEF) — moves the eyes

Summer & Wurtz 2006
Biological views on function

Efference copy = copy of motor instructions, for (e.g.) stability

Superior Colliculus (SC) — “issues motor commands”

Medial Dorsal (MD) of thalamus — “routing”

Frontal Eye Field (FEF) — moves the eyes

Summer & Wurtz 2006
Biological views on function

Neuron

Volume 53, Issue 6, 15 March 2007, Pages 881-890

Article

Remembering Visual Motion: Neural Correlates of Associative Plasticity and Motion Recall in Cortical Area MT

Anja Schlack, Thomas D. Albright
Biological views on function

Area MT

“responsible” for motion perception

medial temporal Cortex (Area MT) Specialized for Visual Motion
Newsome and others

100% correlation 0% correlation 50% correlation
normal threshold = 1-2%
threshold after lesions of MT = 10-20%

Schlack & Albright 2007
Biological views on function

Retained memory responses after training

A

B

Schlack & Albright 2007
Biological views on function

Retained memory responses after training

Schlack & Albright 2007

Area MT
Biological views on function

Retained memory responses after training

Schlack & Albright 2007

Area MT
Prefrontal cortex (PFC) ~ “executive control”, long-range planning, decision making, task switching

short-term memory
Biological views on function

Prefrontal cortex (PFC) ~ “executive control”, long-range planning, decision making, task switching

short-term memory

Mixture of memory, task (“executive control”), and prediction

“Feedback projections from prefrontal cortex to the posterior association cortex appear to serve the executive control of voluntary recall.”
Biological views on function

Top-down signal from prefrontal cortex in executive control of memory retrieval

Hyoe Tomita*, Machiko Ohbayashi*, Kiyoshi Nakahara†, Isao Hasegawa† & Yasushi Miyashita*††

Mixture of memory, task (“executive control”), and prediction

“Feedback projections from prefrontal cortex to the posterior association cortex appear to serve the executive control of voluntary recall.”

“Split-brain paradigm” — transection of posterior corpus callosum — IT neurons in one hemisphere are activated by direct bottom-up inputs only in the contralateral hemifield, but not when the inputs are in the ipsilateral hemifield.
Biological views on function

Mixture of memory, task (“executive control”), and prediction

“Feedback projections from prefrontal cortex to the posterior association cortex appear to serve the executive control of voluntary recall.”

“Split-brain paradigm” — transection of posterior corpus callosum — IT neurons in one hemisphere are activated by direct bottom-up inputs only in the contralateral hemifield, but not when the inputs are in the ipsilateral hemifield.

Ipsilateral presentation *still* activated IT neurons, but later than contralateral.

Neuron’s pattern of responses across stimuli similar regardless of ipsi/contral presentation ($r = \sim 0.8$)
Biological views on function

Top-down signal from prefrontal cortex in executive control of memory retrieval

Hyoe Tomita*, Machiko Ohbayashi*, Kiyoshi Nakahara†, Isao Hasegawa*† & Yasushi Miyashita*††

Mixture of memory, task (“executive control”), and prediction

“Feedback projections from prefrontal cortex to the posterior association cortex appear to serve the executive control of voluntary recall.”

“Split-brain paradigm” — transection of posterior corpus callosum — IT neurons in one hemisphere are activated by direct bottom-up inputs only in the contralateral hemifield, but not when the inputs are in the ipsilateral hemifield.

Ipsilateral presentation *still* activated IT neurons, but later than contralateral.

Neuron's pattern of responses across stimuli similar regardless of ipsi/contra presentation (r = ~0.8)

No such transfer in *full* split.

Pair associated test indicates prospective information from PFC sent to IT.
Biological views on function

Attention

Review

The Normalization Model of Attention

John H. Reynolds ¹, David J. Heeger ²

https://doi.org/10.1016/j.neuron.2009.01.002

Under an Elsevier user license
Biological views on function

\[ R(x, \theta) = \text{ReLU}_T \left[ \frac{E(x, \theta)}{S(x, \theta) + \sigma} \right] \]

\( E = \) feedforward input
\( S = \) suppression
Biological views on function

\[ R(x, \theta) = \text{ReLu}_T \left[ \frac{E(x, \theta)}{\text{Conv}_s(x, \theta) \left[ E(x, \theta) \right] + \sigma} \right] \]

- E = feedforward input
- S = suppression
- x = position
- theta = orientation
Biological views on function

\[ R(x, \theta) = \text{ReLU}_T\left[ \frac{A(x, \theta) E(x, \theta)}{\text{Conv}_s(x, \theta) [A(x, \theta) E(x, \theta)] + \sigma} \right] \]

E = feedforward input
S = suppression
x = position
theta = orientation
A = attention field

\(--\) implemented as *equilibrium* of simple recurrent circuit (Heeger 1993)
Biological views on function

Top-down influence in early visual processing
A Bayesian perspective

Tai Sing Lee
Center for the Neural Basis of Cognition
Department of Computer Science
Carnegie Mellon University, Pittsburgh, PA 15213, U.S.A.
Department of Neuroscience
University of Pittsburgh, Pittsburgh, PA 15213, U.S.A.

\[ P(S_i \mid E, H) = \frac{P(E \mid S_i, H)P(S_i \mid H)}{P(E \mid H)} \]

\( S_i \) = scene i

\( E \) = evidence

\( H \) = prior information
Top-down influence in early visual processing
A Bayesian perspective

Tai Sing Lee
Center for the Neural Basis of Cognition
Department of Computer Science
Carnegie Mellon University, Pittsburgh, PA 15213, U.S.A.
Department of Neuroscience
University of Pittsburgh, Pittsburgh, PA 15213, U.S.A.

Bayesian interaction of two brain areas

\[
P(S_i | E, H) = \frac{P(E | S_i, H)P(S_i | H)}{P(E | H)}
\]

\(S_i\) = scene i output of V1
\(E\) = evidence finished to V1 by retina
\(H\) = prior information generated by V2
Top-down influence in early visual processing
A Bayesian perspective

Tai Sing Lee
Center for the Neural Basis of Cognition
Department of Computer Science
Carnegie Mellon University, Pittsburgh, PA 15213, U.S.A.

Bayesian interaction of two brain areas

\[ P(S_i | E, H) = \frac{P(E | S_i, H)P(S_i | H)}{P(E | H)} \]

\( S_i \) = scene i output of V1
\( E \) = evidence finished to V1 by retina
\( H \) = prior information generated by V2
Biological views on function

Top-down influence in early visual processing
A Bayesian perspective

Tai Sing Lee

Center for the Neural Basis of Cognition
Department of Computer Science
Carnegie Mellon University, Pittsburgh, PA 15213, U.S.A.
Department of Neuroscience
University of Pittsburgh, Pittsburgh, PA 15213, U.S.A.

Bayesian interaction of two brain areas

\[
P(S_i \mid E, H) = \frac{P(E \mid S_i, H)P(S_i \mid H)}{P(E \mid H)}
\]

\(S_i\) = scene i output of V1

\(E\) = evidence finished to V1 by retina

\(H\) = prior information generated by V2

illusory contour sensitivity emerges in \textbf{V2 at 65 ms} (first feedforward wave) but in \textbf{V1 at 100ms}
Direct Feedback Alignment Provides Learning in Deep Neural Networks

Arild Nøkland

(Submitted on 6 Sep 2016 (v1), last revised 21 Dec 2016 (this version, v5))

Random synaptic feedback weights support error backpropagation for deep learning

Timothy P. Lillicrap, Daniel Cownden, Douglas B. Tweed & Colin J. Akerman
Direct Feedback Alignment Provides Learning in Deep Neural Networks

Arild Nøkland

(Submitted on 6 Sep 2016 (v1), last revised 21 Dec 2016 (this version, v5))

Random synaptic feedback weights support error backpropagation for deep learning

Timothy P. Lillicrap, Daniel Cownden, Douglas B. Tweed & Colin J. Akerman
Biological views on function

Task Dependence          Executive control

Adaptive Shape processing

Efferent Copy

Memory

Generalized Attention    Bayesian inference

Implementing Learning
Models

What and where: A Bayesian inference theory of attention

Sharat Chikkerur *, Thomas Serre, Cheston Tan, Tomaso Poggio

McGovern Institute for Brain Research, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, United States

\[ P(O, L, X^1, \ldots, X^N, I) = P(O)P(L)P(I \mid X^1, \ldots, X^N) \prod_{i=1}^{N} P(X^i \mid L, F^i)P(F^i \mid O) \]

N = number of objects

object encoding O (PFC)

feature encoding F (IT)

location encoding L (FEF)

joint location/feature map \( X^i \) (V4)

feedforward input (V1/V2)
What and where: A Bayesian inference theory of attention

Sharat Chikkerur*, Thomas Serre, Cheston Tan, Tomaso Poggio

McGovern Institute for Brain Research, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, United States

Fig. 2. Left: Proposed Bayesian model. Right: A model illustrating the interaction between the parietal and ventral streams mediated by feedforward and feedback connections. The main additions to the original feedforward model (Serre, Kouh, et al., 2005) (see also Supplementary Online Information) are (i) the cortical feedback within the ventral stream (providing feature-based attention); (ii) the cortical feedback from areas of the parietal cortex onto areas of the ventral stream (providing spatial attention); and (iii) feedforward connections to the parietal cortex that serves as a 'saliency map' encoding the visual relevance of image locations (Koch & Ullman, 1985).

Fig. 4. Effect of spatial attention on tuning response. The tuning curve shows a multiplicative modulation under attention. The inset shows the replotted data from McAdams and Maunsell (1999).
What and where: A Bayesian inference theory of attention
Sharat Chikkerur *, Thomas Serre, Cheston Tan, Tomaso Poggio
McGovern Institute for Brain Research, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, United States

$F^i = \text{pools across locations in } X^i$
What and where: A Bayesian inference theory of attention
Sharat Chikkerur*, Thomas Serre, Cheston Tan, Tomaso Poggio

McGovern Institute for Brain Research, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, United States

\[ F^i = \text{pools across locations in } X^i \]

Spatial attention spotlight \( \times \)
What and where: A Bayesian inference theory of attention
Sharat Chikkerur *, Thomas Serre, Cheston Tan, Tomaso Poggio

McGovern Institute for Brain Research, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, United States

\[ F^i = \text{pools across locations in } X^i \]

Spatial attention spotlight \( X \)

Effect read out as \( P(F^i | I) \)
What and where: A Bayesian inference theory of attention

Sharat Chikkerur *, Thomas Serre, Cheston Tan, Tomaso Poggio

McGovern Institute for Brain Research, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, United States

F^i = pools across locations in X^i
Spatial attention spotlight X
Effect read out as P(F^i | I)

L represented in FEF
Feature attention makes P(F^i) high for preferred feature
Location of preferred feature read out as P(L | I)
What and where: A Bayesian inference theory of attention
Sharat Chikkerur *, Thomas Serre, Cheston Tan, Tomaso Poggio
McGovern Institute for Brain Research, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, United States
A mostly complete chart of

Neural Networks

©2016 Fjoerd van Veen - asimovinstitute.org
Unlike feedforward networks, recurrent networks can store state.
Models: RNNs

Simple (unrestricted) RNNs

- Input layer
- Hidden layers: “deep” if > 1
- Recurrent network
- Output layer (class/target)
Recurrent convolutional neural networks suppress occluders and enhance targets in occluded object recognition

Courtney J. Spoerer (courtney.spoerer@mrc-cbu.cam.ac.uk)
Medical Research Council Cognition and Brain Sciences Unit,
15 Chaucer Road, Cambridge, CB2 7EF, UK

Nikolaus Kriegeskorte (nikokriegeskorte@gmail.com)
Medical Research Council Cognition and Brain Sciences Unit,
15 Chaucer Road, Cambridge, CB2 7EF, UK

---

Figure 1: The process for generating stimuli for digit debris. First the target digit is generated. Random crops of all possible targets are taken to create a mask of debris, which is applied to the target as an occluder.
Recurrent convolutional neural networks suppress occluders and enhance targets in occluded object recognition

Courtney J. Spoerer (courtney.spoerer@mrc-cbu.cam.ac.uk)
Medical Research Council Cognition and Brain Sciences Unit, 15 Chaucer Road, Cambridge, CB2 7EF, UK

Nikolaus Kriegeskorte (nikokriegeskorte@gmail.com)
Medical Research Council Cognition and Brain Sciences Unit, 15 Chaucer Road, Cambridge, CB2 7EF, UK
Models

All RNNs executed by **unrolling in time**
Recurrent convolutional neural networks suppress occluders and enhance targets in occluded object recognition

Courtney J. Spoerer (courtney.spoerer@mrc-cbu.cam.ac.uk)
Medical Research Council Cognition and Brain Sciences Unit,
15 Chaucer Road, Cambridge, CB2 7EF, UK

Nikolaus Kriegeskorte (nikokriegeskorte@gmail.com)
Medical Research Council Cognition and Brain Sciences Unit,
15 Chaucer Road, Cambridge, CB2 7EF, UK

Results
Recurrent networks significantly out-performed feedforward networks across varying levels of occlusion. The difference in performance between feedforward and recurrent networks increased as the occlusion increased (Figure 2).

Figure 1: The process for generating stimuli for digit debris. First the target digit is generated. Random crops of all possible targets are taken to create a mask of debris, which is applied to the target as an occluder.

Figure 2: Classification error of the networks across increasing levels of debris (left). Pairwise differences across architectures for different levels of debris are indicated in matrix form (right).
Recurrent convolutional neural networks suppress occluders and enhance targets in occluded object recognition

Courtney J. Spoerer (courtney.spoerer@mrc-cbu.cam.ac.uk)
Medical Research Council Cognition and Brain Sciences Unit,
15 Chaucer Road, Cambridge, CB2 7EF, UK

Nikolaus Kriegeskorte (nikokriegeskorte@gmail.com)
Medical Research Council Cognition and Brain Sciences Unit,
15 Chaucer Road, Cambridge, CB2 7EF, UK

Error for MNIST under varying levels of noise

<table>
<thead>
<tr>
<th>Classification error (%)</th>
<th>No noise</th>
<th>SNR = 1</th>
<th>SNR = (\frac{1}{2})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feedforward</td>
<td>B</td>
<td>B-F</td>
<td>B-K</td>
</tr>
<tr>
<td>Recurrent</td>
<td>BT</td>
<td>BL</td>
<td>BLT</td>
</tr>
</tbody>
</table>

Training images

- Target: 7

Testing images

- No noise: B
- SNR = 1: B-F, B-K
- SNR = \(\frac{1}{2}\): B-L, B-LT

Pairwise tests

- B vs B-F
- B-F vs B-K
- B-K vs B-L
- B-L vs B-LT

Significant difference (two-sided McNemar test, expected FDR = 0.05)
Bridging the Gaps Between Residual Learning, Recurrent Neural Networks and Visual Cortex

by

Qianli Liao and Tomaso Poggio
Center for Brains, Minds and Machines, McGovern Institute, MIT

(A) ResNet with shared weights  (B) ResNet in recurrent form
Models

Bridging the Gaps Between Residual Learning, Recurrent Neural Networks and Visual Cortex

(A) Multi-state (Fully) Recurrent Neural Network

(B) Full model

(C) Simulating our model in time by unrolling

(D) An example ResNet: for comparison

Figure 2: Modeling the ventral stream of visual cortex using a multi-state fully recurrent neural network
Models

Bridging the Gaps Between Residual Learning, Recurrent Neural Networks and Visual Cortex

by

Qianli Liao and Tomaso Poggio
Center for Brains, Minds and Machines, McGovern Institute, MIT

CIFAR-10 Error
Feedback Networks

Amir R. Zamir¹,³  Te-Lin Wu¹  Lin Sun¹,²  William B. Shen¹  Bertram E. Shi²  
Jitendra Malik³  Silvio Savarese¹

¹ Stanford University  ² HKUST  ³ University of California, Berkeley

http://feedbacknet.stanford.edu/

$$L = \sum_{t=1}^{T} \gamma L_t, \text{ where } L_t = -\log \frac{e^{H_t^D[C]}}{\sum_j e^{H_t^D[j]}}.$$
Feedback Networks

Amir R. Zamir$^{1,3,*}$  Te-Lin Wu$^1$  Lin Sun$^{1,2}$  William B. Shen$^1$  Bertram E. Shi$^2$
Jitendra Malik$^3$  Silvio Savarese$^1$

$^1$ Stanford University  $^2$ HKUST  $^3$ University of California, Berkeley

http://feedbacknet.stanford.edu/

$$L = \sum_{t=1}^{T} \gamma^t L_t, \text{ where } L_t = -\log \frac{e^{H^D[C]}}{\sum_j e^{H^D[j]}}.$$
Feedback Networks

Amir R. Zamir\textsuperscript{1,3*}, Te-Lin Wu\textsuperscript{1*}, Lin Sun\textsuperscript{1,2}, William B. Shen\textsuperscript{1}, Bertram E. Shi\textsuperscript{2}, Jitendra Malik\textsuperscript{3}, Silvio Savarese\textsuperscript{1}

\textsuperscript{1} Stanford University, \textsuperscript{2} HKUST, \textsuperscript{3} University of California, Berkeley

http://feedbacknet.stanford.edu/

<table>
<thead>
<tr>
<th>Model</th>
<th>Physical Depth</th>
<th>Virtual Depth</th>
<th>Top1 (%)</th>
<th>Top5 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feedback Net</td>
<td>12</td>
<td>48</td>
<td>71.12</td>
<td>91.51</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>32</td>
<td>69.57</td>
<td>91.01</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>16</td>
<td>67.83</td>
<td>90.12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feedback (ResNet\textsuperscript{19})</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>48</td>
<td>-</td>
<td>-</td>
<td>70.04</td>
<td>90.96</td>
</tr>
<tr>
<td>32</td>
<td>-</td>
<td>-</td>
<td>69.36</td>
<td>91.07</td>
</tr>
<tr>
<td>12</td>
<td>-</td>
<td>-</td>
<td>66.35</td>
<td>90.02</td>
</tr>
<tr>
<td>8</td>
<td>-</td>
<td>-</td>
<td>64.23</td>
<td>88.95</td>
</tr>
<tr>
<td>128*</td>
<td>-</td>
<td>-</td>
<td>70.92</td>
<td>91.28</td>
</tr>
<tr>
<td>110*</td>
<td>-</td>
<td>-</td>
<td>72.06</td>
<td>92.12</td>
</tr>
<tr>
<td>64*</td>
<td>-</td>
<td>-</td>
<td>71.01</td>
<td>91.48</td>
</tr>
<tr>
<td>48*</td>
<td>-</td>
<td>-</td>
<td>70.56</td>
<td>91.60</td>
</tr>
<tr>
<td>32*</td>
<td>-</td>
<td>-</td>
<td>69.58</td>
<td>91.55</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feedforward (VGG\textsuperscript{48})</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>48</td>
<td>-</td>
<td>-</td>
<td>55.08</td>
<td>82.1</td>
</tr>
<tr>
<td>32</td>
<td>-</td>
<td>-</td>
<td>63.56</td>
<td>88.41</td>
</tr>
<tr>
<td>12</td>
<td>-</td>
<td>-</td>
<td>64.65</td>
<td>89.26</td>
</tr>
<tr>
<td>8</td>
<td>-</td>
<td>-</td>
<td>63.91</td>
<td>88.90</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>1001</td>
<td>110</td>
<td>32 fat</td>
<td>4 fat</td>
</tr>
</tbody>
</table>

Table 6. Endpoint performance comparison on CIFAR-100. Baselines denoted with * are the architecture used in the original ResNet paper.

\[ L = \sum_{t=1}^{T} \gamma^t L_t, \text{ where } L_t = -log \frac{e^{H^P[C]}}{\sum_j e^{H^P[j]}}. \]

Figure 5. Evaluation of early predictions. Comparison of accuracy of feedback (FB) model and feedforward (FF) baselines (ResNet \& VGG, with or without auxiliary loss layers).
Feedback Networks

Amir R. Zamir\textsuperscript{1,3}\*, Te-Lin Wu\textsuperscript{1}\*, Lin Sun\textsuperscript{1,2} William B. Shen\textsuperscript{1} Bertram E. Shi\textsuperscript{2}
Jitendra Malik\textsuperscript{3} Silvio Savarese\textsuperscript{1}

\textsuperscript{1} Stanford University \quad \textsuperscript{2} HKUST \quad \textsuperscript{3} University of California, Berkeley

http://feedbacknet.stanford.edu/

\begin{align*}
L = \sum_{t=1}^{T} \gamma^t L_t, \quad \text{where } L_t = -\log \frac{e^{H^P[C]}}{\sum_j e^{H^P[j]}}.
\end{align*}

<table>
<thead>
<tr>
<th>Model</th>
<th>Physical Depth</th>
<th>Virtual Depth</th>
<th>Top1 (%)</th>
<th>Top5 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feedback Net</td>
<td>12</td>
<td>48</td>
<td>71.12</td>
<td>91.51</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>32</td>
<td>69.57</td>
<td>91.01</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>16</td>
<td>67.83</td>
<td>90.12</td>
</tr>
<tr>
<td>Feedforward</td>
<td>48</td>
<td>-</td>
<td>70.04</td>
<td>90.96</td>
</tr>
<tr>
<td>(ResNet[19])</td>
<td>32</td>
<td>-</td>
<td>69.36</td>
<td>91.07</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>-</td>
<td>66.35</td>
<td>90.02</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>-</td>
<td>64.23</td>
<td>88.95</td>
</tr>
<tr>
<td>128*</td>
<td>-</td>
<td>-</td>
<td>70.92</td>
<td>91.28</td>
</tr>
<tr>
<td>110*</td>
<td>-</td>
<td>-</td>
<td>72.06</td>
<td>92.12</td>
</tr>
<tr>
<td>64*</td>
<td>-</td>
<td>-</td>
<td>71.01</td>
<td>91.48</td>
</tr>
<tr>
<td>48*</td>
<td>-</td>
<td>-</td>
<td>70.56</td>
<td>91.60</td>
</tr>
<tr>
<td>32*</td>
<td>-</td>
<td>-</td>
<td>69.58</td>
<td>91.55</td>
</tr>
<tr>
<td>Feedforward</td>
<td>48</td>
<td>-</td>
<td>55.08</td>
<td>82.1</td>
</tr>
<tr>
<td>(VGG[48])</td>
<td>32</td>
<td>-</td>
<td>63.56</td>
<td>88.41</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>-</td>
<td>64.65</td>
<td>89.26</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>-</td>
<td>63.91</td>
<td>88.90</td>
</tr>
<tr>
<td>Highway [53]</td>
<td>19</td>
<td>-</td>
<td>67.76</td>
<td>-</td>
</tr>
<tr>
<td>ResNet v2[20]</td>
<td>1001</td>
<td>-</td>
<td>77.29</td>
<td>-</td>
</tr>
<tr>
<td>Stochastic Depth [24]</td>
<td>110</td>
<td>-</td>
<td>75.02</td>
<td>-</td>
</tr>
<tr>
<td>SwapOut [49]</td>
<td>32 fat</td>
<td>-</td>
<td>77.28</td>
<td>-</td>
</tr>
<tr>
<td>RCNN [37]</td>
<td>4 fat</td>
<td>16</td>
<td>68.25</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 6. Endpoint performance comparison on CIFAR-100. Baselines denoted with * are the architecture used in the original ResNet paper.
ATTENTION FOR FINE-GRAINED CATEGORIZATION

Pierre Sermanet, Andrea Frome, Esteban Real
Google, Inc.
{sermanet, afrome, ereal, }@google.com

Figure 2: Diagram of the model. The grayed-out boxes denote resolutions not in use; in our experiments the context is always a low-resolution patch, while each glimpse can be any combination of the low-, medium-, and high-resolution patches.
ATTENTION FOR FINE-GRAINED CATEGORIZATION

Pierre Sermanet, Andrea Frome, Esteban Real
Google, Inc.
{sermanet, afrome, ereal,}@google.com

Table 1: Results on Stanford Dogs for (a) our RNN model and (b) our GoogLeNet baselines and previous state-of-the-art results, measured by mean accuracy percentage (mA) as described in Chai et al. (2013). The GoogLeNet baseline models were pre-trained on the de-duped ILSVRC 2012 training set and fine-tuned with the Stanford Dogs training set. Results marked with a star indicate use of tight ground truth bounding boxes around the dogs in training and testing.

<table>
<thead>
<tr>
<th># glimpses</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>high res only</td>
<td>43.5</td>
<td>48.3</td>
<td>49.6</td>
</tr>
<tr>
<td>medium res only</td>
<td>70.1</td>
<td>72.3</td>
<td>72.8</td>
</tr>
<tr>
<td>low res only</td>
<td>70.3</td>
<td>70.1</td>
<td>70.7</td>
</tr>
<tr>
<td>high+medium res</td>
<td>70.7</td>
<td>72.6</td>
<td>72.7</td>
</tr>
<tr>
<td>3-resolution</td>
<td>76.3</td>
<td>76.5</td>
<td>76.8</td>
</tr>
</tbody>
</table>

Figure 3: Visualizations of 2-resolution (a) and 3-resolution (b) glimpses on an image from our validation set, with learned fixation points. For each the glimpse images are in order, from top to bottom, and the box diagram corresponds to the second glimpse. The composite image is created from all three glimpses. The context image is not shown but is always the same resolution and size as the low-resolution glimpse patches shown in (b).