Scene Understanding Without Labels

S. M. Ali Eslami

February 2017
Learning paradigms

Supervised Learning

horse

\[ x \rightarrow z \rightarrow y \]
Learning paradigms

Supervised Learning

Reinforcement Learning
Learning paradigms

- **Supervised Learning**
- **Reinforcement Learning**
- **Generative Modelling**
Learning paradigms

- **Supervised Learning**
- **Reinforcement Learning**
- **Generative Modelling**

Diagram:
- Horse image labeled as "not blinking" with coordinates (2.3, -1, 0.5, 3)
- Left image labeled as "not blinking" with coordinates (2.3, -1, 0.5, 3)
- Eye image labeled as "not blinking" with coordinates (2.3, -1, 0.5, 3)
Highly structured

General Purpose Graphics Programming
Vikash Mansinghka, Tejas D. Kulkarni, Yura N. Perov, and Joshua B. Tenenbaum (2013)
Partially structured

A Stochastic Grammar of Images
Song-Chun Zhu and David Mumford (2007)
Partially structured

Fully unstructured

Antti Rasmus et al. (2016)  
Jeff Donahue et al. (2016)
Recurrent Neural Networks for Image Generation

Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, Daan Wierstra (2015)
Variational Inference

Approximate $p(z|x)$ using $q(z|x)$

Parameterise $q(z|x)$ by deep network

Minimise $\text{KL}[ q(z|x) \mid p(z|x) ]$ via SGD
Recurrent Neural Networks for Image Generation

\[ p(x|c) \]

Encoding Inference

Decoding Generation

Hinton (2006)
Recurrent Neural Networks for Image Generation

Hinton (2006)

Gregor et al. (2015)
Recurrent Neural Networks for Image Generation
Recurrent Neural Networks for Image Generation
Towards Conceptual Compression

Karol Gregor, Frederic Besse, Danilo Rezende, Daan Wierstra (2016)
Write $z$

Read $x$

Write $z$

Write $z$

Write $z$

$p(x|c_T)$
76 bits

Original raw image:
24576 bits
112 bits
2364 bits
Attend, Infer, Repeat: Fast Scene Understanding with Generative Models

Attend, Infer, Repeat
blue brick
Cause

Model

Image

blue brick

pile of bricks

not sufficient for grasping
counting
transfer
generalisation
Model

Cause

pile of bricks

blue brick  red brick

Image

x

z

x

z^1  z^2
<table>
<thead>
<tr>
<th>Cause</th>
<th>Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>pile of bricks</td>
<td>blue brick above red brick below</td>
</tr>
</tbody>
</table>

```
blue brick

red brick

blue brick above
red brick below
```

```
x

z

x

z'1

z'2

y'1

y'2

y1

y2
```
Decoder

Model

Inference Network
Recap: Key ideas

1. Build in structure
   Get out meaning

2. Inference networks that are
   a. recurrent
   b. variable-length
   c. attentive

3. End-to-end learning through
   a. discrete, continuous vars
   b. inference and model nets
Demo reel

Scanning policy is left-to-right
OmniGlot
Representational power

Sum?

\[
\begin{array}{c}
5 & \rightarrow & 6 \\
2 & \rightarrow & 9 \\
\end{array}
\]

Increasing order?

\[
\begin{array}{c}
5 & \rightarrow & \text{no} \\
2 & \rightarrow & \text{yes} \\
\end{array}
\]
Additional structure

distributed vector
that correlates
with blue brick

learned
Additional structure

distributed vector that correlates with blue brick

class=brick
colour=blue
position=P
rotation=R
Additional structure
Inverse graphics
Policy learning

MNIST

Table-top
Unsupervised Learning of 3D Structure from Images

Danilo Rezende, S. M. Ali Eslami, Shakir Mohamed, Peter Battaglia, Max Jaderberg, Nicolas Heess (2016)
Motivation

How to recover 3D structure from 2D images?
To form stable representations, regardless of camera position
How to recover **3D structure** from **2D images**?

To form **stable** representations, regardless of camera position

- Inherently ill-posed
  - All objects appear under self occlusion, infinite explanations
  - Therefore build statistical models to know what’s likely and what’s not

- Even with models, inference is intractable
  - Important to capture multi-modal explanations

- How are 3D scenes best represented?
  - Meshes or voxels?

- Where is training data collected from?
Unsupervised Learning of 3D Structure from Images
Unsupervised Learning of 3D Structure from Images
Unsupervised Learning of 3D Structure from Images

Projection operators
Unsupervised Learning of 3D Structure from Images

Unconditional samples
Unsupervised Learning of 3D Structure from Images

Unconditional inference
Unsupervised Learning of 3D Structure from Images

Class-conditional samples

- chair
- table
- vase
- car
- laptop
- airplane
- plant
- bowl
- person
- cone
Unsupervised Learning of 3D Structure from Images

Multi-modality of inference
Unsupervised Learning of 3D Structure from Images

3D structure from multiple 2D images
Unsupervised Learning of 3D Structure from Images

Inferring object meshes
Unsupervised Learning of 3D Structure from Images

Inferring object meshes
Recap

- Deep Supervised Learning
- Deep Reinforcement Learning
- Model-based Methods
- Deep Variational Inference
- Structured / Unstructured Generative Models
Questions

aeslami@google.com
arkitus.com