Overview

Generating scenes from rich and complex semantics is an important step towards understanding the visual world. To generate scenes from complex semantic descriptions, we develop a flexible latent distribution built from reusable, modular components.

Our Method: Probabilistic Neural Programmed Networks (PNP-Net)

- Modularity
- Probabilistic
- Compositional
- Reusability

Problem Setting: Rich semantics to complex scenes

Semantics with different forms:
- sentence
- synset
- scene graph

Typical semantics contain:
- Attributes
- Relations

Challenges: Complex scenes contain varied visual elements, compositional visual concepts, and complicated relations between objects, which makes the generation highly challenging.

- An ideal generative model should encompass three core components:
  - An interpretable generation process that composes complex scenes
  - What are the primitive visual elements and how are they composed
  - The rendering of abstract concepts into their pixel-level realizations

- Monolithic generative models can generate images of single objects well, but have difficulties handling more complex scenes.

Contributions

We propose Probabilistic Neural Programmed Networks (PNP-Net), a programmatic generative model framework for scene generation. Specifically, we:
- Construct a set of visual elements and neural modules/programs for modeling the appearance of these visual elements
- Integrate our probabilistic neural modules into the canonical VAE framework and generate the VAE by empowering it with reusable, composable, interpretable modules
- Demonstrate generalization ability for complicated scene understanding, including zero-shot learning of novel semantic compositions.

PNP-Net: Probabilistic operators

Concept mapping operator

Aggregation operator

PNP-Net: Overall Formulation

We integrate our probabilistic neural modules into the canonical VAE framework, the reusable neural modules compose the prior of the images conditioned by given semantics:

\[ p_{\theta}(\mathbf{z}|\mathbf{y}) = \prod_{i} p_{\theta}(\mathbf{z}_{i}|\mathbf{y}) \]

\[ p_{\theta}(\mathbf{x}|\mathbf{y}) = \prod_{i} p_{\theta}(\mathbf{x}_{i}|\mathbf{z}_{i}, \mathbf{y}) \]

\[ p_{\theta}(\mathbf{y}|\mathbf{x}) = \prod_{i} p_{\theta}(\mathbf{y}_{i}|\mathbf{z}_{i}, \mathbf{x}_{i}) \]

Notes:

- \( \mathbf{y} \): some image
- \( \mathbf{y} \): semantics
- \( \mathbf{y} \): some latent
- \( \mathbf{y} \): data distribution
- \( \mathbf{y} \): prior
- \( \mathbf{y} \): random generation
- \( \mathbf{y} \): compositional prior
- \( \mathbf{y} \): generalization process

Datasets

- an variant of MINC
- an image dataset of at least 2 dimensions, possibly with scene boundaries
- digit with different color, size, and type
- CLEVR with monolithic, composable neural modules

More Details

Please check our project page on Github for more details about model implementation and data.

References