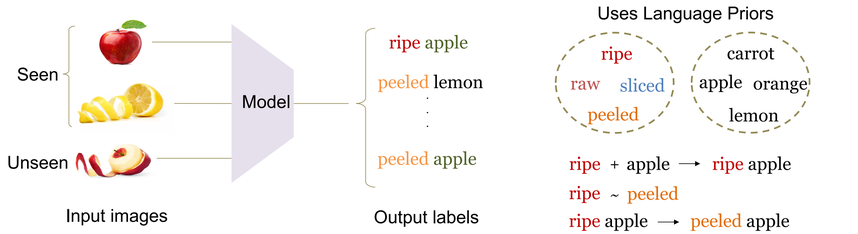
Beyond Seen Primitive Concepts and Attribute-Object Compositional Learning

Nirat Saini Khoi Pham Abhinav Shrivastava

University of Maryland, College Park

CVPR SEATTLE, WA JUNE 17-21, 2024

Compositional Zero-shot Learning (CZSL)



Goal: To compose *seen attributes* and *objects* for recognizing *unseen compositions.*

Method: Semantic correlations in linguistic embedding space help recognizing unseen compositions.

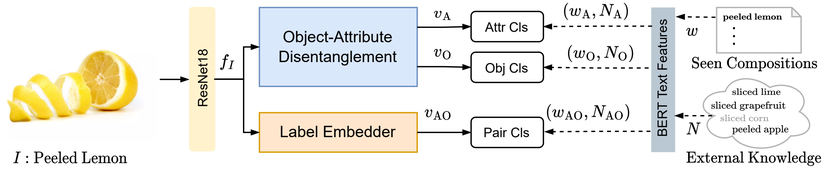
Assumption: All objects and attributes are individually seen during training.

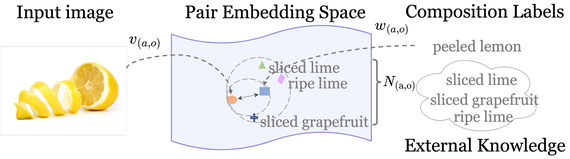
Practical Settings: All objects and attributes are not present in training set. How to learn composition for *unseen attributes and objects*?

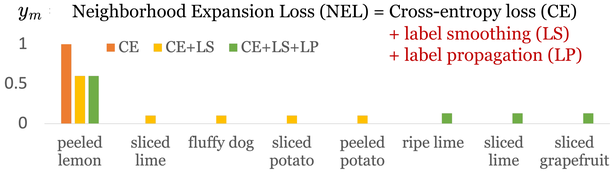
Pipeline

Proposed method is built on top of OADis[1] (baseline for CZSL)

- Images and labels are embedded together (Vision+Language : Pair Embedding Space).
- Visual features are separated for attributes and objects (Object-Attribute Disentanglement).
- Composition classification is a separate head (Label Embedder).
- Each embedding space (attribute, object and pair) is regularized using linguistic losses.
- Neighborhood Expansion Loss transfers knowledge from seen to unseen concepts and compositions.



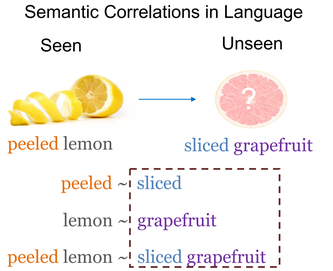


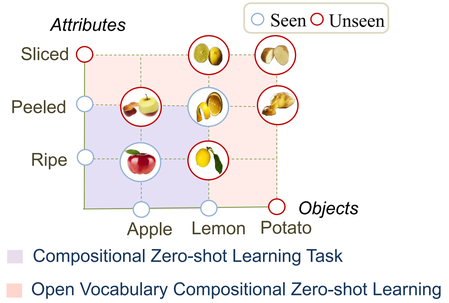


Neighborhood Expansion Loss is combination of:

- Cross Entropy Loss (CE): $H(y, C) = \sum_{m=1}^{M} -y_m \log(C_m)$
- Label Smoothing (LS): Avoids negative bias towards unseen labels
- Label Propagation (LP): Transfer from seen to unseen labels (from External Knowledge Source)

Open Vocabulary Compositional Zero-shot Learning (OV-CZSL)





Goal: To recognize attributes, objects and their compositions beyond seen vocabulary.

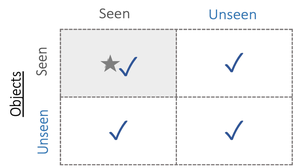
Training: set of compositions of seen attributes and objects

Testing: 4 splits of compositions of seen and unseen attributes and objects

Method: Use External Knowledge source and transfer knowledge from seen to unseen attributes, objects and their compositions.

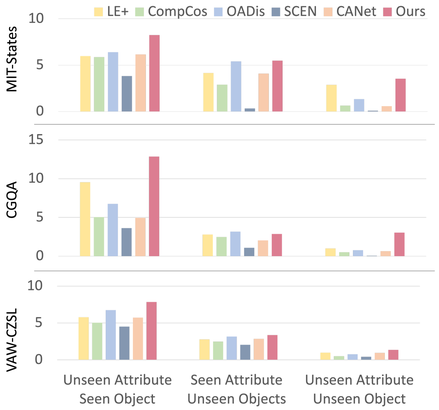
In-the-wild generalization for Open Vocabulary Compositions.

Test splits for CZSL and OV-CZSL

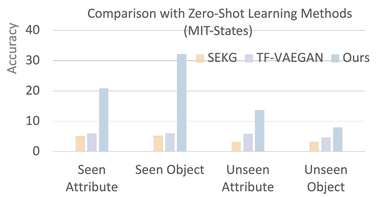


Results

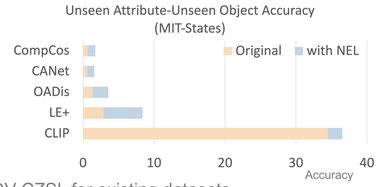
Quantitative Results on Benchmark splits



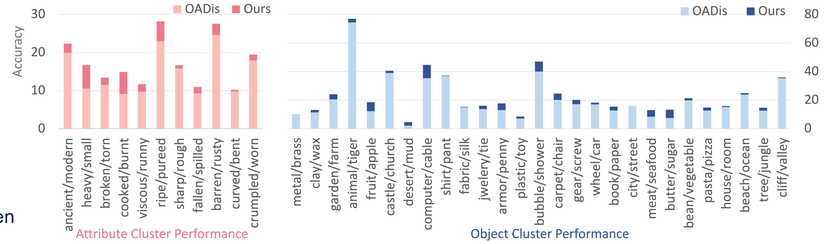
Zero-shot Learning Baselines for Seen and Unseen Attribute and Object Accuracy



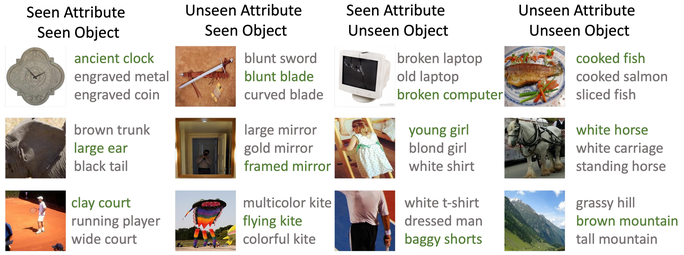
NEL as plug-and-play module improves Unseen Composition Accuracy for baselines



Our approach with NEL shows improvement in generalizing to similar unseen concepts when clustering all Attributes and Objects with seen and unseen concepts



Top 3 predictions for Images



- Experiments are done on a new benchmark split for OV-CZSL for existing datasets.
- We report Top@1 AUC for MIT-states and CGQA, and Top@3 and 5 AUC for VAW-CZSL.
- AUC (%) is computed between seen and unseen compositions with different bias terms and chosen where HM is maximum.

1: Disentangling Visual Embeddings for Attributes and Objects, CVPR 22.