

Grounded Situation Recognition with Transformers - Supplementary Material

BMVC 2021 Submission

This supplementary material provides more details of our model, further analyses on it, additional ablation studies and experimental results, which are left out from the main paper due to the space limit. Section A describes the transformer architecture of GSRTR in detail, Section B performs the ablation studies on GSRTR, Section C provides more qualitative examples of the total prediction of GSRTR, and a more thorough qualitative analysis on attention of GSRTR is illustrated in Section D. Finally, the practical applications of Grounded Situation Recognition are described in Section E.

A Detailed Transformer Architecture

Transformer Encoder-Decoder: The detailed transformer architecture of GSRTR is given in Figure A1. The encoder takes as input a verb token and flattened image features, and then produces a verb feature and image features. Along with image features given by the encoder, the decoder takes as input semantic role queries, and then produces output features corresponding to the semantic roles. The encoder is a stack of six encoder layers and the decoder is a stack of six decoder layers. Each encoder layer consists of a Multi-Head Self-Attention (MHSA) block and a Feed-Forward Network (FFN) block. Each decoder layer consists of a MHSA block, a Multi-Head Attention (MHA) block, and a FFN block. We use Pre-Layer Normalization (Pre-LN) [9], *i.e.*, LayerNorm is used before each MHSA block, MHA block, and FFN block, and also before the verb feature and before the decoder output features corresponding to the semantic roles. The skip connection, using 0.15 dropout rate, is given by:

$$\mathbf{x} + \text{Dropout}(\text{Block}(\text{LayerNorm}(\mathbf{x}))), \quad (\text{A.1})$$

where $\mathbf{x} \in \mathbb{R}^d$ and Block denotes one of the MHSA block, MHA block, and FFN block. Note that we use $d = 512$. The FFN block is 2 fully-connected layers with ReLU activation function and 2048 hidden dimensions, using 0.15 dropout rate, and it is given by:

$$\text{FFN}(\mathbf{x}) = W_2(\text{Dropout}(\max(W_1\mathbf{x} + \mathbf{b}_1, \mathbf{0}))) + \mathbf{b}_2, \quad (\text{A.2})$$

where $\mathbf{x} \in \mathbb{R}^d$, $W_1 \in \mathbb{R}^{2048 \times d}$, $\mathbf{b}_1 \in \mathbb{R}^{2048}$, $W_2 \in \mathbb{R}^{d \times 2048}$, and $\mathbf{b}_2 \in \mathbb{R}^d$. We use Xavier initialization [10] for the learnable parameters in the encoder and decoder.

Multi-Head Attention: MHA takes as input a query sequence $X_Q \in \mathbb{R}^{d \times n_Q}$ and a key-value sequence $X_{KV} \in \mathbb{R}^{d \times n_{KV}}$, where n_Q denotes the query sequence length and n_{KV} denotes the

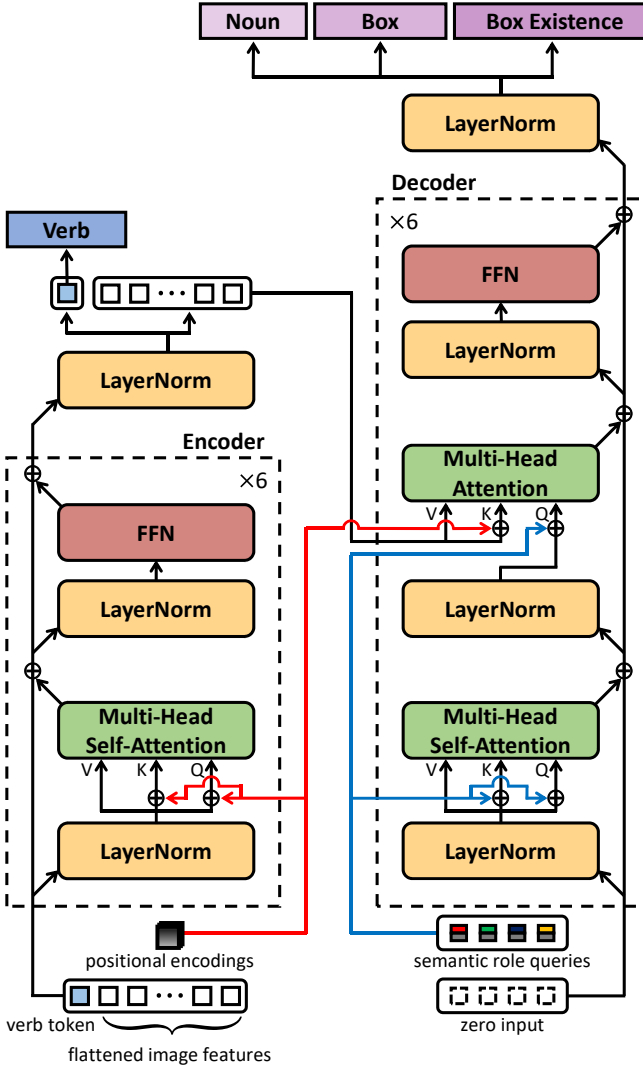


Figure A1: The detailed transformer architecture of GSRTR. A verb token and flattened image features are used for the first encoder layer input (black line in Encoder). Zero input is used for the first decoder layer input (black line in Decoder). Positional encodings are added to the keys and queries of the MHSA block in each encoder layer and the keys of the MHA block in each decoder layer (red line). Semantic role queries are added to the keys and queries of the MHSA block in each decoder layer and the queries of the MHA block in each decoder layer (blue line). We omit Dropout in this diagram.

key-value sequence length. MHSA corresponds to the case when the query sequence is same with the key-value sequence in MHA, *i.e.*, when $X_Q = X_{KV}$ in MHA. MHA is formulated as:

$$\text{MHA}(X_Q, X_{KV}) = W_O [\text{Head}^1(X_Q, X_{KV}); \dots; \text{Head}^H(X_Q, X_{KV})], \quad (\text{A.3})$$

where H is the number of heads, $[\cdot]$ denotes a concatenation and $W_O \in \mathbb{R}^{d \times d}$ denotes an

output projection. Note that we use $H = 8$. Head^m denotes each attention function with linear projections for $m = 1, \dots, H$, and it is given by:

$$\text{Head}^m(X_Q, X_{KV}) = \text{Attn}(W_Q^m X_Q, W_K^m X_{KV}, W_V^m X_{KV}), \quad (\text{A.4})$$

where $W_Q^m, W_K^m, W_V^m \in \mathbb{R}^{d' \times d}$ denotes linear projection of m^{th} head for key, query, and value, respectively. The linear projection matrices are learnable parameters, which are not shared across the MHA and MHSA blocks in the encoder and decoder layers. Note that we use $d' = 64$, where $d' = \frac{d}{H}$. Attn denotes an attention function which transforms a query sequence $Q \in \mathbb{R}^{d' \times n_Q}$ into an output sequence, whose element is a weighted sum of a value sequence $V \in \mathbb{R}^{d' \times n_{KV}}$. For i^{th} query $\mathbf{q}_i \in \mathbb{R}^{d'}$, each weight of the sum is computed by a softmax function (i.e., Softmax) after a scaled dot-product between the i^{th} query \mathbf{q}_i and a key sequence $K \in \mathbb{R}^{d' \times n_{KV}}$. In other words, the i^{th} element of the attention function output from the query sequence Q , key sequence K , and value sequence V is given by:

$$\text{Attn}_i(Q, K, V) = \sum_j \text{Softmax}_j \left(\frac{1}{\sqrt{d'}} \mathbf{q}_i K \right) \mathbf{v}_j, \quad (\text{A.5})$$

where Softmax_j denotes the j^{th} output of the softmax function and $\mathbf{v}_j \in \mathbb{R}^{d'}$ denotes the j^{th} value.

The MHSA block in the encoder: The encoder takes as input a verb token and flattened image features. The positional encodings $P \in \mathbb{R}^{d \times hw}$ are used, where hw denotes the length of flattened image features. The positional encodings P are 2D learnable embeddings, and they are used at the attention function of each MHSA block in the encoder. To be specific, the positional encodings are added to the corresponding image features, which are used as the key and query inputs at the attention function. For the verb token, we append zero to the positional encodings, leading to $P' \in \mathbb{R}^{d \times (1+hw)}$. As a result, the positional encodings P' are added to the key and query inputs of the attention function in each MHSA block of the encoder. Thus, the m^{th} attention function in each MHSA block of the encoder is given by:

$$\text{Head}^m(X_Q, X_{KV}) = \text{Attn}(W_Q^m(X_Q + P'), W_K^m(X_{KV} + P'), W_V^m X_{KV}), \quad (\text{A.6})$$

where $X_Q = X_{KV}$ and $X_Q \in \mathbb{R}^{d \times (1+hw)}$.

The MHSA and MHA blocks in the decoder: Along with the image features given by the encoder, the decoder takes as input a sequence of the semantic role queries. Additionally to Section 3.3, each semantic role query $\mathbf{w}_{(v,r)}$ per semantic role $r \in \mathcal{R}_v$ can formulate a sequence with arbitrary role orders, leading to the semantic role query sequence $S_v \in \mathbb{R}^{d \times |\mathcal{R}_v|}$. Note that the initial decoder input is set to zero. In each MHSA block of the decoder, the semantic role query sequence S_v is added to the query and key inputs of the attention function. In other words, the m^{th} attention function in each MHSA block of the decoder is given by:

$$\text{Head}^m(X_Q, X_{KV}) = \text{Attn}(W_Q^m(X_Q + S_v), W_K^m(X_{KV} + S_v), W_V^m X_{KV}), \quad (\text{A.7})$$

where $X_Q = X_{KV}$ and $X_Q \in \mathbb{R}^{d \times |\mathcal{R}_v|}$. In each MHA block of the decoder, the semantic role query sequence S_v are added to the query inputs of the attention function, and positional encodings P are added to the key inputs of the attention function. In other words, the m^{th} attention function in each MHA block of the decoder is given by:

$$\text{Head}^m(X_Q, X_{KV}) = \text{Attn}(W_Q^m(X_Q + S_v), W_K^m(X_{KV} + P), W_V^m X_{KV}), \quad (\text{A.8})$$

where $X_Q \in \mathbb{R}^{d \times |\mathcal{R}_v|}$ and $X_{KV} \in \mathbb{R}^{d \times hw}$.

Table B1: Ablation studies on our model (GSRTR).

set	model	top-1 predicted verb					top-5 predicted verbs					ground-truth verb			
		verb	value	value-all	grnd value	grnd value-all	verb	value	value-all	grnd value	grnd value-all	value	value-all	grnd value	grnd value-all
dev	GSRTR w/ 4 layers	40.26	31.88	19.20	25.44	10.20	69.34	53.52	30.33	42.29	15.69	74.09	38.88	57.97	19.75
	GSRTR w/ 8 layers	40.49	32.10	19.46	25.69	10.39	69.11	53.34	30.62	42.35	15.88	74.07	39.12	58.27	19.92
	GSRTR w/ Post-LN	40.18	31.50	18.54	25.20	9.89	68.82	52.72	29.30	41.79	15.27	73.30	37.60	57.50	19.34
	GSRTR	41.06	32.52	19.63	26.04	10.44	69.46	53.69	30.66	42.61	15.98	74.27	39.24	58.33	20.19
test	GSRTR w/ 4 layers	40.87	32.21	19.13	25.35	9.83	69.87	53.78	30.25	41.97	15.22	73.89	38.42	57.00	18.88
	GSRTR w/ 8 layers	40.83	32.20	19.17	25.49	10.03	69.47	53.40	30.07	41.99	15.35	73.75	38.54	57.20	19.19
	GSRTR w/ Post-LN	40.31	31.72	18.69	25.03	9.56	69.86	53.57	29.89	41.99	15.14	73.33	37.76	56.70	18.78
	GSRTR	40.63	32.15	19.28	25.49	10.10	69.81	54.13	31.01	42.50	15.88	74.11	39.00	57.45	19.67

B Ablation Studies

We study the effect on the number of layers and the location of LayerNorm in GSRTR. Our experiments are evaluated on the *dev* and *test* splits of SWiG dataset [2], and the results are compared with the proposed model and setting in Section 4.2.

The effect on the number of layers in the encoder and decoder is shown at the first and second row of each set in Table B1. GSRTR w/ 4 layers denotes that each of the transformer encoder and decoder has four layers, and GSRTR w/ 8 layers denotes that each has eight layers. In ground-truth verb setting, the noun and grounded noun accuracies of both models decrease. The top-1 predicted verb and top-5 predicted verbs accuracies of both models marginally fluctuate.

The effect on the location of LayerNorm in GSRTR is shown at the third row of each set in Table B1. GSRTR w/ Post-LN denotes that LayerNorm is placed between skip connections, leading to Post-Layer Normalization (Post-LN) [9] transformer architecture. In all evaluation metrics of each set, the accuracies of GSRTR w/ Post-LN decrease.

C More Qualitative Results of Our Model

In top-1 predicted verb setting on the *test* split of the SWiG dataset, the prediction results of GSRTR are shown in Figure C2, Figure C3 and Figure C4. The SWiG dataset has three noun annotations for each semantic role. The noun prediction is considered correct if the predicted noun matches one of the three noun annotations. The box prediction is considered correct if the model correctly predicts box existence and the predicted box has an Intersection-over-Union (IoU) value of at least 0.5 with the ground-truth box. Note that the grounded noun prediction is considered correct if the predicted noun and predicted box are correct.

Figure C2 shows the correct grounded noun prediction results. Figure C3 shows the failure cases of box prediction. There are incorrect box predictions when bounding boxes have extreme aspect ratios (e.g., the boxes of the role *Tool* in the *Surfing* and the *Coloring* image), or small scales (e.g., the box of the role *Agent* in the *Mowing* image and the box of the role *Tool* in the *Helping* image). Figure C4 shows the failure cases of noun prediction, including incorrect box predictions. Even in the failure cases, there are the cases where GSRTR reasonably predicts nouns. For example, in the *Tilting* image, GSRTR predicts that the noun of the role *Place* is *Outdoors*, which is similar to the first annotation *Outside*. In the *Curling* image, GSRTR predicts that the nouns of the role *Agent* and *Place* are *Person* and \emptyset , which are enough to describe the given image. There is also the case where GSRTR inappropriately predicts nouns. In the *Chasing* image, GSRTR predicts that the noun of the role *Chasee* is *Zebra*, whereas the three noun annotations are *Bull*, *Calf*, and *Cow*.

GT					PRED				
Sketching					Sketching				
Agent	Image	Material	Tool	Place	Agent	Image	Material	Tool	Place
Person	Woman	Paper	Pencil	ø	Person	Woman	Paper	Pencil	ø
Person	Sketch	Paper	Pen	ø					

GT					PRED				
Plowing					Plowing				
Agent	Instrument	Place			Agent	Instrument	Place		
Man	Horse	Farm			Farmer	Horse	Field		
Farmer	Horse	Field							
Horse	Plow	Field							

GT					PRED				
Clawing					Clawing				
Agent	Victim		Place		Agent	Victim		Place	
Cat	Bark		Tree		Cat	Tree		Outdoors	
Cat	Limb		Outdoors						
Kitten	Tree		Outside						

GT					PRED				
Falling					Falling				
Agent	Source	Goal	Place		Agent	Source	Goal	Place	
Man	Horse	Earth	Outdoors		Man	Horse	Land	Field	
Horsemán	Horseback	Land	Field						
Jockey	Horse	Grass	Racetrack						

Figure C2: Correct grounded noun predictions of GSRTTR in top-1 predicted verb setting on the *test* set. For each semantic role, three annotators record noun annotations.


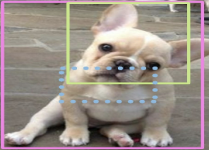
GT					PRED				
Surfing					Surfing				
Agent	Path	Place	Tool		Agent	Path	Place	Tool	
Surfer	Water	Ocean	Surfboard		Surfer	Water	Ocean	Surfboard	
Surfer	Water	Ocean	Surfboard						



GT					PRED				
Mowing					Mowing				
Agent	Item	Tool	Place		Agent	Item	Tool	Place	
Person	Weed	Tractor	Farm		Person	Grass	Land	Mower	
Man	Grass	Tractor	Field						
Tractor	Field	Land	Mower						



GT					PRED				
Coloring					Coloring				
Agent	Item	Tool	Place		Agent	Item	Tool	Place	
Child	Paper	Crayon	Table		Child	Paper	Crayon	Table	
Child	Paper	Crayon	ø						

GT					PRED				
Helping					Helping				
Agent	Entity Helped	Tool	Place		Agent	Entity Helped	Tool	Place	
Man	Person	Hand	Mountain		Person	Person	Hand	Mountain	
Person	Person	Hand	ø						
Person	Person	Hand	Mountain						

Figure C3: Incorrect box predictions of GSRTTR in top-1 predicted verb setting on the *test* set. The dashed box denotes incorrect box prediction.

GT				PRED			
							
Tilting				Tilting			
Agent	Item	Agent Part	Place	Agent	Item	Agent Part	Place
Dog	Head	Neck	Outside	Dog	Head	Neck	Outdoors
Dog	Head	Neck	Street				
Puppy			Patio				

GT				PRED			
							
Extinguishing				Extinguishing			
Agent	Item	Tool	Place	Agent	Item	Tool	Place
Man	Land	Fire Extinguisher	Outside	Man	Fire	Fire Extinguisher	Outdoors
Man	Land	Fire Extinguisher	Outdoors				
Man	Ø						

GT				PRED			
							
Curling				Curling			
Agent	Target	Tool	Place	Agent	Target	Tool	Place
Woman	Hair	Curling Iron	Salon	Person	Hair	Curling Iron	Ø
Woman	Hair	Curling Iron	Inside				
Woman							

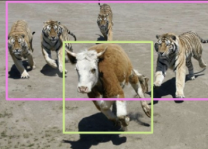
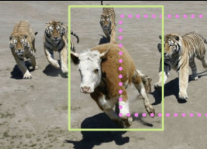
GT				PRED			
							
Chasing				Chasing			
Agent	Chasee	Place		Agent	Chasee	Place	
Tiger	Bull	Outdoors		Tiger	Zebra	Outdoors	
Tiger	Calf	Desert					
Tiger	Cow	Field					

Figure C4: Incorrect noun predictions of GSRTR in top-1 predicted verb setting on the *test* set. The incorrect noun predictions are highlighted in red color. The dashed box denotes incorrect box prediction.

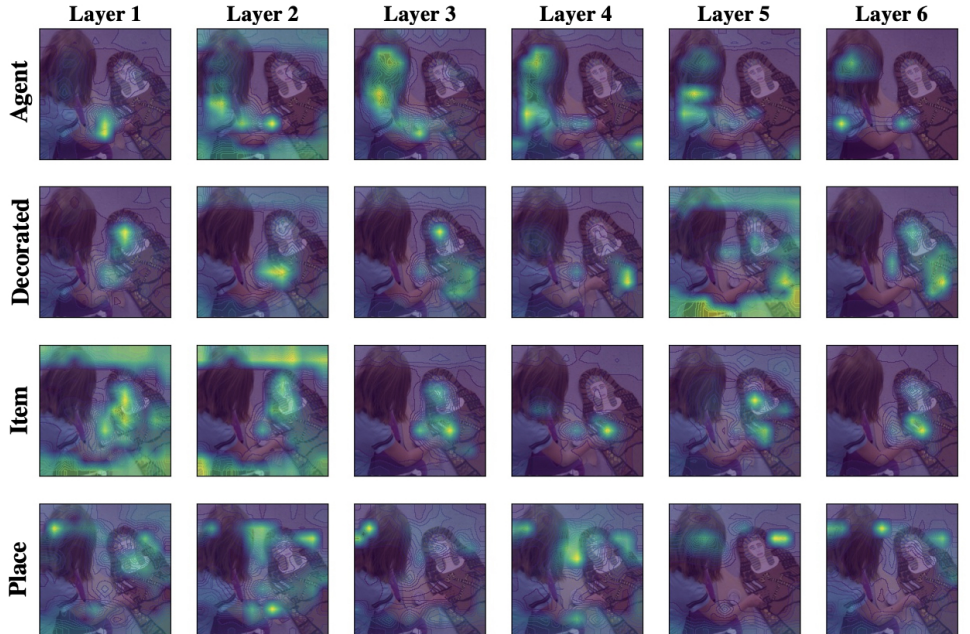


Figure D5: Role Attention Map on Image Features for a *Decorating* image from the MHA block in each decoder layer.

D Qualitative Analysis on Attention

Role Attention Map on Image Features: In Figure D5, Figure D6 and Figure D7, each column shows the difference of attention maps among roles. Each row shows the transition of attention maps through the decoder layers. In Figure D5, the role *Decorated* focuses on the decorated stuff and the role *Item* focuses on the decoration item. Figure D6 shows that GSRTR can understand the given image and distinguish between the role *agent* and the role *victim*. Figure D6 and Figure D7 show that GSRTR can figure out the background for the role *Place* in the given image.

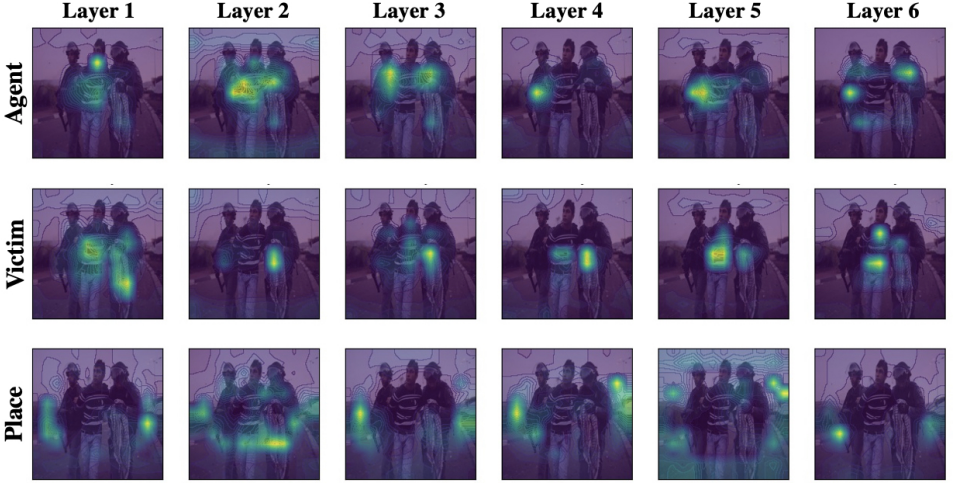


Figure D6: Role Attention Map on Image Features for a *Apprehending* image from the MHA block in each decoder layer.

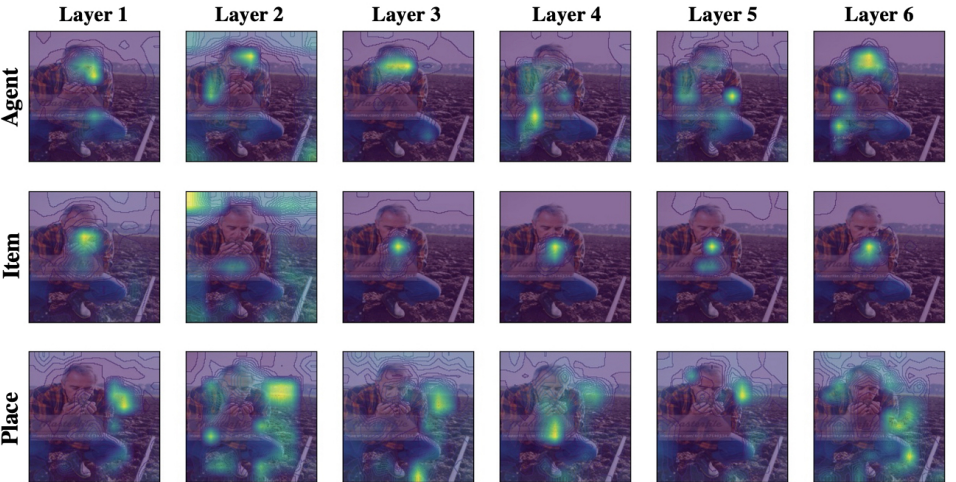


Figure D7: Role Attention Map on Image Features for a *Smelling* image from the MHA block in each decoder layer.

Visualization of Role Relations: GSRTTR captures the relations among roles in the similar way if the situations of the given images are similar. In Figure D8, the role *Vehicle* focuses on the role *Place*, i.e., the runway (*Place*) and the railway station (*Place*) are highly related to the airplane (*Vehicle*) and the train (*Vehicle*) given the verb *Boarding*, respectively. In Figure D9, the role *Obstacle* and the role *Tool* focus on the role *Place*, i.e., the cliff (*Place*) is highly related to the rock (*Obstacle*) and the rope (*Tool*) given the verb *Climbing*.

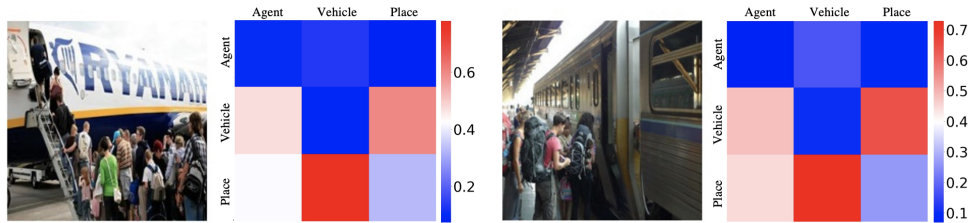


Figure D8: Visualization on Role Relations for two *Boarding* images from the MHSA block in the last decoder layer. Attention scores are represented as column-wise sum to 1.

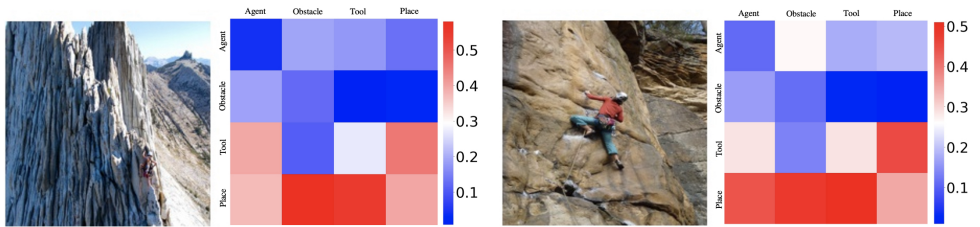


Figure D9: Visualization on Role Relations for two *Climbing* images from the MHSA block in the last decoder layer. Attention scores are represented as column-wise sum to 1.

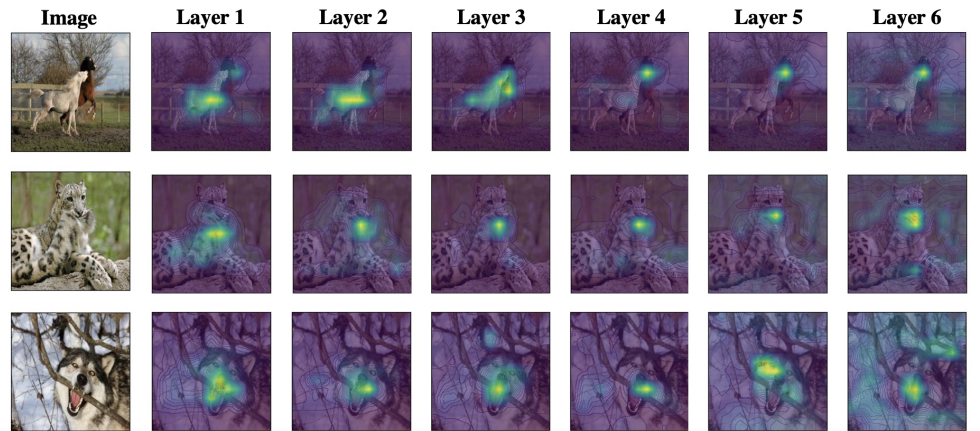


Figure D10: Verb Token Attention Map on Image Features for three *Biting* images. Each row consists of an image and attention maps from the MHSA block in each encoder layer.

Verb Token Attention Map on Image Features: GSRTTR can capture the key feature to infer the salient action. Figure D10 and Figure D11 show that GSRTTR focuses on the bit-ten part and the falling agent, respectively. The rightmost column shows that the semantic regions where the verb token focuses on are similar for the same verb. Each row shows the transition of attention maps through the encoder layers.

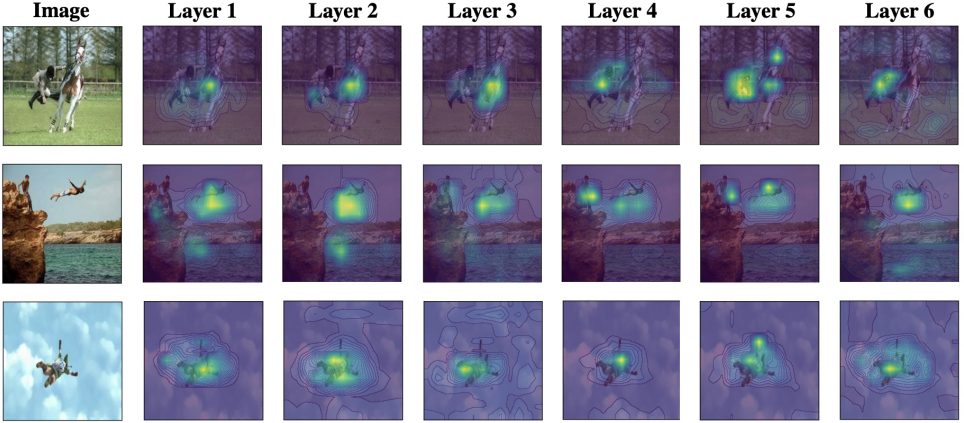


Figure D11: Verb Token Attention Map on Image Features for three *Falling* images. Each row consists of an image and attention maps from the MHSA block in each encoder layer.

E Applications

We present practical applications of Grounded Situation Recognition.



Figure E12: Grounded-Semantic-Aware Image Retrieval results on the *dev* set. The retrieval results have similar semantics and object arrangements with the query image. In this retrieval, the similarity between two images is computed by the results of verb prediction and grounded noun prediction as in [9].

There have been many studies in image retrieval by computing the similarities between the visual representations of images. But, they do not work well for getting the retrieval results which have similar situations with respect to semantics or object arrangements. Grounded-Semantic-Aware Image Retrieval enables image retrieval in the aspects of main activity and key objects with their arrangements, as shown in Figure E12. This retrieval uses the results of verb prediction and grounded noun prediction instead of visual representations. The predictions of main activity (*verb*) and entities (*nouns*) enable image retrieval for similar semantics, and the predictions of entity locations enable image retrieval for similar object arrangements. In this retrieval, we compute the GrSitSim(I, J) [2] as similarity score function between image I and J . For an image I , we compute the top-5 verb predictions $\hat{v}_1^I, \dots, \hat{v}_5^I$. For each verb prediction \hat{v}_i^I , we predict nouns $\hat{n}_{i,1}^I, \dots, \hat{n}_{i,|\mathcal{R}_{\hat{v}_i^I}|}^I$ and bounding boxes $\hat{\mathbf{b}}_{i,1}^I, \dots, \hat{\mathbf{b}}_{i,|\mathcal{R}_{\hat{v}_i^I}|}^I$. Note that we ignore the predicted bounding box if its existence probability is less than 0.5. We calculate the similarity between two images I and J as follows:

$$\text{GrSitSim}(I, J) = \max \left\{ \frac{\mathbb{1}_{[\hat{v}_i^I = \hat{v}_j^J]}}{2 \cdot i \cdot j \cdot |\mathcal{R}_{\hat{v}_i^I}|} \sum_{k=1}^{|\mathcal{R}_{\hat{v}_i^I}|} \mathbb{1}_{[\hat{n}_{i,k}^I = \hat{n}_{j,k}^J]} \cdot \left(1 + \text{IoU}(\hat{\mathbf{b}}_{i,k}^I, \hat{\mathbf{b}}_{j,k}^J) \right) \mid 1 \leq i, j \leq 5 \right\}. \quad (\text{E.9})$$

GrSitSim(I, J) is computed by the results of verb prediction and grounded noun prediction for image I and J . The similarity is not zero when at least one verb is shared in the top-5 verb predictions for image I and J . The similarity is maximized when the top-1 verb predictions and noun predictions of two images are same, and the sizes and locations of predicted bounding boxes are same. For this reason, we can get the retrieval result which has similar semantics and object arrangements in Grounded-Semantic-Aware Image Retrieval. Thus, we can apply this image retrieval to the applications where semantics and object arrangements are important, *e.g.*, search engine using semantics and object arrangements of images.

Grounded Situation Recognition models produce complete predictions with respect to the semantic roles corresponding to a verb. Thus, the models can answer the following questions more strictly, “What is the main activity” (*verb*), “Who is participating in the main activity” (role *Agent*), “What does the actor use in the main activity” (role *Tool*), “Where is the actor in the image” (entity location of role *Agent*), etc. For this reason, the models are useful for predetermined questions on situations. Taking advantages of these properties, we can apply the models for industry such as unmanned surveillance system or service robot.

References

- [1] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feed-forward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256, 2010.
- [2] Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. Grounded Situation Recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 314–332, 2020.

-
- [3] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On Layer Normalization in the Transformer Architecture. In *International Conference on Machine Learning (ICML)*, pages 10524–10533. PMLR, 2020.