# Identifying Negative Exemplars in Grounded Language Data Sets

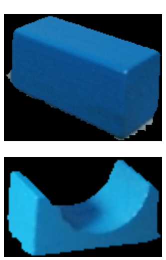

Nisha Pillai, Cynthia Matuszek

{ npillai1 | cmat } @ umbc.edu

# Task



- Learn the meaning of a word from natural conversation despite not having negative examples
  - Learn an association from language to perceived environment
  - Visual percepts ←→ attribute words
  - Joint model of visual percepts and natural language to identify novel object, shape, and color described by tokens (words)[1]
- Obtain important positive terms to learn
- Find appropriate negative examples
  - statistical language comparison metrics



**Blue**
**(positive example)**

**Not blue**
**(negative example)**

[1] Matuszek, FitzGerald, Zettlemoyer, Bo, Fox. ICML 2012.

# Motivation



- Unavailability of negative examples in natural conversation

“this is a lemon”
“this object is an yellow ball”
~~“this is not a carrot”~~

- Difficulty in gathering negative information without prompting
- Lack of positive label may not be a negative!

“this is a lemon” $\not\Rightarrow$ “not yellow”

# Goals



- Choose words to learn
  - Relevant, semantically meaningful, important
- Find an efficient way of obtaining negative examples
- Measure effectiveness of choices for language acquisition

Choose to train "banana" classifier

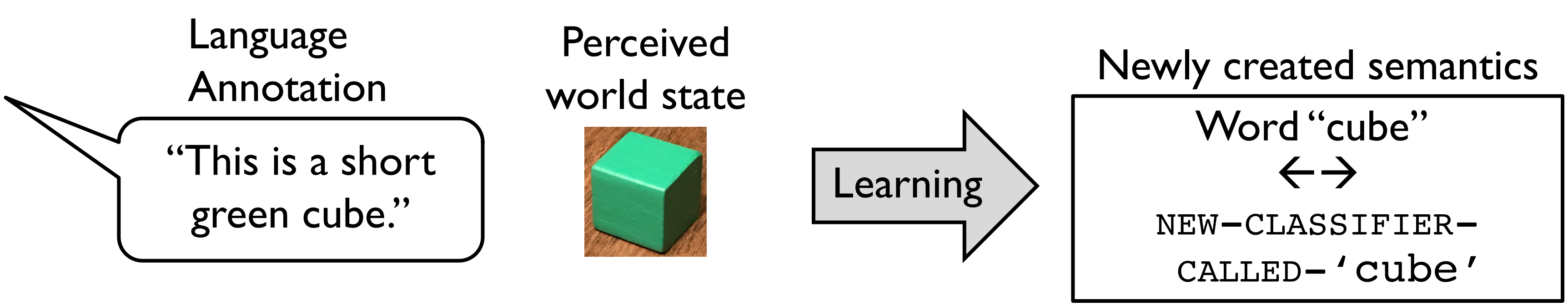
Positive example

Negative example

# Grounding Training



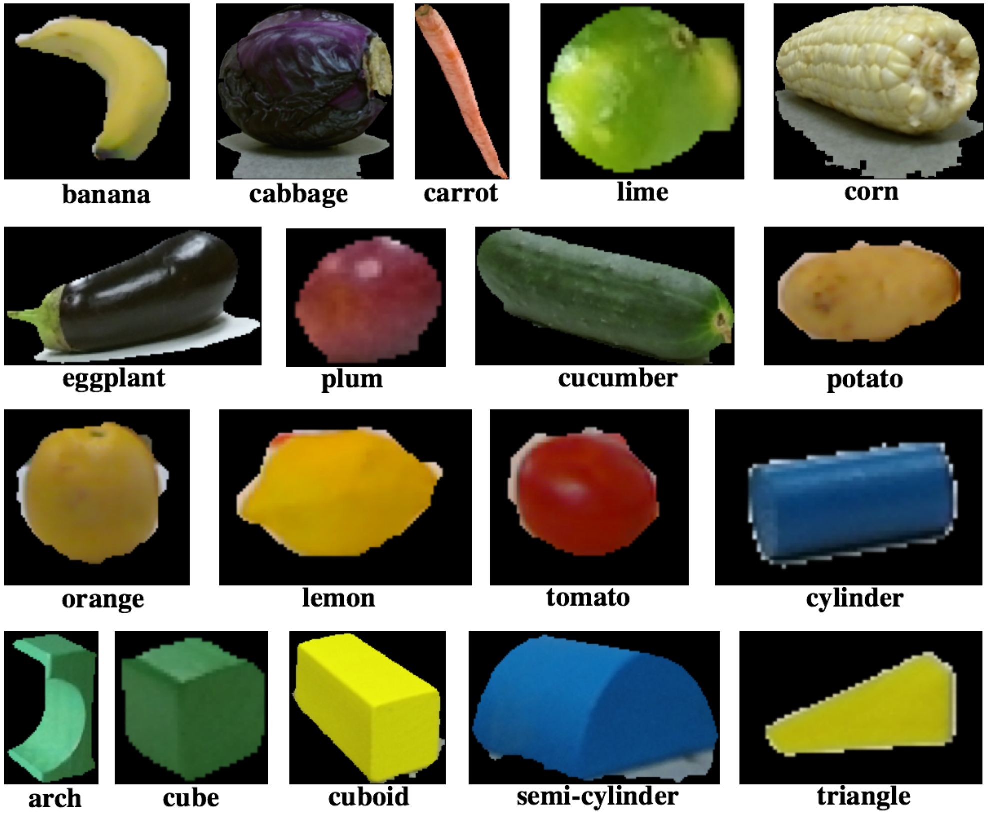
- Training visual classifiers based on percepts
- When new language tokens are encountered:
  - Important tokens selected
  - Visual classifiers created and trained on perceptual context
- As more objects are seen, 'best' classifier emerge
  - E.g., most predictive of data observed so far



# Data Corpus Collection

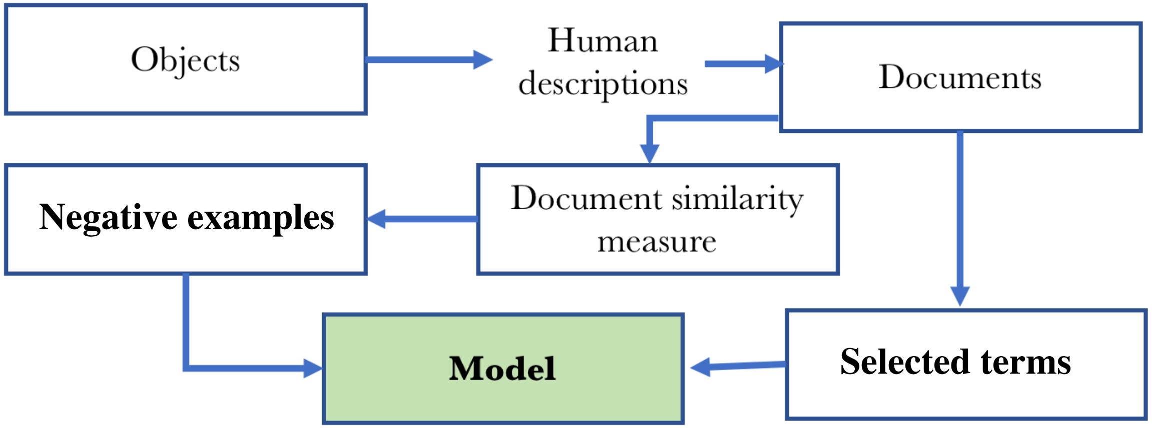


- 72 classes, 18 categories
  - Food objects
  - Children's blocks
- Descriptive Language:
  - 3055 descriptions from Mechanical Turk
  - 19,947 unique words
    - 200-450 words/document
- 230 unique tokens selected for learning



# Approach Overview

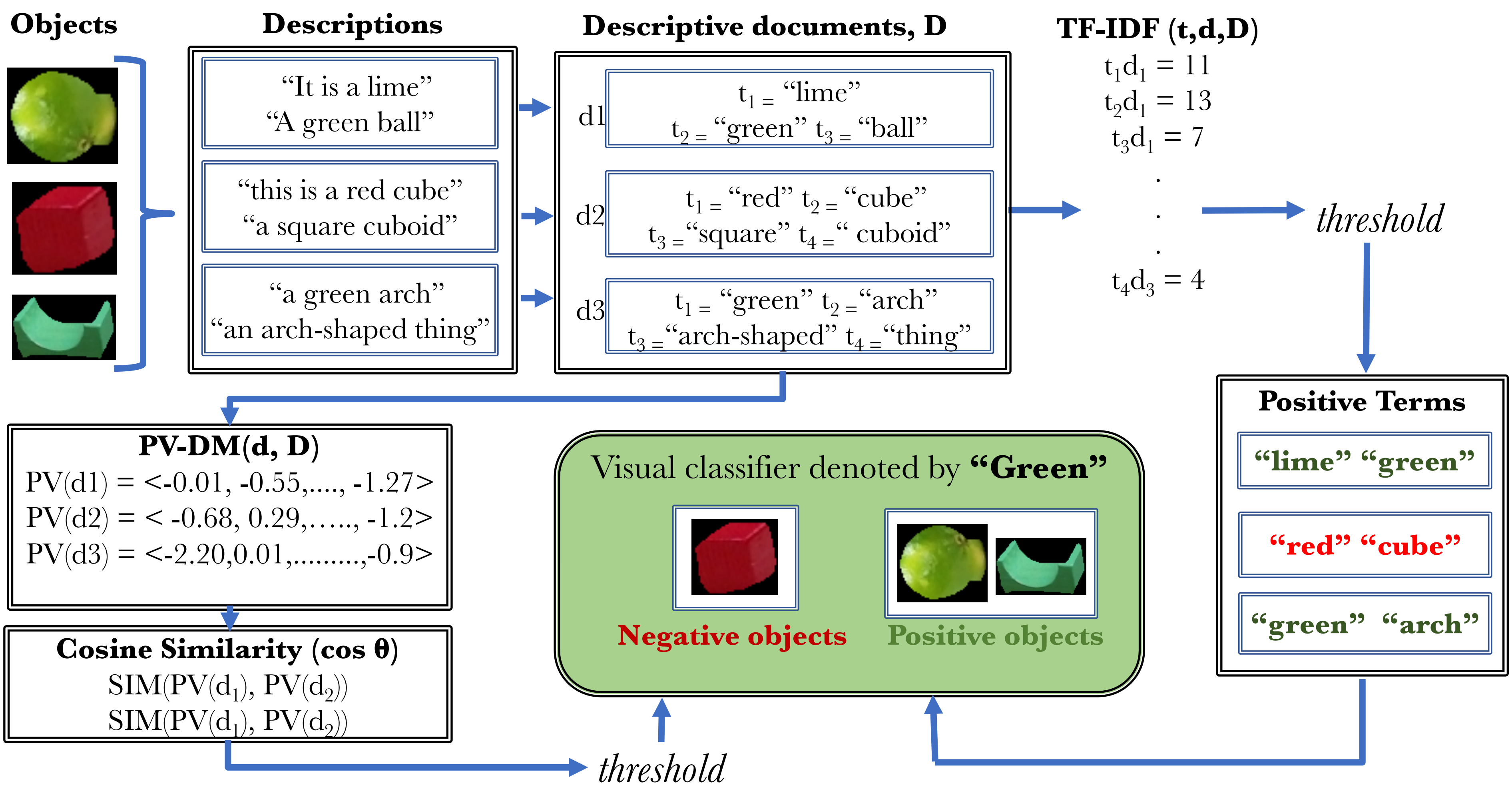
- **Dataset:** real world objects (toys, food)
- **Language:** crowdsourced human descriptions
- **Documents:** set of a descriptions of each object




- **Positive labels:** visually meaningful words worth learning
- **Negative examples:** objects chosen as negatives for them

# Approach Overview 2





# Choosing Words to Learn



- Positive labels: choosing visually meaningful words to train classifiers for

- tf-idf: *term frequency-inverse document frequency*
  - How important a word is to a document
  - ***Increases*** proportionally to the number of times a term appears in the document
  - ***Decreases*** with the number of documents containing that term

$$\boxed{tf\text{-}idf(t,d,D) = tf(t,d) \cdot \log \frac{N}{|\{d \in D : t \in d\}|}}$$

*tf(t,d)* - the number of times a term t appears in document, d.
$|\{d \in D : t \in d\}|$ - the number of documents in which the term t appears.
*N* - the size of the set of documents. |D|

# Document Features



**Negative examples:** semantically distant objects using Paragraph Vector[2] and cosine distance

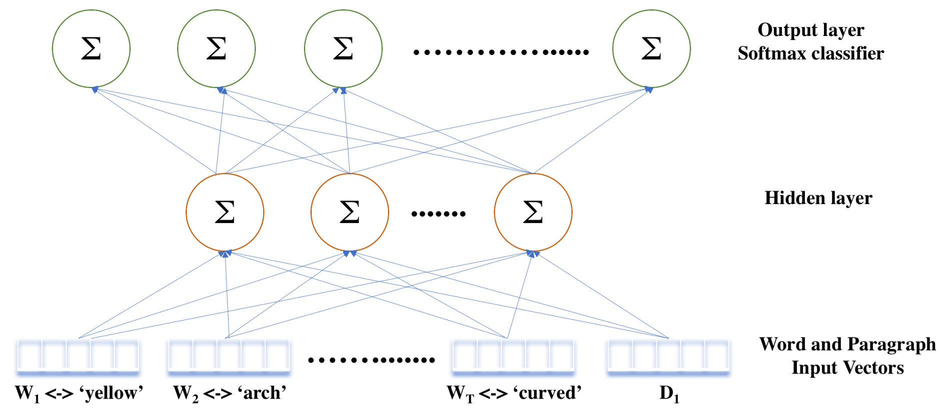
- Log probability vector

  $y = b + Uh(w_{t-k}, \ldots w_{t+k}; W, D)$

  *U, b – Softmax parameters*

  *h – average of W's and D*

  *k – context window parameter*




Learning using softmax classifier

$$p(w_t | w_{t-k}, \ldots, w_{t+k}) = \frac{e^{y_{w_t}}}{\sum_i e^{y_i}}$$

Maximize average log probability: $\frac{1}{T} \sum_{t=k}^{T-k} \log p(w_t | w_{t-k}, \ldots, w_{t+k})$

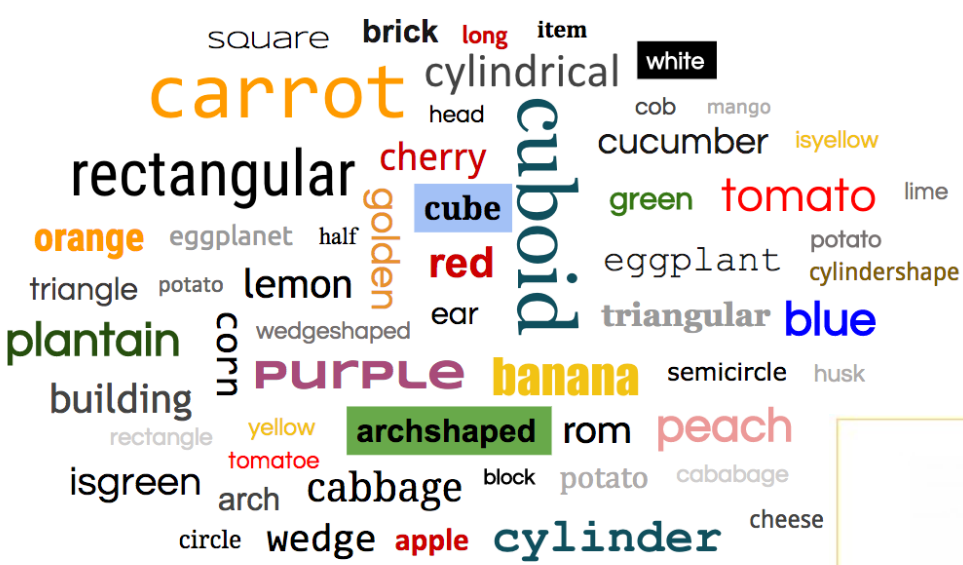
- Cosine similarity:

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|_2 \|\mathbf{B}\|_2}$$ - cosine of angle between documents in vector space

[2] Quoc Le and Tomas Mikolov. ICML 2014.

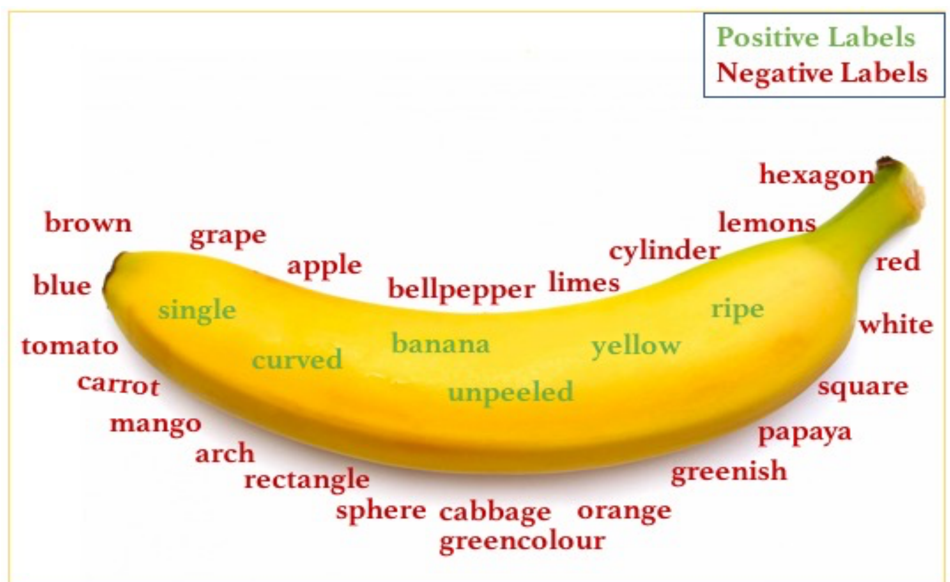
# Term Selection





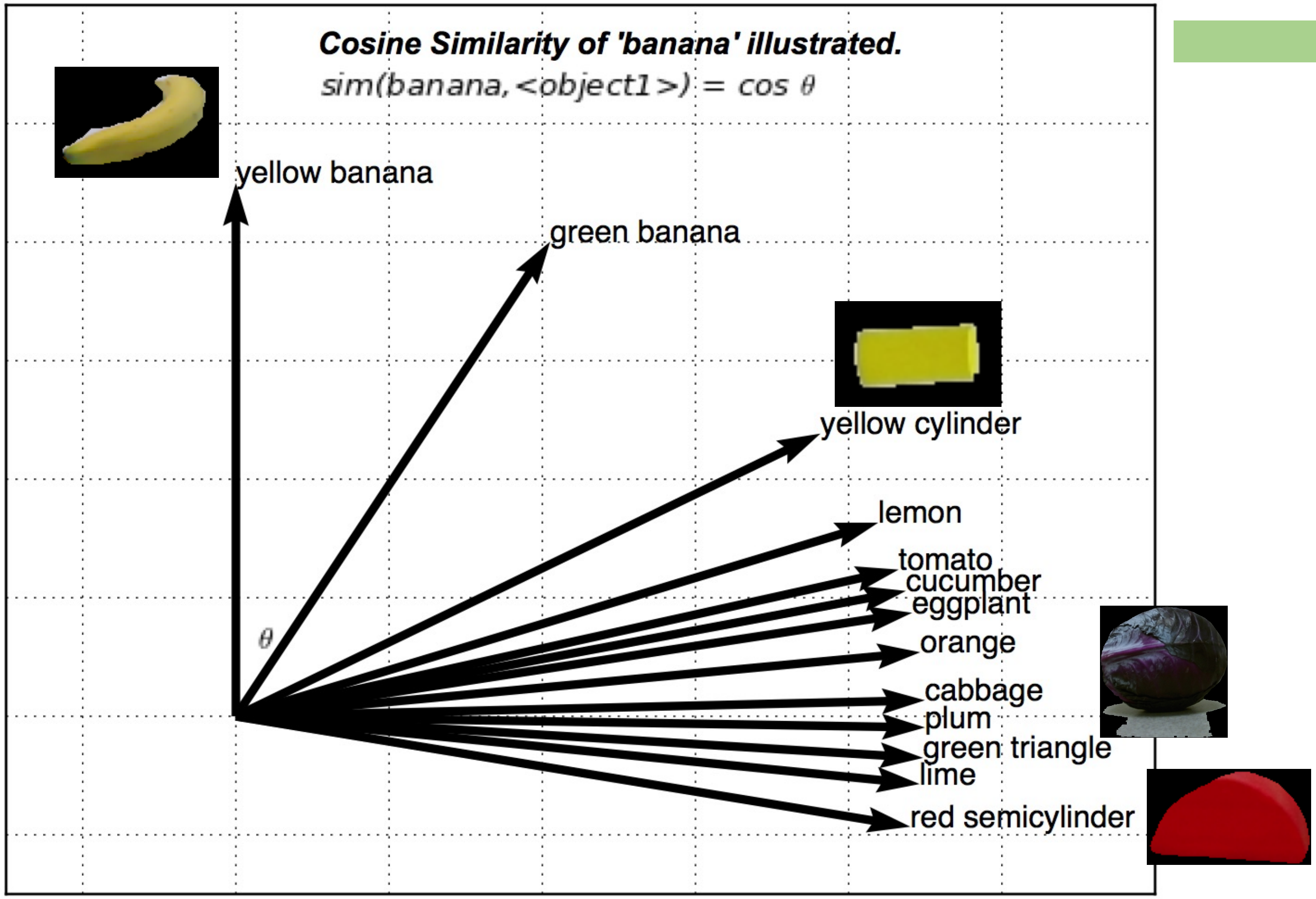

- 57 top terms
- Human errors
  - Tomato / tomatoe
  - Eggplant / eggplanet

- tf-idf positive / negative labels
  - "Arch" is negative ☹
  - PV-DM fixes this



# Choosing Negatives

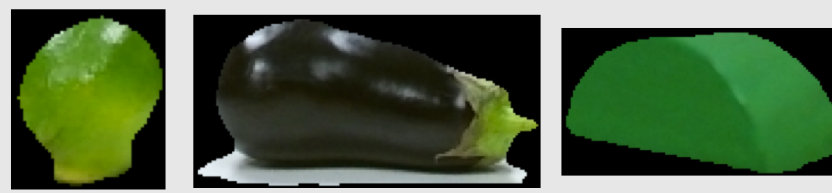
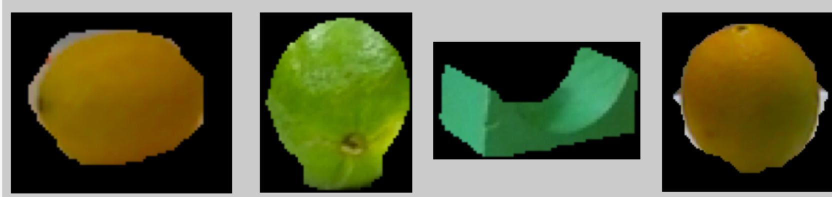





- Vectors → individual objects
- Angle → similarity of descriptions

# Example Results



| Label | Positive Examples | Negative Examples |
|---|---|---|
| "carrot" |  |  |
| "rectang-ular" |  |  |
| "red" |  |  |

# Color and Shape



**Ground truth**

**Color classifier denoted by "term"**

| | yellow | red | green | white | orange |
|---|---|---|---|---|---|
| "yellow" | 0.93 | 0.20 | 0.37 | 0.05 | 0.02 |
| "building" | 0.09 | 0.11 | 0.00 | 0.00 | 0.17 |
| "red" | 0.00 | 0.89 | 0.05 | 0.16 | 0.35 |
| "green" | 0.27 | 0.00 | 0.89 | 0.02 | 0.00 |
| "tomato" | 0.24 | 0.94 | 0.00 | 0.00 | 0.00 |
| "white" | 0.06 | 0.68 | 0.55 | 0.85 | 0.73 |
| "orange" | 0.50 | 0.93 | 0.21 | 0.26 | 0.66 |

**Ground truth**

**Shape classifier denoted by "term"**

| | cube | cylinder | sphere | arch | triangle |
|---|---|---|---|---|---|
| "cylinder" | 0.32 | 0.87 | 0.06 | 0.29 | 0.29 |
| "rectangular" | 0.82 | 0.43 | 0.51 | 0.78 | 0.30 |
| "circle" | 0.25 | 0.25 | 0.75 | 0.26 | 0.21 |
| "archshaped" | 0.29 | 0.27 | 0.12 | 0.82 | 0.33 |
| "triangle" | 0.54 | 0.60 | 0.52 | 0.31 | 0.82 |

- Performance of trained model
  - Ability to correctly classify held-out test set
- Goal: classifiers associated with attribute keywords have strong predictive power (only)

# Results: Object identification



- Object classification:
  - Possibility of learning more complex concepts
  - Good performance on interacting problem

**Ground Truth**

**Object classifier denoted by "term"**

| | corn | semi-cylinder | banana | eggplant | tomato |
|---|---|---|---|---|---|
| "corn" | 0.92 | 0.01 | 0.77 | 0.04 | 0.00 |
| "building" | 0.08 | 0.61 | 0.30 | 0.02 | 0.03 |
| "banana" | 0.00 | 0.15 | 1.00 | 0.00 | 0.04 |
| "tomato" | 0.00 | 0.00 | 0.05 | 0.00 | 0.94 |
| "wedge" | 0.49 | 0.30 | 0.00 | 0.43 | 0.00 |
| "eggplant" | 0.26 | 0.24 | 0.01 | 0.84 | 0.11 |

# Future Work



- A thorough evaluation in positive and negative term selection
  - Use Amazon Mechanical Turk
- Comparison of model with a traditional base model
- Evaluate the model in a more 'real world' problem
  - A more varied set of objects.
  - Additional kinds of classifiers.
  - Complex visual classification tasks.

# Conclusion



- Semantic representations of their perceived environments
- Discovered ground truth labels
- Document similarity metrics in negative exmple selection
  - Efficient in unprompted human interaction scenario
  - Effective for grounded language acquisition tasks

Thank you!
Questions?