Attend, Infer, Repeat: Fast Scene Understanding with Generative Models arXiv:1603.08575v1

Eslami SMA, Heess N, Weber T, Tassa Y, Kavukcuoglu K, Hinton GE. Attend, Infer, Repeat: Fast Scene Understanding with Generative Models. 2016.
http://arxiv.org/abs/1603.08575.

Presented by: Namdar Homayounfar

April 5, 2016

Motivation — Probabilistic Scene Understanding Systems

- The goal is to produce high probability descriptions of scenes conditioned on observed images and videos.
- Given an image, what objects are present, where are they, their relative positions with respect to each other, depth of an object, 3d bounding box, . . .
- Tackled via discriminative or generative approaches.

Motivation — Discriminative Approaches

- Model the dependence of an unobserved variable y on an observed variable x.
- Doesn't have to be probabilistic: SVMs, Decision Trees, Neural Nets, ...
- Or it could be probabilistic: Model $p(y|x)$, Logistic regression, conditional random fields, ...
- Optimization problem: $f(x) = \arg\max_y p(y|x)$. Cannot generate samples
- Examples in vision: Deformable Parts Models, Convolutional Nets

Motivation — Discriminative Approaches in vision, DPM

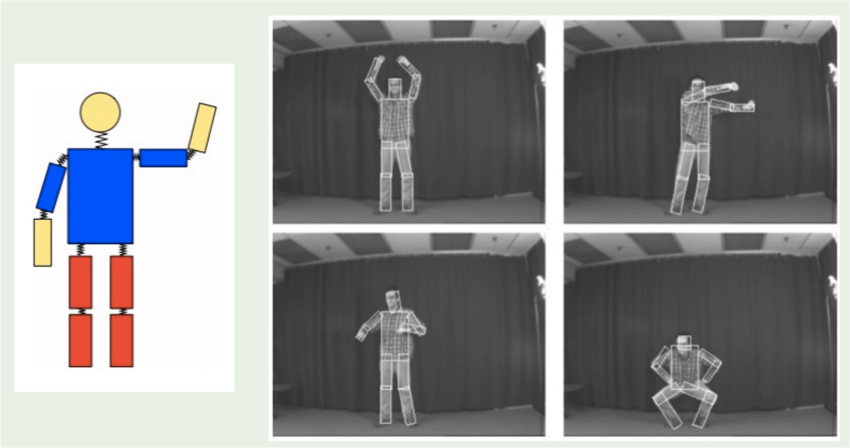

Figure : DPM, http://vision.stanford.edu/teaching/cs231b_spring1213/slides/dpm-slides-ross-girshick.pdf

Motivation — Discriminative Approaches in vision, CNN

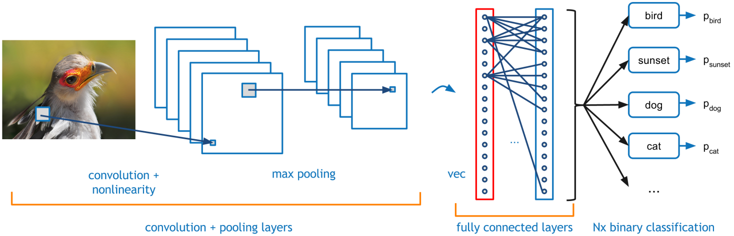

Figure : CNN, `http://code.flickr.net/2014/10/20/introducing-flickr-park-or-bird/`

Motivation — Generative Approaches

- Specify a joint probability distribution $p(x, y)$ over the observed and latent variables:

$$y \sim p(y)$$
$$x|y \sim p(x|y)$$

- Bayes Rule:

$$p(y|x) = \frac{p(x|y)p(y)}{\int p(x|yp(y)dy}$$

- Allows sampling of any variables in the model.

- Focus of this paper

Vision as Inverse Graphics

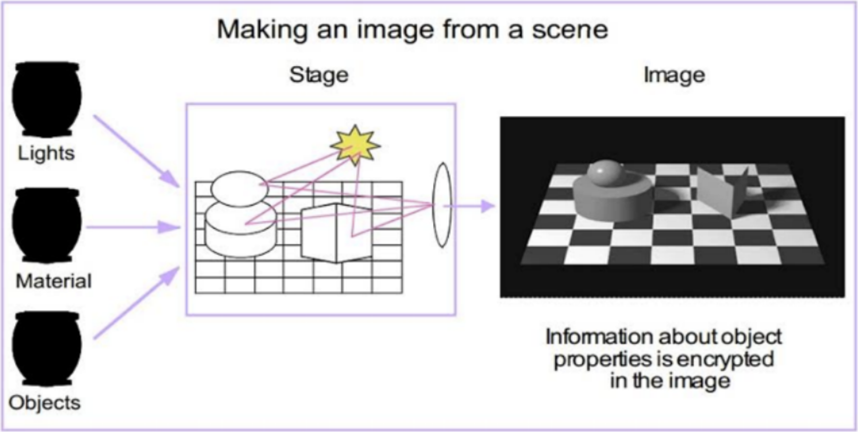

Figure : Kulkarni,
http://cs.wellesley.edu/~vision/slides/tejask_fall2015.pdf

Motivation — Generative vs. Discriminative modeling in Vision

- Discriminative models: Fast bottom up inference methods, data intensive training. Have been very successful in recognition tasks
- Generative Models: Hold the promise of analyzing complex scenes more richly and flexibly by obtaining a joint probability distribution over everything variable in the scene. However they have been less accepted.

Motivation — Challenges of Generative Models and A Solution

Challenges

- In practice, hard to define expressive models that capture the complexity of the scene
- Hard to define models that are subject to tractable inference

Contributions of this Paper:

Model Structure:

- A scene is formed by a variable number of entities, different locations

Efficient Inference:

- Treat Inference as an **iterative** process implemented as an RNN that
  - **attends** to one object at a time
  - **learns** to use an **appropriate number** of inference steps for each image

Approach: Scene Interpretation as Inference in a Generative Model

- Given an image $\mathbf{x}$
- Sample the number of objects n from a prior - max N
- Sample from a scene model

$$\mathbf{z} = (\mathbf{z}^1, \ldots, \mathbf{z}^n) \sim p_\theta(\mathbf{z}|n)$$

- $\mathbf{z}^i = (\mathbf{z}^i_{where}, \mathbf{z}^i_{what})$
- Scene description rendered to form an image:

$$\mathbf{x}|\mathbf{z} \sim p_\theta(\mathbf{x}|\mathbf{z})$$

- Goal: Recover underlying scene description $\mathbf{z}$ by computing the posterior:

$$p_\theta(\mathbf{z}, n|\mathbf{x}) = \frac{p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})p_N(n)}{p_\theta(\mathbf{x})}$$

Approach: Inference

- Variational Auto Encoder: Approximate $p_\theta(\mathbf{z}, n|\mathbf{x})$ by $q_\phi(\mathbf{z}, n|\mathbf{x})$

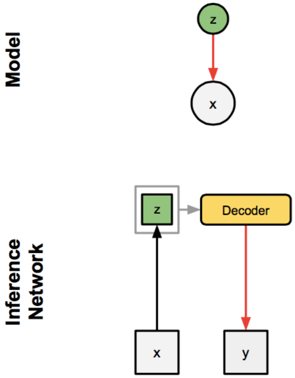

Figure : Eslami et al. 2016

Approach: Inference

- Variational Auto Encoder: Approximate $p_\theta(\mathbf{z}, n|\mathbf{x})$ by $q_\phi(\mathbf{z}, n|\mathbf{x})$

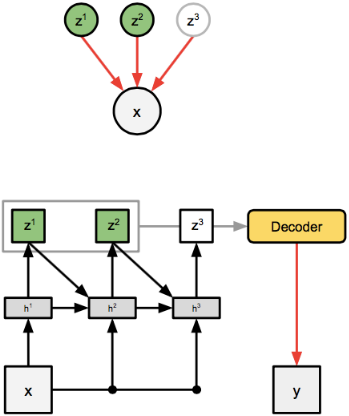

Figure : Eslami et al. 2016

Approach: Learning

- Maximize the ELBO

$$\log p_\theta(\mathbf{x}) \geq E_{q_\phi}\Big[\log \frac{p_\theta(\mathbf{x}, \mathbf{z}, n)}{q_\phi(\mathbf{z}, n|\mathbf{x})}\Big]$$

w.r.t ϕ and θ.

Some Videos