Attend, Infer, Repeat: Fast Scene Understanding with Generative Models

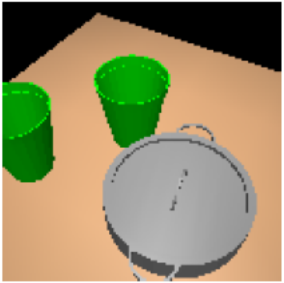
S. M. Ali Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, Koray Kavukcuoglu, Geoffrey E. Hinton

Ryan Dick

CSC2547

February 9, 2018

Motivation



Scenes naturally decompose into *objects* that...

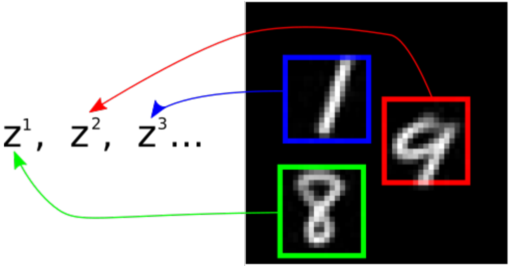
- Are arranged in space
- Have visual properties
- Have physical properties
- Have functional relationships with each other

High-Level Approach

- Generative model
  - Goal is good representations not reconstructions
- Partly-specified latent structure
  - Must have structure without being overly rigid

Main Contribution

- **Variable dimensionality** of latent space (list of vectors)



- Treats inference as an **iterative** process, using an RNN to attend to one object at a time
- Learn the appropriate number of iterative steps (and thus the appropriate number of object latent variable representations)

A Bayesian Approach

$$p(\mathbf{z}|\mathbf{x}) = \frac{p_\theta^x(\mathbf{x}|\mathbf{z})\, p_\theta^z(\mathbf{z})}{p(\mathbf{x})}$$

Given image $\mathbf{x}$ and model $p_\theta^x(\mathbf{x}|\mathbf{z})\, p_\theta^z(\mathbf{z})$, we want to recover the underlying scene description, $\mathbf{z}$, by calculating $p(\mathbf{z}|\mathbf{x})$.

- $p_\theta^z(\mathbf{z})$ captures our model's assumptions about the underlying scene
- $p_\theta^x(\mathbf{x}|\mathbf{z})$ models how an image is generated from a scene description

Handling a Variable-Length Scene Descriptor

- Assume that $\mathbf{z}^i$ is a group of variables that describes (type, appearance, pose, etc.) a single object in the scene
- $\mathbf{z}$ then becomes a latent, variable-length, scene descriptor, $\mathbf{z} = \left(\mathbf{z}^1, \mathbf{z}^2, ..., \mathbf{z}^n\right)$
- Since the number of objects in the scene will vary, we assume the following:

$$p_\theta(\mathbf{x}) = \sum_{n=1}^{N} p_N(n) \int p_\theta^z\left(\mathbf{z}|n\right) p_\theta^x\left(\mathbf{x}|\mathbf{z}\right) d\mathbf{z}$$

- But... this is **intractable**

Inference

- Let's learn $q_\phi(\mathbf{z}, n|\mathbf{x})$ an approximation of the true posterior that minimizes the divergence $\mathrm{KL}[q_\phi(\mathbf{z}, n|\mathbf{x})||p_\theta^z(\mathbf{z}, n|\mathbf{x})]$
- Two new challenges:
  - **Trans-Dimensionality**: the size of the latent space, n, is a random variable itself
  - **Symmetry**: symmetry arises from different assignments of objects to $\mathbf{z}^i$

Inference (cont'd)

- Overcome these challenges by formulating inference as an **iterative** process performed by an RNN
- To simplify, parameterize the number of objects, n, as a variable length vector, $\mathbf{z}_{\text{pres}}$, consisting of n ones followed by a single zero.

$$q_\phi(\mathbf{z}, \mathbf{z}_{\text{pres}} | \mathbf{x}) = q_\phi(z_{\text{pres}^{n+1}} = 0 | \mathbf{z}^{1:n}, \mathbf{x}) \prod_{i=1}^{n} q_\phi(\mathbf{z}^i, z^i_{\text{pres}} = 1 | \mathbf{z}^{1:i-1}, \mathbf{x})$$

Learning

Can now jointly optimize the parameters θ of the model and ϕ of the inference network by maximizing a lower bound on the marginal likelihood of an image under the model:

$$\log p_\theta(\mathbf{x}) \geq \mathcal{L}(\theta, \phi) = \mathbb{E}_{q_\phi}\left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z}, n)}{q_\phi(\mathbf{z}, n|\mathbf{x})}\right]$$

AIR Implementation

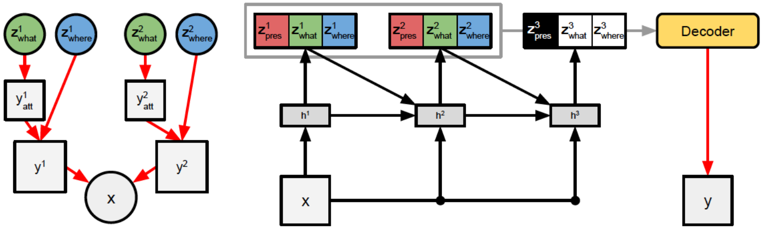


Figure: *Left:* Assumed generative model. *Right:* AIR inference model.

AIR Implementation (cont'd)

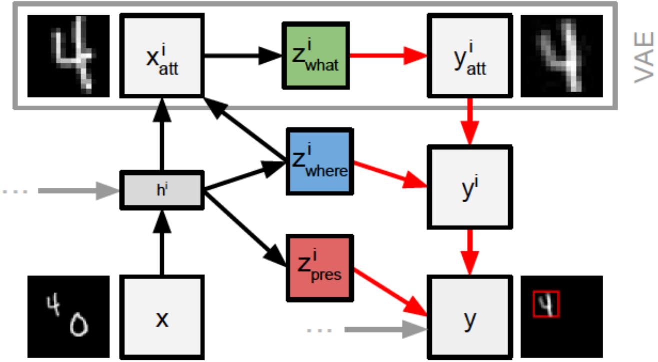


Figure: Interaction between inference and generative models.

A slight variation: DAIR

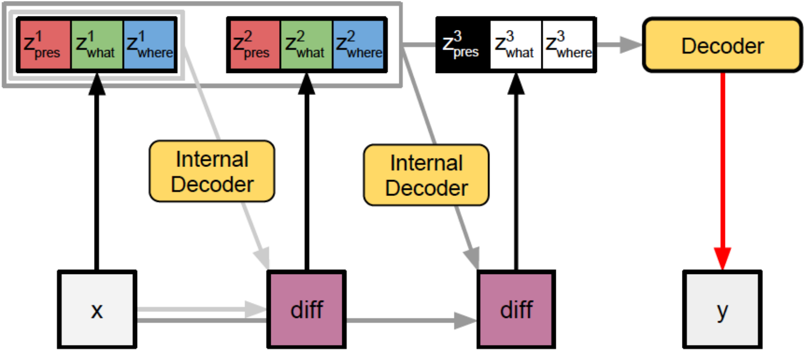


Figure: The difference-AIR (DAIR) model

Evaluation: Multi-MNIST

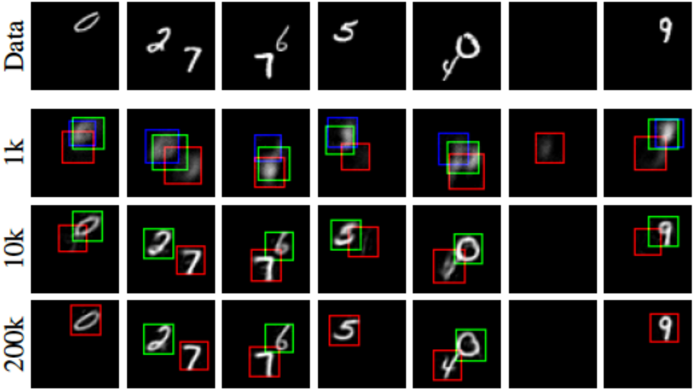
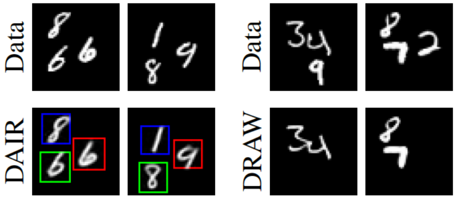


Figure: Multi-MNIST results with attention windows shown

Evaluation: Multi-MNIST Generalization



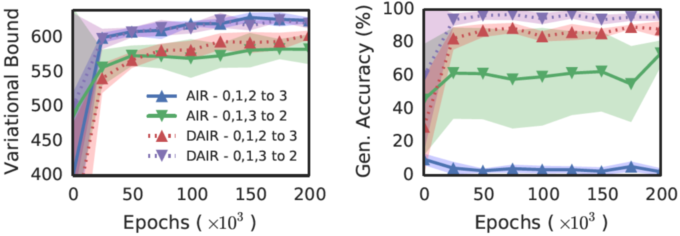
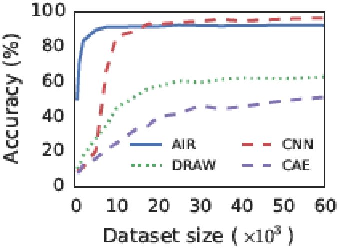


Figure: Generalization to numbers of digits not seen in training

Evaluation: Representational Power



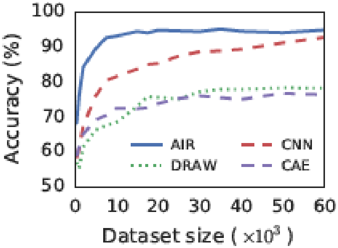


Figure: *Left:* Predicting sum of two digits. *Right:* Determine if digits appear in ascending order.

An extension: 3D Scenes

- Replace generative network with a 3D graphics renderer
- $\mathbf{z}_{\mathrm{what}}$ becomes a discrete variable identifying the object from a small set of possibilities
- $\mathbf{z}_{\mathrm{where}}$ now represents position and orientation

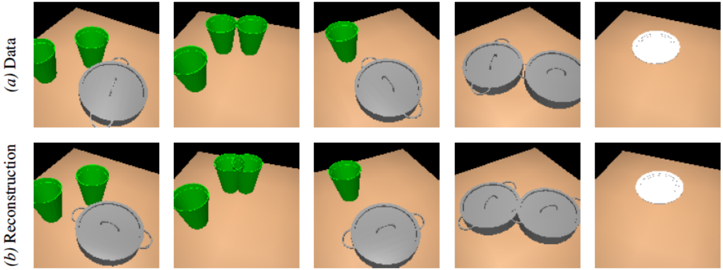


Figure: 3D reconstruction samples

Takeaways

- Model structure can provide an inductive bias that results in interpretable latent representations
- Variable-sized latent spaces can be achieved through iterative inference that learns when to 'stop'

References

Eslami, S. M. A., Heess, N., Weber, T., Tassa, Y., Kavukcuoglu, K., and Hinton, G. E. (2016).
Attend, infer, repeat: Fast scene understanding with generative models.
CoRR, abs/1603.08575.

Gregor, K., Danihelka, I., Graves, A., and Wierstra, D. (2015).
DRAW: A recurrent neural network for image generation.
CoRR, abs/1502.04623.