Semantics to Space(S2S): Embedding semantics into spatial space for zero-shot verb-object query inferencing

Sungmin Eum[*†]
eum_sungmin@bah.com

Heesung Kwon[*]
heesung.kwon.civ@mail.mil

[*]U.S. Army Research Laboratory
[†]Booz Allen Hamilton

ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

DEVCOM ARMY RESEARCH LABORATORY Booz | Allen | Hamilton

Semantics to Space(S2S): Embedding semantics into spatial space for zero-shot verb-object query inferencing

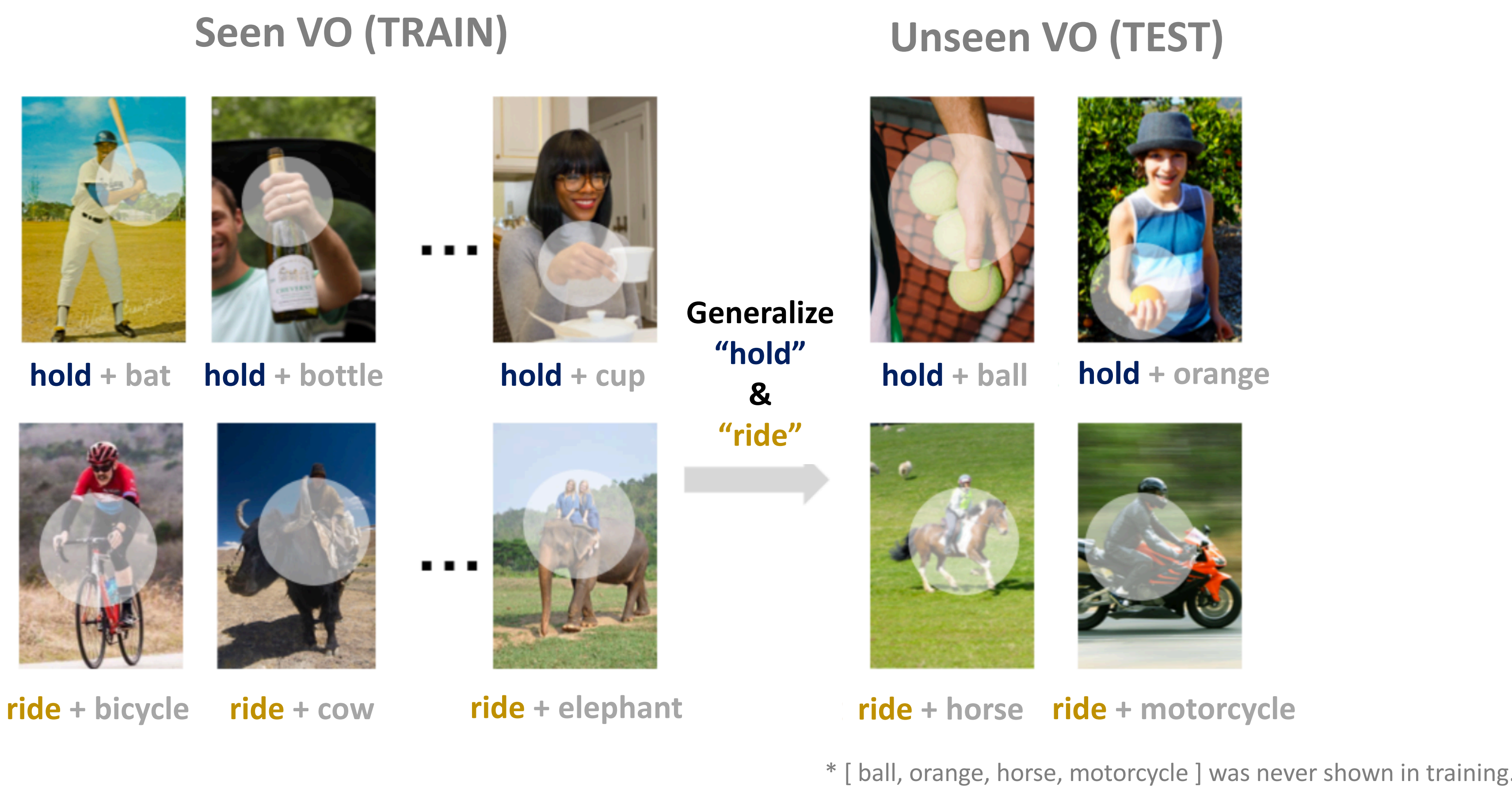
Verb-object query inferencing?

Recognize images by using verb-object expressions, e.g., “ride a horse”, ”hold ball”

Zero-shot?

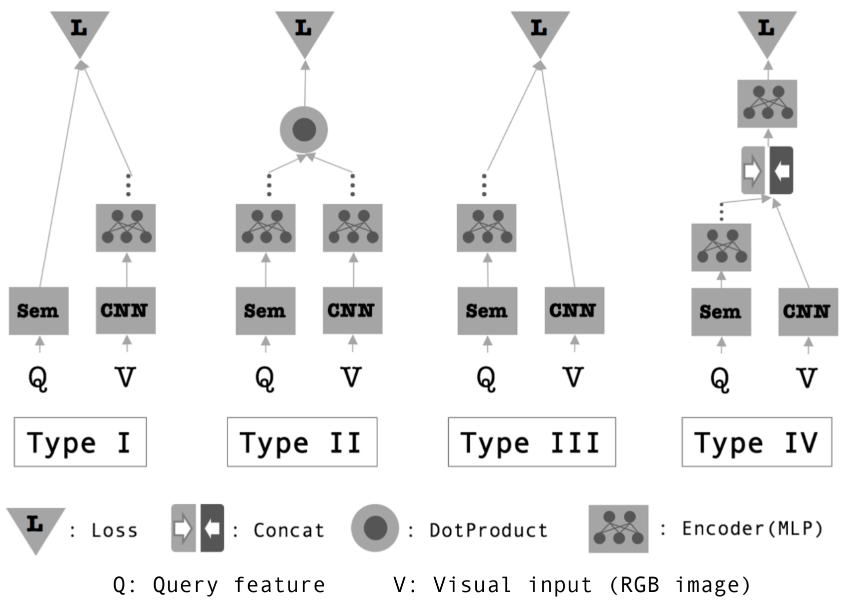
Be able to classify/recognize images that fall under “unseen” category

TASK: Zero-shot Verb-Object (VO) Inferencing



Comparison with previous work:

Previous work



Competition: What is the best joint embedding subspace for co-learning visual features with query/semantic features?

Ours

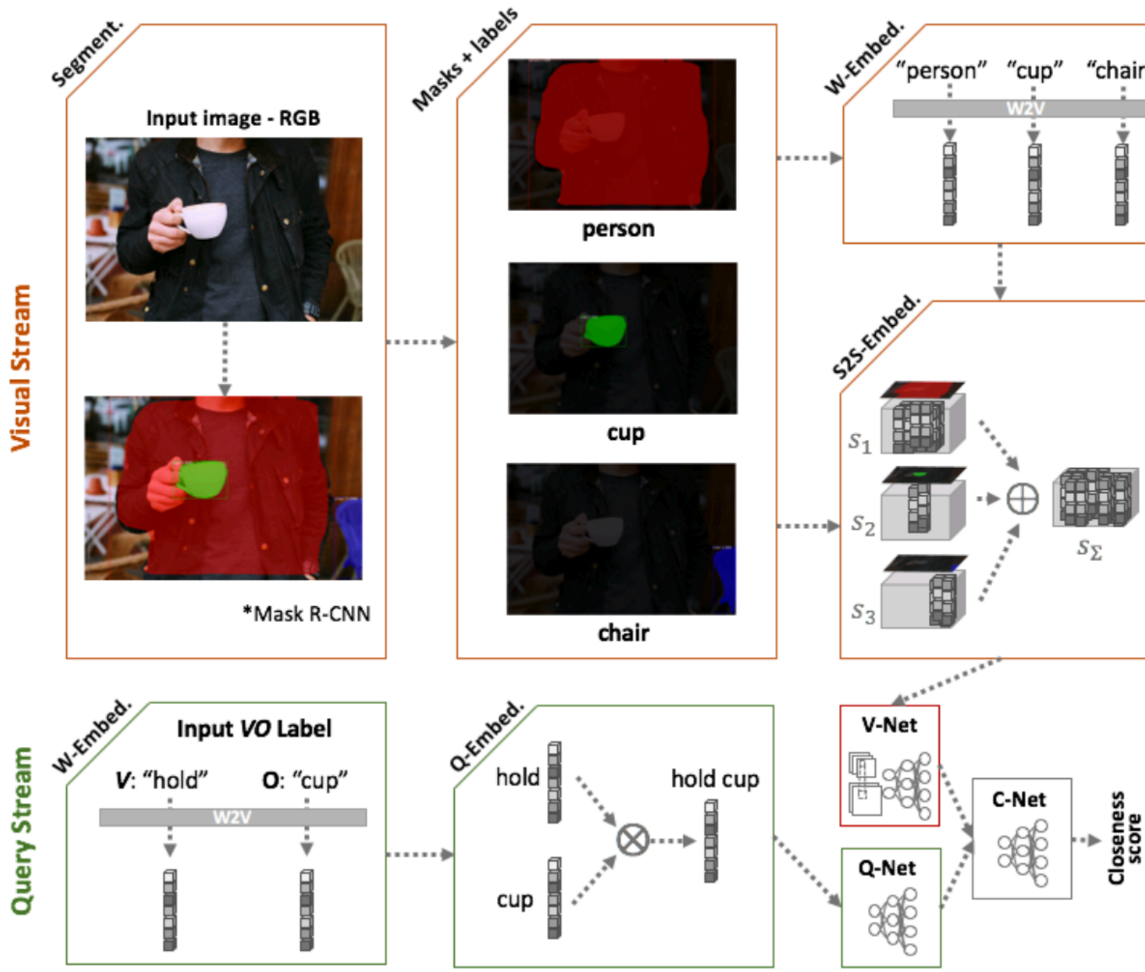
✓ Do NOT compete with previous approaches in searching for the best joint embedding subspace

Q1 "Why are the *semantics* only exploited for constructing query features but *not* for generating visual features when the end goal is to train a module to match the two?"

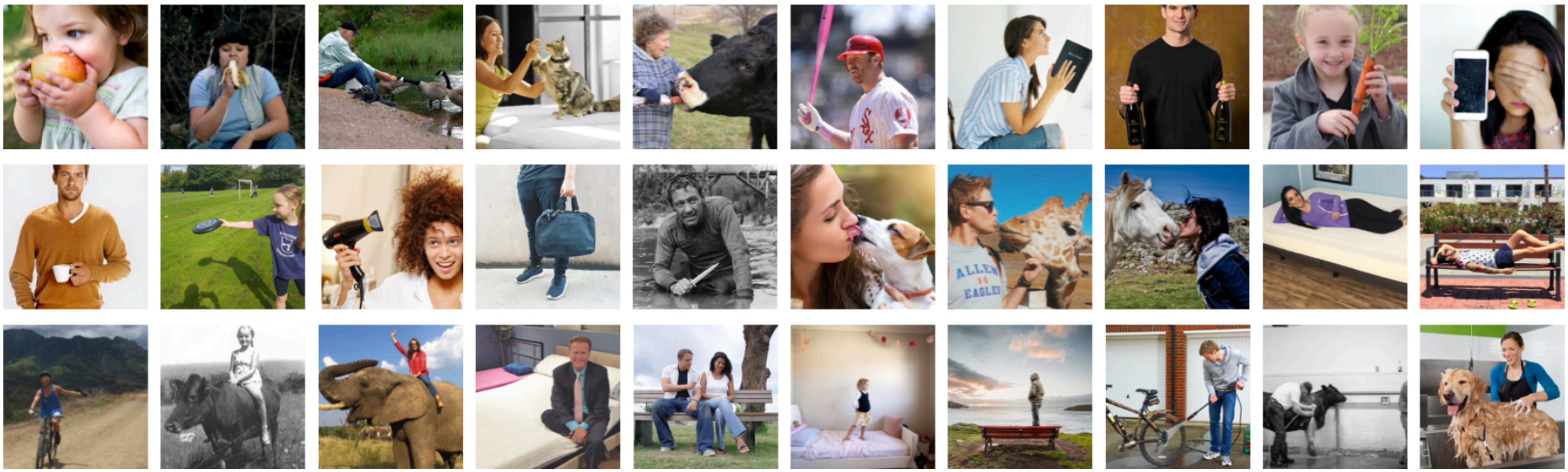
Q2 "Will the semantics be effective if embedded into both the visual and the semantic stream?"

✓ Directly embed the semantics into the spatial space of the visual stream (Contribution)

APPROACH: Semantics-to-Space (S2S)



DATASET: Verb-Transferability 60



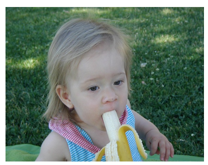
Verbs	Train	Test
eat	apple, banana	broccoli, donut
feed	bird, cat, cow	dog, giraffe, horse, sheep
hold	baseball bat, book, bottle, carrot, cell phone, cup, frisbee, hair dryer, handbag, knife	orange, scissors, skateboard, sports ball, surfboard, tennis racket, toothbrush, vase, wine glass
kiss	dog, giraffe, horse	bird, cat, cow
lie on	bed, bench	couch, surfboard
ride	bicycle, cow, elephant	horse, motorcycle, sheep
sit on	bed, bench	chair, couch
stand on	bed, bench	chair, couch
wash	bicycle, cow, dog	elephant, horse, motorcycle

EXPERIMENT: Scenario 1. Verb-Transferability

Evaluate how well the network transfers the “seen” verbs paired with “unseen" objects

Seen (Train)

eat + apple



eat + banana

+

eat	apple, banana
feed	bird, cat, cow
hold	baseball bat, book, bottle, carrot, cell phone, cup, frisbee, hair dryer, handbag, knife
kiss	dog, giraffe, horse
lie on	bed, bench
ride	bicycle, cow, elephant
sit on	bed, bench
stand on	bed, bench
wash	bicycle, cow, dog

Unseen (Test)

eat + broccoli?
feed + broccoli?
hold + broccoli?
kiss + broccoli?
lie on + broccoli?
ride + broccoli?
sit on + broccoli?
stand on + broccoli?
wash + broccoli?

EXPERIMENT: Scenario 1. Verb-Transferability

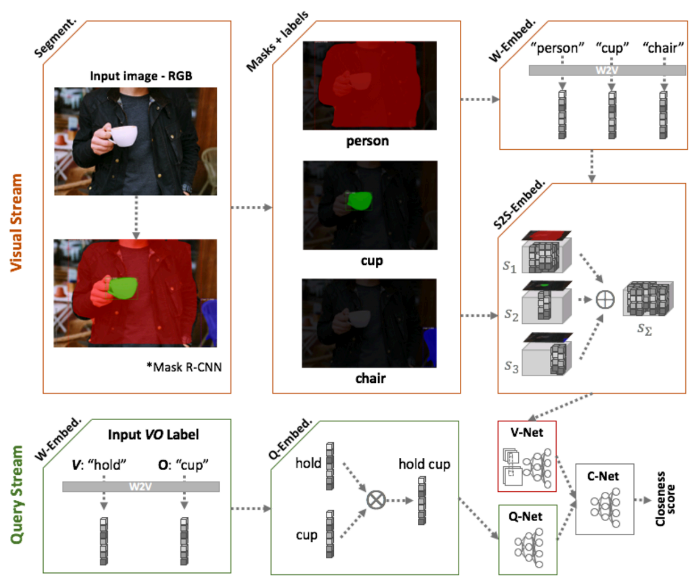
Evaluate how well the network transfers the “seen” verbs paired with “unseen" objects

Verb transferability evaluation. BASELINE: SUNG ET AL. [8], ORTHOVEC2S: ORTHONORMAL VECTORS-TO-SPACE, S2S: SEMANTICS-TO-SPACE. ALL THE NUMBERS INDICATE RECOGNITION ACCURACY IN [%].

Architecture	Sung et al.	OrthoVec2S	S2S
ResNet18	33.53	40.40	**46.27**
ResNet34	38.00	43.53	**48.87**
ResNet50	41.73	44.60	**50.47**

CONCLUSION:

Semantics to Space(S2S): Embedding semantics into spatial space for zero-shot verb-object query inferencing



- Introduced a simple, yet powerful semantics embedding approach for two-stream ZSL approach
- Augmented visual information by directly embedding the semantics in a spatial sense
- Validated that S2S can be used as a general module to enhance the performances of various ZSL baseline architectures

Please check out our manuscript!!