# Unsupervised learning of Visual Object Relations with Graph-Level Analogy

Daniel Zeng, dazeng@stanford.edu | CS 231N Spring 2022, Stanford University

## Introduction

- Visual relations form the basis of understanding our compositional world, as relationships between visual objects capture key information in a scene.
- We tackle learning and discovering relations without supervision, where relation types and labels are not known *a priori*
- In contrast, when relations are learned with predefined labels in a supervised context, this limits us to settings which depend on those seen relations.

## Problem Statement

- Input: *[16 x 16 x 9 x n]* images, and no knowledge of objects, nor relation types, nor underlying graph per example
- The goal: Infer the global relation types
- The graph $E(g_i)$ of each image $x_i$
- The relational graph structure $E(G_t)$ for each task $t$
  - Requires identifying relation $e_{k,l}$ between object pairs
  - Metric: Relation Classification Accuracy %

## Dataset: BabyARC

- Collection of images (observations) $x_i$
- Each image $x_i$ belongs to some known task $t \in T$.
- Each task $t$ has an unknown unique task graph $G_t = (V, E)$, where nodes V are objects and edges E are the corresponding relations
- All of its corresponding images share this common relational subgraph $E(G_t)$
- Objects in $x_i$ part of relational subgraph: "core" objects
  Not part of relational subgraph: "distractor" objects

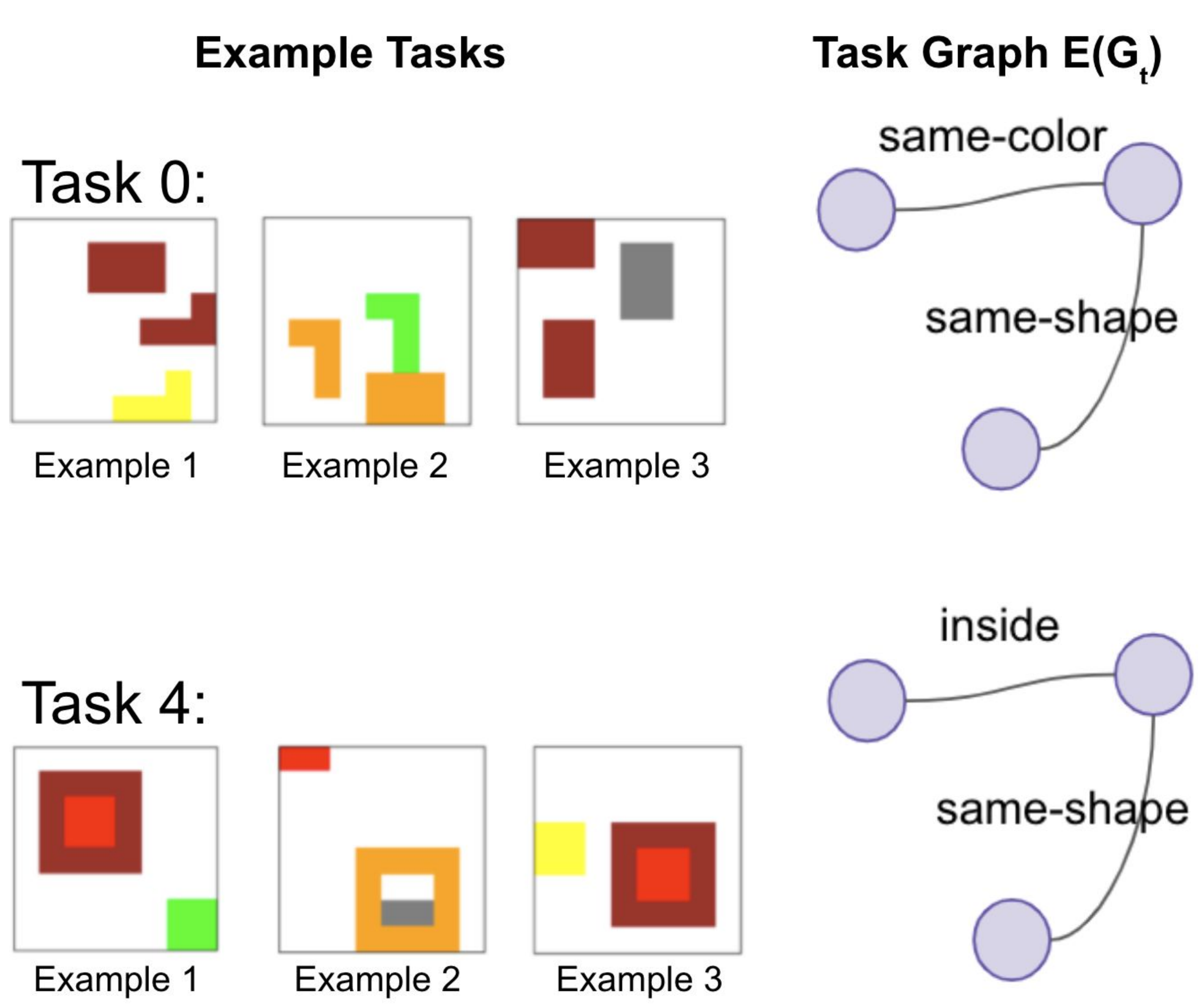



*Figure 1. The relations in dataset, 'same-shape', 'same-color', 'inside'*

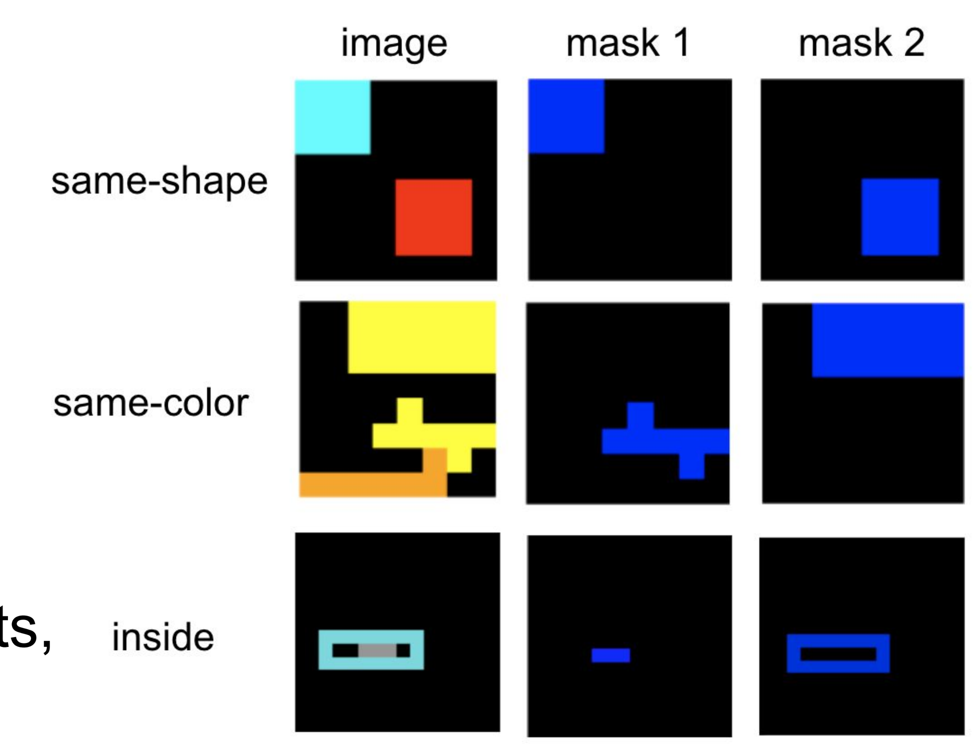
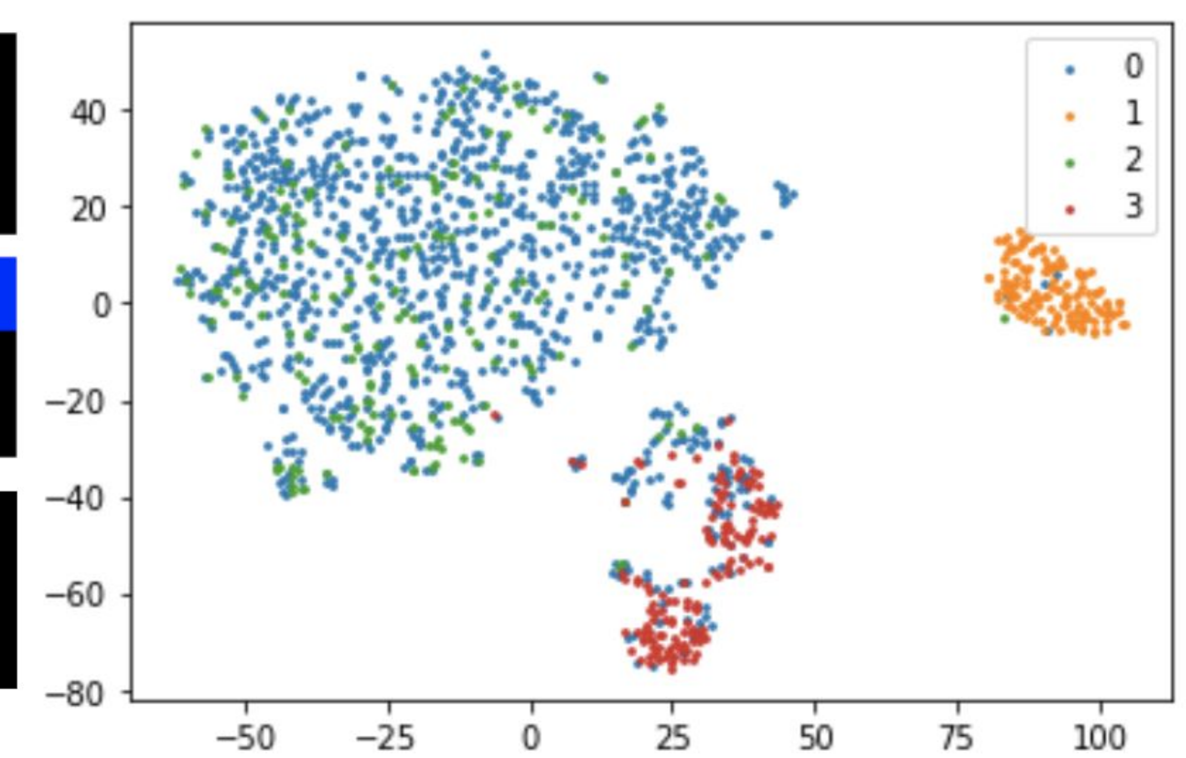
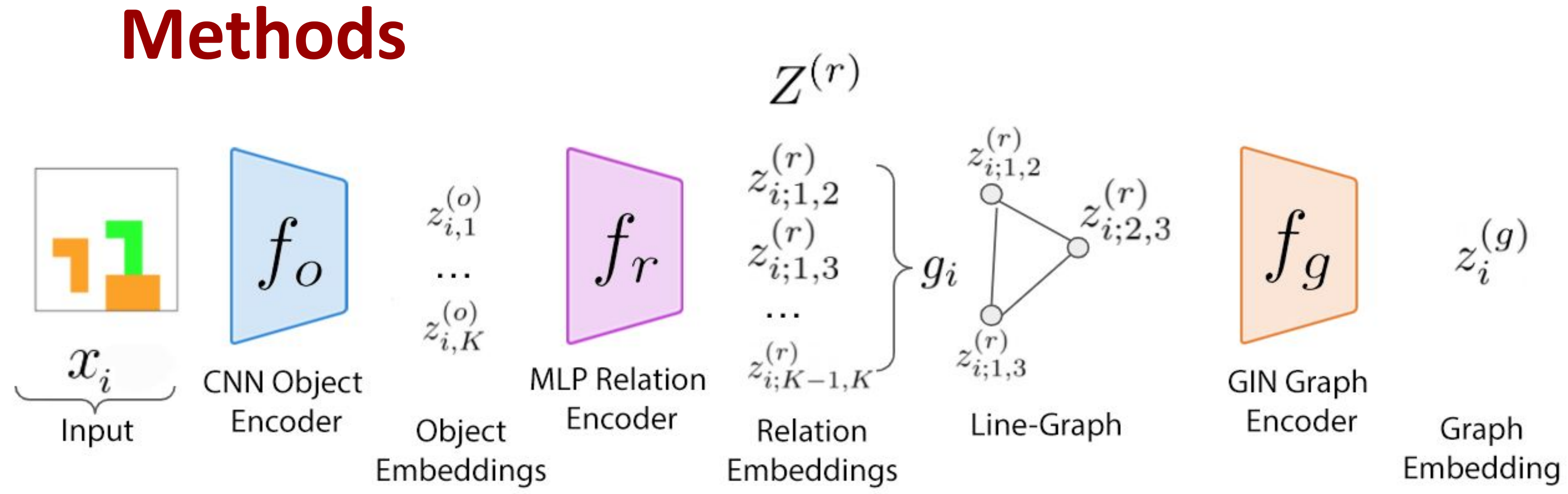



*Figure 2. t-SNE visualization of relation embeddings. 0 is 'none', 1 is 'inside', 2 is 'same-color', 3 is 'same-shape'*




## Methods




- Each image $x_i$ is represented as latent graph $g_i$ after CNN + MLP, becomes $z_i^{(g)}$ after graph isomorphism network (GIN)
- Two types of loss functions to train CR-GNN architecture (above)
  - Contrastive objective

$$\mathcal{L}_{\text{contrastive}} = \sum_{i,j \in \text{same-task}} ||f_g(g_i) - f_g(g_j)||_2 + \sum_{k,m \in \text{diff-task}} \max(0, \eta - ||f_g(g_k) - f_g(g_m)||_2$$

    - graph representation $f_g(g_i)$ should be similar within the task (intra-task loss), and should be different between different tasks (inter-task loss). $\eta$ is margin hyperparameter
  - Classification objective

$$\mathcal{L}_{\text{classify}} = \sum_{\forall i \in n} \mathcal{L}_{CE}(\text{Linear}(f_g(g_i)), y)$$

    - standard cross-entropy loss, between the true task ID $y$ against the predicted task ID
- Additional regularizer: Information Bottleneck (IB)
  - Constraints information between input and relation embedding

$$\mathcal{L}_{\text{IB}} = \mathcal{I}(X; Z^{(r)})$$

## Results

*Table 1. Relation classification accuracy for 2-3 core objects*

| Method | # Distractors | | |
|---|---|---|---|
| | 0 | 1 | 0-2 |
| Classify | 0.923 | 0.926 | 0.946 |
| Classify + IB | 0.919 | 0.918 | 0.901 |
| Contrastive | **0.959** | 0.961 | 0.954 |
| Contrastive + IB | 0.952 | **0.963** | **0.957** |
| Best | **0.959** | **0.963** | **0.957** |

*Table 2. Relation classification accuracy for 2-4 core objects*

| Method | # Distractors | | |
|---|---|---|---|
| | 0 | 1 | 0-2 |
| Classify | 0.956 | 0.955 | 0.965 |
| Classify + IB | 0.960 | 0.962 | 0.959 |
| Contrastive | **0.965** | 0.971 | 0.965 |
| Contrastive + IB | 0.960 | **0.973** | **0.971** |
| Best | **0.965** | **0.973** | **0.971** |

- Apply *k*-means clustering to assign cluster labels to each of the learned relation embeddings
- Model performance is similar between both objectives
  - Contrastive objective performing slightly better
- No accuracy degradation due to varying the number of introduced distractor objects, in all cases
- Overall slight improvement with more tasks in 2-4 core objects setting compared to 2-3 core objects
- Model infers the global relation types, as the *t*-SNE visualization shows clustering of relation embeddings by same relation label, shown in Figure 2

## Conclusion / Future Work

- Our method achieves above 95% accuracy in relation classification, discovers the relation graph structure for most tasks, and further generalizes to unseen tasks with more complicated relational structures
- Limitation: model only learns the necessary relation representations needed to distinguish between the given tasks, such as overlapping clusters in Figure 2
- Future work is expanding towards more datasets, potentially CLEVR generation with graph structure