RelViT: Concept-guided Vision Transformer for Visual Relational Reasoning

Xiaojian Ma¹, Weili Nie², Zhiding Yu², Huaizu Jiang³, Chaowei Xiao²,⁴, Yuke Zhu²,⁵, Song-Chun Zhu¹, Anima Anandkumar²,⁶

¹UCLA ²NVIDIA ³Northeastern University ⁴ASU ⁵UT Austin ⁶Caltech
Motivation

Visual Relational Reasoning: the niche of visual intelligence!

Visual Relationship Recognition

Visual Question Answering

Abstract Visual Reasoning
Motivation

What makes Visual Relational Reasoning so challenging?

-> How do our humans perform visual reasoning?

On the Binding Problem in Artificial Neural Networks, *arXiv*, 2020
Motivation

- What makes Visual Relational Reasoning so challenging?

- Object-centric (disentangled) representations
- Relational inductive bias
- Systematic generalization

Object-Centric Learning with Slot Attention, in NeurIPS, 2020
A simple neural network module for relational reasoning, in NeurIPS, 2017
Compositionality decomposed: how do neural networks generalise?, in JAIR, 2020
Background

Transformers and Vision transformers

Transformer: explicitly capture the **pairwise relations** among input entities.

Vision transformers: **image patches (object candidates)** as input entities.
Background

Given the appealing nature of Vision transformers (ViTs) on **object-centric learning** and **relational inductive bias**, we choose to start with this model and see if we can make it better.

We propose to use **self-supervised contrastive learning** to achieve this goal.

Vision transformers: **image patches (object candidates)** as input entities.
Background

Contrastive learning (CL) tasks for ViT

Global CL: Contrasting the **global features** of input images.

Local CL: Contrasting the **(spatially) local image features** of the input images.

Dense Contrastive Learning for Self-Supervised Visual Pre-Training, in CVPR, 2021
Background

How can vision transformer benefit from contrastive learning?

Global CL: Contrasting the **global descriptions** of input images.
-\> boosting **relational meaning and reasoning** via instance contrasting

Local CL: Contrasting the **(spatially) local descriptions** of the input images.
-\> boosting **object-centric representation** via unsupervised **correspondence learning**
Background

An issue: the original global and local CL are concepts/semantics-free -- image are treated as isolated samples. Therefore, these CL methods will promote:

-representations that fail to capture the semantic similarity of different objects
-relational deduction that fails to exploit these semantics for more efficient / lifted reasoning.
Figure 1: An overview of our method. **Red+Green**: the learning pipeline of DINO (Caron et al., 2021) and EsViT (Li et al., 2021); **Red+Blue**: our pipeline.
RelViT

• Concept-guided Vision Transformer

// Pseudo code

Input: image x, concept c

1. $x_1, x_2 = \text{aug1}(x), \text{aug2}(x)$
2. $f_1, f_2 = \text{backbone}(x_1), \text{backbone}(x_2)$
3. $\text{fn2} = \text{dequeue_n_enqueue}(f_2, c)$
4. $\text{loss}_\text{global} = \text{DINO}_\text{GLOBAL}(f_1[0], \text{fn2}[0])$ // $f_*[0]$ is [CLS]
5. $\text{loss}_\text{local} = \text{DINO}_\text{LOCAL}(f_1[1:], \text{fn2}[1:])$
6. $(\text{loss}_\text{local} + \text{loss}_\text{global}).\text{backward}()$
Experiments

We evaluate RelViT on two datasets:

-HICO: Human-object-interaction recognition

Formula: $I \Rightarrow \langle \text{object, interaction} \rangle$

-GQA: Relational visual question answering

Formula: $\langle Q, I \rangle \Rightarrow \text{answer category}$
Experiments

Concept in HICO

-> H.O.I category (#concepts=600)

-> Interaction category (#concepts=117)

-> Object category (#concepts=80)

<Ride, Horse>
Experiments

Concept in GQA

We propose to parse the question into concept tokens.
Experiments

**Systematic generalization** test for HICO:

- We make several HOI category **unseen** during training, e.g. \( \text{TV, sit} \) only appears in testing data.

- We ensure the training data includes all the objects and interactions (e.g. TV and sit).

- Testing **systematicity** of systematic generalization.
Experiments

Systematic generalization test for GQA:

- Each GQA question is also labelled with a reasoning program.
- We make training data only contain questions with shorter reasoning programs -- testing **productivity** of systematic generalization.

![Question]

What color is the food on the red object left of the small girl that is holding a hamburger, yellow or brown?

Select: hamburger $\rightarrow$ Relate: girl, holding $\rightarrow$ Filter size: small $\rightarrow$ Relate: object, left $\rightarrow$ Filter color: red $\rightarrow$ Relate: food, on $\rightarrow$ Choose color: yellow | brown

#reasoning hops: 7

(b) Productivity
Experiments (HICO)

- We largely improve the current learners on both standard test and systematic generalization test, without any oracle object-centric representations (bboxes).

- Our model makes significant progress on unseen categories.

<table>
<thead>
<tr>
<th>Method</th>
<th>Ext. superv.</th>
<th>Backbone</th>
<th>Orig.</th>
<th>Systematic-easy</th>
<th>Systematic-hard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mallya &amp; Lazebnik (2016)*</td>
<td></td>
<td>ResNet-101</td>
<td>33.8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Girdhar &amp; Ramanan (2017)*</td>
<td>bbox</td>
<td>ResNet-101</td>
<td>34.6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Fang et al. (2018)*</td>
<td>pose</td>
<td>ResNet-101</td>
<td>39.9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ViT-only</td>
<td></td>
<td>PVTv2-b2</td>
<td>35.48</td>
<td>31.06</td>
<td>11.14</td>
</tr>
<tr>
<td>EsViT (2021)</td>
<td></td>
<td>PVTv2-b2</td>
<td>38.23</td>
<td>35.15</td>
<td>11.53</td>
</tr>
<tr>
<td>RelViT (Ours)</td>
<td></td>
<td>PVTv2-b2</td>
<td>39.4</td>
<td>36.99</td>
<td>12.26</td>
</tr>
<tr>
<td>RelViT + EsViT (Ours)</td>
<td></td>
<td>PVTv2-b2</td>
<td>40.12</td>
<td>37.21</td>
<td>12.51</td>
</tr>
</tbody>
</table>
Experiments (GQA)

- We largely improve the current learners on both standard test and systematic generalization test, without any oracle object-centric representations (bboxes).

- This can be way impressive for VQA tasks -- object-detection play crucial role in almost all state-of-the-art VQA learners but not with our method.

<table>
<thead>
<tr>
<th>Method</th>
<th>Bbox feat.*</th>
<th>Backbone</th>
<th>Orig.</th>
<th>Sys.</th>
</tr>
</thead>
<tbody>
<tr>
<td>BottomUp (2018)</td>
<td>✓</td>
<td>ResNet-101</td>
<td>53.21</td>
<td>-</td>
</tr>
<tr>
<td>MAC (2018b)</td>
<td>✓</td>
<td>ResNet-101</td>
<td>54.06</td>
<td>-</td>
</tr>
<tr>
<td>MCAN-Small (2019)</td>
<td>✓</td>
<td>ResNet-101</td>
<td>58.35</td>
<td>36.21</td>
</tr>
<tr>
<td>ViT-only</td>
<td></td>
<td>PVTv2-b2</td>
<td>56.62</td>
<td>31.39</td>
</tr>
<tr>
<td>EsViT (2021)</td>
<td></td>
<td>PVTv2-b2</td>
<td>56.95</td>
<td>31.76</td>
</tr>
<tr>
<td>RelViT (Ours)</td>
<td></td>
<td>PVTv2-b2</td>
<td>57.87</td>
<td>35.48</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GQA overall accuracy</th>
<th>MCAN-Small (w/ bbox)</th>
<th>RelViT (PVTv2-b2)</th>
<th>RelViT (PVTv2-b3)</th>
<th>RelViT (Swin-base)</th>
</tr>
</thead>
<tbody>
<tr>
<td>original</td>
<td>58.35</td>
<td>57.87</td>
<td>61.41</td>
<td><strong>65.54</strong></td>
</tr>
<tr>
<td>systematic</td>
<td>36.21</td>
<td>35.48</td>
<td>36.25</td>
<td><strong>37.51</strong></td>
</tr>
</tbody>
</table>
Takeaway

ViT is a promising architecture that offers \textbf{object-centric representations} and \textbf{relational inductive bias}.

Using \textbf{concept-guided contrastive learning} as an \textbf{auxiliary task} to further exploit the visual relational reasoning data could significantly boost the performance of ViTs on these tasks, especially on \textbf{systematic generalization}.

\textbf{RelViT: Concept-guided Vision Transformer for Visual Relational Reasoning}