The Operating System of Vision
Recent Advances in General AI

Xavier Boix
Computer Vision Recap
What is this?
What is this?

a cat
Convolutional Neural Networks

from LeCun et al. 98
Learning

What is in the image?

- Bike
- Train
- Bird
Learning to recognise objects

Training on images + ground truth given

Testing on images never shown before
Learning Paradigm

Error in prediction

Backpropagation:
correction of the synaptic weights in the direction leading to lower prediction error

The true image category is needed for each training image!
The Revolution of Depth

- **VGG (2014)** - 4.8%
  - Baidu (2015) - 5.33%
  - Spatial pyramid pooling
  - Optimized PReLU
  - Improved (random) initialization

- **GoogleNet (2014)** - 6.67%
  - Inception module
  - Multi-scale convolutions (including 1x1 filters)
  - Minimal dense layers
  - Auxiliary classifiers

- **AlexNet (2012)** - 15.3%
  - Clarifai (2013) - 11.7%

- **MSRA (2015)** - 4.94%
from Zeiler & Fergus, ECCV 13
General AI
How many chairs are at the table?

Is there a pedestrian in my lane?

Is the person with the blue hat touching the bike in the back?

Is there a matte cube that has the same size as the red metal object?

What color is the thing with the same size as the blue cylinder?
What color is the thing with the same size as the blue cylinder?

1. Any task

2. Sequential tasks

What color is the thing with the same size as the blue cylinder?

find blue cylinder \[ \rightarrow \] compare size \[ \rightarrow \] describe color

The visual OS
what color is the thing with the same size as the blue cylinder?

find blue cylinder → compare size → describe color

how many things are the same size as the ball?

find ball → compare size → count

What color is the thing with the same size as the blue cylinder?

find blue cylinder

compare size

describe color

green

How many things are the same size as the ball?

find ball

compare size

count

four

Does the blue cylinder have the same material as the big block on the right side of the red metallic thing?
What is behind the foot of the bed?
Neural Module Networks (and variants)
Question

Network layout (Section 4.1)

find[city] → and → relate[in] → lookup[Georgia]

Module inventory (Section 4.2)

lookup  relate

and  find

Atlanta

find city

relate in

lookup Georgia

Question

Network layout (Section 4.1)

find[city] \rightarrow and \rightarrow relate[in] \rightarrow lookup[Georgia]

Atlanta

Module inventory (Section 4.2)

lookup \hspace{1cm} relate

and \hspace{1cm} find

Deep nets can’t solve all problems

- **Inputs and outputs must be of fixed dimensionality**
  - Great for images: input is a big image of a fixed size, output is a 1-of-N encoding of category

- **Bad news for machine translation and speech recognition**
Recurrent Neural Networks (RNNs)

- RNNs can work with sequences

Key idea: each timestep is a different layer with the same weights

Recurrent Neural Networks (RNNs)

- Neural networks that can process sequences well
  - Very expressive models

- Use backpropagation
  - Fun fact: recurrent neural networks were trained in the original backpropagation paper in 1986

- Sadly RNNs are hard to train with backpropagation
  - Unstable
  - Has trouble learning “long-term dependencies”
  - Vanishing gradient problems (Hochreiter 1991; Bengio et al., 1994)

- There are ways to learn RNNs but they are hard to use

The heart of the LSTM

- RNNs overwrite the hidden state
- LSTMs add to the hidden state

- Addition has nice gradients
  - All terms in a sum contribute equally

- LSTM is good at noticing long-range correlations
  - Because of the nice gradients of addition

from Sutskever, et al. "Sequence to sequence learning with neural networks." *NIPS 2014*
Language translation with LSTMs

Question: Are there more cubes than yellow things?  
Answer: Yes

Program Generator

Predicted Program

Execution Engine

What color is the bird?

Attention

\[ \text{attend} : \text{Image} \rightarrow \text{Attention} \]
Re-attention

re-attend: Attention $\rightarrow$ Attention
Combination

\[
\text{combine}: \text{Attention} \times \text{Attention} \rightarrow \text{Attention}
\]
Classification

classify : Image × Attention → Label
Measurement

\[ \text{measure: } \text{Attention} \rightarrow \text{Label} \]
Question

Network layout (Section 4.1)

find[city] and relate[in] lookup[Georgia]

Module inventory (Section 4.2)

lookup relate and find

Atlanta

and

find city relate in lookup Georgia

What color is his tie?

from Andreas et al., “Neural Module Networks” CVPR 16
Is there a red shape above a circle?
<table>
<thead>
<tr>
<th>Question</th>
<th>Code</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>How many different lights in various different shapes and sizes?</td>
<td>measure<a href="attend%5Blight%5D">count</a></td>
<td>four (four)</td>
</tr>
<tr>
<td>What is the color of the horse?</td>
<td>classify<a href="attend%5Bhorse%5D">color</a></td>
<td>brown (brown)</td>
</tr>
<tr>
<td>What color is the vase?</td>
<td>classify<a href="attend%5Bvase%5D">color</a></td>
<td>green (green)</td>
</tr>
</tbody>
</table>

from Andreas et al., “Neural Module Networks” CVPR 16
<table>
<thead>
<tr>
<th>what is stuffed with toothbrushes wrapped in plastic?</th>
<th>where does the tabby cat watch a horse eating hay?</th>
<th>what material are the boxes made of?</th>
</tr>
</thead>
<tbody>
<tr>
<td>classify<a href="attend%5Bstuff%5D">what</a></td>
<td>classify<a href="attend%5Bwatch%5D">where</a></td>
<td>classify<a href="attend%5Bbox%5D">material</a></td>
</tr>
<tr>
<td>container (cup)</td>
<td>pen (barn)</td>
<td>leather (cardboard)</td>
</tr>
</tbody>
</table>

from Andreas et al., “Neural Module Networks” CVPR 16
There is more, much more...
Using an external memory

Illustration of the DNC architecture