Supplementary Material for VLGrammar: Grounded Grammar Induction of Vision and Language

Yining Hong[1], Qing Li[1], Song-Chun Zhu[2,3,4], Siyuan Huang[1]
[1] University of California, Los Angeles
[2] Beijing Institute for General Artificial Intelligence, [3] Tsinghua University, [4] Peking University

1. Production Rules

$S \rightarrow$ Chair | Table | Bed | Bag
Chair → Upper Part, Support System | Support System
Upper Part → Chair Head, Chair Back | Chair Back
Chair Back → Back Bars | Back Surface | Back Bars, Back Surface
Support System → Seating Area, Base
Seating Area → Seat, Arms | Seat
Arms → Arms, Arm | Arm | Arm Vertical Bars, Arm Horizontal Bars
Arm Vertical Bars → Arm Vertical Bars, Arm Vertical Bar
Arm Horizontal Bars → Arm Horizontal Bars, Arm Horizontal Bar
Base → LegBase | Pedestal Base
LegBase → Central Support, Leg System | Leg System
Leg System → Leg | Leg Bar | Leg System, Leg | Leg System, Leg Bar
Pedestal Base → Central Support, Pedestal
Table → Upper Part, Table Base
Upper Part → Functional Part, Panels | Functional Part
Functional Part → TableTops, Storage Part | TableTops
TableTops → TableTops, TableTop
StoragePart → Drawers | Cabinets | Drawers, Cabinets
Panels → Side Panels | Bottom Panels | Side Panels, Bottom Panels
Side Panels → Side Panel | Side Panels, Side Panel
Bottom Panels → Bottom Panels, Bottom Panel | Bottom Panel
Drawers → Drawers, Drawer | Drawer
Cabinets → Cabinets, Cabinet | Cabinet
Base → LegBase | Pedestal Base
LegBase → Central Support, Leg System | Leg System
Leg System → Leg System, Leg Bars | Legs | Legs, Shelves
Legs → Legs, Leg | Leg
Shelves → Shelves, Shelf | Shelf
Leg Bars → Leg Bars, Leg Bar | Leg Bar
Pedestal Base → Central Support, Pedestal
Bed → Functional Part, Base
Functional Part → Upper Bed, Single Functional Part | Single Functional Part
Upper Bed → Single Functional Part, Bed Posts | Single Functional Part, Ladder
Bed Posts → Bed Posts, Bed Post | Bed Post
Single Functional Part → Sleeping Area, Side Panels | Sleeping Area
Sleeping Area → Bed Sleep Area | Sleeping Area, Frame Horizontal Surface| Headboard, Sleeping Area
Side Panels → Side Panel | Side Panels, Side Panel
Base → Legs | Surface Base
Legs → Legs, Leg | Leg
Bag → Main Body, Shoulder Straps | Main Body
Main Body → Bag Body, Handles
Handles → Handles Handles | Handle
Shoulder Straps → Shoulder Straps, Shoulder Strap | Shoulder Strap

Table 1: Production rules of vision grammar.

Table 1 lists the production rules of vision grammar.

2. Implementation Details

We adopt parameter settings suggested by the authors for the baseline model. For Compound PCFGs, We adopt most of the parameters from [3]. For language Compound PCFG, the parsing model has 20 nonterminals and 30 preterminals. For vision Compound PCFG, the parsing model has 10 nonterminals and 13 preterminals. Each of them is represented by a 256-dimensional vector. The inference model $q_{\phi_w}(\mathbf{z}|\boldsymbol{w})$ uses a single-layer BiLSTM. It has a 512-dimensional hidden state and relies on 512-dimensional word embeddings. We apply a max-pooling layer over the hidden states of the BiLSTM and then obtain 64-dimensional mean vectors $\boldsymbol{\mu}_{\phi_w}(\boldsymbol{w})$ and log-variances $\log \sigma_{\phi_w}(\boldsymbol{w})$ using an affine layer to obtain z. We apply a convolutional layer with out channels 16 augmented with position embedding, which is a $W \times H \times 4$ feature vector encoding the distance to the borders of the feature map. We apply an average pooling layer over the features output by the convolutional layer and then obtain 64-dimensional mean vectors $\boldsymbol{\mu}_{\phi_v}(\boldsymbol{v})$ and log-variances $\log \sigma_{\phi_v}(\boldsymbol{v})$ using an affine layer to obtain z. The clustering module uses a ResNet-18, replaces the last layer so that is outputs a feature of dimension 512, followed by a fully connected layer that maps the feature to the preterminals. The clustering module is trained with SimCLR for 100 epochs, and SCAN for another 100 epochs. The parameters are the best parameters from the original papers. The vision-language alignment module projects both vision constituents and language constituents as 128-dimensional vectors. Specifically, the language span representation model is another single-layer BiLSTM, with the same hyperparameters as in the inference model. The vision constituent representation model is the same the ResNet-18 used in the clustering module. For the total loss:

$$\mathcal{L} = \lambda_w \mathcal{L}_{\mathcal{G}}(\mathcal{W}; \phi_w, \theta_w) + \lambda_v \mathcal{L}_{\mathcal{G}}(\mathcal{V}; \phi_v, \theta_v) + \lambda_C \mathcal{L}_C(\mathcal{W}, \mathcal{V}) \quad (1)$$

λ_w and λ_v are set to 1.0, and λ_C is set to 0.01.

We use Adam optimizer with a learning rate of 0.01 and β_1 is 0.75 and β_2 is 0.999. The batch size is 8.

3. Contrastive Learning

To reduce computation we estimate the contrastive loss using only the $min(2n, \frac{n(n-1)}{2})$ shortest spans for a sentence of length n. This is reasonable since the phrase that can be grounded in a part in the image is typically very short. The tendency to focus on short spans is also in [3, 2, 1]

4. Parsing

In inference, the parser can be directed used without the alignment between images and sentences.

$$t_w^* = \text{argmax} \int_{\mathbf{z}} p_\theta(t \mid \boldsymbol{w}, \mathbf{z}) p_\theta(\mathbf{z} \mid \boldsymbol{w}) d\mathbf{z} \qquad (2)$$

which becomes intractable because of $\mathbf{z}$. The MAP inference is inferenced by:

$$t_w^* \approx \text{argmax} \int_{\mathbf{z}} p_\theta(t \mid \boldsymbol{w}, \mathbf{z}) \delta\left(\mathbf{z} - \boldsymbol{\mu}_{\phi_w}(\boldsymbol{w})\right) d\mathbf{z} \qquad (3)$$

For vision:

$$t_v^* \approx \text{argmax} \int_{\mathbf{z}} p_\theta(t \mid \boldsymbol{v}, \mathbf{z}) \delta\left(\mathbf{z} - \boldsymbol{\mu}_{\phi_v}(\boldsymbol{v})\right) d\mathbf{z} \qquad (4)$$

5. Annotation

Fig. 1 shows the website on Mechanical Turk for the collection of our dataset. Fig. 2 shows the examples we provide for workers. Fig. 3 shows the instructions for workers.

6. Robustness Analysis

We conduct extended experiments by calculating the standard deviations to analyze the model performance and robustness. The results were run on a fixed seed due to GPU and time limit. We evaluate three models: VLGrammar, L-PCFG, and V-PCFG. We report the results with standard deviations using four random seeds in Table 2. Our model outperforms baselines a lot.

	Chair		Table		Bed		Bag	
	C	I	C	I	C	I	C	I
V-PCFG	43.3±6.2	51.2±6.4	44.5±3.1	56.6±3.9	37.5±1.8	49.8±4.0	82.8±3.5	91.8±2.4
VLG(V)	50.2±7.1	59.3±6.6	52.1±4.2	67.7±5.8	39.1±2.6	54.0±5.2	91.3±2.9	96.9±0.8
L-PCFG	33.8±10.9	35.1±8.5	46.3±6.6	46.5±5.8	54.2±3.3	53.7±4.2	68.3±2.1	68.4±2.2
VLG(L)	40.8±15.5	44.7±14.8	52.7±9.3	52.5±8.5	55.1±5.9	54.5±6.0	71.8±4.8	72.5±5.5

Table 2: Results with standard deviations. VLG is VLGrammar.

References

[1] Noriyuki Kojima, Hadar Averbuch-Elor, Alexander M. Rush, and Yoav Artzi. What is learned in visually grounded neural syntax acquisition. In *ACL*, 2020. 2

[2] Haoyue Shi, Jiayuan Mao, Kevin Gimpel, and Karen Livescu. Visually grounded neural syntax acquisition. *ArXiv*, abs/1906.02890, 2019. 2

[3] Yanpeng Zhao and Ivan Titov. Visually grounded compound pcfgs. *ArXiv*, abs/2009.12404, 2020. 1, 2

Figure 1: Annotating interface for our dataset.

Examples:

A. **Chair**:

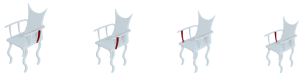
Parts:

chair_back:

chair_seat:

arm_horizontal_bar:

arm_vertical_bar



Leg:

Write Your Sentence here: This chair has an irregular back, a seat, four vertical bars and two horizontal bars to form its arms, as well as four curved legs.

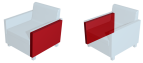
B. **Chair:**

Parts:

chair_back:

chair_seat:

arm_sofa_style:



leg:

Write Your Sentence Here: This is a sofa with a back, a thick seat, two sofa-style arms and four short legs.

Figure 2: Examples we provide for turkers.

Instructions:

1. Use only **one sentence** to describe the object. This means you cannot write two grammarly correct sentences and combine them into one using “,” and “;”. e.g.,
   Correct Example:
   This high chair consists of a back, a seat, two arms which have one horizontal bar and two vertical bars in each arm, as well as four long legs connected with three leg bars.
   Wrong Examples:
   This is a high chair. It consists of a back and a seat. It also has two arms. Each arm has one horizontal bar and two vertical bars.
   This is a high chair consisting of a back and a seat, it also has two arms, each arm has one horizontal bar and two vertical bars.
2. When writing the sentence, you are given images of an object and its parts and need
   a. Describe the object’s category:
      i. chair  , Sofa  ,computer/office/desk chair  , folding chair  , rocking chair  , ball/bowl chair  , cuddler chair  , high chair  , stool  etc.
   b. Describe **all parts of the object**:
      i. chair head  , chair back  , chair back vertical bars  , chair back horizontal bars  , chair seat  , chair arms  , sofa arms  , chair arm horizontal bars  , chair arm vertical bars  , chair legs  , chair leg bars  , central support  , pedestal 
   c. Describe each part’s attributes:
      i. **Count**: e.g., how many legs? How many bars to connect the legs? How many arms? How many horizontal bars in the arm (back)? How many vertical bars in the arm (back)?
      ii. **Shape**: round, square, triangle, curved, tilted, irregular shape, etc.
      iii. **Size**: thick, thin, large, small
      iv. **Length**: long, short

Figure 3: Instructions we provide for turkers.