Spatially Invariant Attend, Infer, Repeat

ERIC.CRAWFORD@CS.MCGILL.CA, JPINEAU@CS.MCGILL.CA

Motivation

When reasoning about the physical world, object-like representations are powerful: good for generalization and transfer, very interpretable. But first need a way of extracting objects from visual stream. Want to be able to do this without labels, since they are expensive to obtain, and for complex images containing many objects.

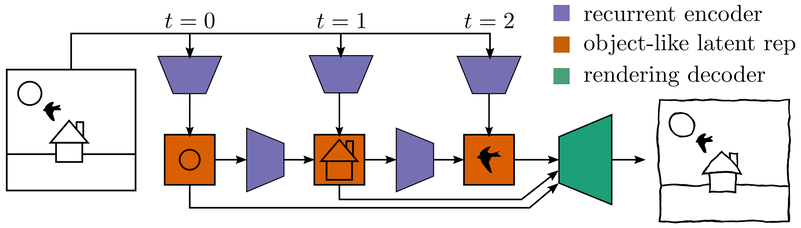
Problem Statement

Unsupervised Object Detection: Discover objects that are common in a dataset, and learn to detect them. Like supervised object detection, but model *not* provided with images annotated with object bounding boxes.

Prior Work

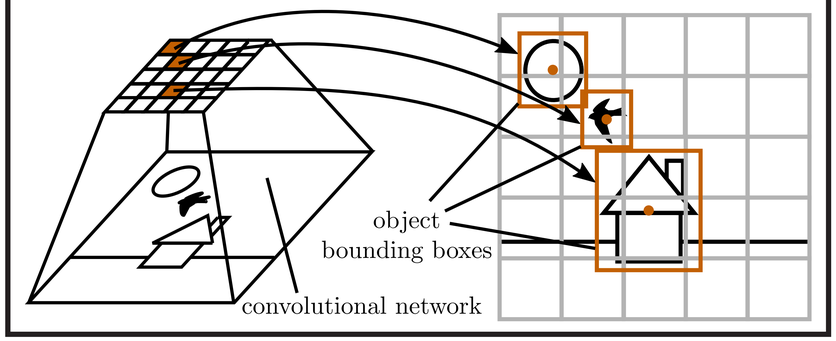
Attend, Infer, Repeat (AIR)

- Addresses unsupervised object detection.
- Variational autoencoder with recurrent encoder, object-like latent representation, and a "rendering" decoder.
- Performance collapses for large images with many objects.



You Only Look Once (YOLO)

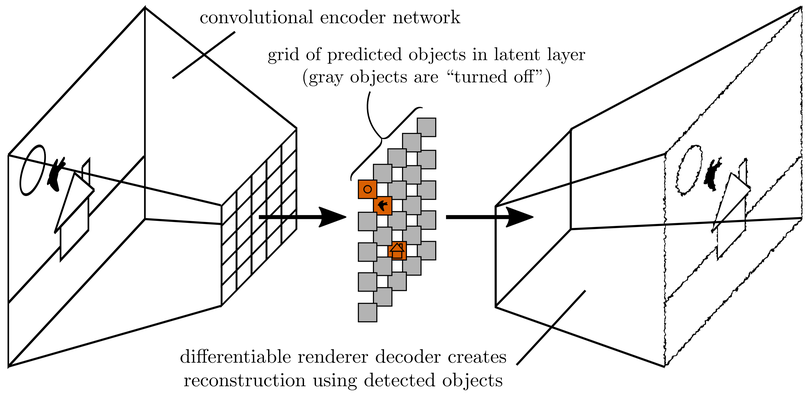
- Addresses *supervised* object detection: requires images labelled with bounding boxes for objects.
- Single shot object detection: map from image directly to detected objects.
- Uses convolutions to exploit structure of objects in images (i.e. spatial invariance) and learn efficiently.



Proposed Approach

Spatially Invariant Attend Infer Repeat (SPAIR)

- Combine spatial invariance properties of YOLO with unsupervised trainability of AIR.
- Variational autoencoder with a convolutional encoder network that exploits spatial invariance, object-like latent representation, and decoder that implements a differentiable rendering process.



Object at spatial location (i, j) represented by variables:

$$z^{ij}_{\text{what}} \in \mathbb{R}^A \qquad z^{ij}_{\text{depth}} \in \mathbb{R} \qquad z^{ij}_{\text{pres}} \in \{0, 1\} \qquad z^{ij}_{\text{where}} \in \mathbb{R}^4$$

Let z_{pres} be the vector consisting of all values of z^{ij}_{pres}. We place a Geometric prior on z_{pres} which favours having few non-zero entries, encouraging the network to explain each scene using few objects. Additionally, we place a prior on z^{ij}_{where} which favours small, self-contained objects.

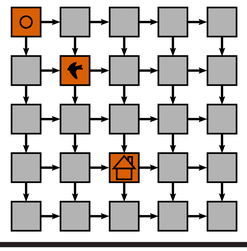
Training. Maximize the Evidence Lower Bound:

$$\mathcal{L}(\theta, \phi) := E_{z \sim q_\phi(z|x)} \left[\log p_\theta(x|z)\right] - D_{KL}(q_\phi(z|x) \parallel p(z))$$

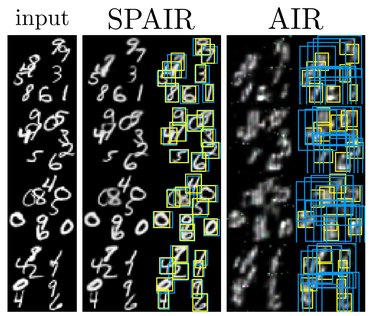
where q_ϕ / p_θ are the encoder/decoder networks and $p(z)$ is the prior.

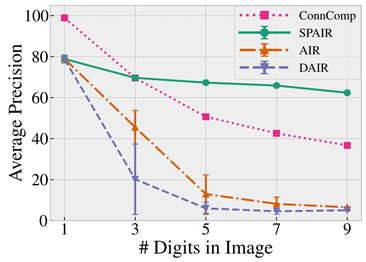
Scaling. Use of convolutional encoder and specifying objects with respect to grid cells (as in YOLO) divides up the task of explaining the image, making it easier to parse complex images. Also allows the network to process images that have different size than what it was trained on.

Lateral Connections. Learnable lateral connections allow objects to condition on on nearby objects that have already been generated. Can show empirically that these are important for performance.



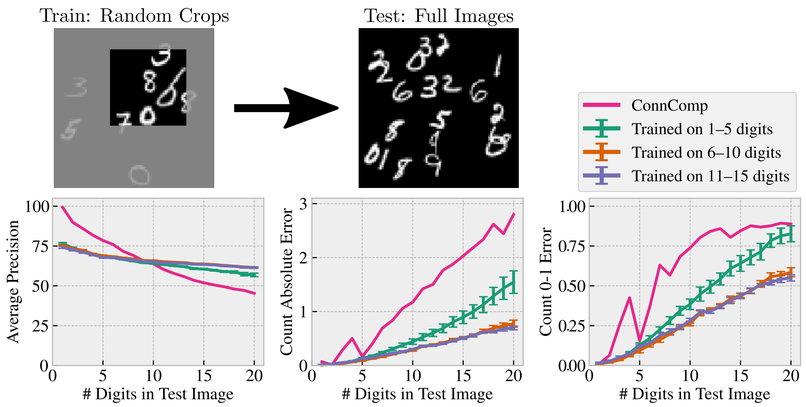
Comparison with AIR





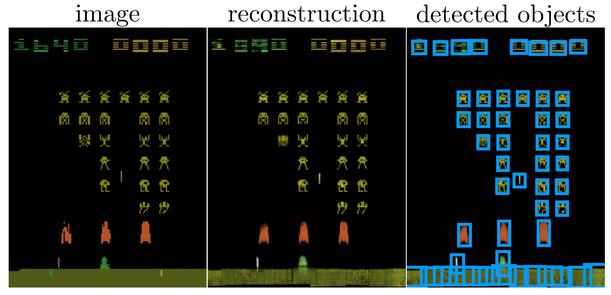
Testing abilities of AIR and SPAIR at discovering and detecting MNIST digits in cluttered scenes. SPAIR outerpforms AIR when many digits are present.

Generalization



Testing SPAIR's ability to generalize to images that are larger and have more objects than those seen during training. SPAIR was trained on small random crops, but tested on full images.

Scaling



Qualitative results on relatively large images (size 210×160) from Space Invaders. Trained on small random crops, tested on full images. SPAIR can handle objects of different sizes, from small bullets to larger rocks.