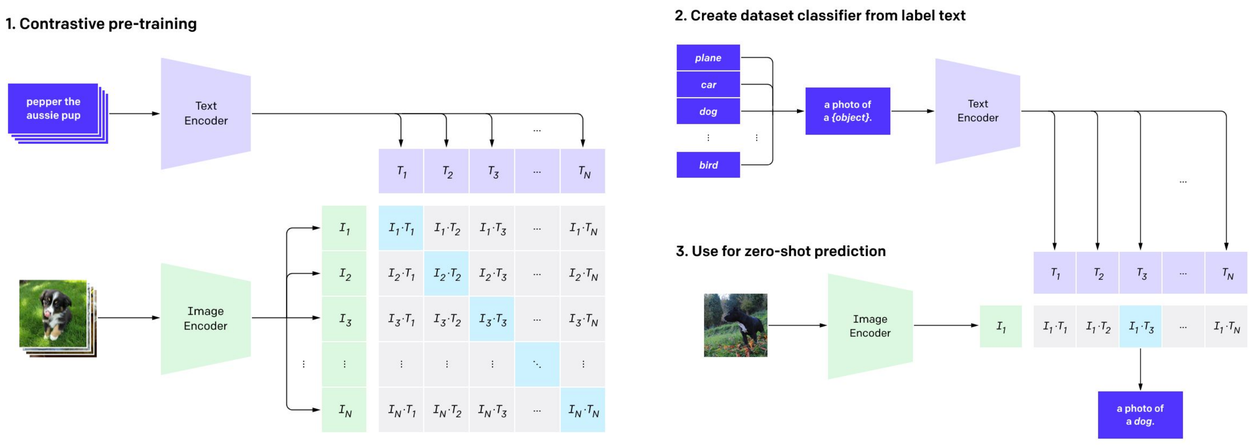
# Scalable Performance Analysis for **Vision-Language Models**

Santiago Castro*, Oana Ignat*, and Rada Mihalcea

## Image Understanding through Language

Background: CLIP (Radford et al., 2021)

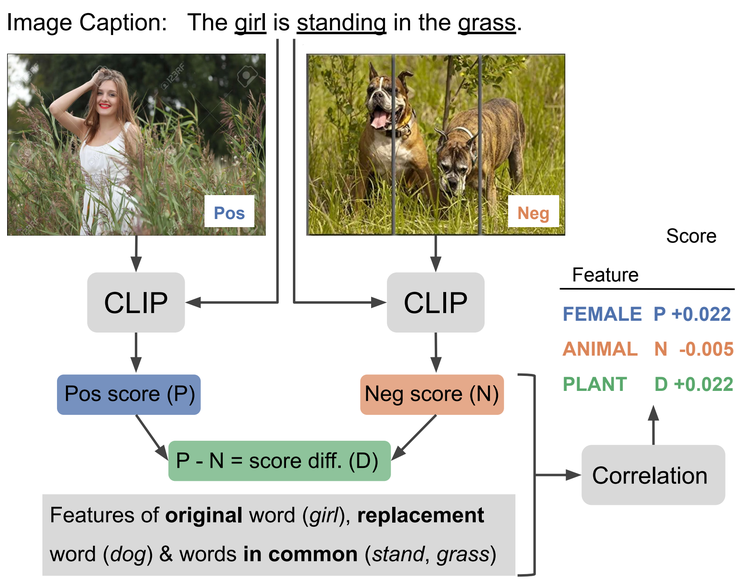



## Motivation

CLIP works really well for zero-shot prediction, esp. for object understanding.

What are its limitations in image-sentence understanding?

## Our Framework




## Dataset: SVO-Probes

- 48,000 image-sentence pairs (differ in exactly one of S, V, or O)
- 14,000 images
- 100 subjects
- 421 verbs
- 275 objects

## Features

- **Levin** verb classes:
  - broad (e.g., *change of state*, *social interaction*)
  - fine-grained (e.g., *roll*, *run*, *hug*)
- **LIWC** 2015 psycholinguistic markers for words
  - E.g., *female*, *family*, *social*, *religion*, *health*
- **General Inquirer** word classes
  - E.g., *power*, *strong*, *legal*, *vehicle*
- WordNet **hypernyms**
  - E.g., *building* is a hypernym of *house* and *school*
- **Word** presence
- **Sentence length**
- **Semantic similarity** (Sentence-BERT)
  - To the (hidden) sentence from the negative image
- **Concreteness** score (1-5)
  - E.g., *beauty* score is 2.93, *table* score is 4.9
- Word **ambiguity** (number of WordNet synsets associated)
- Word **frequency** (in LAION-13M)

Feature Importance:

For binary: **diff. of means** when true *vs.* when false

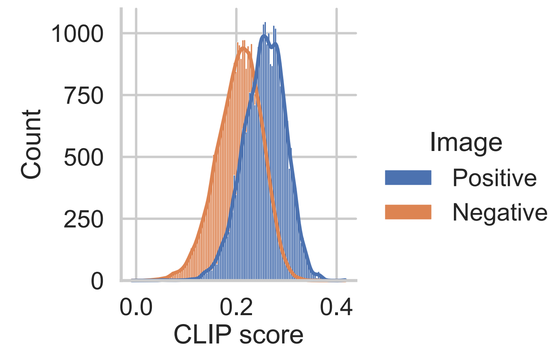
For numerical: Pearson's **correlation**

t-test for significance (95% confidence level)

## Findings:

## CLIP Behaves Like a Bag-of-Words Model

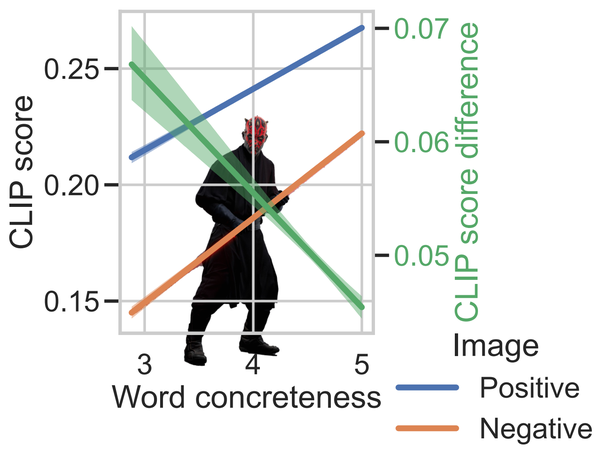
A word from the text being represented in the image **increases** the sentence-image **score**, **regardless if the image is positive or negative**.



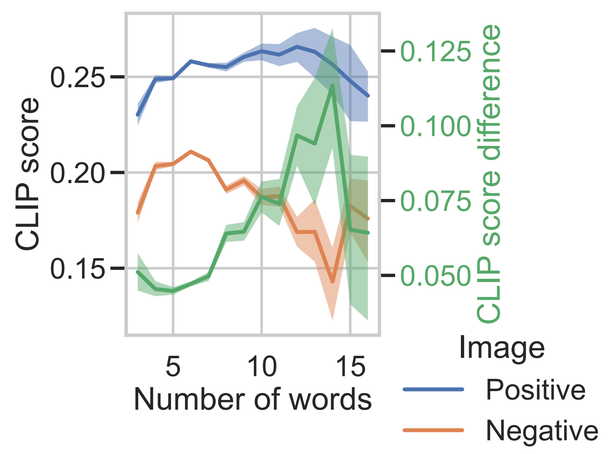

## CLIP Performs Better with Nouns Than with Verbs

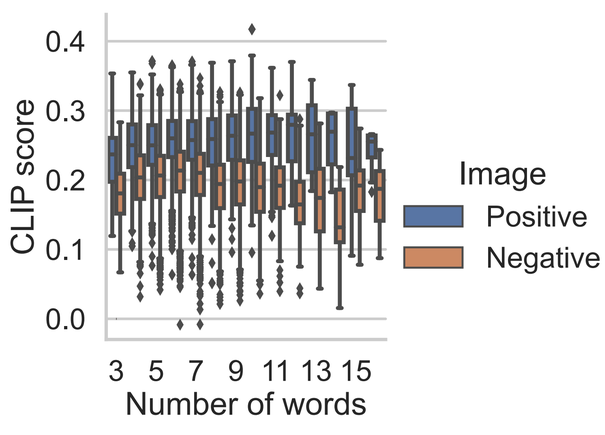
| | accuracy (%) |
|---|---|
| verbs | 81.45 |
| subjects | 86.87 |
| objects | 88.78 |

## CLIP Gets Confused by Concrete Words

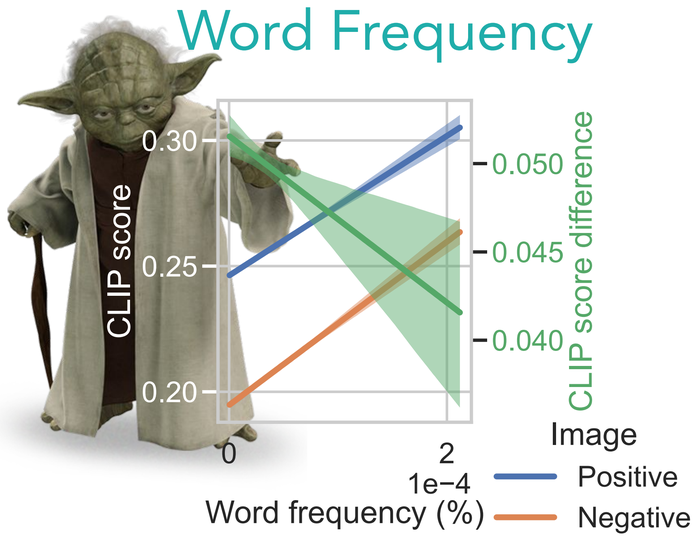



## CLIP Prefers Average-Length Sentences

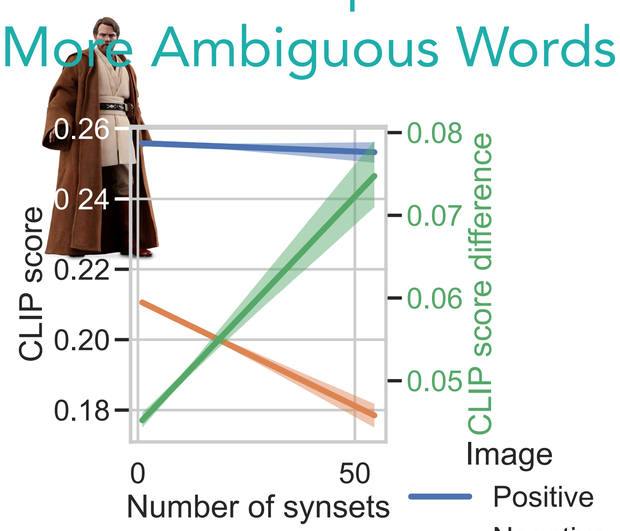


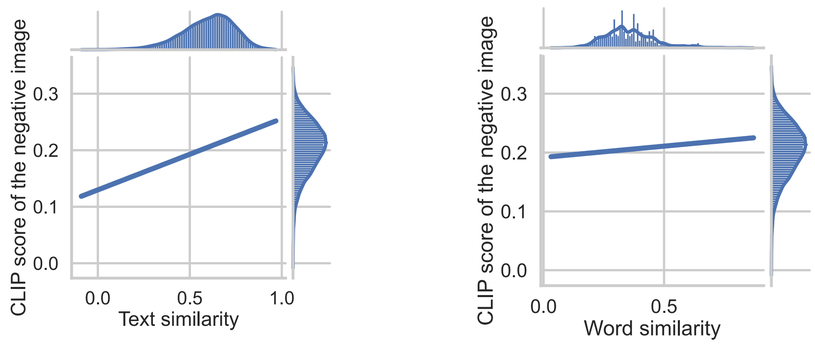

## CLIP is Affected by Word Frequency




## The Score Improves for More Ambiguous Words




## Similar situations confuse CLIP




CLIP performs relatively better on nature-related and personal care concepts and relatively worse on furniture, transportation, herbivores, sports, academia.

Top/bottom examples:

| feature | diff. of means |
|---|---|
| Hypernym physical_phenomenon.n.01 (original) | 0.038 |
| Presence of word "sofa" (in common) | -0.032 |