- Attend Infer Repeat: Fast Scene Understanding with Generative Models
- S. M. Ali Eslami, Geoffrey E. Hinton

Assumption

Here we assume that, naturally, a scene with multiple objects are generated as follows:

- The number of objects n is drawn from $p(n)$
- For $i = 1, \ldots, n$, determine z^i from $p_\theta(z)$
- Generate the scene using $p_\theta(x|z)$.

That is

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

Inference

Two difficulties:

- Trans-dimensionality: the number of z^i's is itself a random variable
- Symmetry: z^i should be permutation-invariant

This is resolved with recurrent neural networks. First, we will denote z_{pres} as indicator for n. For given n, z_{pres} is a vector of n 1's, followed by 0's. Given this, we can model q_ϕ as

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

Several notes here:

- The condition part should really include $z_{prse}^i = 1$. But since this is always true, we can just omit this during modeling.
- In essence, we are assuming an infinite number of z^i and z_{pres}^i.
- Conditioning on $z^{1:i-1}, x$ is modeled with hidden states of the RNN.

Learning

Just trivial. Different gradient estimation for discrete and continuous variables.

Models and Experiments

First, for 2D experiments, there are three types of z:

- z_{pres}^i: presence of object i
- z_{where}^i: 3-D, position and scale
- z_{what}^i: identity

Here we must specify two things:

- The exact form of $p(x|z)$
- The exact form of $q(z^i|x, z^{1:i-1})$.

For the first, we assume that at each time step, a y^i is generated, and they are summed to x. Each y^i is generated as follows:

- from z^i_{what}, we generate the digit y^i_{att}
- from z^i_{where} and y^i_{att}, we generate the component y^i.

Inference goes in the opposite direction. This is best illustrated with this figure:

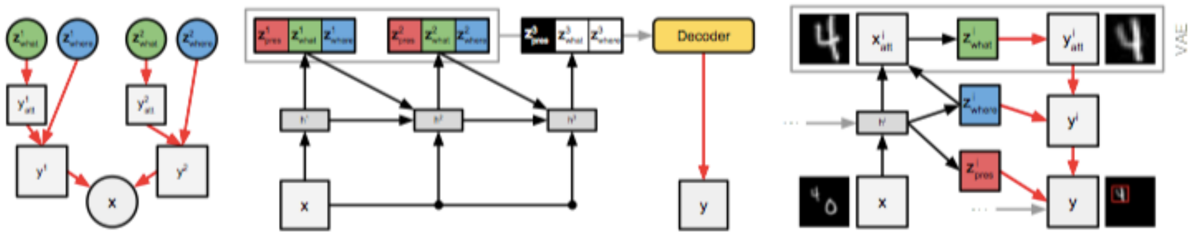

Figure 2: **AIR in practice:** *Left:* The assumed generative model. *Middle:* AIR inference for this model. The contents of the grey box are input to the decoder. *Right:* Interaction between the inference and generation networks at every time-step. In our experiments the relationship between $\mathbf{x}^i_{\text{att}}$ and $\mathbf{y}^i_{\text{att}}$ is modeled by a VAE, however any generative model of patches could be used (even, e.g., DRAW).

Experiments:

- Multi-MNIST: correctly infers the number of digits
- Strong generalization: interpolation
- Representation power: for downstream tasks
- 3D scenes: when the generative model is specified using a differential renderer, this network can be used to infer the pose and identity of the objects.