## Tell, Draw, and Repeat: Generating and Modifying Images Based on Continual LINGUISTIC Instruction

 *Work was performed during an internship with Microsoft Research
iUniversity of Guelph ${ }^{*}$ Microsoft Research ${ }^{\text {s/MILA }}$ IVector Institute ${ }^{\text {'U }}$ Université de Montréal ${ }^{\wedge}$ CIFAR

UNVERSTY Miraest


- MOTIVATION
- Intelligent systems that can generate visual outputs can be used for education, entertainment, graphic design,
and the creative arts.
- Inteligent systems that can generate visual outputs can be used for education, entertainment, graphic design,
and the creative arts.
An iterative natural language interface to graphic design tools would make computer vision technology accessiAn iterative natural lang
ble to a larger population.

2 - Task: Generative Neural Visual Artist (GeNeVA)

- Teller: provides linguistic instruction to Drawer - Drawer: generates image conditioned on prev-
ous image and history of Teller instructions. ous image and history of Teller instructions. - Teller provides new linguistic instruc
tion,.... process repeats until Teller stops. Drawer learns how to map complex linguistic in Drawer learns how to map complex linguistic in
structions to reeilistic objects on a canvas, maintain
object properties \& relationships between objects object properties \& rela
(e.g., relative location).

4-GeNeVA-GAN ModeL


- instructions $Q=\left\{q_{1}, \ldots, q_{n}\right\}$
Generated image:


$\begin{aligned} d_{t} & =\operatorname{BiGRU}\left(q_{t}, d_{t-1}\right)\end{aligned}$ Context-aware condition: $h_{t}=$ GRU
Context-free condition:
- Loss function for discriminator:
 - $L_{D_{\text {pit }}}, L_{D_{\text {mones }}}$,the generator loss $L_{G} L_{D_{\text {ra }}}$ - Binary cross entropy loss for $L_{\text {aux: }}$ detects
object added at current time-step
- object added at current time-step
layer in $G$
- We try both concatenation and subtraction
- We try both concatenation and subtraction
as fusion options in $D$
- is conditioned using $h_{t}$ by projection.

5- Evaluation and Quantitative Results Study


Precision, Recall, and F1-Score are cald Iated over object tetections. Relational Sisilarity (rsim) capt
left-right and front--ehind positions.
$\operatorname{rsim}\left(E_{G_{q},}, E_{G_{\text {gen }}}\right)=\operatorname{recall} \times \frac{\left|E_{G_{\text {gen }}} \cap E_{G_{G_{a}}}\right|}{}$
$\underbrace{\left|E_{G_{\mathrm{g}}}\right|}$
i-CLEVR

## 3 - Datasets

Collaborative Drawing (CoDraw) [1]: 9,993 scenes of varying lengths.

- Scenes of children playing in a park, with dififerent poses and expressions.
- Scenes include objects as trees, tables, animals, etc; 58 object types in total.
Interactive CLEVR (i-CLEVR): 10,000 sequences of 5 instructions, Interactive CLEVR (i-CLEVR): 10,000 sequences of 5 instructions, images.
- We modify CLEVR [ $[2]$ to create an interactive version with image sequences. - We modify CLEVR $[2]$ to create an interactive version with image sequences
- For requiring textual context, we refer to most recently added object by it. - Instructions are templated. E.g. a template for the second instruction:
Add a [object color [object shapel) [relative position: depth]


[^0]| Model | $L_{D_{\text {muong }}}$ | $f_{G_{t-1}}$ | $L_{\text {aux }}$ | D Fusion |  | Precision, Recall, and F1-Score are calculated over object detections. |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  |  | concat | subtract |  |  |  |  |
| Baseline | $x$ | $x$ | $x$ | $x$ | $x$ | left-right | and fron | (rsim) | caps. |
| Mismatch | $\checkmark$ | $x$ | $x$ | $x$ | $x$ |  |  |  |  |
| $G$ prior | $\checkmark$ | $\checkmark$ | $x$ | $x$ | $x$ |  |  |  |  |
| Aux | $\checkmark$ | $\checkmark$ | $\checkmark$ | $x$ | $x$ |  |  |  | $\underline{\left\|E_{G_{\text {gem }}} \cap E_{G_{g q}}\right\|}$ |
| $D$ Concat | $\checkmark$ | $\checkmark$ | $\checkmark$ | $\checkmark$ | $x$ |  |  |  | ${ }^{\text {\| }}$ E ${ }_{\text {Gax }}$ |
| $D$ Subtract | $\checkmark$ | $\checkmark$ | $\checkmark$ | $x$ | $\checkmark$ |  |  |  |  |
|  |  |  | CoDraw |  |  |  |  | LEVR |  |
| Model | Precision | Recall | F1 |  | rsim | Precision | Recall | F1 | rsim |
| Non-iterative | 50.60 | 43.42 | 44.9 |  | 22.33 | 25.49 | 20.95 | 22.63 | 11.52 |
| Baseline | 55.61 | 42.31 | 48.05 |  | 25.31 | 69.09 | 56.38 | 62.08 | 45.19 |
| Mismatch | 62.47 | 48.95 | 54.8 |  | 32.74 | 71.15 | 60.57 | 65.44 | 50.21 |
| $G_{\text {G prior }}$ | 60.78 | 49.37 | 54.4 |  | 33.60 | 82.80 | 77.22 | 79.91 | 63.93 |
| Aux | 54.78 | 51.51 | 53.1 |  | 33.83 | 83.63 | 75.63 | 79.43 | 55.36 |
| ${ }^{D}$ Concat | 66.38 | 51.27 | 57.85 |  | 33.57 | 88.47 | 83.35 | 85.83 | 70.22 |
| $D$ Subtract | 66.64 | 52.66 | 58.8 |  | 35.41 | 92.39 | 84.72 | 88.39 | 74.02 |

6-Qualitative results


Figure: Example generated $128 \times 128$ image sequences for the i -CLEVR dataset


7- References



[^0]:    Code: https://github.com/Maluuba/GeNeVA

