KU LEUVEN

What Can We Learn from the Structures Found in Visual and Language Data and their Correlations?

Opening the Discussion on Joint Visual and Linguistic Structures

Victor Milewski

European Research Council
Established by the European Commission

Introduction

- Structures, interactions, and relations in the physical world are varied and complex
- Humans quickly identify and understand relations between objects
- Humans describe the world with natural language or structured representations
- Improvements in foundation models allow for better generation of text and images

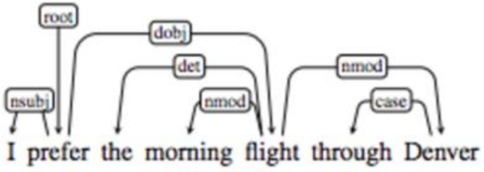
Milewski, V., Trusca, M.M., Moens, M.F. (2023). What Can We Learn from the Structures Found in Visual and Language Data and their Correlations?. Fourteenth International Workshop Modelling and Reasoning in Context (MRC)

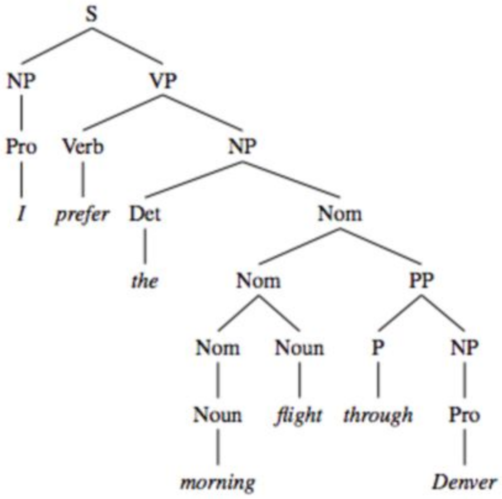
Outline

- The Structure of Language
- The Structure of Visual Data
- Investigating correlations between the Structures
- Discussion and Open Questions

Language Grammars

- Several types of structures available
- Most based on Natural Language with grammar rules:
  - Constituency Trees
    - Phrase structured hierarchical tree
    - Described by CFG
  - Dependency Trees
    - Grammatical relationships between words
    - Universal grammar across languages

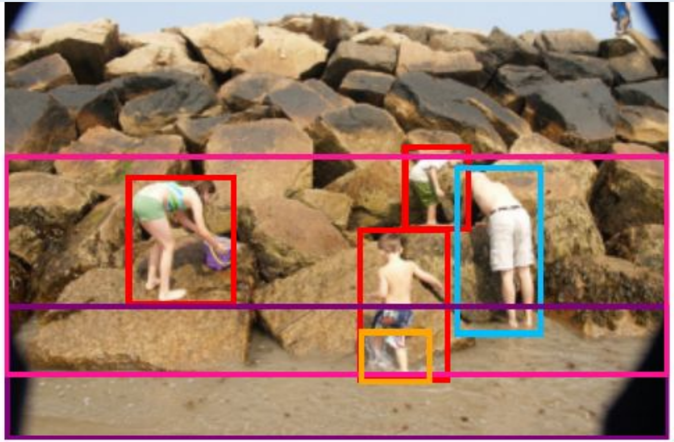


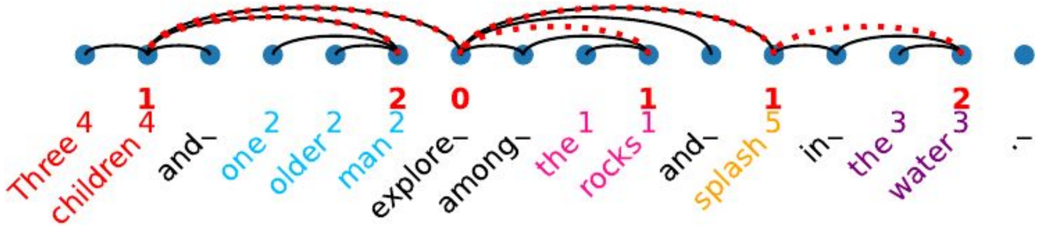


Scene Trees

- Designed as structured tree over objects in the image
- Follows the dependency tree, but truncated to entities that appear in the image

- Created to evaluate encoded structures in pretrained multimodal-BERT models



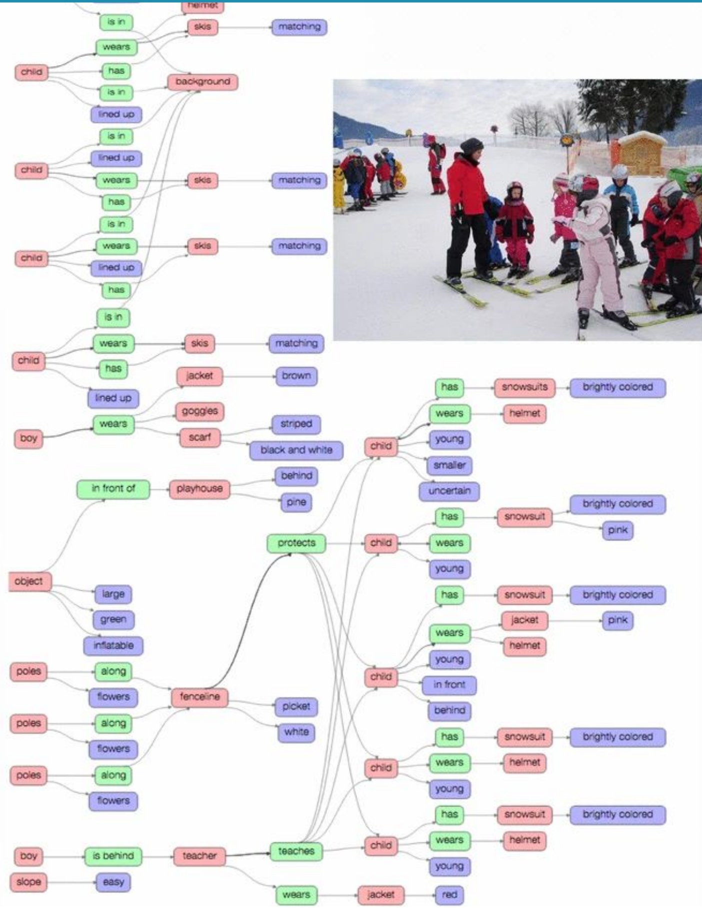


Milewski, V., de Lhoneux, M., & Moens, M. (2022). Finding Structural Knowledge in Multimodal-BERT. Annual Meeting of the Association for Computational Linguistics (ACL).

Scene Graphs

- A visual graph has its own unique language to describe the world
  - Common example is Visual Genome (VG)

- Goals and Arguments:
  1. Explaining relations is cognitive in nature
  2. Better distinguish between images
  3. Ground visual concepts to language
  4. Formalized representations of image components



Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L., Shamma, D.A., Bernstein, M.S., & Fei-Fei, L. (2016). Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. International Journal of Computer Vision, 123, 32 - 73.

VG Creation

3 people on laptops
3 people in bed
a pair of feet
3 apple laptops
a bed headboard in dark wood
a framed picture
loose blue pants
a striped duvet cover
a man wearing shorts
"woman with a laptop"
"girl with laptop"
" man with laptop"
" bed with brown headboard"
"bed with tan stripe sheets"
"white wall with picture"
"girls bare feet"
" computer wires"
three gray laptops
brown stripes in a sheet
a woman's kneecaps
a man's hairy legs
blue pants on a child
the bottom of a child's bare feet

"woman with white top and blue skirt"
" man wearing brown shirt and tan shorts"
the Apple logo on the back of a laptop
brown striped sheets on the bed

black shirt on a man
a child's bare toes
a tan speckled wall
a black charging cord
three Apple logos on laptops
three people relaxing in bed
three laptops in people's laps
a black charger cord
a brown wooden head board
a picture frame on the wall
a window curtain
a round hole in the head board
a young girls feet
the knee of a woman
logo on a laptop
a silver laptop
the toes of a girl
a black computer cord
a woman sitting in bed
a man sitting in bed
a young girl sitting in bed
a white wall behind the bed
a young girl wearing eyeshadow

VG Creation

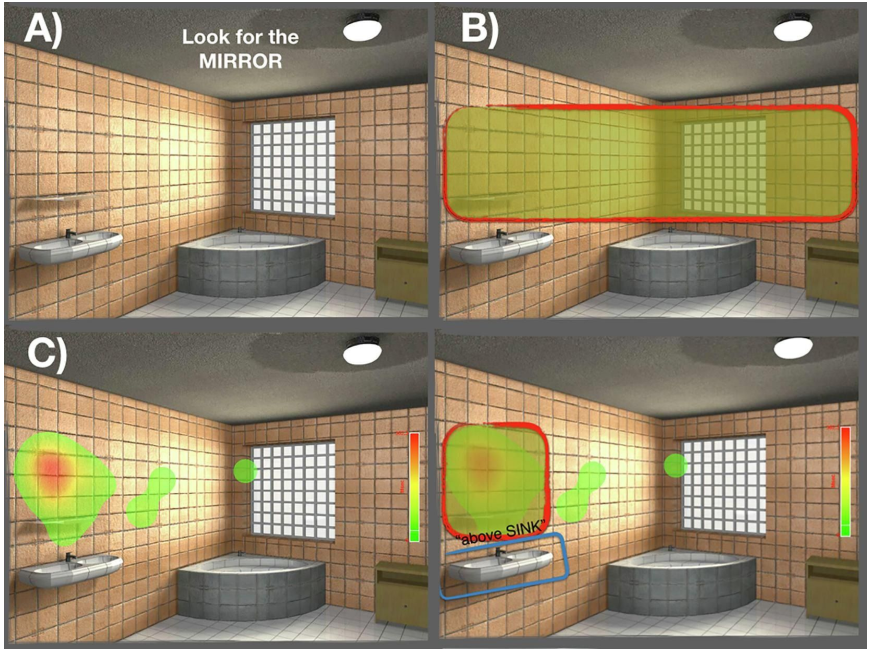
1. Humans create many short descriptions
2. Humans convert these into objects and relations
3. These are merged into graphs
4. The graphs are joined

KU LEUVEN

erc European Research Council

Structure of Visual Data

- How do humans perceive and process the physical world?
- People have *semantic* and *episodic* knowledge
- Investigated through visual search:
  - Humans perform search tasks tracking their eye movement
  - They use prior knowledge while searching



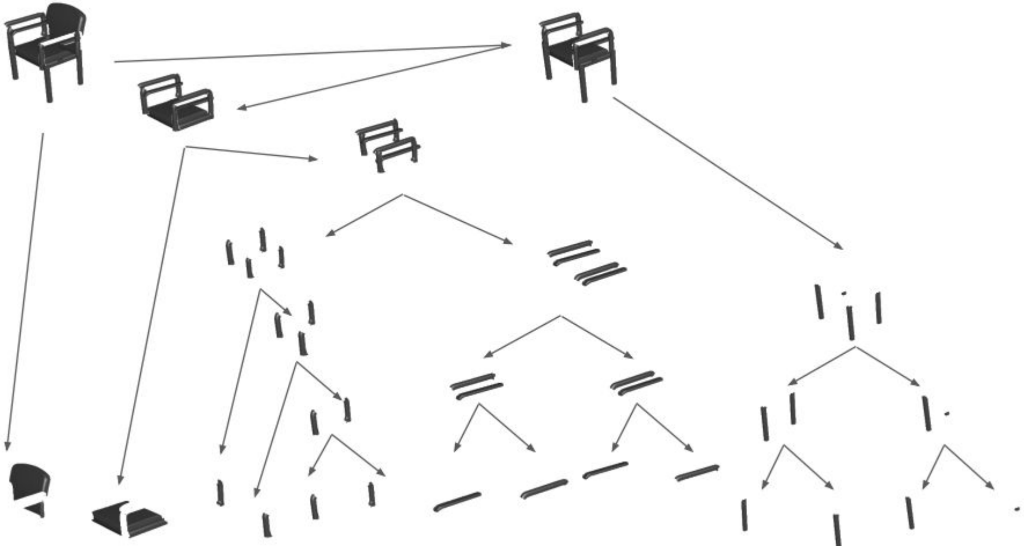
Võ, M.L. (2021). The meaning and structure of scenes. Vision Research, 181, 10-20.

Structure of Visual Data

- Episodic knowledge is about familiarity with the room

- Semantic knowledge is a common pattern or structure of scenes
  - This can form a grammar

- e.g.
  - furniture, roads, rooms



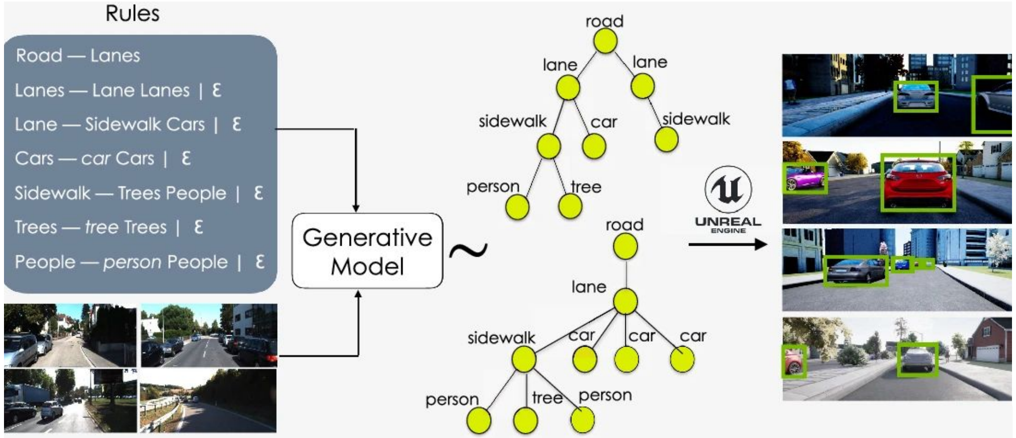
Hong, Y., Li, Q., Zhu, S., & Huang, S. (2021). VLGrammar: Grounded Grammar Induction of Vision and Language. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 1645-1654.

The Structure of Visual Data

- Episodic knowledge is about familiarity with the room

- Semantic knowledge is a common pattern or structure of scenes
  - This can form a grammar

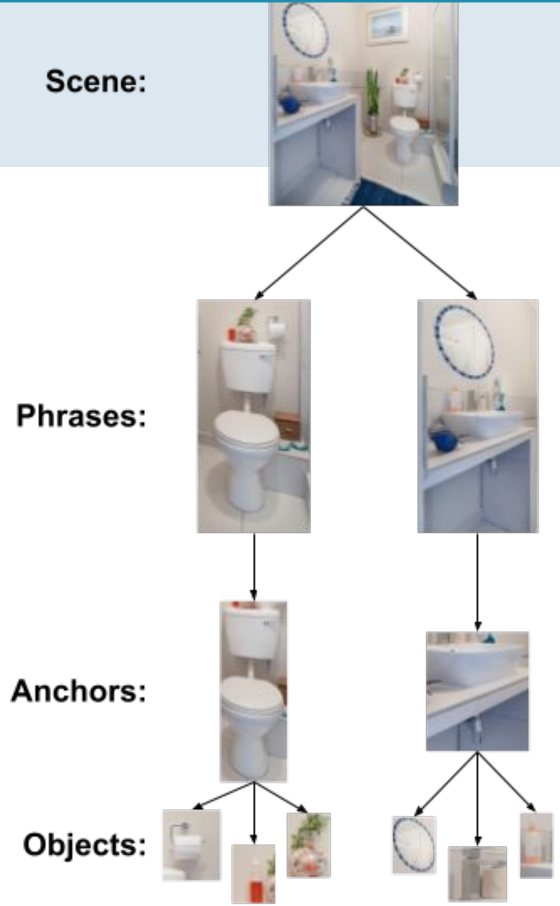
- e.g.
  - furniture, roads, rooms



Devaranjan, J., Kar, A., & Fidler, S. (2020). Meta-sim2: Unsupervised learning of scene structure for synthetic data generation. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16 (pp. 715-733). Springer International Publishing.

Structure of Visual Data

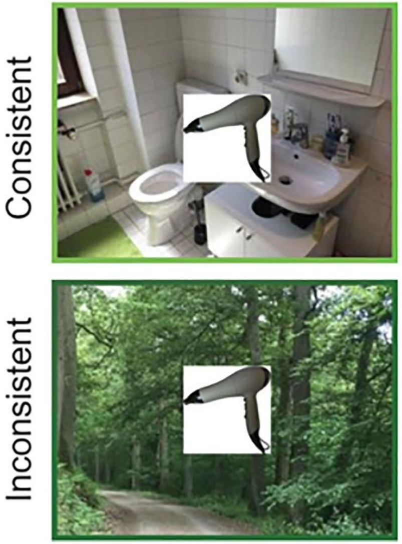
- Episodic knowledge is about familiarity with the room
- Semantic knowledge is a common pattern or structure of scenes
    - This can form a grammar
- e.g.
    - furniture, roads, rooms



Võ, M.L., Boettcher, S.E., & Draschkow, D. (2019). Reading scenes: how scene grammar guides attention and aids perception in real-world environments. Current opinion in psychology, 29, 205-210 .

Human responses to errors

- Using EEG studies able to compare brain responses
- Seeing semantic inconsistencies cause similar responsive between language and visual data
- Such studies indicate similar responses, but there is no evidence that the processing is equal



Võ, M.L. (2021). The meaning and structure of scenes. Vision Research, 181, 10-20.

Experiments

- **Dataset:**
  - 145 images from overlap between Flickr30k-entities and VG
  - 483 captions
  - Spacy Parser with Berkley neural parser for creating dependency and constituency trees

- **Metrics:**
  - representational similarity analysis (RSA)
    - computes a dissimilarity matrix (distances in graphs) and performs Spearman rank correlation between matrices

Milewski, V., Trusca, M.M., Moens, M.F. (2023). What Can We Learn from the Structures Found in Visual and Language Data and their Correlations?. Fourteenth International Workshop Modelling and Reasoning in Context (MRC)

Experiments

- Compare the visual distances of **object regions** with **objects/nouns** in the language graph

	Q1	Median	Q3
Const. Tree	-0.03	0.55	0.81
Dep. Tree	-0.03	0.53	0.81
Scene Tree	0.00	0.69	0.89
Scene Graph	0.88	0.99	1.00

- Positive correlation in almost all experiments
- Head nouns in text can be further apart
- The scene tree is reduces to only nouns, making it flatter
- Scene graphs describe direct relations between object

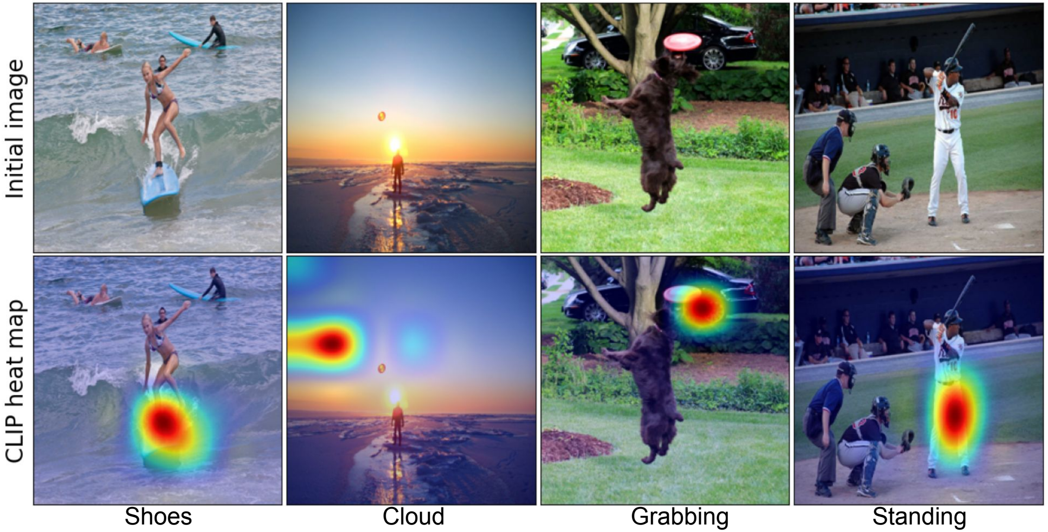
Discussion and Open Questions

- Most studies on visual grammars are from psychological studies or very domain specific
  - What can we learn from more data driven studies?
  - Can we improve our understanding of human processing?
  - Can we find correlations in structured processing between modalities?

- We showed correlations between language and the physical world
  - Did language influence how humans see the world? or vice versa?

Discussion and Open Questions

- Visio-linguistic models can find regions without objects present or show appropriate regions for verbs
  - Similar capabilities to humans
- What structures did CLIP learn?
  - Semantic or Episodic knowledge?



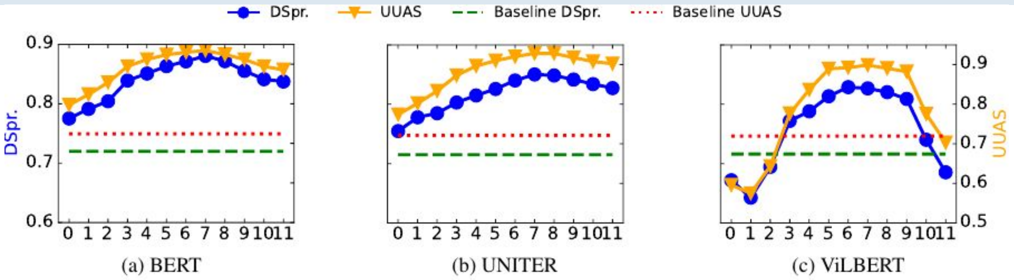
Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. International Conference on Machine Learning.

Questions?

This work is part of the CALCULUS project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 788506).

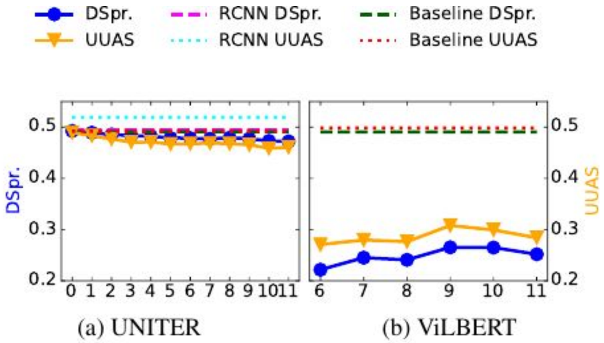
Latent trees in visio-linguistic models

- Dependency trees are encoded in BERT models



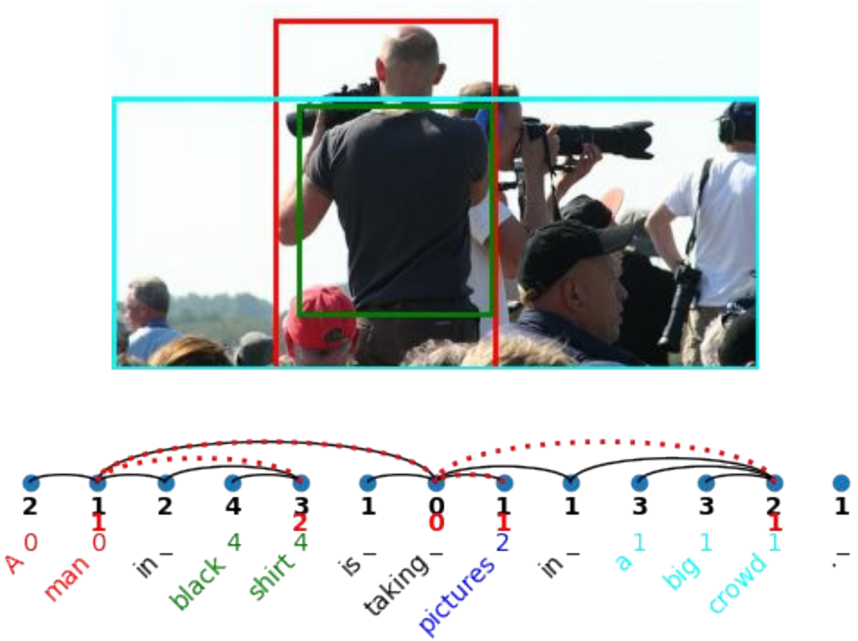
Comparison for the distance probe on the Flickr30k test set, with textual embeddings.

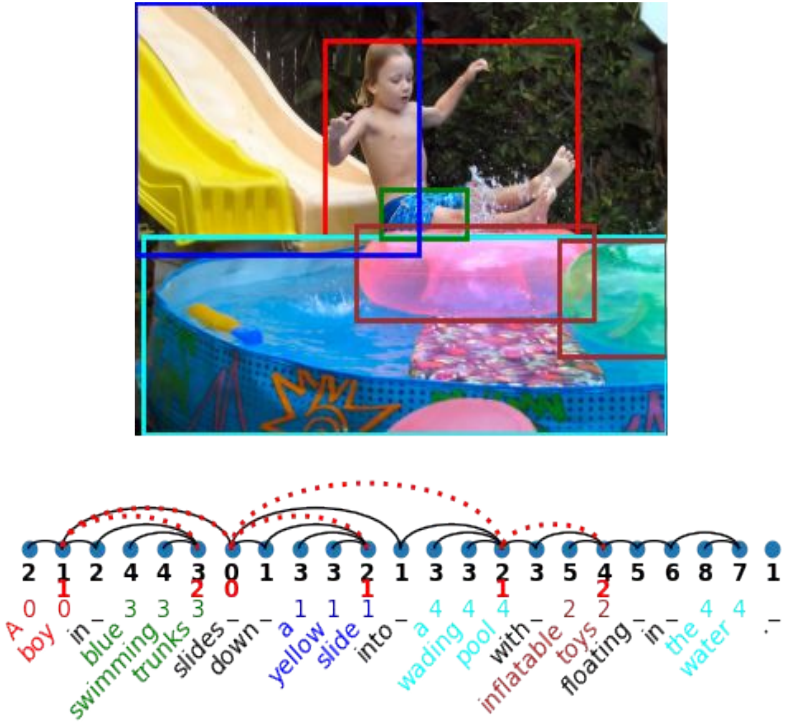
- The scene tree is not encoded in BERT models
    - The training paradigm does not encourage learning of structure



Comparison for the distance probe on the Flickr30k test

Scene Tree Examples





2 - VG Creation

1. Humans create many short descriptions

2. Humans convert these into objects and relations

3. These are merged into graphs

4. The graphs are joined

3 - Structure of Visual Data

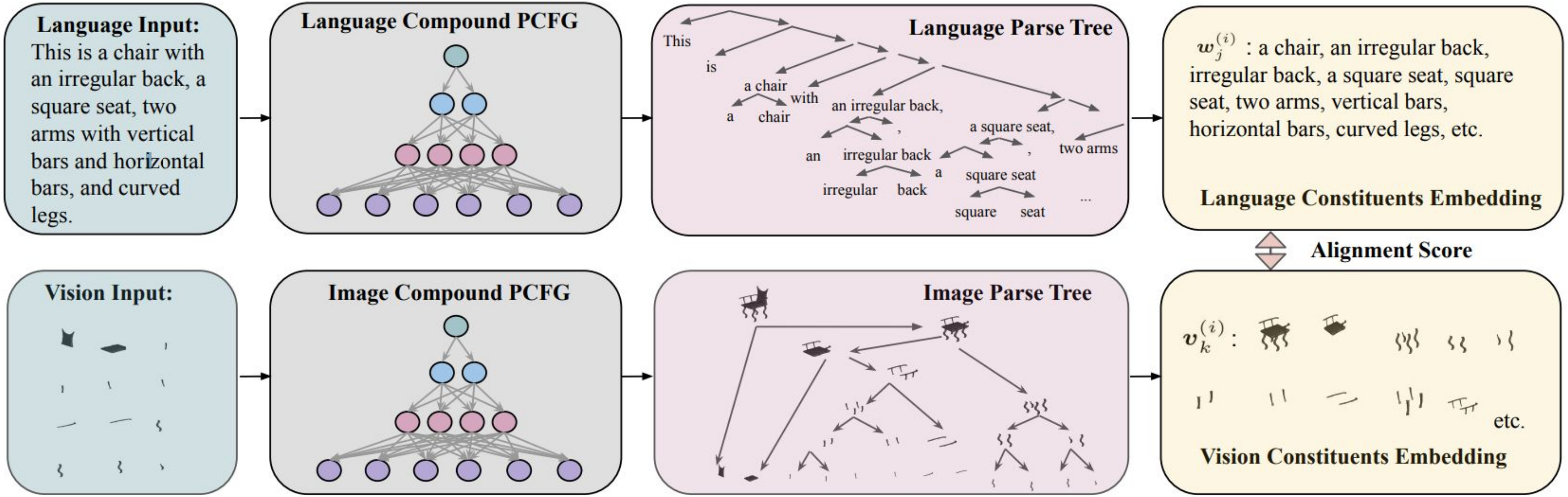


Figure 3: **Our proposed VLGrammar framework.** We implement image grammar induction and language grammar induction via compound PCFGs. Parse trees are derived from the grammars. We compute alignment scores between the vision and language constituents in the parse trees to guide the joint learning procedure.

6 - Discussion and Open Questions

- Scene trees are based on language, making it difficult to study visual structures
  - The simple captions with the reduction to head nouns creates a flat tree

- While scene graph distances are strongly correlated with the visual distances, they can be very dense
  - Parts can be derived from common knowledge
  - No rules and restrictions on ordering relations or labels used
  - Difficult to study graph patterns and hierarchical nature of objects