A COMPUTATIONAL ACQUISITION MODEL FOR MULTI MODAL WORD LEARNING

[1] The Hebrew University of Jerusalem

[2] University of Melbourne

Uri Berger[1], Gabriel Stanovsky[1], Omri Abend[1], Lea Frermann[2]

Introduction

Tasks learned in early infant language acquisition:

1 Noun identification: as an early cue for syntactic structure

"I am throwing the ball to the dog" → Noun identification → "I am throwing the ball to the dog" → 3 nouns → Main verb is transitive

"Mommy is eating" → Noun identification → "Mommy is eating" → 1 noun → Main verb is non-transitive

Can predict word concreteness as an approximation
(most concrete words are nouns, and children learn concrete nouns first)

2 Semantic clustering of words

- Used to estimate word similarity
- Used as context:
In experiments, when preceded by a word from the same cluster, a target word was processed faster

3 Object recognition

- Identify the location of objects in an image
- Cluster the objects to gradually learned clusters

Objective

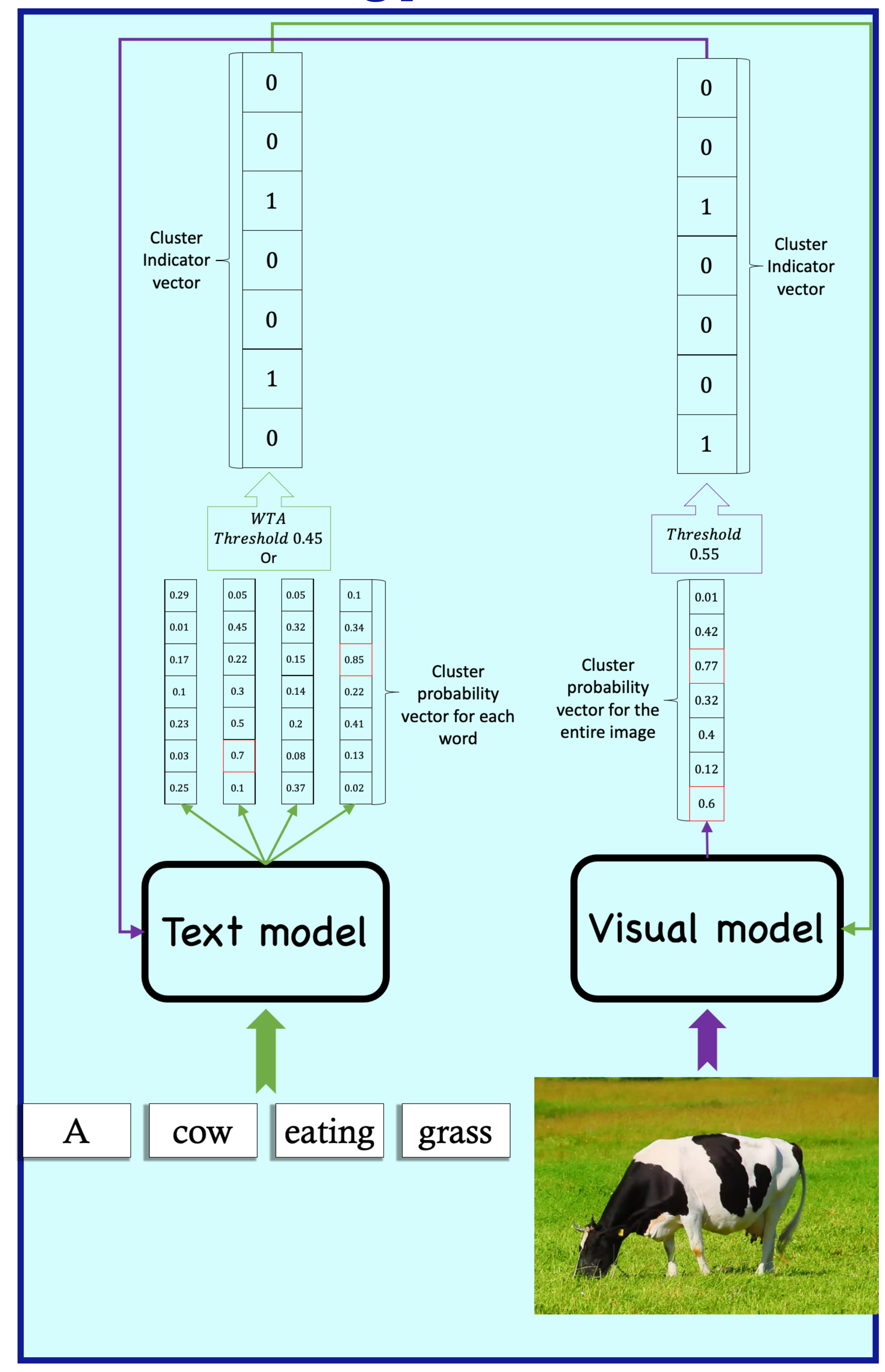
- What information can be acquired from **raw image + caption** pairs, without **any pre-training** or **external supervision**.

- Can **concreteness prediction**, **word semantic clustering** and **object detection** be learned?

Previous work

- Used pre-trained visual model or structured input instead of raw images
- Unrealistic as a cognitive setting

Methodology



Text model:
1. Compute $p(cluster|word)$ for each word in the input and each cluster
2. For each word select cluster with highest probability if it exceeds threshold θ_t, otherwise select none
3. The final prediction the union of all predicted clusters

Visual model:
1. Compute $p(cluster|image)$ for each cluster
2. Select clusters for which $p(cluster|image)$ exceeds threshold θ_v

Learning:

Visual model learning:
- Text output vector supervises the visual model
- Visual loss function compares visual probability vector and text output vector

Text model learning:
- Visual output vector supervises the text model
- We use a simple word-cluster co-occur count model

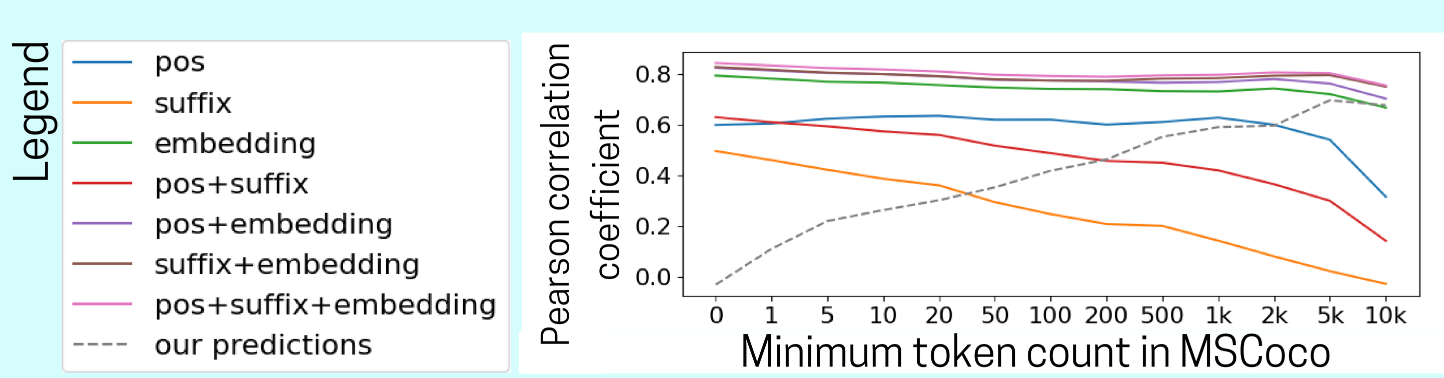
Experiments

Trained the model on **MSCOCO**, a dataset with image-caption pairs (captions created by human annotators):
- **55,700 images**
- **278,628 captions** (~5 captions per image)
- **65 ground-truth classes**

Results Analysis

Concreteness prediction

- By taking maximum over the cluster probability vector
- Compared to a **supervised baseline** with linguistic features (POS, frequent suffixes, pre-trained embedding)
- Evaluated the **Pearson coefficient** of predictions with ground-truth values (annotated by humans)
- Used filtered validation sets: Tokens that occur more than X times in the MSCOCO training set



Word semantic clustering

- Evaluated on a categorization dataset: **purity 40.05, collocation 0.3565, F1 0.3772** (chance-level is 0.193, 0.135, 0.159)
- Induced clusters are **associative clusters** (rather than similarity clusters)
- The words in each cluster are words that are **likely to appear in the same scene**
- Not necessarily semantic similar words

"salad" scene: salad, tomato, fork, healthy, plate, lettuce, sauce, spoon

"animal on tree" scene: bird, climb, koala, lynx, parrot, tree, owl, pheasant, quail, squirrel, branch

"ball game" scene: baseketball, soccer, goal, player, ball, sport, running, football

Visual object recognition

- Mapped ground-truth classes to clusters using the **name of the class**
- Identified each cluster by the names of the ground-truth classes mapped to it
- Given an image, predicted the clusters
- Using **class activation mapping** (CAM), extracted a heat map of salient pixels

cluster 41: skateboard | cluster 42: toilet, sink | cluster 23: truck | cluster 99: bicycle, motorcycle

Conclusion

- Learned concreteness prediction, **without explicit training** for this task
- Performed better on **"familiar" tokens**
- On very frequent tokens performed better than a **supervised baseline that also uses a pre-trained POS tagger**
- The induced word clusters are **associative** (unlike the classic semantic clusters)
- Learned to **recognize objects** without direct supervision

Future work: Word concreteness can be used as a building block for more complex tasks: constituency parsing and semantic role labelling