Compositional Structures in 3D Vision and Language

Siyuan Huang

Beijing Institute of General Artificial Intelligence

University of California, Los Angeles

Oct 16, 2021

ICCV Workshop on

Structural and Compositional Learning on 3D Data

Vision is not just about recognition

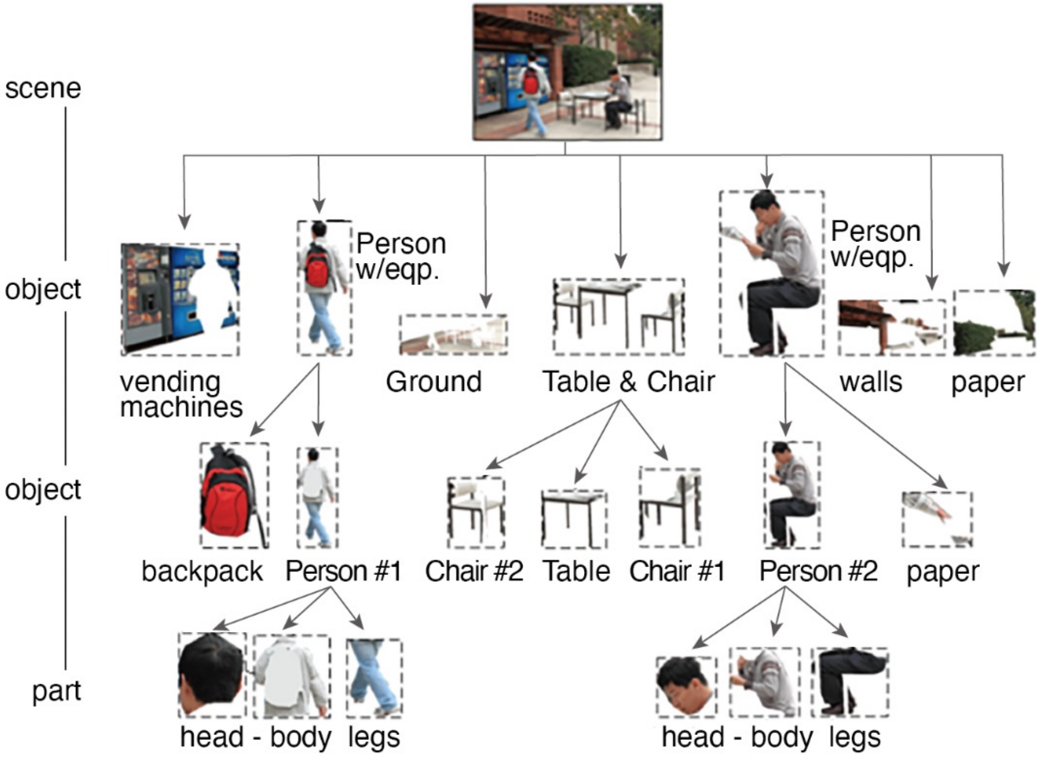
- It is the tool to model the world from visual perspective
- It is the key component in building the intelligence
  - Understanding the physical and social world
  - Problem solving and task planning
  - Connecting to language and mind

Compositionality is the key ingredient for true intelligence!

Compositional Structures in AI

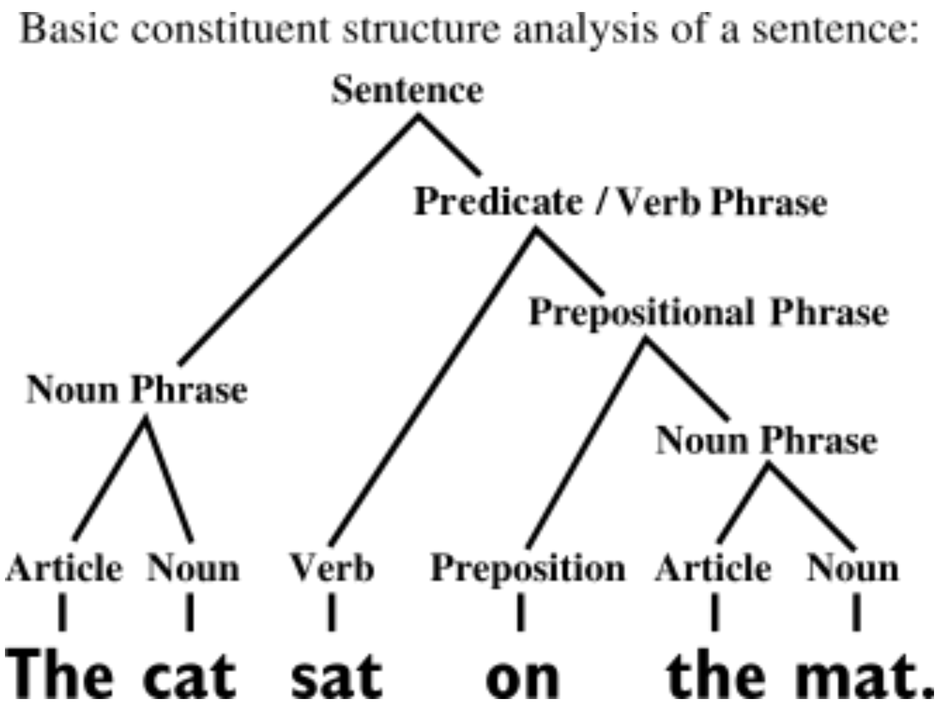
Vision

Scenes, Objects, Events



Language

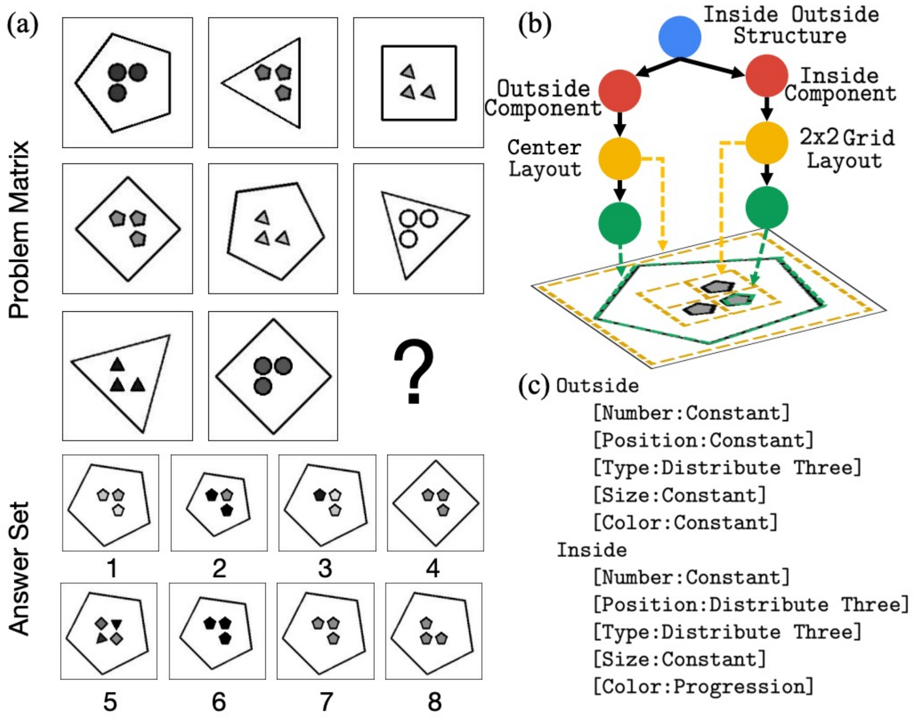
Syntax, Semantics, Pragmatics



Compositional Structures in AI

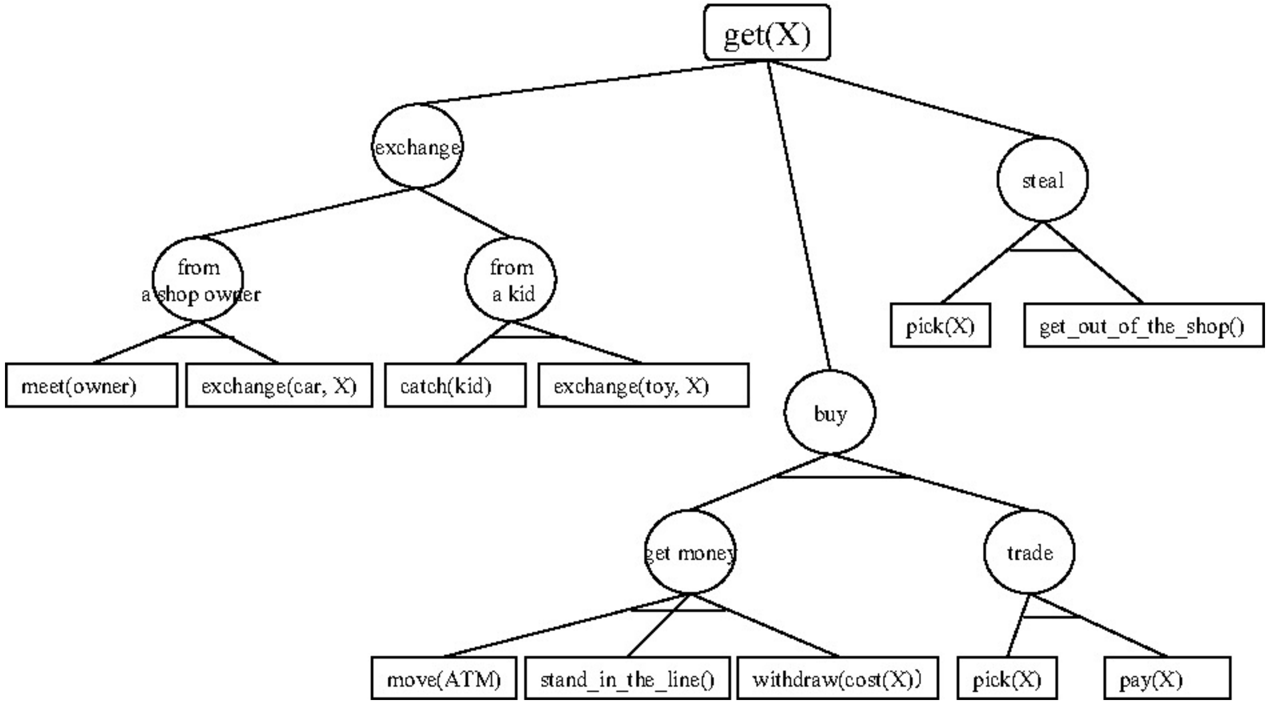
Reasoning

Abstraction, Generalization



Planning

Actions, Goals, States

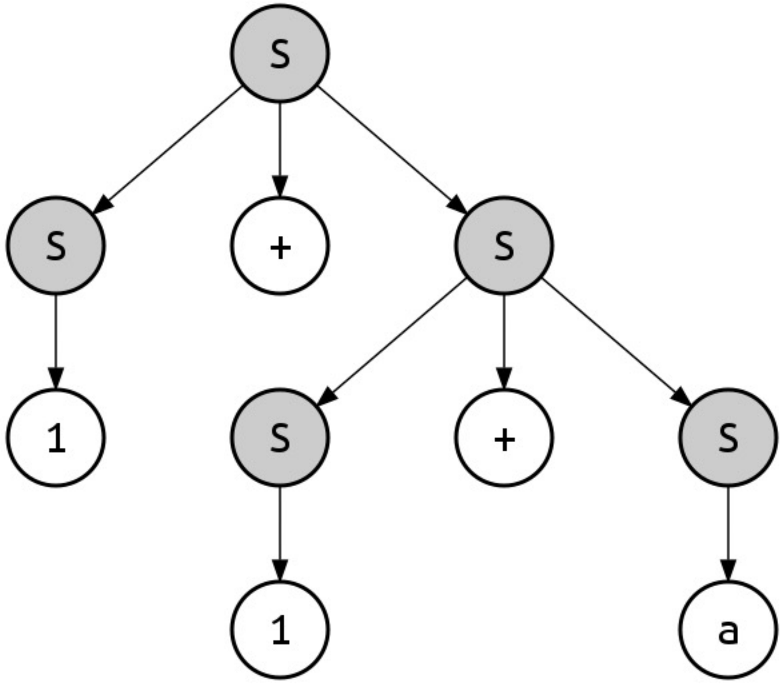


Modeling the Compositionality

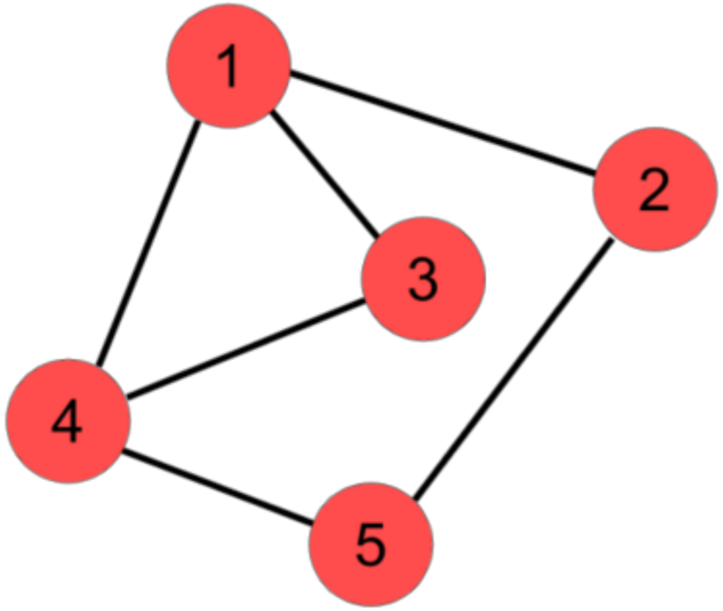
Logics & Programs

Sentence	$\rightarrow$	*AtomicSentence* | *ComplexSentence*
AtomicSentence	$\rightarrow$	*Predicate* | *Predicate*(*Term*, ...) | *Term* = *Term*
ComplexSentence	$\rightarrow$	(*Sentence*) | [*Sentence*]
	|	$\neg$ *Sentence*
	|	*Sentence* $\land$ *Sentence*
	|	*Sentence* $\lor$ *Sentence*
	|	*Sentence* $\Rightarrow$ *Sentence*
	|	*Sentence* $\Leftrightarrow$ *Sentence*
	|	*Quantifier Variable*, ... *Sentence*
Term	$\rightarrow$	*Function*(*Term*, ...)
	|	*Constant*
	|	*Variable*
Quantifier	$\rightarrow$	$\forall$ | $\exists$
Constant	$\rightarrow$	*A* | X_1 | *John* | $\cdots$
Variable	$\rightarrow$	*a* | *x* | *s* | $\cdots$
Predicate	$\rightarrow$	*True* | *False* | *After* | *Loves* | *Raining* | $\cdots$
Function	$\rightarrow$	*Mother* | *LeftLeg* | $\cdots$
OPERATOR PRECEDENCE	:	$\neg, =, \land, \lor, \Rightarrow, \Leftrightarrow$

Grammars



Graph Neural Network



How to bridge vision and language for a better structural understanding of the world?

VL Grammar: Grounded Grammar Induction of Vision and Language

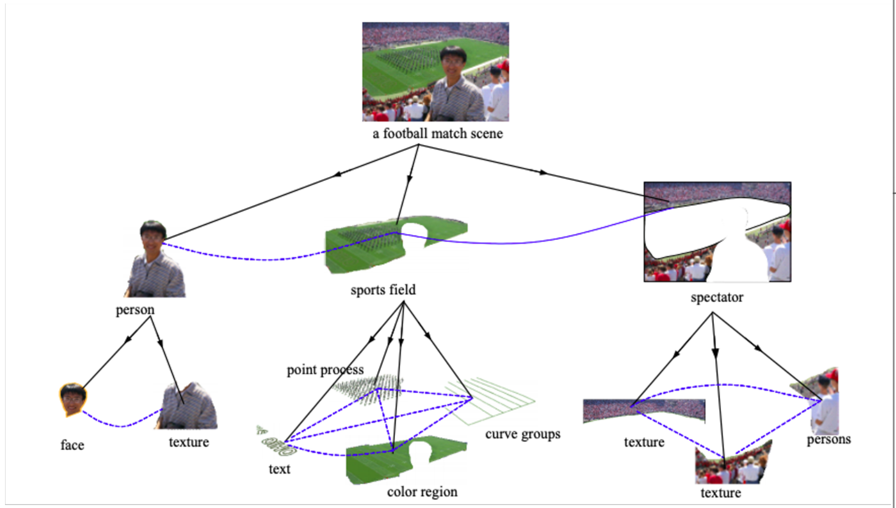
Yining Hong, Qing Li, Song-Chun Zhu, Siyuan Huang

University of California, Los Angeles   Beijing Institute of General Artificial Intelligence   Peking University   Tsinghua University

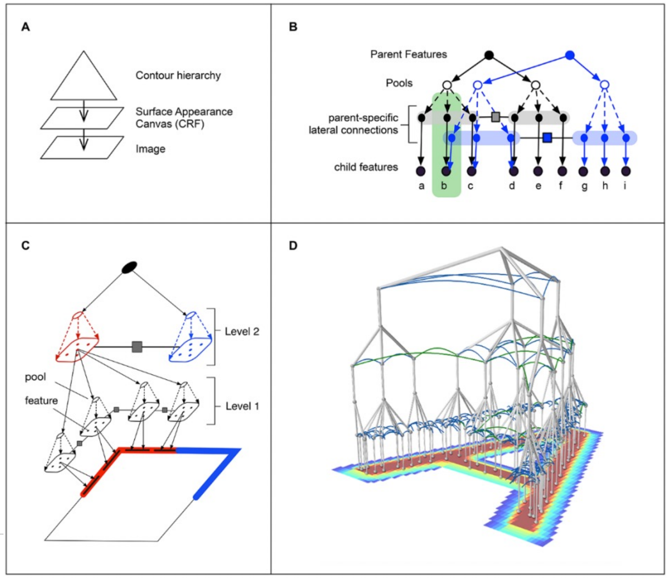



Visual Structure Learning

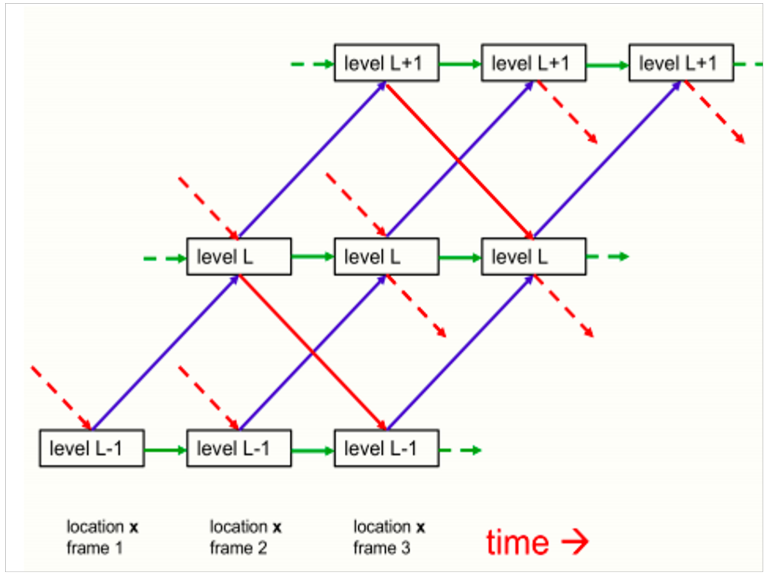
Inducing the underlying structures and grammars (especially part-whole hierarchies) from raw data (images) is a long standing challenge



[1] S.-C. Zhu and D. Mumford. A Stochastic Grammar of Images.



[2] George et al. A generative vision model that trains with high data efficiency and breaks text-based CAPTCHASs. Science, 2017



[3] Geoffrey Hinton et al. How to represent part-whole hierarchies in a neural network. Preprint

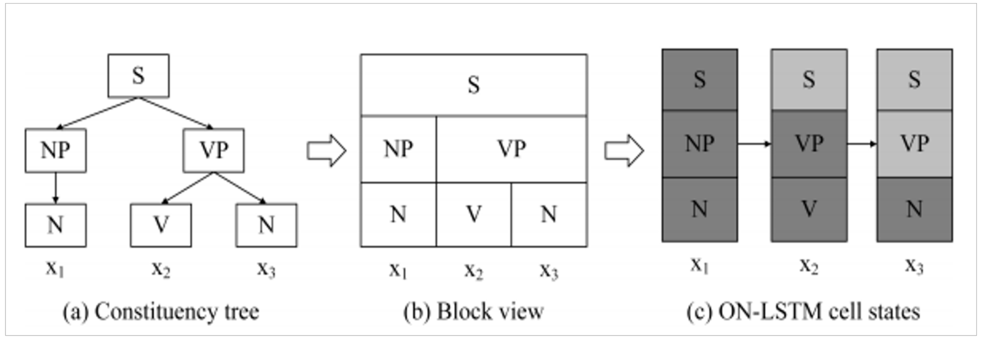
Challenges of Visual Structure Learning

How to represent flexible part-whole hierarchies that vary with images using an identical model?

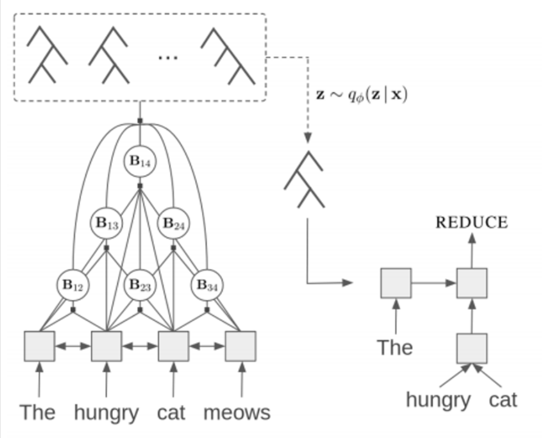
How to learn structure automatically without pre-defined templates?

How to avoid ambiguities in structure learning?

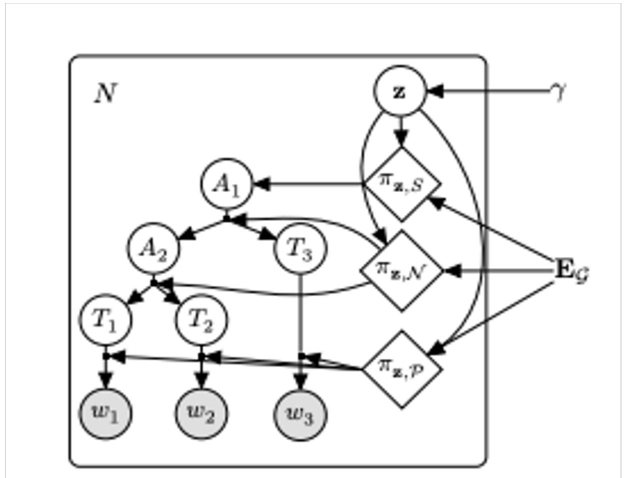
Grammar Induction in Natural Language



[3] Yikang Shen, Shawn Tan, Alessandro Sordoni, and Aaron C. Courville. Ordered neurons: Integrating tree structures into recurrent neural networks.



[4] Yoon Kim, Alexander M. Rush, L. Yu, Adhiguna Kuncoro, Chris Dyer, and Gabor Melis. Unsupervised recurrent neural network grammars.



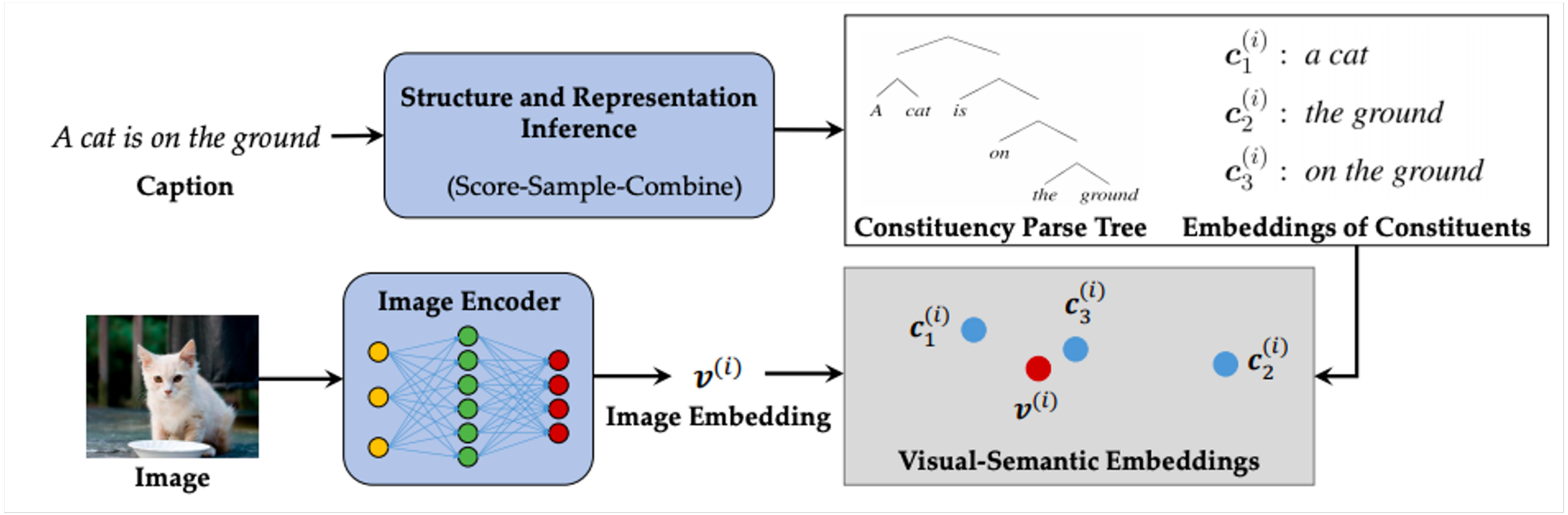
[5] Yoon Kim, Chris Dyer, and Alexander M. Rush. Compound probabilistic context-free grammars for grammar induction.

Cognitive Grammar

We should analyze grammatical units with reference to their semantics, which is grounded and structured by patterns of perception, such as vision.

[6] Ronald W. Langacker. Foundations of cognitive grammar.
[7] Ronald W. Langacker. An introduction to cognitive grammar.

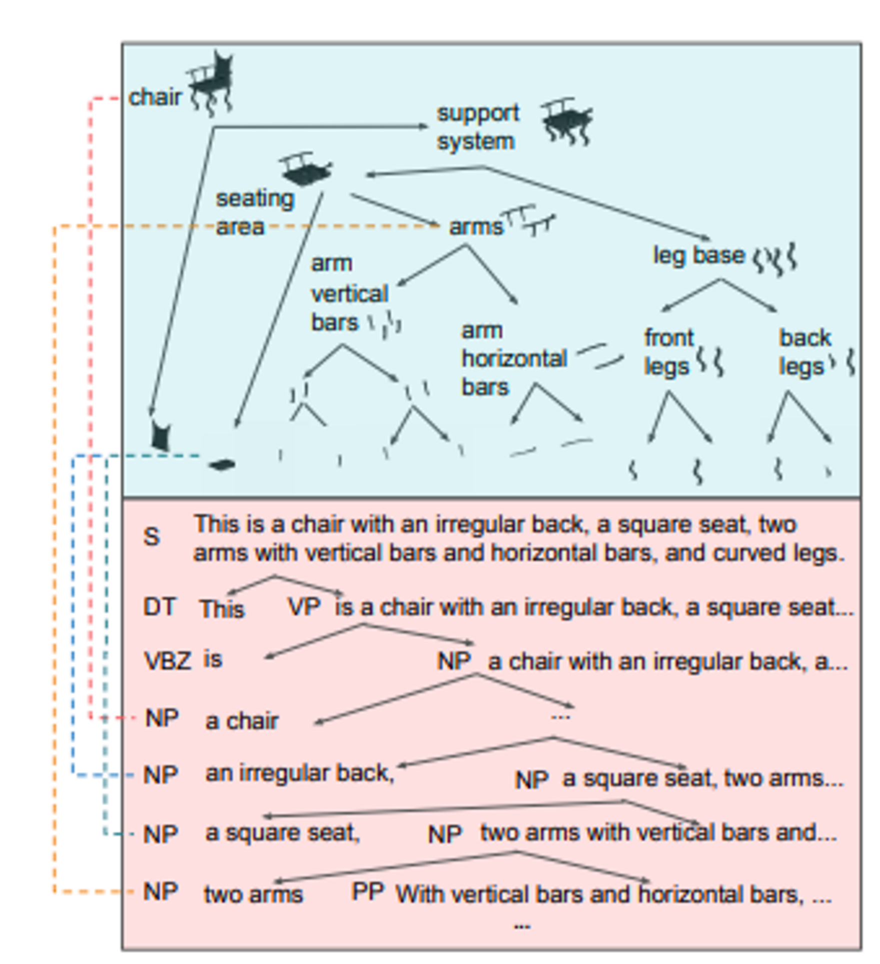
Visually-Grounded Language Grammar Induction



[8] Haoyue Shi, Jiayuan Mao, Kevin Gimpel, and Karen Livescu. Visually grounded neural syntax acquisition

Cognitive Grammar

A constituent's semantic value does not reside in one individual image base, but rather in the relationship between the substructure and the base.

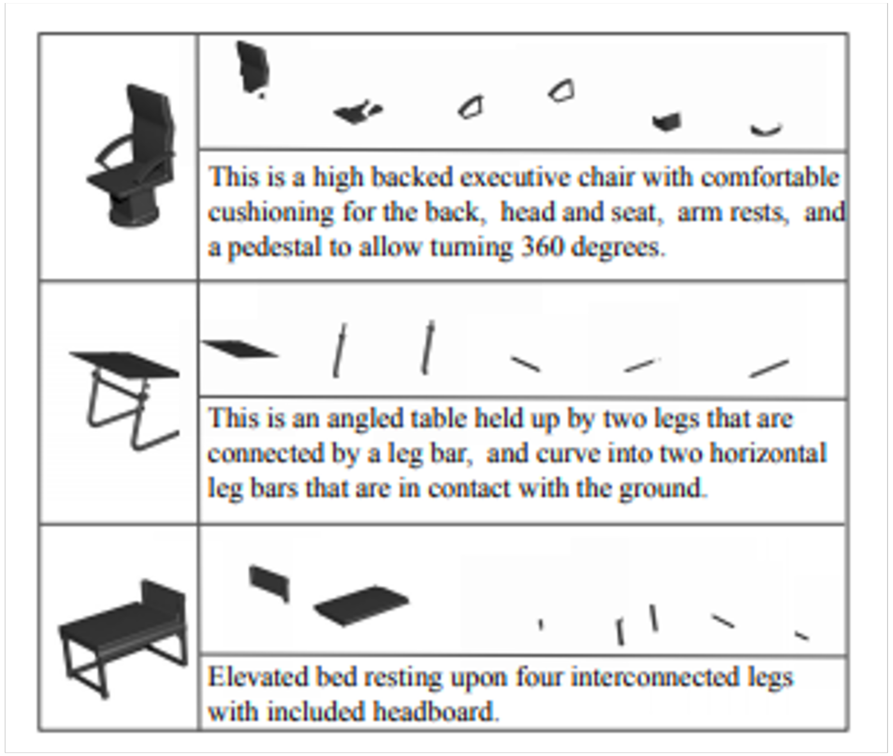


The PartIt Dataset

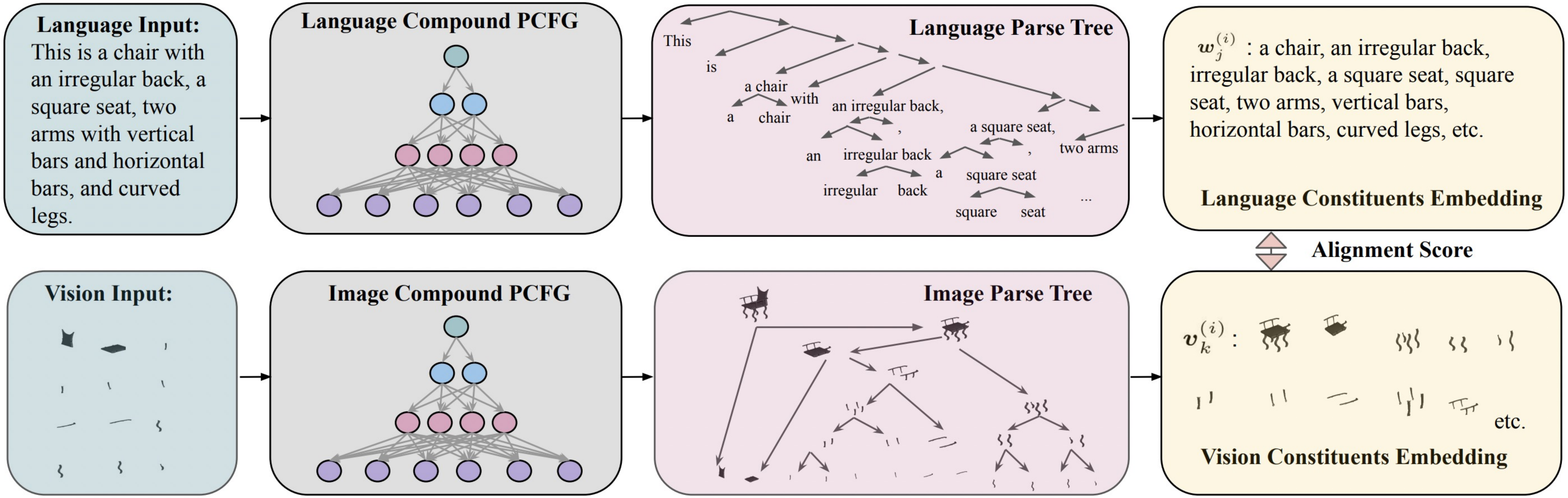
The first dataset with annotated natural language sentences that describe both object semantics and fine-grained part semantics paired with images

Also suitable for other tasks, e.g., , image captioning, language-guided part segmentation, 3D reconstruction

	All	Chair	Table	Bed	Bag
$\#I$	10613	5031	5290	185	109
$\#PS$	120110	13	10	8	3
$\#G$	75	23	34	18	4
P_{med}	7	8	6	9	3
P_{max}	136	38	136	28	6
G_{med}	8	8	8	7	3
G_{max}	18	12	18	15	4
LG_{med}	16	19	13	19	15
LG_{max}	98	98	68	42	21
$Vocab$	2007	1634	903	176	61



Overall Framework



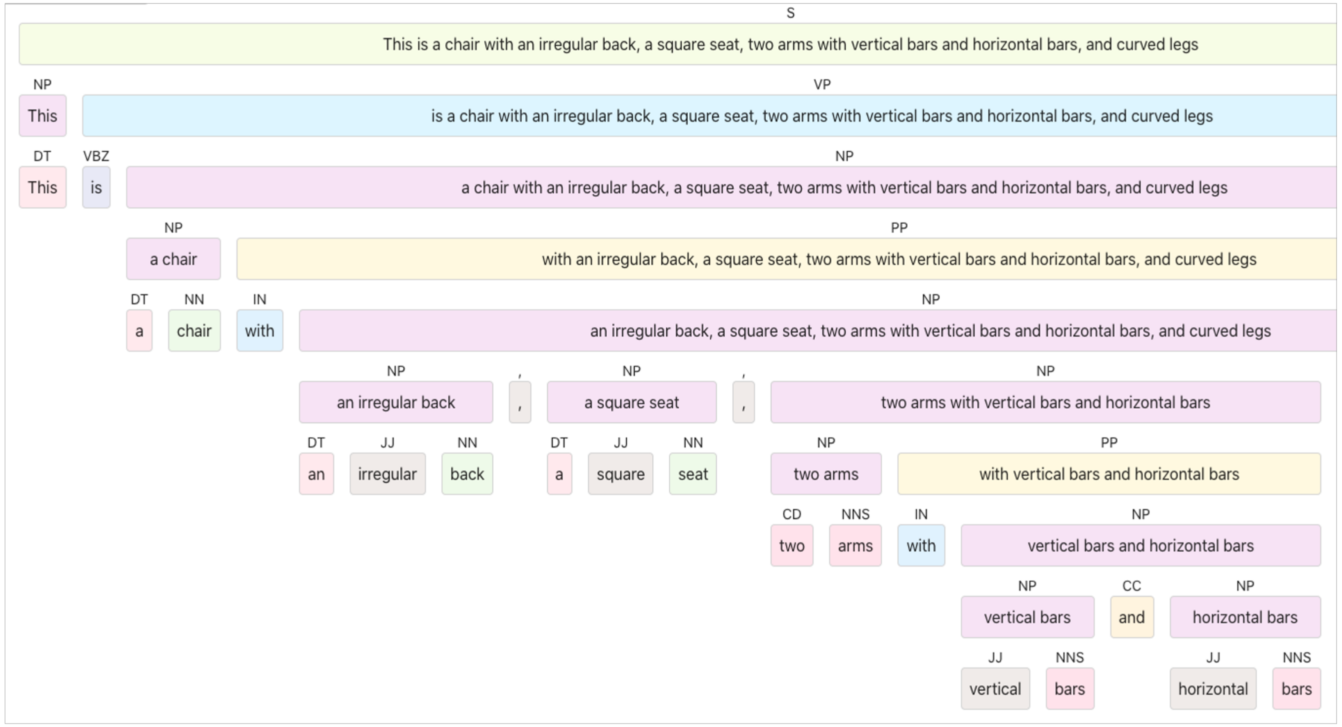
$$\mathcal{L} = \lambda_w \mathcal{L}_{\mathcal{G}}(\mathcal{W}; \phi_w, \theta_w) + \lambda_v \mathcal{L}_{\mathcal{G}}(\mathcal{V}; \phi_v, \theta_v) + \lambda_C \mathcal{L}_C(\mathcal{W}, \mathcal{V})$$

Context-free Grammar (CFG)

A context-free grammar (CFG) can be defined as a 5-tuple $\mathcal{G} = (S, \mathcal{N}, \mathcal{P}, \Sigma, \mathcal{R})$, where S is the start symbol, $\mathcal{N}$ is a finite set of nonterminal nodes, $\mathcal{P}$ is a finite set of preterminal nodes, Σ is a finite set of terminal nodes, and $\mathcal{R}$ is a set of production rules in the Chomsky normal form:

$$
\begin{aligned}
S &\rightarrow A, & A &\in \mathcal{N} \\
A &\rightarrow BC, & A &\in \mathcal{N}, B, C \in \mathcal{N} \cup \mathcal{P} \\
T &\rightarrow w, & T &\in \mathcal{P}, w \in \Sigma
\end{aligned}
$$

In natural language, nonterminals $\mathcal{N}$ are constituent labels and preterminals $\mathcal{P}$ are part-of-speech tags. A terminal node w is a word from a sentence, and Σ is the vocabulary. During implementation, we do not have the ground truth constituent labels and part-of-speech tags. Therefore, nonterminals and preterminals are sets of nodes (or clusters) which implicitly represent their functions.



Compound PCFG for Language

Context Free Grammar (CFG):

$$
\begin{array}{rr}
S \to A, & A \in \mathcal{N} \\
A \to BC, & A \in \mathcal{N}, B, C \in \mathcal{N} \cup \mathcal{P} \\
T \to w, & T \in \mathcal{P}, w \in \Sigma
\end{array}
$$

Probabilistic Context Free Grammar (PCFG):

$$\sum_{r:A\to\gamma} \pi_r = 1$$

Compound Probabilistic Context Free Grammar (Compound PCFG)

$$\pi_{S\to A} = \frac{\exp\left(\mathbf{u}_A^T f_s\left([\mathbf{w}_S;\mathbf{z}]\right)\right)}{\sum_{A'\in\mathcal{N}} \exp\left(\mathbf{u}_{A'}^T f_s\left([\mathbf{w}_S;\mathbf{z}]\right)\right)}$$

$$\pi_{A\to BC} = \frac{\exp\left(\mathbf{u}_{BC}^T [\mathbf{w}_A;\mathbf{z}]\right)}{\sum_{B',C'\in\mathcal{N}\cup\mathcal{P}} \exp\left(\mathbf{u}_{B'C'}^T [\mathbf{w}_A;\mathbf{z}]\right)} \qquad \pi_r = g_r(\mathbf{z};\theta), \quad \mathbf{z} \sim p(\mathbf{z})$$

$$\pi_{T\to w} = \frac{\exp\left(\mathbf{u}_w^T f_t\left([\mathbf{w}_T;\mathbf{z}]\right)\right)}{\sum_{w'\in\Sigma} \exp\left(\mathbf{u}_{w'}^T f_t\left([\mathbf{w}_T;\mathbf{z}]\right)\right)}$$

Maximum Likelihood with ELBO

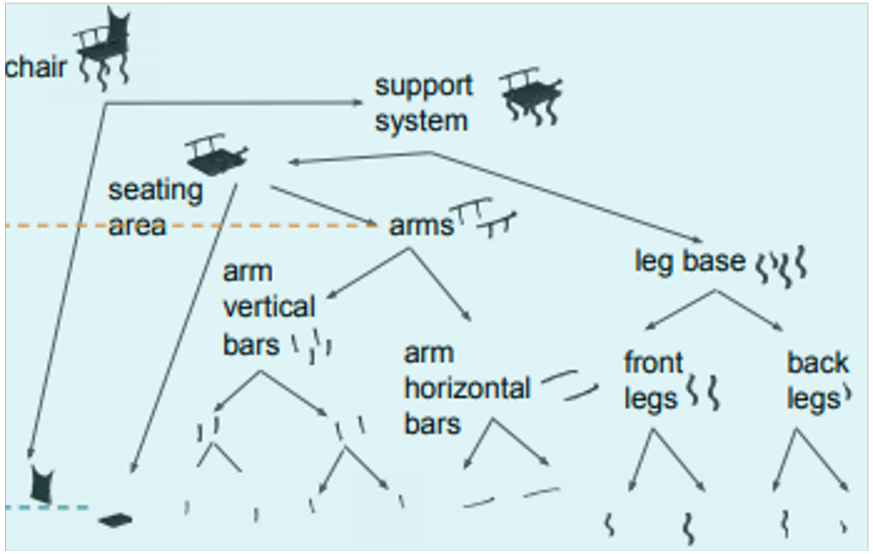
$$
\begin{aligned}
\mathcal{L}_g(\boldsymbol{w};\phi,\theta) &= -\operatorname{ELBO}(\boldsymbol{w};\phi,\theta) \\
&= -\mathbb{E}_{q_\phi(\mathbf{z}\mid\boldsymbol{w})}\left[\log p_\theta(\boldsymbol{w}\mid\mathbf{z})\right] + \operatorname{KL}\left[q_\phi(\mathbf{z}\mid\boldsymbol{w}) \| p(\mathbf{z})\right]
\end{aligned}
$$

Compound PCFG for Image

Context Free Grammar (CFG):

$$
\begin{array}{rr}
S \rightarrow A, & A \in \mathcal{N} \\
A \rightarrow BC, & A \in \mathcal{N}, B, C \in \mathcal{N} \cup \mathcal{P} \\
T \rightarrow w, & T \in \mathcal{P}, w \in \Sigma
\end{array}
$$

Compound PCFGs can be naturally extended to image grammar. In a compound PCFG for image, S denotes an object, *e.g.*, a chair. Nonterminals $\mathcal{N}$ are types of middle-level coarse parts. Preterminals $\mathcal{P}$ are types of fine-grained leaf-parts. The middle-level parts can be further decomposed into sub-parts which are either middle-level parts or leaf-parts; for example, the base of a chair is decomposed into the central support and the leg system, and the leg system is further decomposed into several legs.



Compound Probabilistic Context Free Grammar (Compound PCFG):

$$\pi_{S \rightarrow A} = \frac{\exp\left(\mathbf{u}_A^T f_s\left([\mathbf{w}_S; \mathbf{z}]\right)\right)}{\sum_{A' \in \mathcal{N}} \exp\left(\mathbf{u}_{A'}^T f_s\left([\mathbf{w}_S; \mathbf{z}]\right)\right)}$$

$$\pi_{A \rightarrow BC} = \frac{\exp\left(\mathbf{u}_{BC}^T [\mathbf{w}_A; \mathbf{z}]\right)}{\sum_{B', C' \in \mathcal{N} \cup \mathcal{P}} \exp\left(\mathbf{u}_{B'C'}^T [\mathbf{w}_A; \mathbf{z}]\right)}$$

$$\pi_{T \rightarrow w} = \frac{\exp\left(\mathbf{u}_w^T f_t\left([\mathbf{w}_T; \mathbf{z}]\right)\right)}{\sum_{w' \in \Sigma} \exp\left(\mathbf{u}_{w'}^T f_t\left([\mathbf{w}_T; \mathbf{z}]\right)\right)}$$

Compound PCFG for Image

Compound Probabilistic Context Free Grammar (Compound PCFG)

$$\pi_{S\to A} = \frac{\exp\left(\mathbf{u}_A^T f_s\left([\mathbf{w}_S;\mathbf{z}]\right)\right)}{\sum_{A'\in\mathcal{N}}\exp\left(\mathbf{u}_{A'}^T f_s\left([\mathbf{w}_S;\mathbf{z}]\right)\right)}$$

$$\pi_{A\to BC} = \frac{\exp\left(\mathbf{u}_{BC}^T\left[\mathbf{w}_A;\mathbf{z}\right]\right)}{\sum_{B',C'\in\mathcal{N}\cup\mathcal{P}}\exp\left(\mathbf{u}_{B'C'}^T\left[\mathbf{w}_A;\mathbf{z}\right]\right)}$$

Bottom-Up Perception

$$s(T, v_i) = \mathbf{u}_T^T f_t\left(\psi(v_i)\right)$$

Bottom-up perception module

Clustering module

$$\pi_{T\to v_i} = \frac{exp(s(T, v_i))}{\sum_{v'\in\Sigma} exp(s(T, v'))}$$

Maximum Likelihood with ELBO

$$\begin{aligned}\mathcal{L}_g(\boldsymbol{w};\phi,\theta) &= -\operatorname{ELBO}(\boldsymbol{w};\phi,\theta)\\ &= -\mathbb{E}_{q_\phi(\mathbf{z}\mid\boldsymbol{w})}\left[\log p_\theta(\boldsymbol{w}\mid\mathbf{z})\right] + \operatorname{KL}\left[q_\phi(\mathbf{z}\mid\boldsymbol{w})\|p(\mathbf{z})\right]\end{aligned}$$

Joint Learning by Alignment

Alignment Score between a Part and Language Constituent

$$s(\boldsymbol{w}_j, \boldsymbol{v}_k) \triangleq cos(\boldsymbol{w}_j, \boldsymbol{v}_k)$$

Alignment Score between an Image and Sentence

$$\begin{aligned}\mathcal{S}(\boldsymbol{w}, \boldsymbol{v}) &= \sum_{\substack{t_w \in \mathcal{T}_{\mathcal{G}_w}(\boldsymbol{w}) \\ t_v \in \mathcal{T}_{\mathcal{G}_v}(\boldsymbol{v})}} p(t_w | \boldsymbol{w}) p(t_v | \boldsymbol{v}) \sum_{\substack{\boldsymbol{w}_j \in t_w \\ \boldsymbol{v}_k \in t_v}} s(\boldsymbol{w}_j, \boldsymbol{v}_k) \\ &= \sum_{\substack{\boldsymbol{w}_j \in [\boldsymbol{w}] \\ \boldsymbol{v}_k \in [\boldsymbol{v}]}} \sum_{\substack{t_w \in \mathcal{T}_{\mathcal{G}_w}(\boldsymbol{w}) \\ t_v \in \mathcal{T}_{\mathcal{G}_v}(\boldsymbol{v})}} \mathbb{1}_{\{\boldsymbol{w}_j \in t_w\}} \mathbb{1}_{\{\boldsymbol{v}_k \in t_v\}} p(t_w | \boldsymbol{w}) p(t_v | \boldsymbol{v}) s(\boldsymbol{w}_j, \boldsymbol{v}_k) \\ &= \sum_{\substack{\boldsymbol{w}_j \in [\boldsymbol{w}] \\ \boldsymbol{v}_k \in [\boldsymbol{v}]}} p(\boldsymbol{w}_j | \boldsymbol{w}; \mathcal{G}_w) p(\boldsymbol{v}_k | \boldsymbol{v}; \mathcal{G}_v) s(\boldsymbol{w}_j, \boldsymbol{v}_k)\end{aligned}$$

Experiment: Grammar Induction

Table 2: **The performance of grammar induction.** "C" and "I" denote corpus-level and instance-level F1 scores, respectively. "VLG w/o SCAN" denotes that we do not use SCAN to pretrain the unsupervised clustering module of VLGrammar.

Model	**Vision Grammar**										**Language Grammar**									
	All		Chair		Table		Bed		Bag		All		Chair		Table		Bed		Bag	
	C	I	C	I	C	I	C	I	C	I	C	I	C	I	C	I	C	I	C	I
Left-Branch	16.4	20.2	9.9	11.5	21.1	26.3	**38.8**	59.4	54.2	60.0	16.2	17.6	19.2	19.8	13.7	15.8	10.5	12.0	8.4	8.9
Right-Branch	40.8	49.1	42.8	48.0	39.1	50.2	12.8	20.8	81.0	97.5	49.2	**53.5**	43.7	48.6	54.2	**58.1**	43.7	46.2	68.3	69.3
ON-LSTM	/	/	/	/	/	/	/	/	/	/	30.7	33.4	32.5	34.4	28.9	32.4	27.3	29.0	39.4	38.5
L-PCFG-P	/	/	/	/	/	/	/	/	/	/	47.8	49.4	41.4	44.9	53.6	53.5	44.9	44.3	63.7	63.5
L-PCFG	/	/	/	/	/	/	/	/	/	/	48.4	50.3	42.2	46.2	53.6	53.5	55.3	**55.1**	71.2	71.4
V-PCFG	47.5	59.3	51.6	59.0	43.3	59.2	36.2	48.2	82.4	91.3	/	/	/	/	/	/	/	/	/	/
L-PCFG-VG	/	/	/	/	/	/	/	/	/	/	49.0	49.6	42.3	44.0	**54.6**	54.3	56.0	54.6	73.0	73.0
V-PCFG-LG	44.2	52.7	42.0	47.5	45.6	56.6	**38.8**	54.3	88.2	95.7	/	/	/	/	/	/	/	/	/	/
VLGrammar	**51.4**	**63.4**	**56.4**	**65.9**	46.3	60.5	38.1	**59.7**	**94.1**	**98.0**	**51.3**	51.9	**47.8**	**49.4**	54.0	53.8	**56.2**	54.8	**73.6**	**73.6**
VLG w/o SCAN	44.7	55.5	30.5	33.6	**57.9**	**75.4**	29.0	56.4	88.2	95.7	49.0	49.8	43.4	45.3	53.7	53.5	55.1	54.0	72.6	72.6

Experiments

Unsupervised part clustering

Table 3: **The accuracy of the unsupervised part clustering.**

Model	All	Chair	Table	Bed	Bag
SCAN	41.3	43.5	37.5	59.3	88.9
V-PCFG	61.6	68.3	58.3	69.9	88.9
V-PCFG-LG	65.4	66.8	63.2	71.8	**90.5**
VLGrammar	**69.1**	**71.6**	66.0	**75.1**	**90.5**
VLG w/o SCAN	64.4	62.0	**66.2**	60.4	**90.5**

Generalization

Table 5: The performance of image grammars on all categories, while being trained on only `chair` and `table`.

Model	Seen				Unseen			
	Chair		Table		Bed		Bag	
	C	I	C	I	C	I	C	I
V-PCFG	43.9	52.7	38.1	54.5	20.7	33.1	82.4	91.3
V-PCFG-LG	44.3	**54.1**	38.5	54.8	25.6	**50.4**	**88.2**	**95.7**
VLGrammar	**44.8**	53.4	**41.1**	**56.7**	**29.4**	44.2	**88.2**	**95.7**

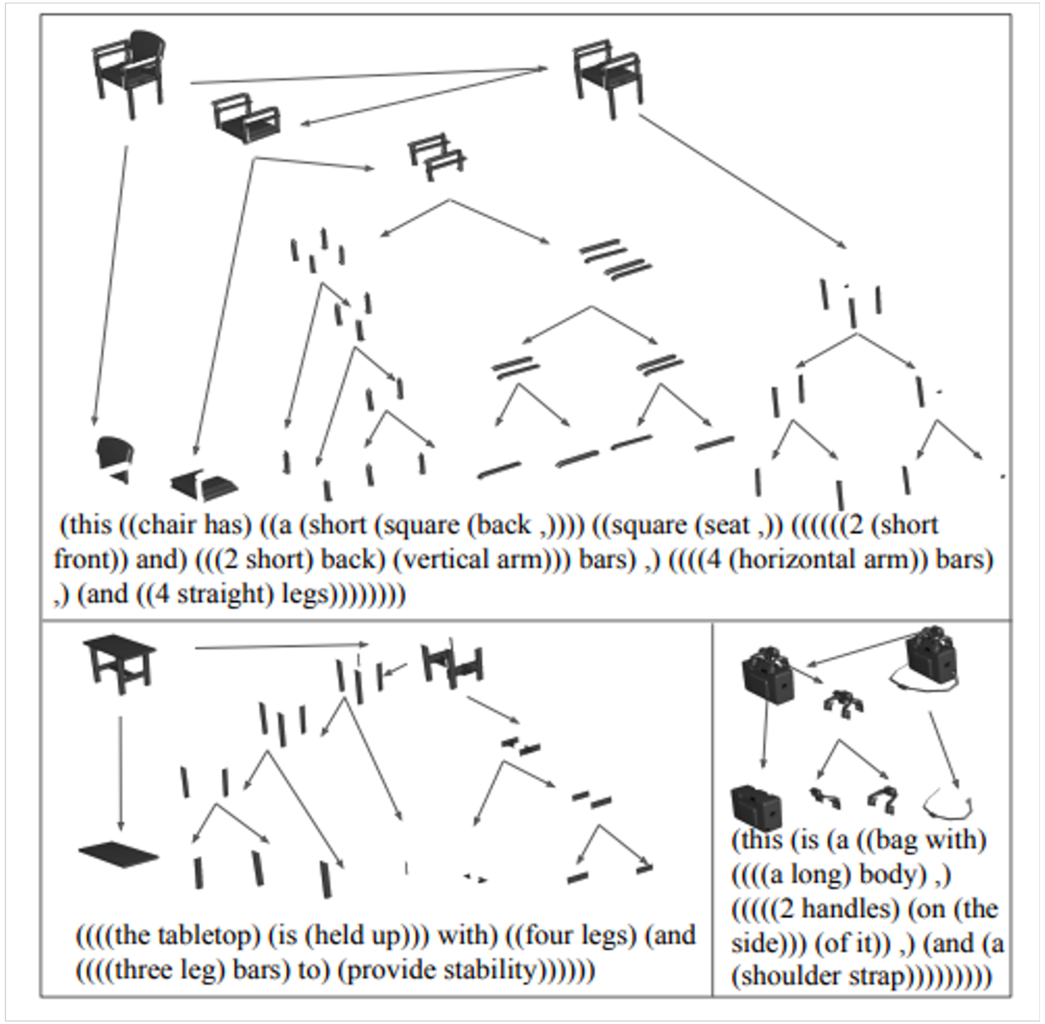
Retrieval

Table 4: **The accuracy of image-text retrieval.** "IR" stands for text-to-image retrieval and "TR" is for image-to-text retrieval.

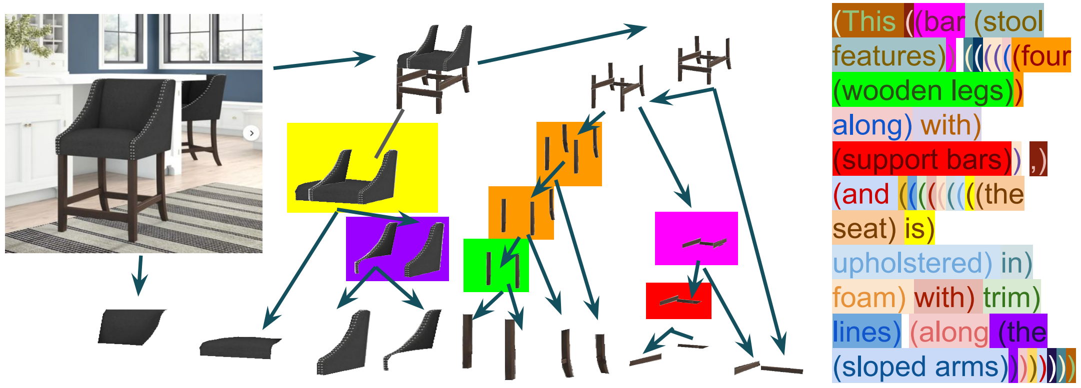
Model	Chair		Table		Bed		Bag	
	IR	TR	IR	TR	IR	TR	IR	TR
Baseline	24.1	28.5	29.8	31.2	20.1	20.1	19.1	24.5
L-PCFG-VG	**34.5**	36.9	39.3	42.0	35.5	**38.4**	23.0	28.7
V-PCFG-LG	25.9	27.8	38.8	41.8	29.6	25.7	23.8	24.9
VLGrammar	33.2	**39.0**	**39.8**	**42.5**	**39.6**	38.2	**24.6**	**29.3**

Qualitative Results

Results on PartIt Dataset



Transfer to Real Images



Future Directions

- Extend the PartIt dataset to fully 3D with detailed part information
- Learn a 2D grammar that can capture more sophiscated spatial relations

Thank you!