DeepMind

Constellation: Learning relational abstractions over objects for compositional imagination

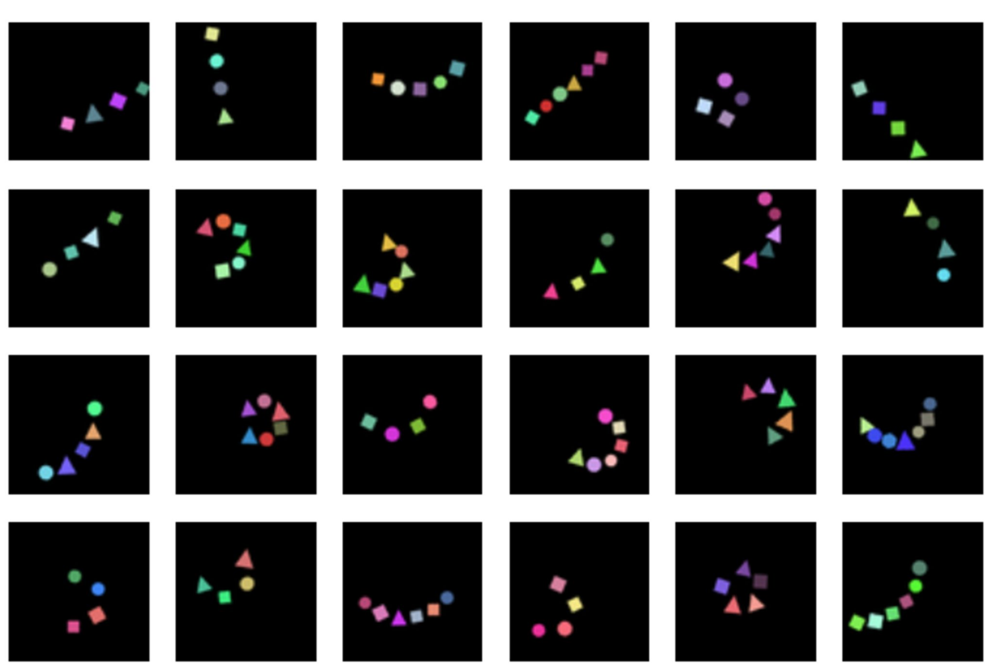
James Whittington, Rishabh Kabra, Loic Matthey, Christopher P. Burgess , Alexander Lerchner

Introduction

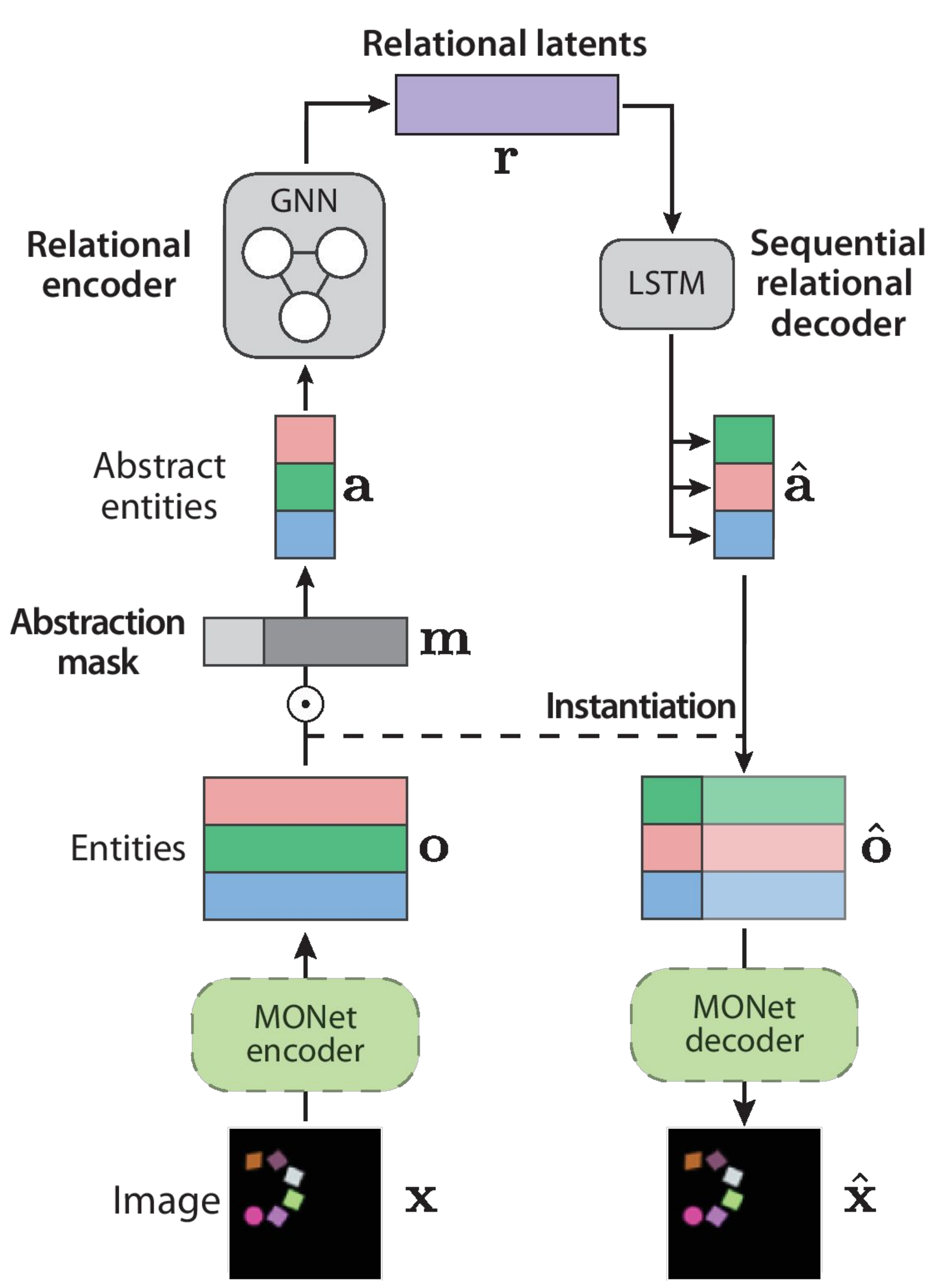
Learning structured representations of visual scenes is currently a major bottleneck to bridging perception with reasoning. While there has been exciting progress with slot-based models, which learn to segment scenes into sets of objects, learning configurational properties of entire groups of objects is still under-explored. To address this problem, we introduce Constellation, a network that learns relational abstractions of static visual scenes, and generalises these abstractions over sensory particularities, thus offering a potential basis for abstract relational reasoning. We further show that this basis, along with language association, provides a means to imagine sensory content in new ways. This work is a first step in the explicit representation of visual relationships and using them for complex cognitive procedures.

Dataset

Random objects in 'super-structures', where the super-structure is defined by generative factors of #objects, length, curviness, position



Constellation architecture



Model features

Use MONet to provide slot based latents for each object

Only tries to reconstruct subset of object latents - governed by learnable abstraction mask

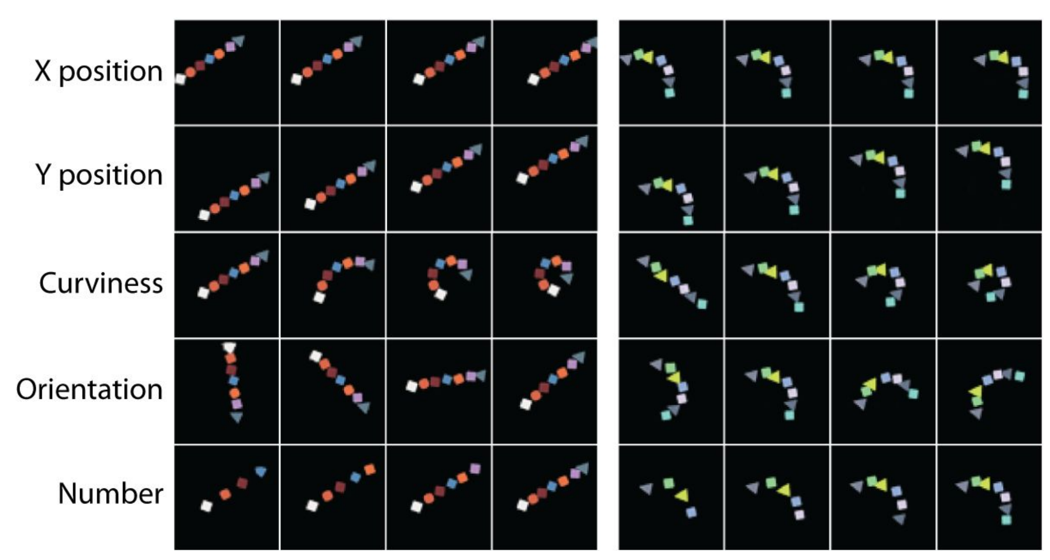
Extracts relational features from objects via a permutation invariant encoder (GNN)

LSTM decoder to reconstruct objects

'Filling in' procedure (instantiation) to replace abstracted object latents with MONet encodings

Disentangled relational representations

Latent space traversals on relational latents, after an image is encoded



Model training

Permutation invariant reconstruction error
Use hungarian algorithm to match LSTM predictions with MONet latents

$$L_{rec} = \frac{1}{2} \sum_{(i,j)-pairs} || \mathbf{a}_i - \hat{\mathbf{a}}_j ||^2$$

Disentangling pressure
Use the β-VAE KL regularisation

$$L_{reg} = \beta D_{\mathrm{KL}} \left(Q(\mathbf{r} \mid \mathbf{x}) \,||\, P(\mathbf{r}) \right)$$

Mask entropy
Entropy loss to avoid collapse onto single latent

$$L_{entropy} = -\sum_j m_j \ln m_j$$

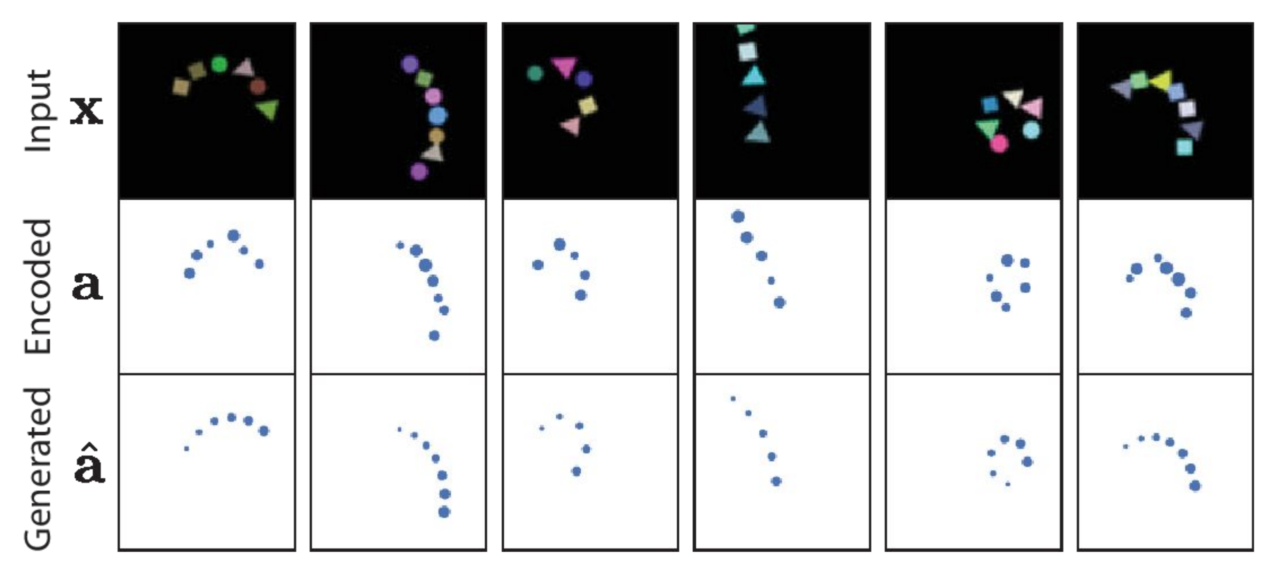
Re-ordering loss
Loss to minimise distance between successively generated objects

$$L_{reorder} = \sum_{i=1} || \hat{\mathbf{a}}_i - \hat{\mathbf{a}}_{i-1} ||^2$$

Conditioning loss
Additional loss to stabilise gradients (see paper)

Constellation reorders objects according to learned structure

Top: Input image
Middle: MONet latent dimensions corresponding to x, y
Bottom: Constellation generated latents x, y
Size of point corresponds to order in the sequence



Imagination from language

Learn associations between language symbols (e.g. 'left circle') and relational latents encoded from an appropriate image

Then can encode an image, and 're-imagine' the objects in novel ways according to inputted language