Propojování vidění, motoriky a přirozeného jazyka v kognitivní robotice

Mgr. Gabriela Šejnová
gabriela.sejnova@cvut.cz
Fakulta elektrotechnická, Katedra kybernetiky
České vysoké učení technické v Praze

CZECH INSTITUTE OF INFORMATICS ROBOTICS AND CYBERNETICS CTU IN PRAGUE

Developmental Robotics Group at CIIRC CTU

Mgr. Michal Vavrečka, Ph.D.
Head of the group, Assistant Professor
Topics: biologically inspired models, vision & language grounding

Mgr. Karla Štěpánová, Ph.D.
Postdoc researcher
Topics: imitation learning, language acquisition

Ing. Megi Mejdrechová
PhD student
FBMI CTU

Nikita Sokovnin
Bachelor student
FEL CTU

incognite
Developmental Robotics

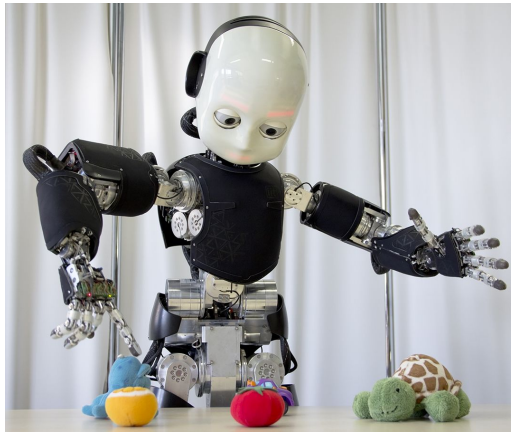
CZECH INSTITUTE OF INFORMATICS ROBOTICS AND CYBERNETICS CTU IN PRAGUE

Talk Outline

- our general approach
- vision & language mapping
  - neural module networks
- visuomotor mapping
  - myGym
  - variational autoencoders
- connecting vision, language and motorics
  - future plans

Developmental Robotics

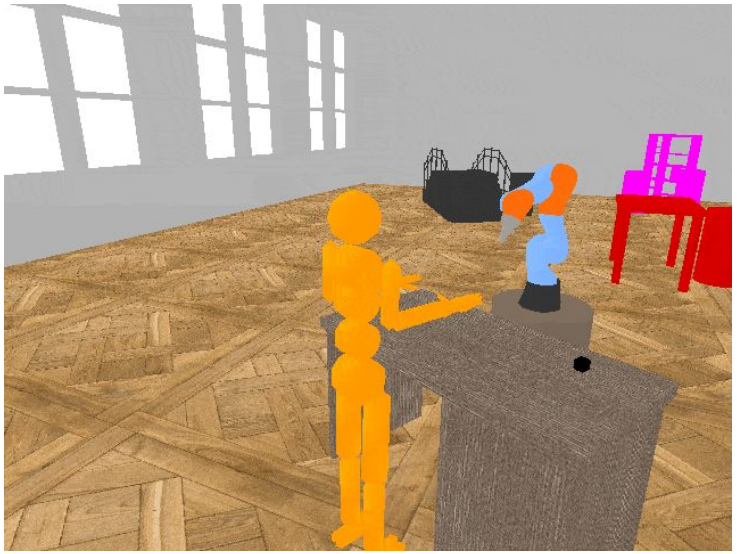
- cognitive robotics - make AI/robots that are autonomous and able to learn, plan complex tasks and communicate
- developmental robotics - inspired by human brain development in early childhood
  - motor babbling ➜ object manipulation
  - understanding speech ➜ actively using it



Developmental Robotics

General concepts:

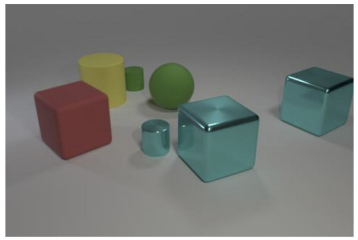
- robots that learn universal skills



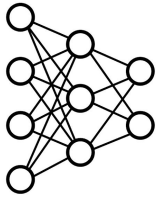
Developmental Robotics

General goals:

- compositionality

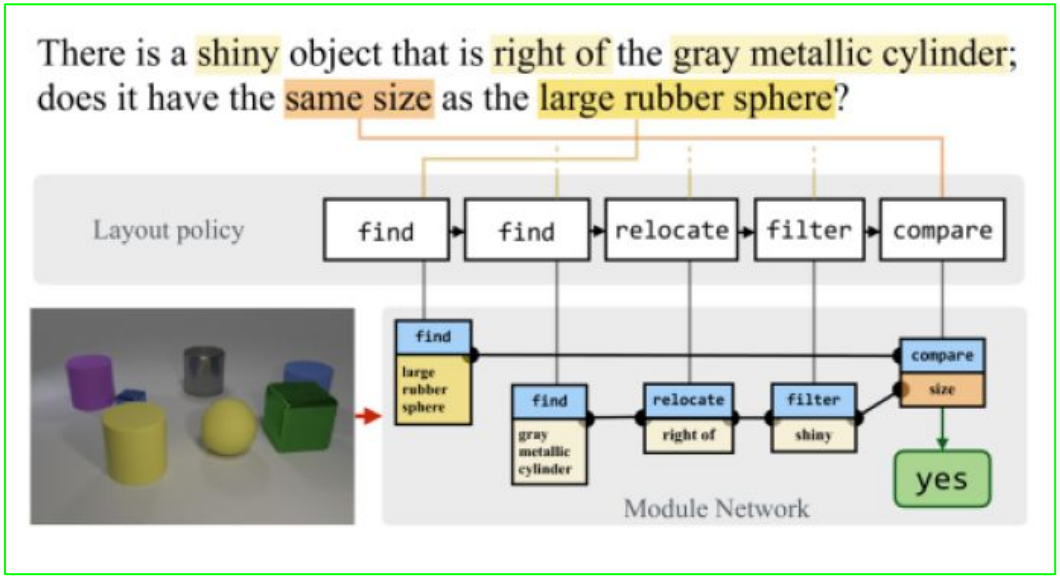


What color is the thing with the same size as the blue cylinder?



blue

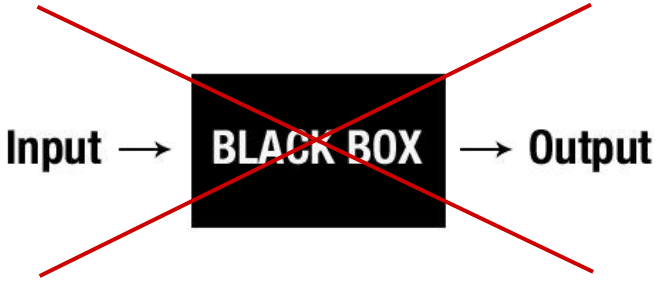
vs.

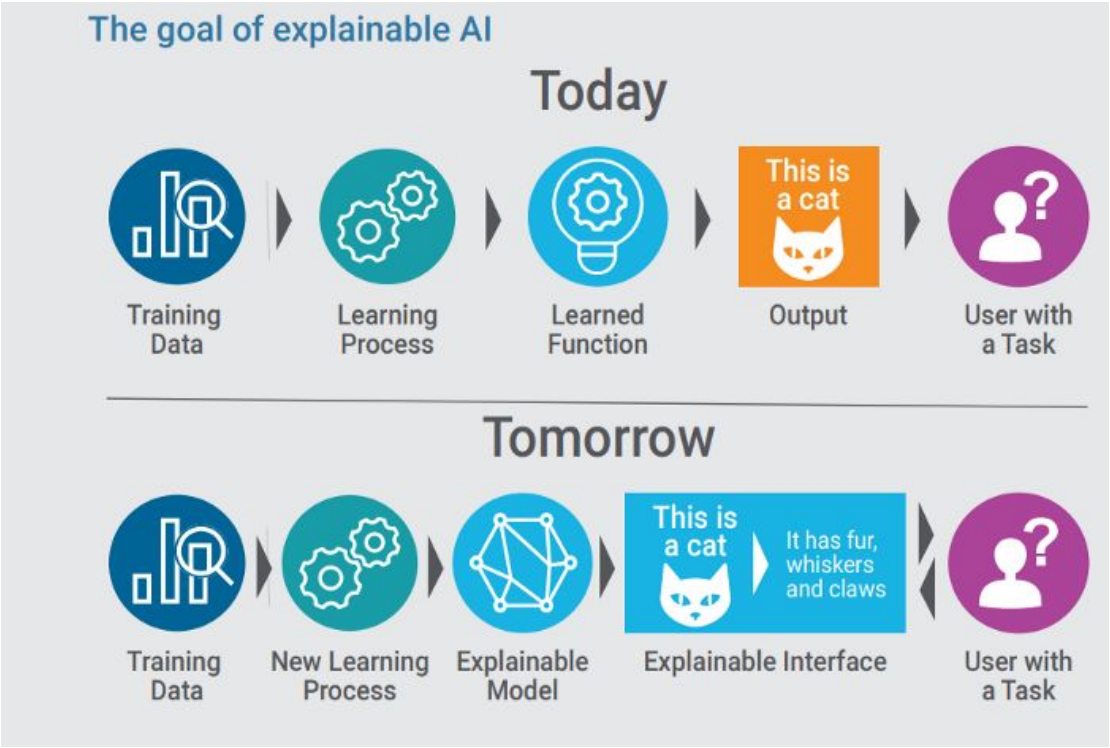


Developmental Robotics

General goals:

- explainability

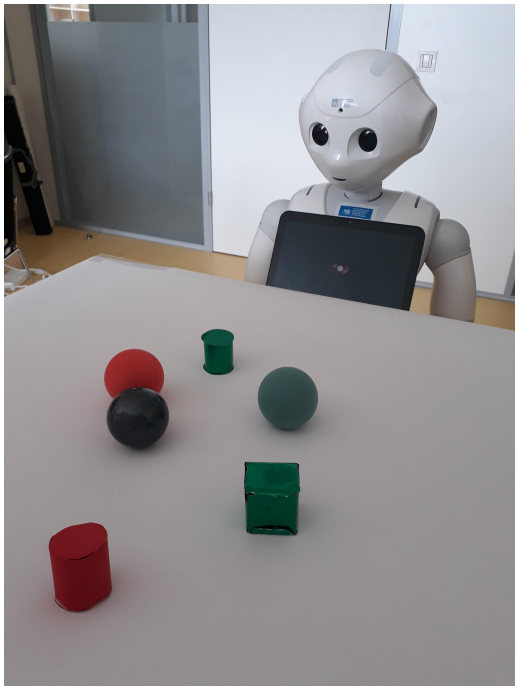




Developmental Robotics

General goals:

- intrinsic motivation & world exploration



Vision & Language Mapping

Visual Question Answering (VQA)

- benchmark task for visual reasoning
- VQA real-world dataset
  - requires understanding of image content ?

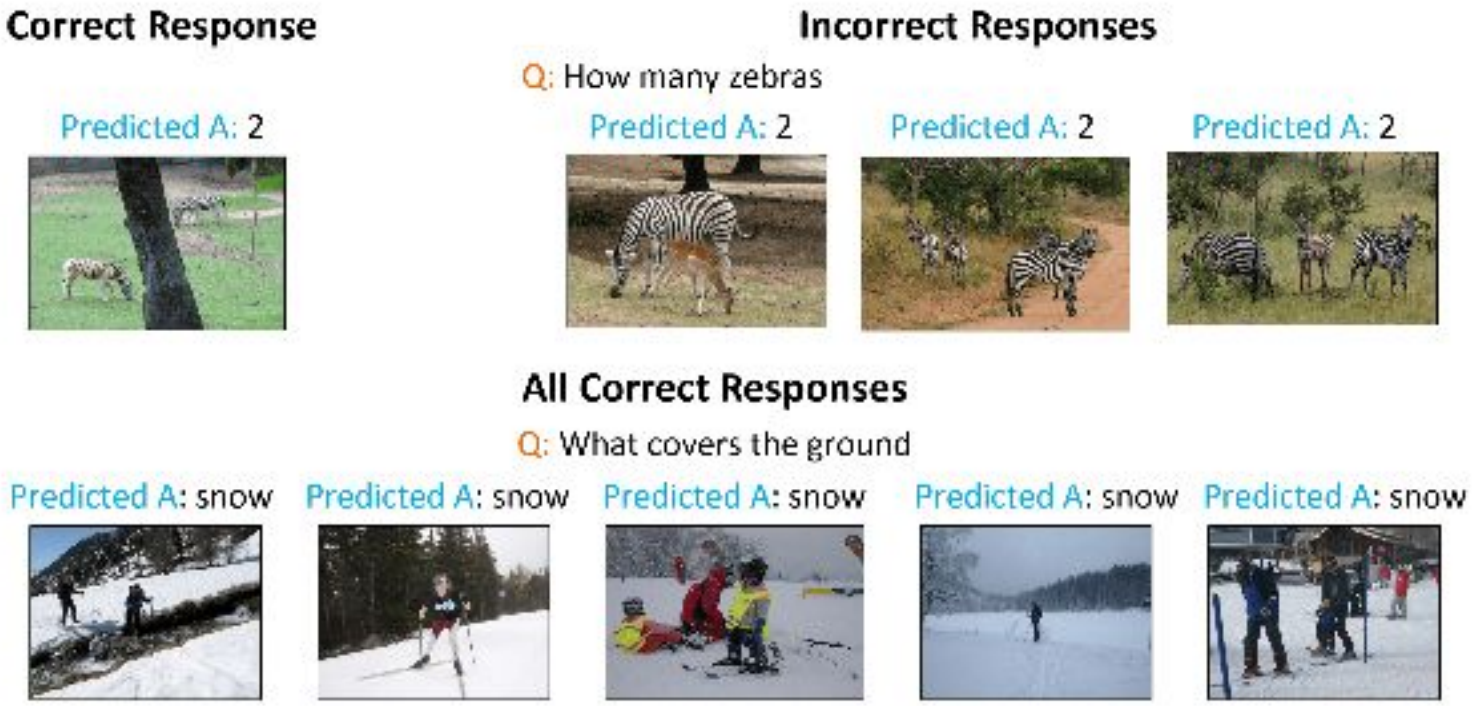
What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

Images from Antol et al, 2015

VQA dataset - problem

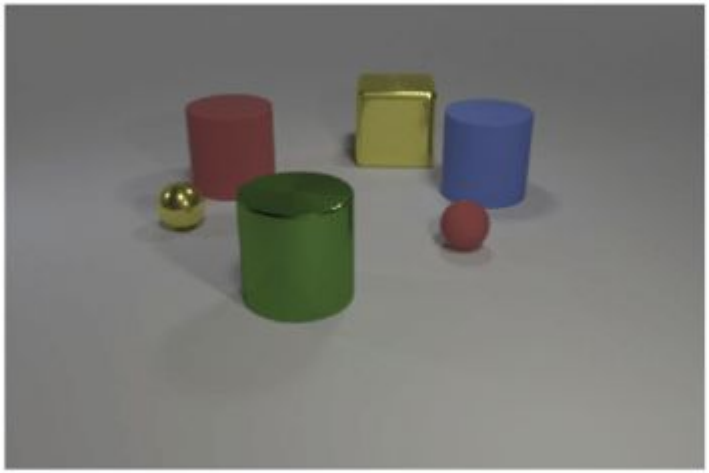
- dataset bias - no image needed, same answers for different questions, jumping to conclusions



Solution: **CLEVR** dataset

= synthetic, well-annotated dataset with minimal bias

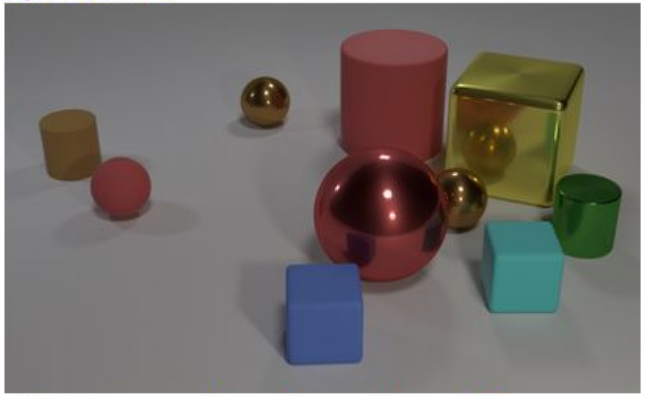
A



B


```
{“color”: “green”,
 “size”: “large”,
 “rotation”: 156.34024,
 “shape”: “cylinder”,
 “3D_coords”: [...],
 “material”: “metal”,
 “pixel_coords”: [...]}
```

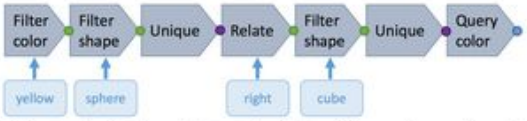
Questions in CLEVR test various aspects of visual reasoning including **attribute identification**, **counting**, **comparison**, **spatial relationships**, and **logical operations**.



Q: Are there an **equal number** of **large things** and **metal spheres**?
Q: **What size** is the **cylinder** **that is left of** the **brown metal** thing **that is left of** the **big sphere**?
Q: There is a **sphere** with the **same size as** the **metal cube**; is it **made of the same material as** the **small red sphere**?
Q: **How many** objects are **either** **small cylinders** or **red** things?

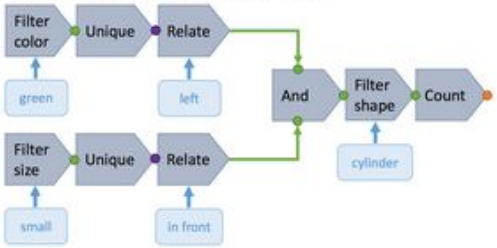
Each question in CLEVR is represented both in **natural language** and as a **functional program**. The functional program representation allows for precise determination of the reasoning skills required to answer each question.

Sample chain-structured question:



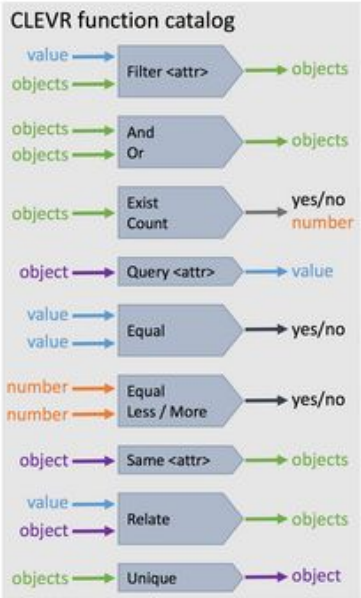
What color is the cube to the right of the yellow sphere?

Sample tree-structured question:



How many cylinders are in front of the tiny thing and on the left side of the green object?

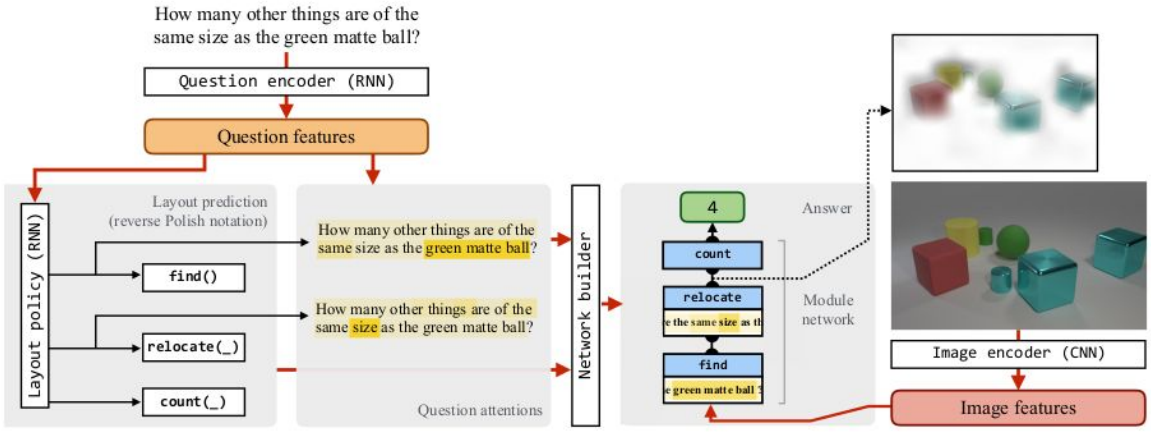
CLEVR function catalog



Images taken from Johnson, J. et al., 2017

Neural Module Networks

- compositional and explainable
- question translated into a sequence of logical operations
- individual network for each operation



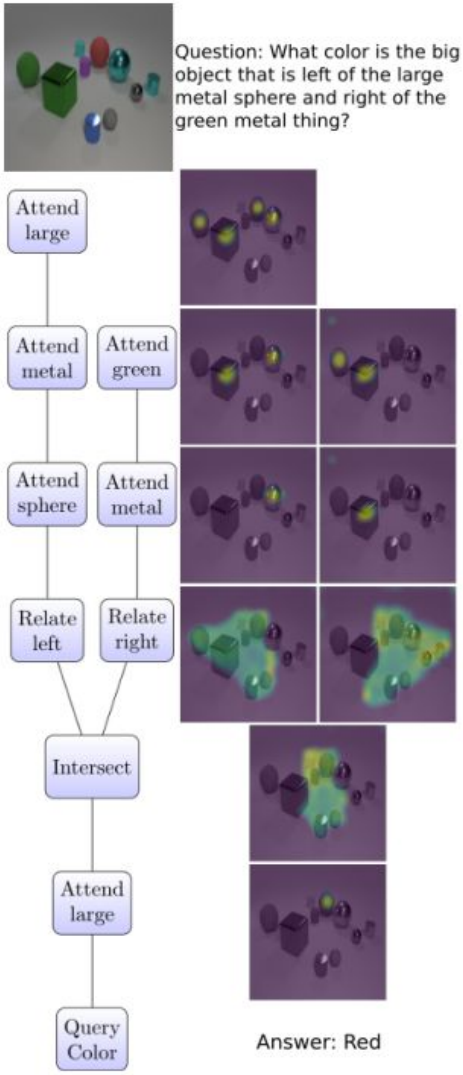
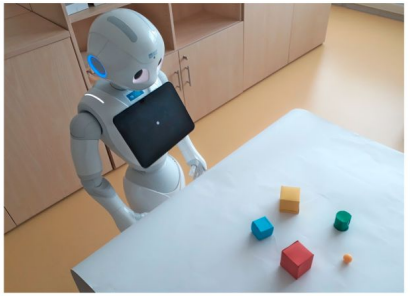


Image from Hu et al., 2017

Image from Mascharka et al., 2018

Neural Module Networks

+ compositional
+ interpretable
- logically inconsistent (counting mainly)



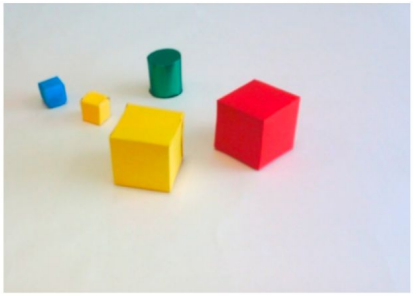


Table 1. Comparison of results on the CLEVR dataset. N2NMN is the original model from [9], N2NMN - COUNT and CONSISTENCY refers to our measured data for virtual/real world scenario. The measured values are the percentage of correctly answered questions from given category.

Method	Overall	Count	Shape	Material	Color	Size
LSTM	47.0	42.5	33.2	50.8	12.2	49.9
CNN+LSTM+SA	68.5	52.2	85	88	81	87
NMN	72.1	52.5	84.2	82.6	68.9	80.2
N2NMN	83.7	68.5	90.6	91.5	84.8	93.1
HUMAN	92.6	86.3	94.0	94.0	95.0	97.0
N2NMN - COUNT - VIRTUAL	75.1	81.3	81.2	65.3	79.3	68.5
N2NMN - COUNT - ROBOT	41.9	39	65	39.5	49.9	16
N2NMN - COSISTENCY - VIRTUAL	0.7	-	58.3	46.1	5.0	49.8
N2NMN - CONSISTENCY - ROBOT	0	-	48.7	20.5	0	5.1

Sejnova et al., 2018

Improving Consistency

- trained on custom, systematic set of questions for each image (How many...?)
- trained on original CLEVR dataset (CLEVR-COUNT) and adapted GYM dataset (GYM-COUNT)
- tested also on real-world data

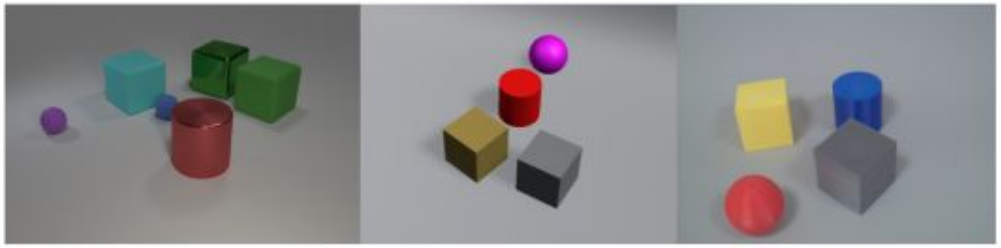
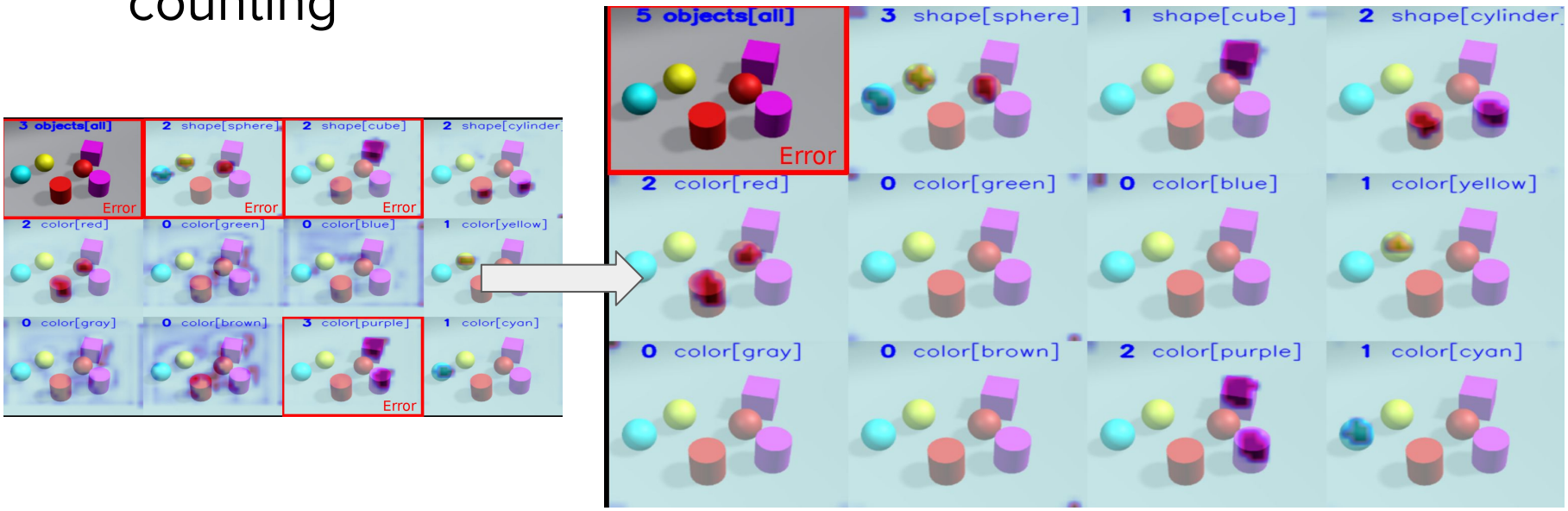


Fig. 2. Comparison between the datasets used in our study. The original CLEVR dataset (*left*) with fixed viewport, a sample from our nine viewport dataset generated using MuJoCo OpenAI environment and Unity render (*center*) and an example of real-world scenes collected by IIWA Kuka LBR 7 robotic manipulator (*right*).

Sejnova et al., 2019

Improving Consistency

- retraining on our adapted version of CLEVR (COUNT dataset) increased accuracy and consistency for counting



Train	Test	Count objects	Count shapes	Count color
CLEVR	COUNT	64.5 (56.1)	94.9 (57.1)	99.6 (62.9)
CLEVR	GYM(3)	69.5 (35.6)	95.8 (63.1)	92.2 (37.6)
CLEVR	ROBOT(3)	51.2 (17.2)	84.5 (41.1)	87.3 (19.0)
COUNT	COUNT	**99.6 (97.4)**	**98.4 (97.5)**	**100 (99.4)**
COUNT	GYM(3)	**97.0 (87.8)**	**99.1 (95.6)**	**98.4 (88.9)**
COUNT	ROBOT(3)	**85.9 (55.2)**	**90.0 (71.8)**	**95.2 (64.4)**

Sejnova et al., 2019

CTU
CZECH TECHNICAL UNIVERSITY IN PRAGUE

Visuomotor Mapping

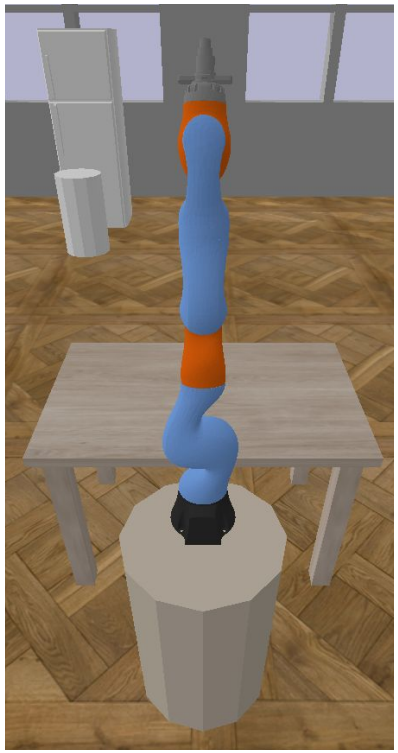
myGym

- modular toolkit for visuomotor robotic tasks
- real-time simulation with PyBullet Physics
- modular across workspaces, robots, tasks, reward types, objects and baselines
- deep reinforcement learning of (visuo)motor skills - supervised, semi-supervised and unsupervised

https://github.com/incognite-lab/myGym

myGym

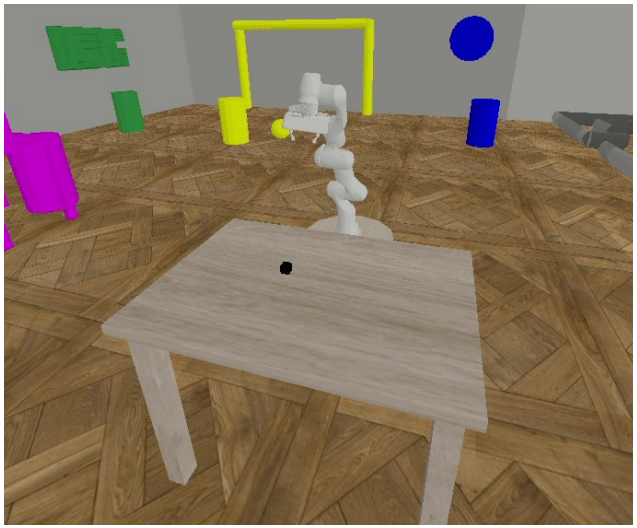
- robots:

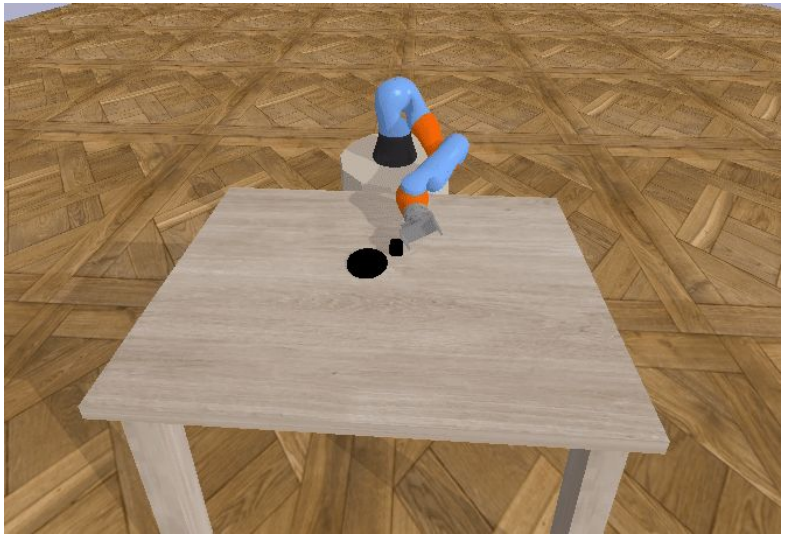


Robot	Type	Gripper	DOF
Kuka IIWA	arm	magnetic	7
Franka-Emica	arm	two finger	6
Jaco arm	arm	two finger	6
UR-3	arm	tactile gripper	6
UR-5	arm	tactile gripper	6
UR-10	arm	tactile gripper	6
Gummiarm	arm	passive palm	6
Reachy	arm	passive palm	8
Leachy	arm	passive palm	8
ReachyLeachy	dualarm	passive palms	16
ABB Yumi	dualarm	two finger	24
Pepper	humanoid	–	–
Thiago	humanoid	–	–
Atlas	humanoid	–	–

myGym

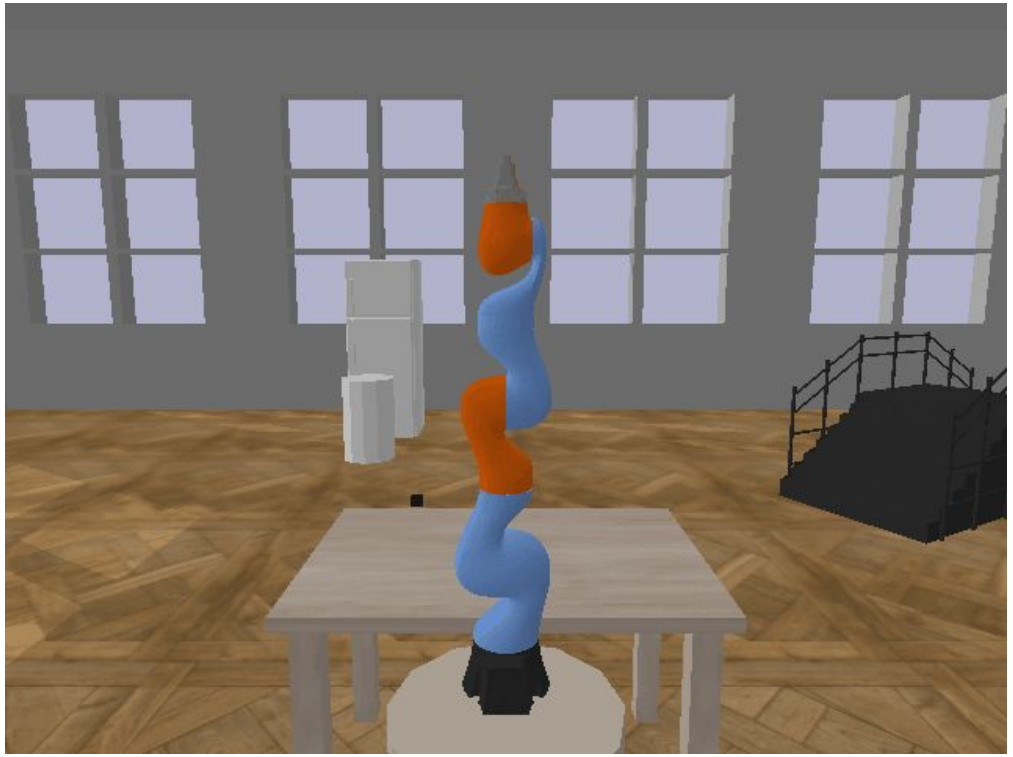
- manipulation tasks:
  - reach
  - push
  - pick and place
  - throw
  - catch
  - navigate...





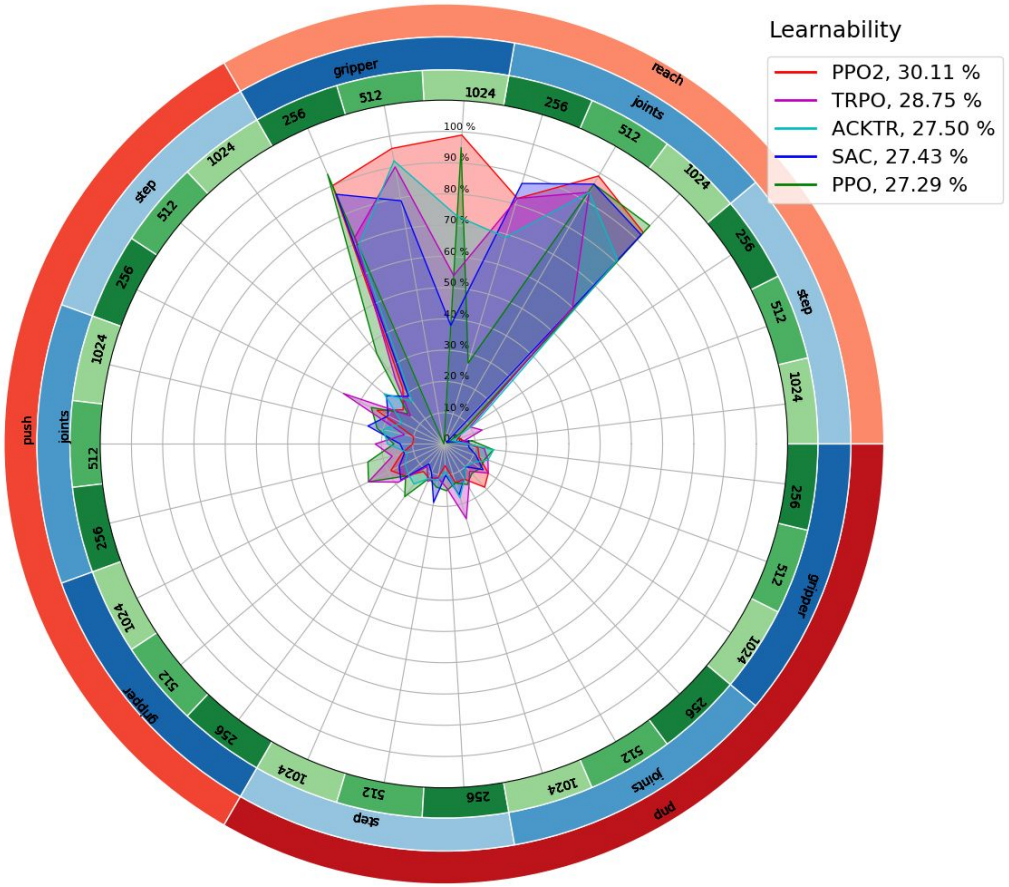
myGym

- RL network baselines:
  - PPO
  - PPO2
  - SAC
  - HER
  - TRPO
  - DDPG
  - ...



myGym

- RL network baselines:
  - PPO
  - PPO2
  - SAC
  - HER
  - TRPO
  - DDPG
  - ...

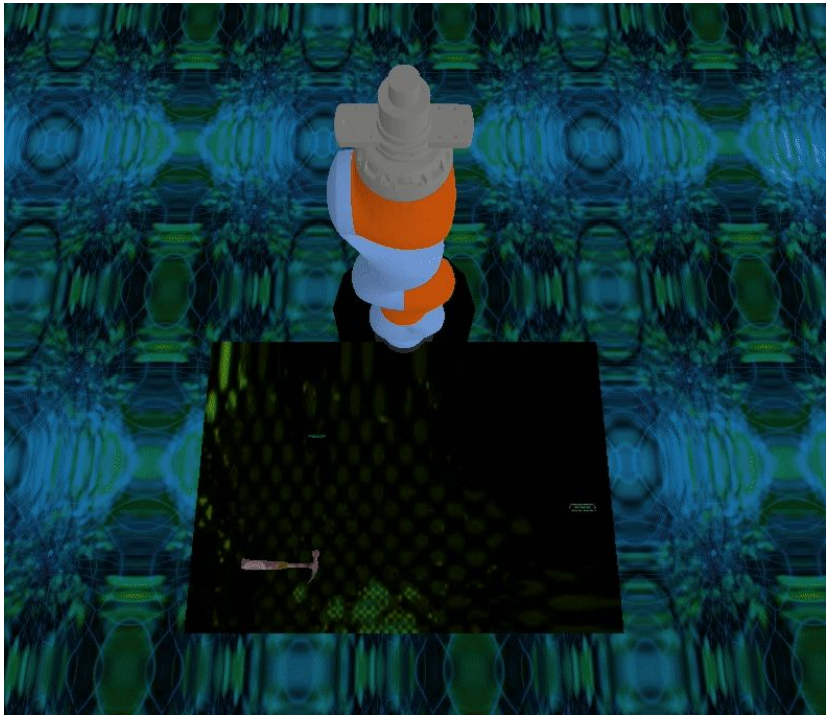


myGym

- pretrained visual networks
  - **YOLACT** - convolutional network for instance segmentation
  - detected objects positions - observation and reward calculation

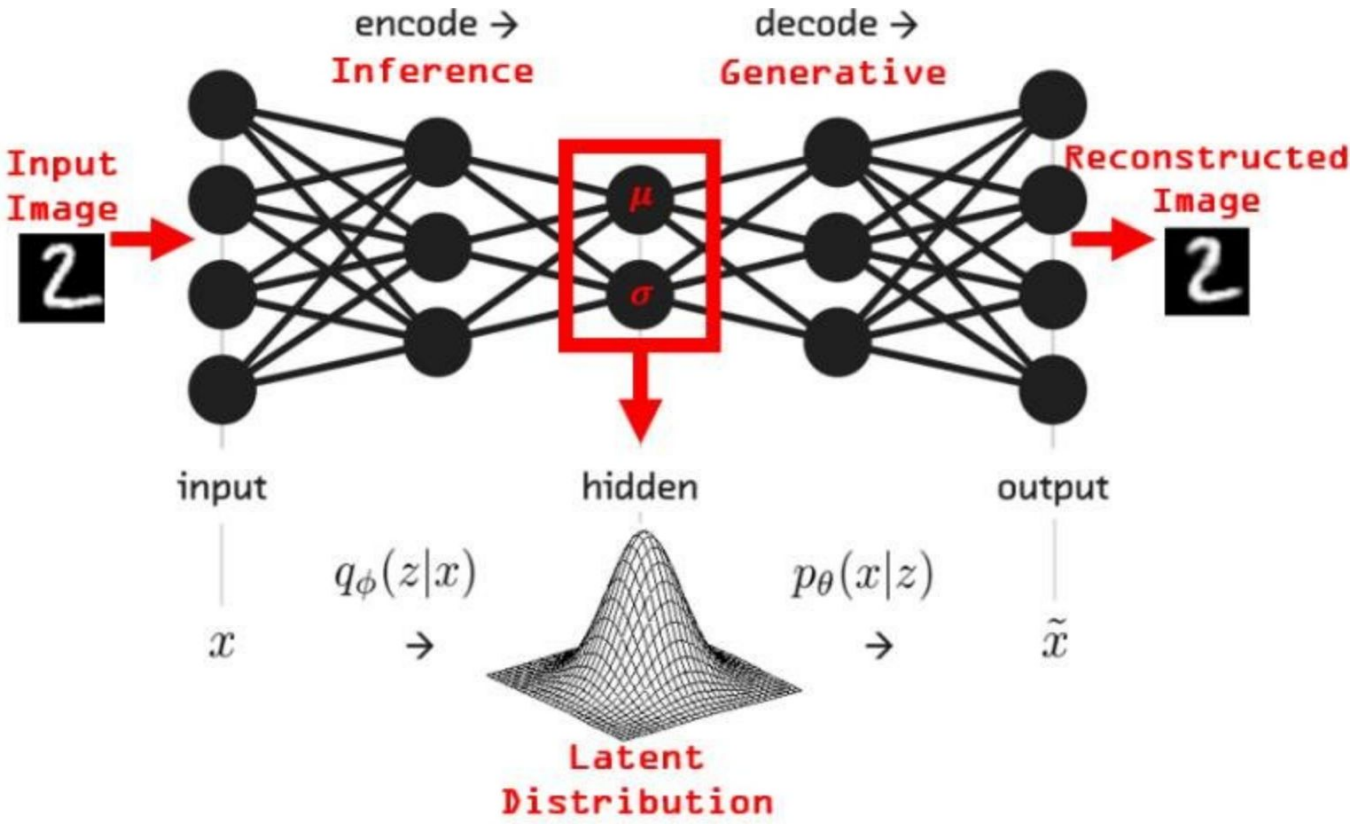
myGym

- pretrained visual networks
  - **YOLACT** - convolutional network for instance segmentation
  - possible to train on own dataset - code for generation included



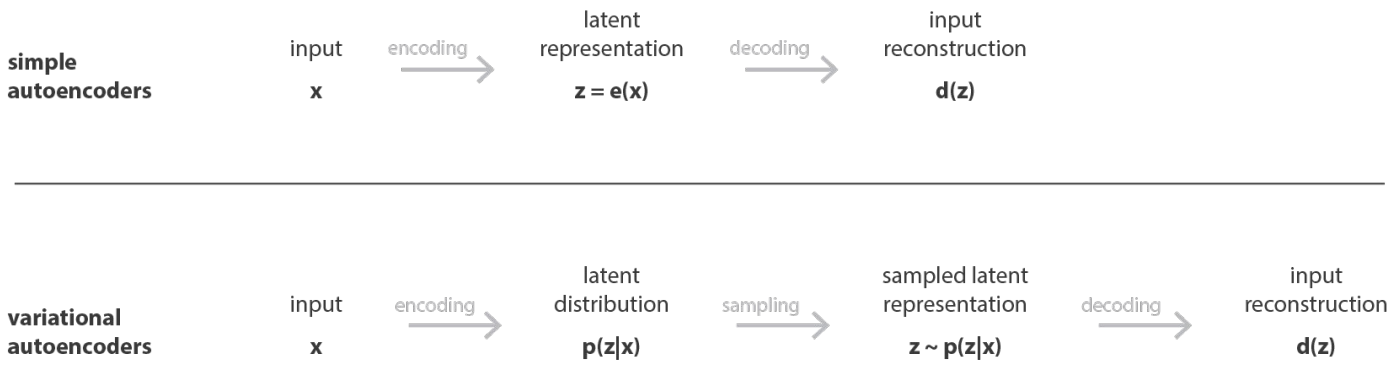
myGym

- pretrained visual networks
  - **variational autoencoder (VAE)** - a probabilistic version of autoencoders

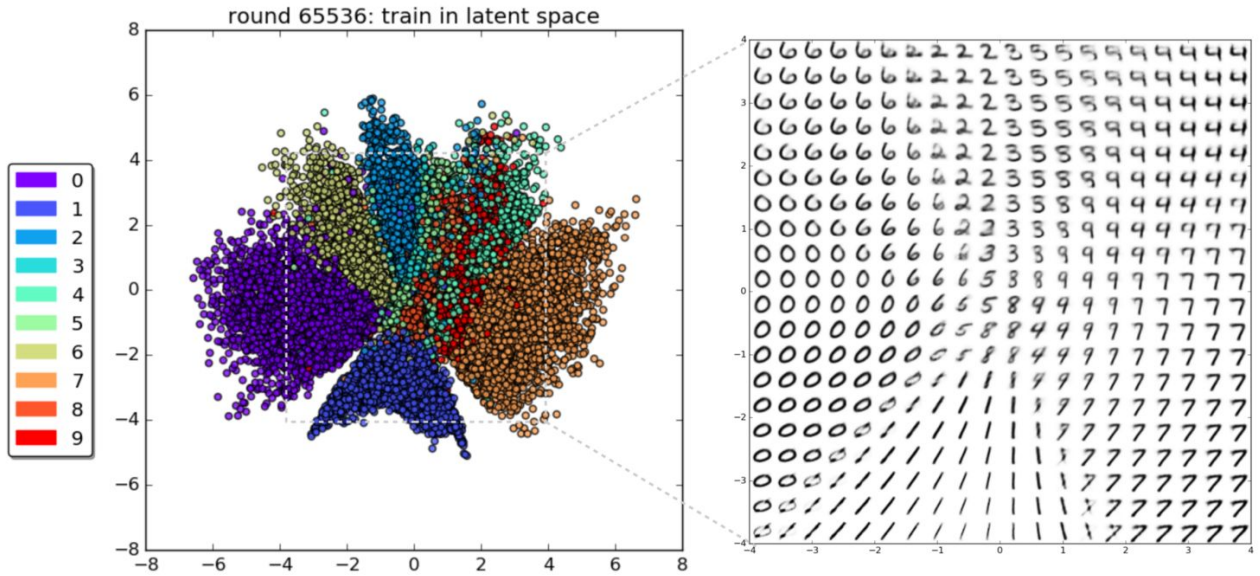


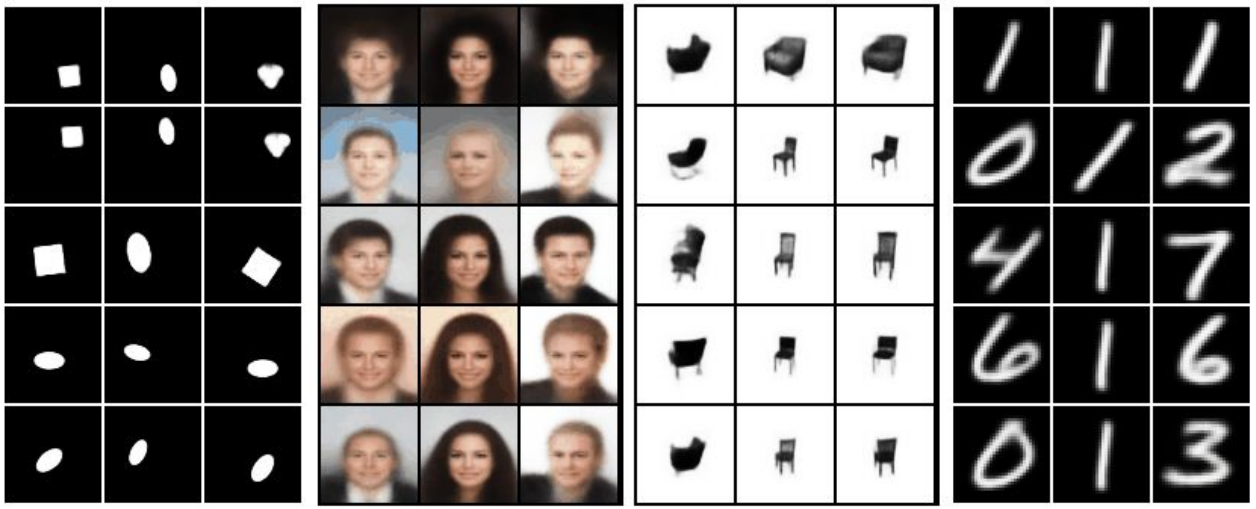
Variational Autoencoder

- continuous latent space - possible to generate new samples



Variational Autoencoder

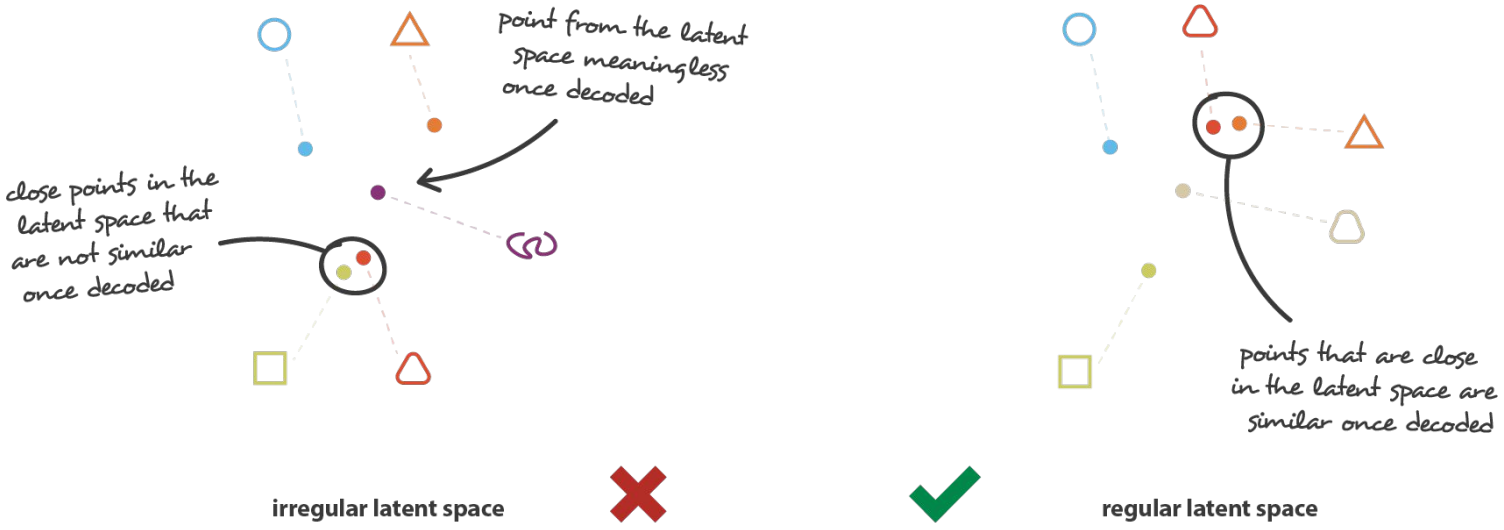




source

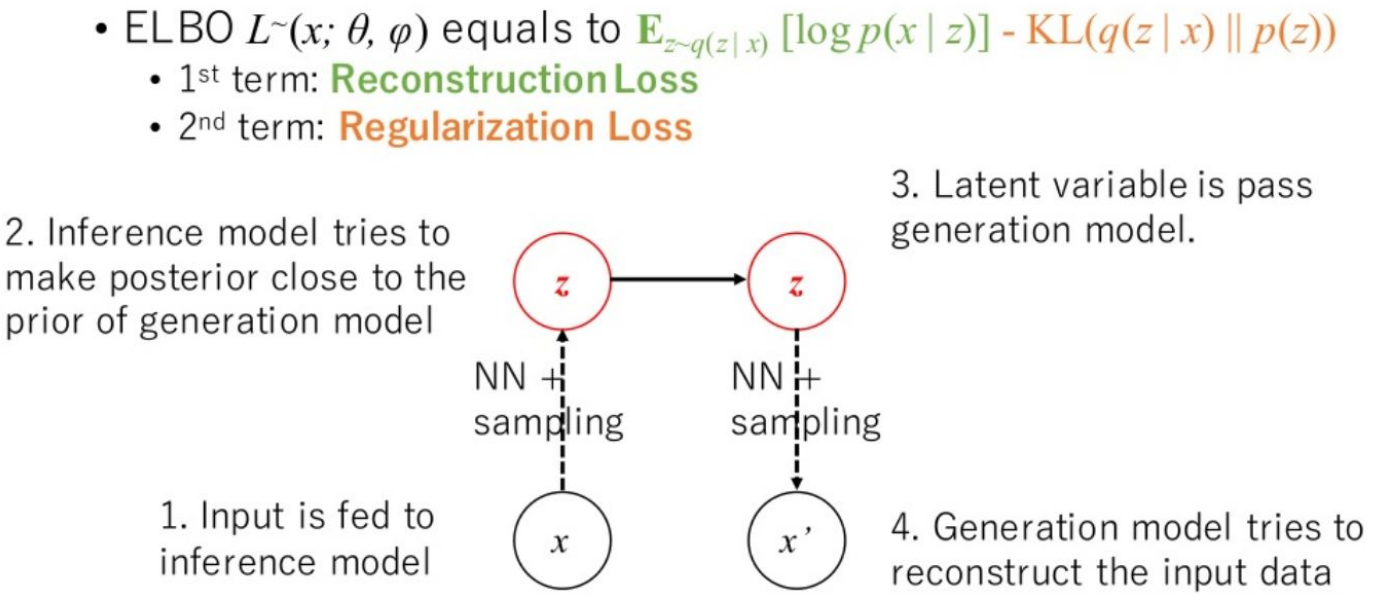
Variational Autoencoder

- VAEs can generate new data
  - only in case the latent space is regularised (organised)
- what is **regularity**?
  - **completeness** and **continuity** of the latent space



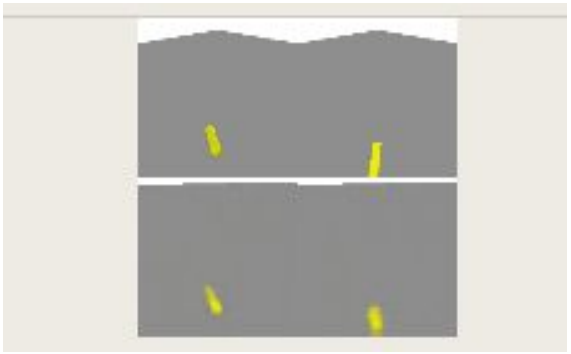
Variational Autoencoder

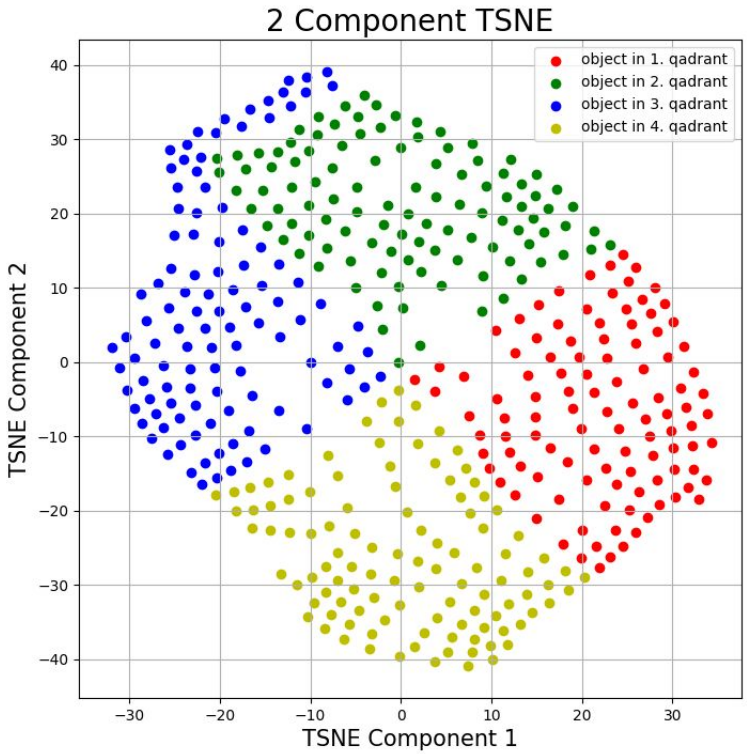
- training process - we maximize the Evidence Lower Bound (ELBO)



Variational Autoencoder

- usage:
  - compressed image representation
  - reward calculation - Euclidean distance between latent vectors
  - goal generation - unsupervised learning

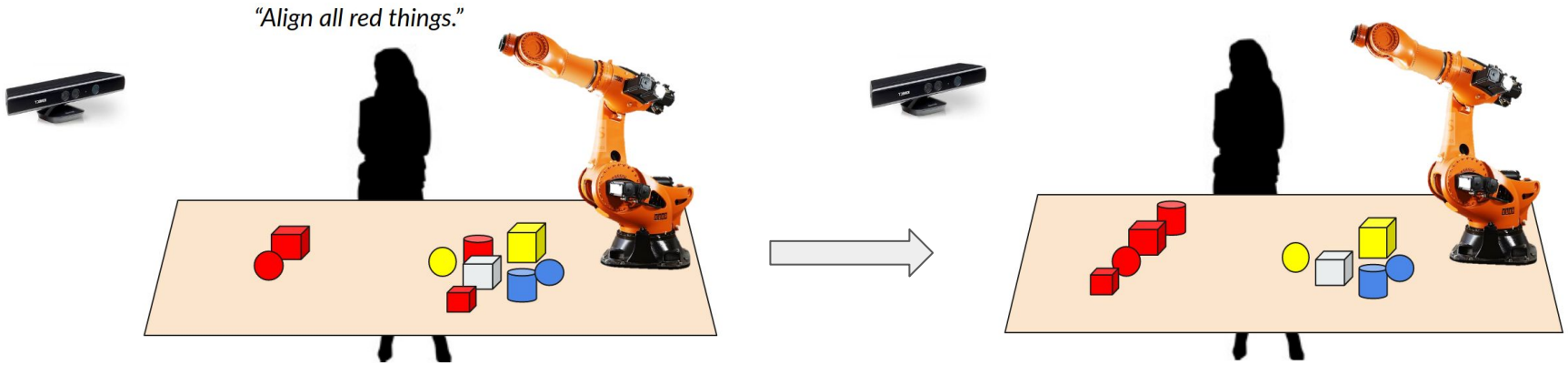




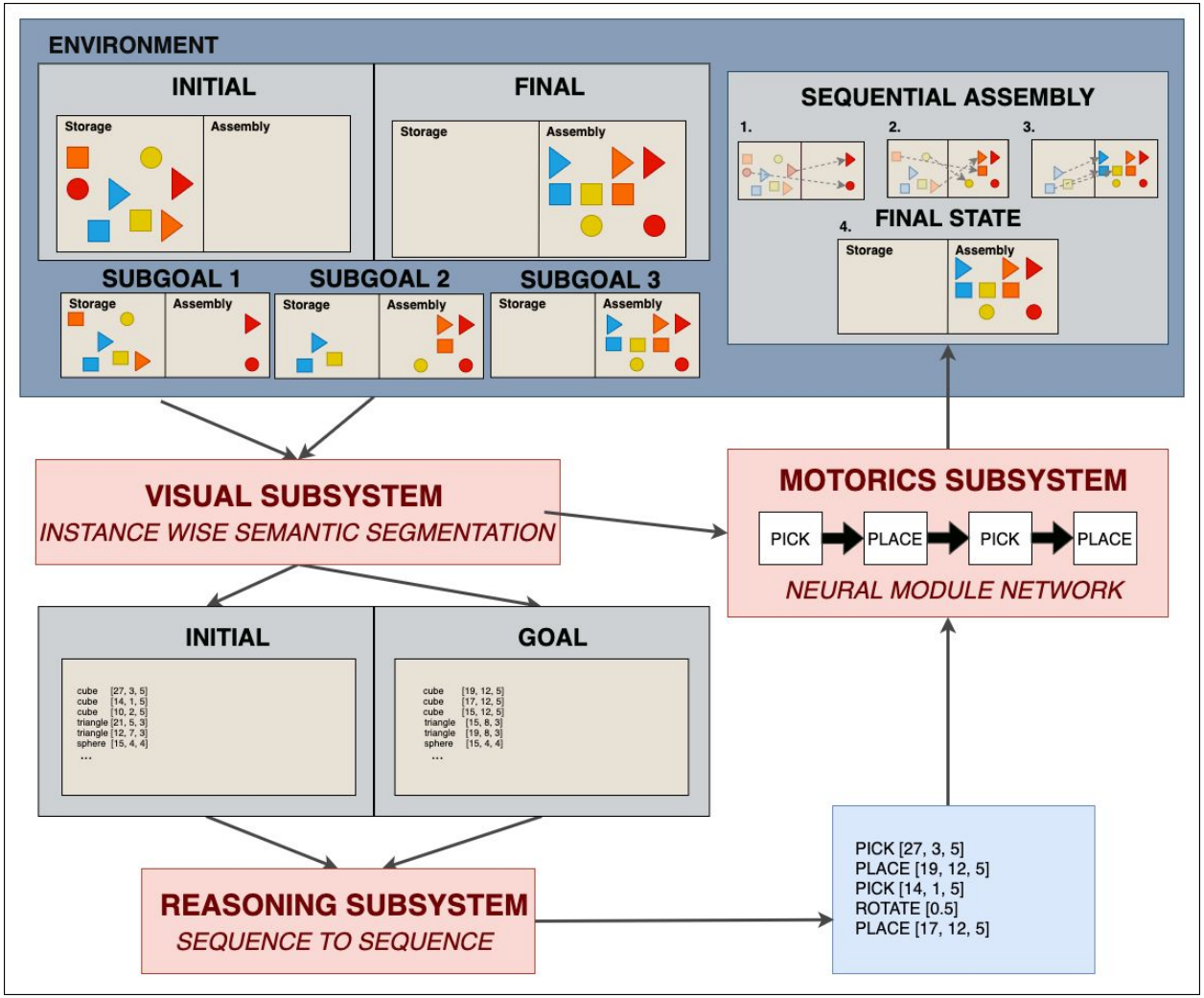
Connecting Vision, Language and Motorics

Compositional Actions

- inputs: natural language command, scene image
- output: robot motion



Compositional Actions



Conclusion

- learning individual motor skills (reach, push, pull, rotate...)
- chaining learned action primitives based on natural language command/goal image and scene image
  - compositional
  - explainable?
  - image / sentence representation using a variational autoencoder -> possible goal generation (unsupervised scenario)

Thank You!

References

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision* (pp. 2425-2433).

R. Hu, J. Andreas, M. Rohrbach, T. Darrell, K. Saenko, *Learning to Reason: End-to-End Module Networks for Visual Question Answering*. in ICCV, 2017.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., & Girshick, R. (2017). Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2901-2910).

Mascharka, D., Tran, P., Soklaski, R., & Majumdar, A. (2018). Transparency by design: Closing the gap between performance and interpretability in visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4942-4950).

Sejnova, G., Tesar, M., & Vavrecka, M. (2018). Compositional models for VQA: Can neural module networks really count?. *Procedia computer science*, *145*, 481-487.

Sejnova, G., Vavrecka, M., Tesar, M., & Skoviera, R. (2019, November). Exploring logical consistency and viewport sensitivity in compositional VQA models. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 2108-2113). IEEE.