Hierarchical models
place cells
grid cells

face cells

invariant repr.
complex motion

‘Gabor filters’
an absolute depth judgment with respect to fixation, while fine stereopsis requires the judgment of relative depth, i.e., comparing depth across space; (2) the particular coarse stereopsis task used requires the monkey to discriminate a signal in noise, while the fine task does not; (3) the range of disparities is quite different.

Chowdhury and DeAngelis (2008) replicate the finding that monkeys initially trained on coarse stereopsis show impaired coarse depth discrimination when muscimol is injected into MT. Remarkably, the same animals, after a second round of training on fine stereopsis, are unimpaired at either fine or coarse depth discrimination by similar injections. Moreover, recordings in MT show that neuronal responses are not altered by learning the fine stereopsis task. Given the differences between the tasks and the large number of visual areas containing disparity-sensitive neurons, one might not be surprised to find different areas involved in the two tasks. But it is quite unexpected that merely learning one task would change the contribution of areas previously involved in the other. Chowdhury and DeAngelis conclude that the change in outcome reflects a change in neural decoding—decision centers that decode signals to render judgments of depth, finding MT signals unreliable for the fine stereopsis task, switch their inputs to select some better source of disparity information. Candidates include ventral stream areas V4 or IT, where relative disparity signals have been reported (Orban, 2008) and which contain far more neurons than MT.

Figure 1. A Scaled Representation of the Cortical Visual Areas of the Macaque

Each colored rectangle represents a visual area, for the most part following the names and definitions used by Felleman and Van Essen (1991). The gray bands connecting the areas represent the connections between them. Areas above the equator of the figure (reds, browns) belong to the dorsal stream. Areas below the equator (blues, greens) belong to the ventral stream. Following Lennie (1998), each area is drawn with a size proportional to its cortical surface area, and the lines connecting the areas each have a thickness proportional to the estimated number of fibers in the connection. The estimate is derived by assuming that each area has a number of output fibers proportional to its surface area and that these fibers are divided among the target areas in proportion to their surface areas. The connection strengths represented are therefore not derived from quantitative anatomy and furthermore represent only feedforward pathways, though most or all of the pathways shown are bidirectional. The original version of this figure was prepared in 1998 by John Maunsell.
an absolute depth judgment with respect to fixation, while fine stereopsis requires the judgment of relative depth, i.e., comparing depth across space; (2) the particular coarse stereopsis task used requires the monkey to discriminate a signal in noise, while the fine task does not; (3) the range of disparities is quite different.

Chowdhury and DeAngelis (2008) replicate the finding that monkeys initially trained on coarse stereopsis show impaired coarse depth discrimination when muscimol is injected into MT. Remarkably, the same animals, after a second round of training on fine stereopsis, are unimpaired at either fine or coarse depth discrimination by similar injections. Moreover, recordings in MT show that neuronal responses are not altered by learning the fine stereopsis task. Given the differences between the tasks and the large number of visual areas containing disparity-sensitive neurons, one might not be surprised to find different areas involved in the two tasks. But it is quite unexpected that merely learning one task would change the contribution of areas previously involved in the other. Chowdhury and DeAngelis conclude that the change in outcome reflects a change in neural decoding—decision centers that decode signals to render judgments of depth, finding MT signals unreliable for the fine stereopsis task, switch their inputs to select some better source of disparity information. Candidates include ventral stream areas V4 or IT, where relative disparity signals have been reported (Orban, 2008) and which contain far more neurons (Figure 1).

When challenged afresh with the coarse depth task, these same decision centers may now find that their new sources of information can solve the coarse task as well as the old ones. MT is no longer critical. Perhaps in other monkeys MT would never have a role in stereopsis at all. Chowdhury and DeAngelis' monkeys were trained simultaneously or previously to discriminate motion, which engages MT. Faced with a qualitatively similar random dot stimulus, it might make sense for the cortex to try to solve the new problem of stereopsis with existing decoding strategies. But if the animals were initially trained on a different task—say, a texture discrimination—MT might never be engaged at all. It would also be interesting to see the outcome if monkeys were trained on depth tasks that were less different and could be interleaved in the same sessions, for example noise-limited depth judgments using similar absolute or relative disparity...
If our model accurately describes the information captured (and dis
generation of metameric stimuli
ment of receptive field sizes in the

Thus, we sought the largest value of the scaling constant at which
images as a function of the scaling constant used in their generation.

As a separate control for the validity of this procedure, we examined
studies of area V2 (refs.

Finally, local correlations are compatible with models of cortical com-
model responses should appear to be

From Freeman & Simoncelli (2011)
(data from Gattass et al. 1981; 1988)
Pareidolia
Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position

Kunihiko Fukushima
NHK Broadcasting Science Research Laboratories, Kinuta, Setagaya, Tokyo, Japan
Fig. 5. An example of the interconnections between cells and the response of the cells after completion of self-organization
Neocognitron: activation rule

\[
u_{Sl}(k_l, n) = r_l \cdot \varphi \left[ 1 + \sum_{k_{l-1}=1}^{K_{l-1}} \sum_{v \in S_l} a_l(k_{l-1}, v, k_l) \cdot u_{Cl-1}(k_{l-1}, n + v) \right]
\]

where

\[
\varphi[x] = \begin{cases} x & x \geq 0 \\ 0 & x < 0 \end{cases}
\]

\[
v_{Cl-1}(n) = \sqrt{\sum_{k_{l-1}=1}^{K_{l-1}} \sum_{v \in S_l} c_{l-1}(v) \cdot u_{Cl-1}^2(k_{l-1}, n + v)}
\]
Neocognitron: learning rule

Let cell $u_{Sl}(\hat{k}_l, \hat{n})$ be selected as a representative.

$$\Delta a_l(k_{l-1}, v, \hat{k}_l) = q_l \cdot c_{l-1}(v) \cdot u_{cI-1}(k_{l-1}, \hat{n} + v),$$

Hebbian learning

From each S-column, every time when a stimulus pattern is presented, the S-cell which is yielding the largest output is chosen as a candidate for the representatives. Hence, there is a possibility that a number of candidates appear in a single S-plane. If two or more candidates appear in a single S-plane, only the one which is yielding the largest output among them is selected as the representative from that S-plane. In
Neocognitron: performance

**Fig. 6.** Some examples of distorted stimulus patterns which the neocognitron has correctly recognized, and the response of the final layer of the network.

**Fig. 7.** A display of an example of the response of all the individual cells in the neocognitron.
‘AlexNet’
(Krizhevsky, Sutskever & Hinton 2012)

image → feature extraction and pooling → classification
Deep networks appear to predict responses of V4 and IT neurons (Yamins & DiCarlo 2016)

**Diagram a:**
- Scatter plot showing IT single-site neural predictivity (% explained variance) against categorization performance (balanced accuracy).
- Models include HCNN, HMAX, PLOS09, V2-like, V1-like, SIFT, Pixels.
- Axis labels: Categorization performance (balanced accuracy).
- Legend includes categories: Ideal observer, All variables, V1-like, HMAX, V2-like.

**Diagram b:**
- HCNN top hidden layer response prediction.
- Test images (sorted by category).
- IT neural response.
- Comparison of IT and V4 single-site neural response predictivity for various models.

**Diagram c:**
- Monkey V4 (n = 128) and Monkey IT (n = 168).
- Single-site neural predictivity (% explained variance).
- Categories: Ideal observers, Control models, HCNN layers.

**Diagram d:**
- Human IT (fMRI) vs HCNN model.
- Categories: Human IT, HCNN model.

**Diagram e:**
- Human V1–V3 and Human IT.
- RDMA voxel correlation (Modalities τ_A).
- Scores for Convolutional and Fully connected layers.

- τ_A = 0.38
This isn’t a good model of perception
an absolute depth judgment with respect to fixation, while fine stereopsis requires the judgment of relative depth, i.e., comparing depth across space; (2) the particular coarse stereopsis task used requires the monkey to discriminate a signal in noise, while the fine task does not; (3) the range of disparities is quite different.

Chowdhury and DeAngelis (2008) replicate the finding that monkeys initially trained on coarse stereopsis show impaired coarse depth discrimination when muscimol is injected into MT. Remarkably, the same animals, after a second round of training on fine stereopsis, are unimpaired at either fine or coarse depth discrimination by similar injections. Moreover, recordings in MT show that neuronal responses are not altered by learning the fine stereopsis task. Given the differences between the tasks and the large number of visual areas containing disparity-sensitive neurons, one might not be surprised to find different areas involved in the two tasks. But it is quite unexpected that merely learning one task would change the contribution of areas previously involved in the other. Chowdhury and DeAngelis conclude that the change in outcome reflects a change in neural decoding—decision centers that decode signals to render judgments of depth, finding MT signals unreliable for the fine stereopsis task, switch their inputs to select some better source of disparity information. Candidates include ventral stream areas V4 or IT, where relative disparity signals have been reported (Orban, 2008) and which contain far more neurons (Figure 1). When challenged afresh with the coarse depth task, these same decision centers may now find that their new sources of information can solve the coarse task as well as the old ones. MT is no longer critical. Perhaps in other monkeys MT would never have a role in stereopsis at all. Chowdhury and DeAngelis' monkeys were trained simultaneously or previously to discriminate motion, which engages MT. Faced with a qualitatively similar random dot stimulus, it might make sense for the cortex to try to solve the new problem of stereopsis with existing decoding strategies. But if the animals were initially trained on a different task—say, a texture discrimination—MT might never be engaged at all. It would also be interesting to see the outcome if monkeys were trained on depth tasks that were less different and could be interleaved in the same sessions, for example noise-limited depth judgments using similar absolute or relative disparity.
Relative spatial relationships are important
The invariant representations produced by deep convolutional neural networks (convnets) are... easily fooled and brittle.

A recent study revealed that Deep neural networks (DNNs) have recently been achieving state-of-the-art performance on a variety of pattern-recognition tasks, most notably visual classification. Convolutional neural networks (convnets) are trained to perform well on either the ImageNet training set distribution or by hard-negative mining, in computer vision, consists of identifying training set examples (or portions thereof) which are given low probabilities by the model, but which should be high probability instead, cf. [1]. We make the connection with hard-negative mining explicitly, as it is close in spirit: hard-negative mining, in computer vision, consists of identifying training set examples (or portions thereof) which are given low probabilities by the model, but which should be high probability instead, cf. [1].

Our results shed light on interesting differences between human vision and current DNNs, and raise questions about the generalities of DNN computer vision. A recent study revealed that deep neural networks (DNNs) believe to be recognizable objects with 99.99% confidence (e.g. labeling with certainty that white noise is a lion). Specifically, we take convolutional neural networks trained to perform well on either the ImageNet training set distribution or by hard-negative mining, in computer vision, consists of identifying training set examples (or portions thereof) which are given low probabilities by the model, but which should be high probability instead, cf. [1]. We make the connection with hard-negative mining explicitly, as it is close in spirit: hard-negative mining, in computer vision, consists of identifying training set examples (or portions thereof) which are given low probabilities by the model, but which should be high probability instead, cf. [1].

- king penguin
- starfish
- baseball
- electric guitar
- freight car
- remote control
- peacock
- African grey
- school bus
- hen
- temple
- ostrich
Images are not bags of features (BagNet - Brendel & Bethge 2019)
Simple example: translation via Fourier phase-shifting

\[ I(\bar{x}) = \delta(\bar{x} - \Delta \bar{x}) \ast S(\bar{x}) \]
Amplitude spectrum is invariant to shift, but excessively so.

\[ I(\tilde{x}) \]

Structural information is contained in phase.
Factorization is required to extract it.

\[ |\tilde{I}(\omega)| \]

\[ \angle \tilde{I}(\omega) \]
Reducing the Dimensionality of Data with Neural Networks

G. E. Hinton* and R. R. Salakhutdinov

Fig. 1. Pretraining.

Fig. 2. Unrolling.

Fig. 3. Fine-tuning.

Decoding:

Encoder:

$$E(v, h) = - \sum_{i \in \text{pixels}} b_i v_i - \sum_{j \in \text{features}} b_j h_j - \sum_{i,j} v_i h_j w_{ij}$$

$$\Delta w_{ij} = \varepsilon (\langle v_i h_j \rangle_{\text{data}} - \langle v_i h_j \rangle_{\text{recon}})$$
Application to hand-written digits

2D PCA

784-1000-500-250-2 autoencoder
Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations

Honglak Lee
Roger Grosse
Rajesh Ranganath
Andrew Y. Ng

Computer Science Department, Stanford University, Stanford, CA 94305, USA

Convolutional Deep Belief Networks (convolutional RBM) with probabilistic max-pooling-CRBM as follows:

The detection units in the block Gibbs sampling using the following conditional

\[ E_{\theta}(h_{ij}) = \log \sum_{v} \exp \left( \sum_{k} \sum_{c} W_{k}(h_{ij}) v_{c} \right) \]

\[ E_{\phi}(v) = \sum_{c} \sum_{k} \sum_{ij} W_{k}(h_{ij}) v_{c} \]

where \( W \) is a learnable weight matrix, \( h \) is a hidden group, and \( v \) is a visible unit.

The detection layer corresponds to a simplified structure with only visible layer and one value for each of the detection units being on, and forces the following constraints: at most one of the detection units is on, which applies the same information elsewhere. Thus, our model can even when applied to large images. Second, objects must scale gracefully and be computationally tractable.

Deep architectures consist of feature detector units arranged in groups of units, and each group has the same information elsewhere. Thus, our model can even when applied to large images. Second, objects must scale gracefully and be computationally tractable.

While DBNs have been successful in controlled domains such as handwritten digits (Hinton et al., 2006) and human motion (Taylor et al., 2007). We build upon the structure in a wide variety of domains, including handwriting recognition, object recognition, and scene understanding.

We formally define the energy function of this simplified structure with only visible layer and one value for each of the detection units being on, and forces the following constraints: at most one of the detection units is on, which applies the same information elsewhere. Thus, our model can even when applied to large images. Second, objects must scale gracefully and be computationally tractable.
Figure 3. Columns 1-4: the second layer bases (top) and the third layer bases (bottom) learned from specific object categories. Column 5: the second layer bases (top) and the third layer bases (bottom) learned from a mixture of four object categories (faces, cars, airplanes, motorbikes).

filling-in
Hierarchical Bayesian inference in visual cortex
(Lee & Mumford, 2003)

\[ P(x_0 | x_1) \cdot P(x_1 | x_2) / Z_1 \]
\[ P(x_1 | x_2) \cdot P(x_2 | x_3) / Z_2 \]
\[ P(x_2 | x_3) \cdot P(x_3 | x_4) / Z_3 \]

image data
\[ x_0 \]
\[ x_1 \]
\[ x_2 \]
\[ x_3 \]
\[ \ldots \]

\[ P(x_1 | x_2) \]
\[ P(x_2 | x_3) \]
\[ P(x_3 | x_4) \]

‘V1’
‘V2’
‘V3’
What do you see?

How do neurons in V1 encode this?

(easy version)
BOLD signal in V1 and LOC
How to form invariant object representations?
Reference frame effects in perception

Diamond or square?
Reference frames require *structured representations*

The meaning of the triangular symbol in fig. 1 is quite complex. It stands for two rules:

1. Multiply the activity level in the retina-based unit by the activity level in the mapping unit and send the product to the object-based unit.

2. Multiply the activity level in the retina-based unit by the activity level in the object-based unit and send the product to the mapping unit.

Hinton (1981)
Dynamic routing
(Olshausen, Anderson, Van Essen 1993)

- Control
- Saliency

Analysis/Recognition

Feature vector

Object-centered reference frame (position and scale invariant)

Window of Attention

High level cortical areas

Early/intermediate cortical areas (form processing)

Retina
Dynamic routing circuit

a. [Diagram showing the routing circuit with output nodes labeled as \( l=3 \) \( N=5 \), \( l=2 \) \( N=21 \), \( l=1 \) \( N=29 \), and \( l=0 \) \( N=33 \).]

b. [Diagram showing the window of attention for layers at different scales.]

c. [Diagram showing the window of attention for layers at different scales.]
Dynamic routing in deep networks

Layer Above Reconstruction
Max Unpooling
Unpooled Maps
Rectified Linear Function
Rectified Unpooled Maps
Convolutional Filtering \{F^1\}
Reconstruction

Pooled Maps
Max Pooling
Rectified Feature Maps
Rectified Linear Function
Feature Maps
Convolutional Filtering \{F\}
Layer Below Pooled Maps

Unpooling
Max Locations “Switches”
Unpooled Maps
Rectified Feature Maps
Pooled Maps
Pooling

(Zeiler & Fergus, 2013)
Figure 8. Visualization of features in a fully trained model. For layers 2-5 we show the top 9 activations in a random subset of feature maps across the validation data, projected down to pixel space using our deconvolutional network approach. Our reconstructions are not samples from the model: they are reconstructed patterns from the validation set that cause high activations in a given feature map. For each feature map we also show the corresponding image patches. Note: (i) the strong grouping within each feature map, (ii) greater invariance at higher layers and (iii) exaggeration of discriminative parts of the image, e.g. eyes and noses of dogs (layer 4, row 1, cols 1). Best viewed in electronic form.
### Bilinear models

(Tenenbaum & Freeman 2000)

<table>
<thead>
<tr>
<th>A</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>B</td>
<td>C</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B</th>
<th>Extrapolation</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>C</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>A</td>
<td>B</td>
</tr>
</tbody>
</table>

| ? | ________ | ? | F | G | H |

### Equation

\[ y_{k}^{sc} = \sum_{i=1}^{I} \sum_{j=1}^{J} w_{ijk} a_{i}^{s} b_{j}^{c} \]
Transforming Auto-encoders
(Hinton, Krizhevsky & Wang 2011)

Consider the feedforward neural network shown in figure 1. The network is

\[ \frac{p}{x} + \frac{\Delta x}{y} \]

...
Dynamic routing between capsules
(Sabour, Frosst & Hinton 2017)

This type of “routing-by-agreement” should be far more effective than the very primitive form of routing implemented by max-pooling, which allows neurons in one layer to ignore all but the most active feature detector in a local pool in the layer below. We demonstrate that our dynamic routing mechanism is an effective way to implement the “explaining away” that is needed for segmenting highly overlapping objects.

For low level capsules, location information is “place-coded” by which capsule is active. As we ascend the hierarchy, more and more of the positional information is “rate-coded” in the real-valued components of the output vector of a capsule.
Dynamic routing between capsules
(Sabour, Frosst & Hinton 2017)

\[ c_{ij} \sim e^{\hat{u}_{j|i} \cdot v_j} \]

\[ v_j = \frac{||s_j||^2}{1 + ||s_j||^2} \frac{s_j}{||s_j||} \]

\[ s_j = \sum_i c_{ij} \hat{u}_{j|i} \]

\[ \hat{u}_{j|i} = W_{ij} u_i \]
My idea

\[ I(\bar{x}, t) = \sum_{\bar{x}'} T(\bar{x}, \bar{x}'; c(t)) I_0(\bar{x}) \]

\[ T(\bar{x}, \bar{x}'; c(t)) = 1 + \sum_{k} c_k(t) \psi_k(\bar{x}, \bar{x}') \]