Received: 27 February 2019 | Revised: 3 June 2019 | Accepted: 29 July 2019

DOI: 10.4218/etrij.2019-0093

ORIGINAL ARTICLE

ETRI Journal WILEY

Improving visual relationship detection using linguistic and spatial cues

Jaewon Jung[1] | **Jongyoul Park**[2]

[1]Artificial Intelligence Laboratory, University of Science and Technology, ETRI SCHOOL, Daejeon, Rep. of Korea

[2]Artificial Intelligence Laboratory, Electronics and Telecommunications Research Institute, Daejeon, Rep. of Korea

Correspondence
Jongyoul Park, Artificial Intelligence Laboratory, Electronics and Telecommunications Research Institute, Daejeon, Rep. of Korea.
Email: jongyoul@etri.re.kr

[Correction added on 7th January 2020, after first online publication: The corresponding author was updated to Jongyoul Park and email address for corresponding author was corrected]

Funding information
This research was supported in part by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government (19ZS1100, Core Technology Research for Self-Improving Artificial Intelligence System), and in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. B0101-15-0266, Development of High Performance Visual BigData Discovery Platform for Large-Scale Realtime Data Analysis).

Detecting visual relationships in an image is important in an image understanding task. It enables higher image understanding tasks, that is, predicting the next scene and understanding what occurs in an image. A visual relationship comprises of a subject, a predicate, and an object, and is related to visual, language, and spatial cues. The predicate explains the relationship between the subject and object and can be categorized into different categories such as prepositions and verbs. A large visual gap exists although the visual relationship is included in the same predicate. This study improves upon a previous study (that uses language cues using two losses) and a spatial cue (that only includes individual information) by adding relative information on the subject and object of the extant study. The architectural limitation is demonstrated and is overcome to detect all zero-shot visual relationships. A new problem is discovered, and an explanation of how it decreases performance is provided. The experiment is conducted on the VRD and VG datasets and a significant improvement over previous results is obtained.

KEYWORDS
deep learning, image retrieval, image understanding, predicate, visual relationship

1 | INTRODUCTION

A meaningful task in computer vision field involves detecting visual relationships in an image. Figure 1 shows the examples of the visual relationships. The task is based on outperformed object classification [1,2] and object detectors [3–6] and acts as a bridge to higher image understanding tasks such as image captioning [7,8], scene graph [9–11], and VQA [12]. The recognition of visual relationships allows a computer to understand what occurs in an image. The aforementioned ability is more advanced than the ability to understand an image by only using an object list. The achievement of excellent performance in deep learning, object classification, and detection performance is sufficient for real data, and thus

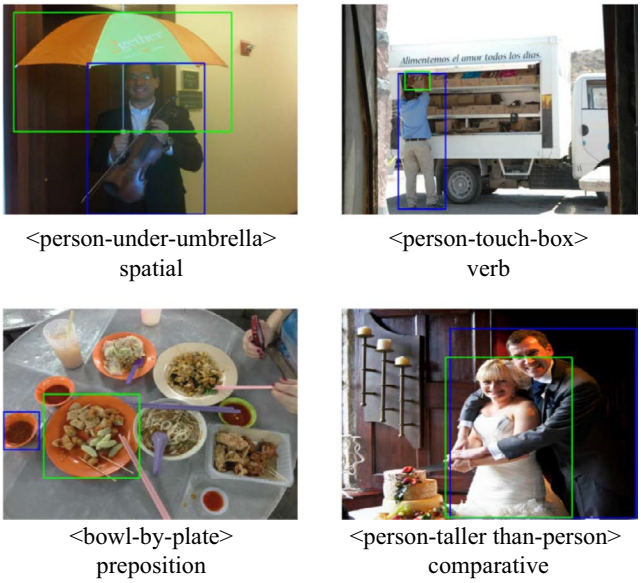


FIGURE 1 An example of visual relationships

researchers can investigate the next level of image understanding tasks.

Higher image understanding tasks focus on interpreting the structure of an image, relationship between objects, and what happens in an image among others. Image captioning is a task wherein the input corresponds to an image, and the output corresponds to a summary of the image. The task adopts a CNN and LSTM [13]. The CNN encodes the image as features, and the LSTM decodes the features to obtain a natural language summary for the image. A scene graph task transforms an image to a scene graph, which is similar to a graph in computer science. A vertex is considered an object and attribute of each object is a sub-vertex. An edge corresponds to a relationship between two objects. Specifically, VQA corresponds to a question and answer problem although this task deals with the visual aspect of an image. "How many children in an image?" is a visual question, and "4" or "4 children" are possible answers.

Visual relationship detection can only detect the relationship between two objects in an image and corresponds to an intermediate-level task as opposed to a higher-level task. A significant improvement on this task can support higher-level tasks because a visual relationship denotes the basis of a structure of an image. The visual relationship includes three components, namely the subject, predicate, and object. The subject and object play the same role as that of a sentence in natural language. However, the predicate corresponds to a type of regular verb but its scope is wider. It is considered as a predicate if it can describe a relationship between two objects.

A visual relationship can be detected via two methods. The first method considers each visual relationship as a class. {person,use,telephone}, {person,use,cellphone} correspond to different classes. Visual phrase [14] follows the first method, and the solution space is significant and given by the number of objects squared times the number of predicates. The other approach involves a significantly smaller solution space. The solution space denotes the sum of the number of objects and predicates, and two objects and predicate are separately detected. Two types of architectures follow this process. The first architecture employs an object detector and constructs a predicate classifier. The object detector determines all objects in an image and object pairs are generated. The predicate classifier classifies the predicate via visual, semantic, and spatial cues. The second architecture adopts RPN [5] to generate feature maps. All components are independently detected in the convolution feature level. A few techniques are applied on the feature level such as region of interest (ROI) pooling and bilinear interpolation.

There are two significant hurdles with respect to the aforementioned tasks. The visual gap is high in two ways. Each visual relationship contains various appearances. {person, ride, motorcycle}'s appearance is based on person, motorcycle, pose, among others. The second, classifying a predicate based on two objects. The predicate can include many subjects and objects. "person" and "monkey" can correspond to subjects for the predicates "eat," and "banana" can correspond to an object of the predicate "eat." Therefore, the visual gap is significant even for the same predicate. The two datasets, namely VRD and VG [15,16], that are commonly used for this type of research exhibit a long-tail distribution. Most visual relationships involve a person because a person is common in many images. The spatial predicate corresponds to a significant portion of the dataset. This is because the spatial property is present in every image and collecting visual relationship is expensive and labor-intensive for the other category's predicate. Most visual relationships are not present in the datasets. Therefore, the ability to detect unseen (zero-shot) performance is necessary.

A visual relationship is linked to language, visual, and spatial cues. The study looks into a language cue wherein certain predicates are reasonable only for certain subjects and objects. For example, "monkey" and "banana," "talk," "wear" are not suitable but "eat" is acceptable. A spatial cue between two objects can be found. Typically, "banana" is closer to the hand or mouth of "monkey." A previous study [15] used the language cue termed as language prior, which is obtained by two losses termed as $\mathcal{L}$ and K in [15]. The $\mathcal{L}$ loss provides a likelihood based on the number of visual relationships in a training dataset. The K loss requires a language module to semantically understand close visual relationships using two word vectors [17] for the subject and object. Another study [18] proposes a spatial vector to recognize the spatial property of a visual relationship. Based on the aforementioned studies, the proposed model improves on performance without using the $\mathcal{L}$ and K losses. However, the objectives are simply obtained, and a more sophisticated spatial vector is proposed. The structural problem of the two losses in [15] is determined and solved by spatial cues to detect all types of zero-shot visual

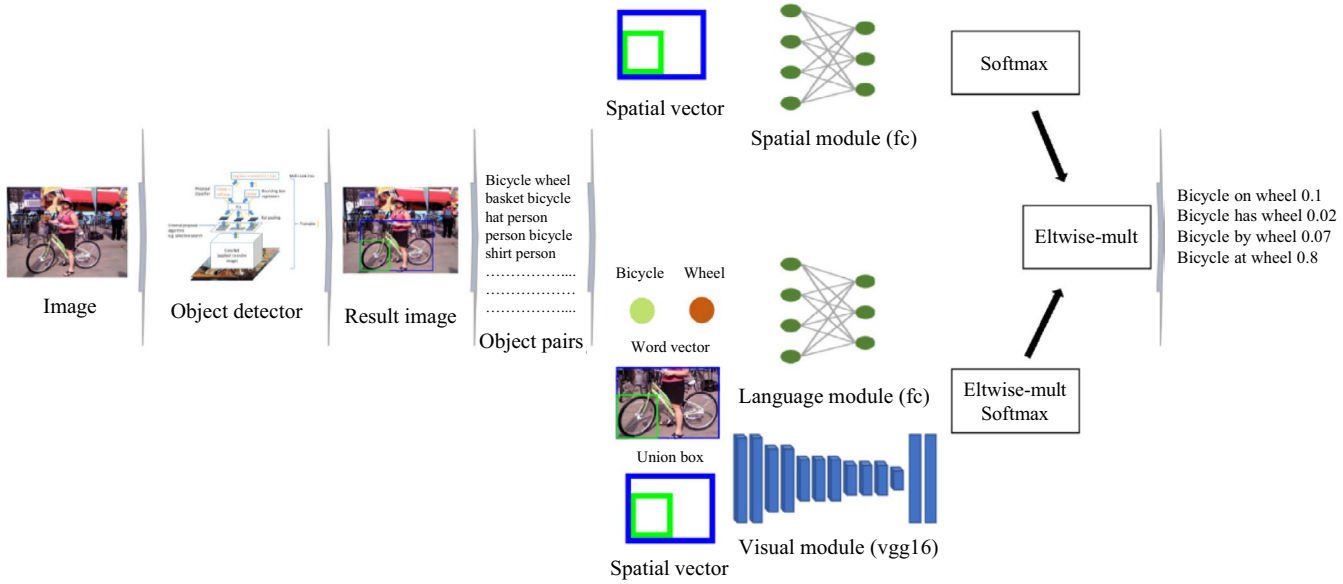


FIGURE 2 Overview of a visual relationship detection model

relationships. The third problem in the field is termed as class-overlapping, which is revealed and explained in detail herein. The study uses the second method and first type of architecture. Faster-RCNN [5] corresponds to an object detector, and VGG16 [1] is used as a predicate classifier. Figure 2 shows an overview of the study.

All experiments are performed on the visual relationship detection (VRD) and visual genome (VG) datasets [15,16]. The study is based on the aforementioned two studies and another study [19] that used the region module to [15].

2 | RELATED WORK

Object classification is a fundamental task in the field of computer vision. In deep learning, researchers stack several convolution and fully connected layers and use them to classify objects in an image while maintaining balance between the number of filters and feature maps. VGG16 [1] is designed via this philosophy and involves many parameters from the visual geometry group at Oxford. RestNet [2] includes fewer parameters, more layers, and performs significantly better than VGG16 by residual learning. A shortcut wherein an output of certain layer is added to a significantly deeper layer in a network is employed to prevent a gradient from disappearing during training in a deeper network, and the network learns residual terms. The VGG16 works as a predicate classifier in the study. The networks are termed as backbone networks and form the base of object detection.

Object detection corresponds to an advanced task and is based on an object classification task. An object detector classifies and localizes objects in an image. Faster-RCNN [5] denotes the representative one in this task. The backbone network extracts the object's feature in an image. Region proposal network (RPN) offers regions that exhibit a high probability of existing objects. ROI pooling is applied on the selected region to obtain meaningful features. A classifier and a bounding box coordinator detect objects. This is used as an object detection module.

Visual phrase [14] adopts the first method and uses three detectors for the subject, phrase, and object. The results from each detector are delivered to a decoding algorithm. The phrase detector denotes the main and is supported by the other two detectors. The visual phrase dataset is built although it exhibits three thousand fewer images for both training and test.

Chao and others [20] focused on human-object interaction, which is a type of visual relationship. The subject always corresponds to a human, and the predicate can be intransitive. In the study, they use three streams for the human, object, and pairwise and present the HICO-DET dataset. Binary masks are fed to the CNN to understand the relationship between a human and an object. The result from the three results undergo element-wise summation to detect human-object interaction.

Gkiozari and others [21] extend Faster-RCNN [5] to detect interaction between a human and an object. A human-centric and interaction branch are coupled on the Faster-RCNN. The image-centric training method and cascade inference wherein the time complex is linear are invented. The proposed model conducts experiments on HICO-DET [20] and V-COCO [22], and the highest score is obtained when compared with other studies.

Bo and others [23] proposed DR-net that uses a revised conditional random field (CRF) in a deep neural network. A predicate is considered as a feature combined with features from a spatial module and features from a backbone network

for a subject and an object. A subject and an object are represented by combining a feature from an object detector, an embedded feature, and the predicate feature.

Li and others [24] used RPN to obtain candidate regions for the subject, predicate, object on the feature level, and a visual relationship is detected. The study demonstrates that training RPN to observe visual relationships improves object detection performance. Triplet NMS reduces meaningless pair candidates and PMPS via convolution or fully connected layers shares features from a subject and an object to a predicate and vice versa.

Zhang and others [25] used RPN to detect visual relationships on the convolution feature level and embedded visual relationship in the relation space to avoid visual variants by setting an equation {"subject + predicate = object"}. As opposed to ROI pooling, bilinear interpolation is applied on selected candidates to produce a visual feature. Three components, namely a class probability vector, a location vector that encodes differences between a subject and an object, and the visual feature, are used to infer a predicate based on the subject and object.

Plummer and others [26] attempted to understand an image with a sentence in the natural language, which is termed as a caption. Single phrase cues (SPCs) and phrase-pair cues (PPCs) are proposed, a language parser is employed to divide the caption, and an object detector is used. By using several components, a higher image understanding task is realized. In the study, it is possible to detect a visual relationship by using part of some components. The performance is improved although ten predicates are selected for only one object pair.

Liang and others [27] developed an object detector, namely an action graph, which was similar to a scene graph, and a deep reinforce network. The object detector detected objects in an image, and the action graph was constructed by the objects and transformed into attribute, predicate, and object states. The states corresponded to the inputs of the deep reinforce network. The network adopted RPN and used features of the image, subject, object, and phrase. The states were used to reinforce features of the subject, predicate, and object prior to the final decision.

Shang and others [28] developed a new method termed as video visual relationship detection and proposed a new video dataset. The original objects and predicates were re-arranged for the video task. The two steps involved in the task included relation feature extraction and relation modeling. The relation feature extraction step was based on combining class vectors, hand features, and relativity features. The relation modeling step established a trajectory for the given subject and object during video segmentation. A metric and task in experiment was changed for the video task. The proposed model obtained the highest score in the experiments.

Lu and others [15] proposed a model operated by visual and language modules. The study obtained a language prior via the language module that was used for classifying a predicate by $\mathcal{L}$ and K losses. Specifically, $\mathcal{L}$ loss denotes likelihood loss functions and provides the likelihood based on the number of visual relationships by ranking loss. The K loss allowed the language module to understand the distance between visual relationships based on word vectors by reducing the variance in sampled visual relationships. The visual module obtained predicate classification results via a linear operation using a CNN feature. Element-wise multiplication was applied to the results obtained from the two modules, and the final visual relationship was detected via the C loss, which corresponded to a type of ranking loss.

Yu and others [18] included linguistic knowledge distillation from internal and external resources to detect visual relationships. The proposed spatial vector contained only the location and size of each object. A union box, word vector, and spatial vector were exploited on their model and a teacher-student network (jointly) detected visual relationships. The method improved the performance of general and zero-shot detection performance.

Zhu and others [19] added a region module and applied a sigmoid function on all modules. The other parameters are identical to those in [15]. The results from three modules are merged into one via element-wise multiplication, and the C loss was extended with the region module. The performance exhibited improvements over the original [15] although the zero-shot performance decreased.

3 | NEW PROBLEM: CLASS-OVERLAPPING

Class overlapping is a problem involved in detecting visual relationships. Figure 3 shows several examples of the same. Each predicate boundary is invaded by several predicates. This is mainly because some predicates exhibit almost identical meanings, and the annotations of each visual relationship are not distinguishable irrespective of categories.

FIGURE 3 Class-overlapping

Furthermore, a predicate is determined by the position of the objects. Thus, {person, next to, bus} can be changed to {person, (in front of or behind), bus} based on the position and rotation of any two objects, and the annotation of the visual relationship varies based on the view angle of an image.

4 | PROBLEM FROM THE BASELINES

In [15], $\mathcal{L}$ loss assigns an appropriate likelihood based on the frequency of a training dataset. This implies that zero likelihood is assigned to the visual relationship, which never occurs in the training dataset. However, the visual relationship consists of the same subject and object in a manner similar to visual relationships that occurred in the training dataset. This implies that a predicate establishes a zero-shot visual relationship. The K loss imposes the language module produces a close likelihood to semantically similar visual relationships in which two subjects or/and objects word vector's distance is close in the two visual relationships and works based on word vector embedding space. If the subject or object or both word vectors in a visual relationship do not include any other close word vector in the word vector space, then the model does not assign a likelihood to the visual relationship. Thus, the model is unable to detect the zero-shot visual relationships via the two losses.

The previous spatial vector in [18] contains only a normalized location and size for a subject and object. The vector does not include relative information on the subject and object and includes more-precise coordinates and size for the normalized bounding box. It can only differentiate between two locations in a model. However, it is insufficient in terms of making the model consider spatial property based on the overlap ratio between two bounding boxes and each property of the predicates.

5 | PROPOSED MODEL

5.1 | Overall pipeline

In Figure 2, an image is provided as input. Faster R-CNN [5] based on VGG16 [1] detects object(s) and pairs of objects termed as candidates are generated from the detected objects. Each pair of objects is fed to the predicate classifier and the spatial module with a pair of word vectors, a spatial vector, and a union box that includes the subject and object. The language and visual module produce results and multiplies it element-wise, and the softmax function is applied to obtain the final result. The result can be calibrated via the spatial module in several ways.

5.2 | Visual module

VGG16 operates as a visual module. This network is trained to classify a predicate based on two objects. The input corresponds to a union box that includes a subject and an object in a visual relationship, and the output corresponds to a predicate classification result as shown in Figure 4.

5.3 | Language module

$$f(R,W,b) = W \times [\text{word vector(subject), word vector(object)}] + b \text{ where } R \text{ is a triplet subject, predicate, object.} \quad (1)$$

Equation (1) describes a language module that corresponds to a fully connected layer, and the input corresponds to two word vectors and the output corresponds to the likelihood for all predicates. W denotes the parameter and the dimension of W denotes the number of predicates × the sum of two word vectors' dimension, and b denotes the parameter and the dimension of b denotes the number of predicates dimension.

5.3.1 | $\mathcal{L}$ loss

$$L(W,b) = \Sigma \max\{f(R',W,b) - f(R,W,b) + m, 0\}, \forall R, R'. \quad (2)$$

Equation (2) expresses the $\mathcal{L}$ loss in [15]. W and b correspond to parameters of the language module, m corresponds to a margin, and R and R' correspond to different visual relationships, which include a subject and an object. If R is more frequent than R' in a training dataset, then R exhibits a higher likelihood than R'. Therefore, the language module produces the likelihood based on the frequency given the minimization of the loss function. {person, eat, pizza} exhibits higher likelihood than {dog, ride, surfboard}. The latter is rare in images.

From Figure 5, all subjects are "person" and all objects are "pizza." {person, eat, pizza} denotes the most visual relationship, {person, next to, pizza} denotes the least visual relationship. {person, hold, pizza} exhibits a moderate visual relationship. Training the language module with a softmax loss realizes the objective of the $\mathcal{L}$ loss. The language module observes the

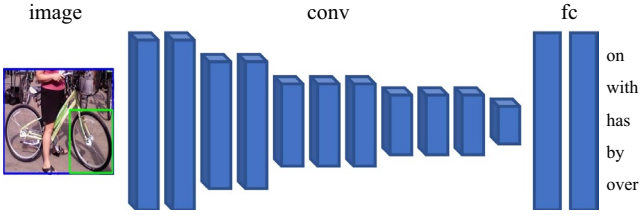


FIGURE 4 Visual module

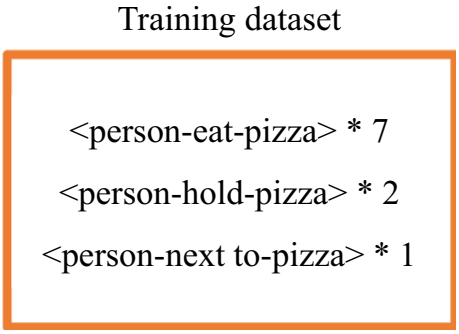


FIGURE 5 Example: Training dataset

number of predicates based on the frequency. For example, "eat," "hold," and "next to" during training. Irrespective of the training batch order, the degree of each likelihood from the model depends on the number of data in the training dataset and the sum of all likelihood is always one due to the softmax loss function.

5.3.2 | *K* loss

$$K(W,b) = \mathrm{var}\left(\frac{\{f(R',W,b) - f(R,W,b)\}^2}{d(R,R')}, \forall R, R'\right). \quad (3)$$

Equation (3) describes the *K* loss in [15]. var denotes the variance, and $d(R, R')$ denotes the sum of the cosine distance between the two visual relationships. The language module assigns similar likelihoods to close visual relationships and different likelihoods to far visual relationships by reducing the variance. Specifically, 50k visual relationships were sampled for the loss function. The likelihood of {person, drive, car} and {person, drive, truck} correspond to close visual relationships.

The language module can assign similar likelihoods to close visual relationships without *K* loss. The *K* loss is achieved when the two word vectors for the subject or/and object are close. For example, the language module produces nearly the same likelihood for {person, eat, pasta} and {person, eat, pizza} if "pasta" and "pizza" are close in the word vector space. This indicates that the model considers "pasta" and "pizza" as nearly the same. Figure 6 shows the aforementioned result.

5.4 | Spatial vector and module

5.4.1 | Spatial vector

$$\left[\mathrm{IOU}, x, y, \frac{S_{\mathrm{subject}}}{S_{\mathrm{image}}}, \frac{S_{\mathrm{object}}}{S_{\mathrm{image}}}, \mathrm{cflag}_{\mathrm{subject}}, \mathrm{cflag}_{\mathrm{object}}\right]. \quad (4)$$

The spatial vector in [18] contains concatenated individual location and size of a subject and an object in an image.

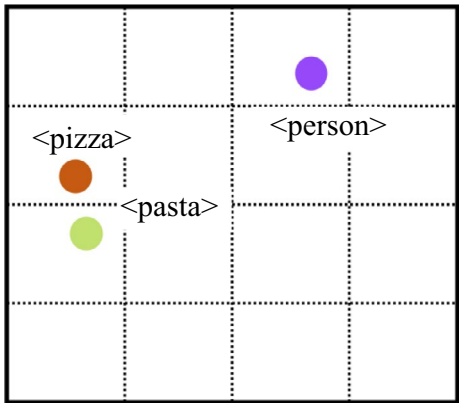


FIGURE 6 Word vector space

However, related information such as IOU, cflag, and normalized object location (x, y) based on the subject in an image are included in the proposed model as shown in (4). The cflag corresponds to a flag, which is one when an object includes another object, and otherwise it corresponds to zero. S_{subject} and S_{object} denote the size of each bounding box. S_{image} denotes the size of an image.

5.4.2 | Spatial module

The proposed spatial module is similar to the language module to obtain a spatial prior. The module with a spatial vector as the input and a predicate classification result as the output is individually trained from the visual and language modules. The classification result is used to calibrate the model result. When a model observes some type of zero-shot visual relationship (as mentioned in the section entitled "Problem from the Baselines"), the likelihood is termed as an unreliable result because the maximum value among all likelihood is low, and the likelihood of most predicates is nearly identical. The aforementioned types of visual relationships can be determined by using a spatial module as shown in Figure 7. The

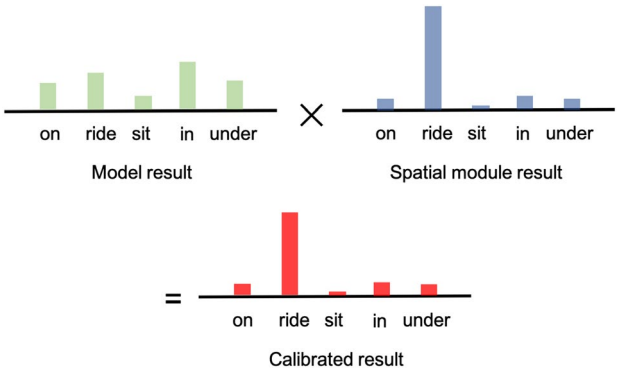


FIGURE 7 Spatial module calibration

spatial module only considers the spatial property for a given subject and object. The result from the module supports the result from the model via calibration.

5.5 | Model explanation

A model can consist of each module or both modules. This implies that a language or visual module can correspond to a model and language and visual modules can collectively constitute a model. All models produce predicate classification result. The spatial vector can be concatenated to the language module. Subsequently, the language model (L), and language and spatial model (LSV) are available and termed as the language-based model. The spatial and/or word vector can be concatenated to a visual module. The visual model (U), and visual and word vector model (UW), and spatial and visual model (SVU), and spatial, visual and word vector model (SVUW) are available and termed as visual-based model. The two based models can work jointly, that is, LSV + SVUW. The results of each model are element-wise multiplied, and softmax is applied to obtain the final detection result as shown in Figure 2. With respect to the language based models, word vectors or concatenated word and spatial vectors are fed to a fully connected layer. With respect to the visual based models, visual feature from VGG, spatial and/or word vectors are fed to a fully connected layer.

6 | EXPERIMENT

A metric, namely top recall(R@n), was adopted as in [15], and "n" corresponds to 50, 100. The method that considers 1, 10, and 70 predicates as the detected predicates was employed based on [18]. "z" denotes the zero-shot performance. Three experiments were conducted on the VRD and VG datasets as follows [15,18]: predicate prediction, phrase detection, and relationship detection. The predicate prediction input corresponded to an image with two localized objects with classes, and the output corresponded to the proper predicate. The phrase detection input corresponded to an image, and output corresponded to a visual relationship with a bounding box that includes subject, predicate, and object. The relationship detection input corresponded to an image, and the output corresponded to a visual relationship with two boxes for the subject and object. The latter two tests require object detection results. Therefore, the object detection result from [15] was used and test results are compared with the test results obtained by [15,19], and the object detection module was trained and test results are compared with the test results obtained by [18]. Additional experiments were conducted to verify the effectiveness of the proposed spatial vector and module in terms of predicate prediction.

6.1 | Predicate prediction

As shown in Table 1, the L + U model exhibits better performance than that of the VRD in which the language prior is obtained by the $\mathcal{L}$ and K losses. This indicates that the proposed simple method is better than using two losses to obtain the language prior. The SVUW model outperforms the model with linguistic knowledge distillation and the model in [18] that uses three components. The result indicates that expensive distillation is not necessary and the proposed spatial vector is more suitable for the task than the model in [18]. The LSV + SVUW exhibits better performance than all other models except the performance improvement for R@50, $k = 1$, and zero-shot is larger than VRD and the model that uses linguistic knowledge distillation in [18]. This is because the objective of the K loss is achieved as mentioned in [15]. The highest scores for the general and zero-shot are not obtained from the single model in [18]. However, the LSV + SVUW model obtains the highest score for general and zero-shot performance.

On the VG dataset, the proposed model significantly outperforms the model in [18] without the language prior as shown in Table 2. The VG dataset used in [18] is not open to the public. A task that extracts the same category visual relationship in the VRD dataset was applied on the latest VG dataset.

6.2 | Phrase detection

As shown in Table 3, the proposed model outperforms all baselines. Specifically, the zero-shot score is twice that of the baselines. An important goal involves determining the merged box. This task is slightly more difficult than relationship detection.

6.3 | Relationship Detection

As shown in Table 4, the proposed model achieves the highest score for relationship detection. The object detection module determines the wrong bounding boxes termed as false positives. This degrades the performance of the task.

6.4 | Spatial vector analysis

Table 5 details whether the new spatial vector is better than the spatial vector in [18]. Thus, the previous spatial vector was adopted in the model in the present study. The performance decreased for the SVUW and LSV + SVUW models for all metrics. Additionally, the zero-shot performance decreased by approximately 3 and 6 percent points.

TABLE 1 Predicate prediction on the VRD dataset

	R@50 $k = 1$	R@100 $k = 1$	R@50 $k = 70$	R@100 $k = 70$	R@50 $k = 1, z$	R@100 $k = 1, z$	R@50 $k = 70, z$	R@100 $k = 70, z$
VRD [15]	47.87	47.87	N/A	N/A	8.45	8.45	N/A	N/A
U + W + SF [18]	41.33	41.33	72.29	84.89	14.13	14.13	48.13	69.41
U + W + L:S [18]	42.98	42.98	71.83	84.94	13.89	13.89	51.37	72.53
U + W + L:T [18]	52.96	52.96	83.26	88.98	7.81	7.81	32.62	40.15
U + SF + L:S [18]	41.06	41.06	71.27	84.81	14.33	14.33	48.32	69.01
U + SF + L:T [18]	51.67	51.67	83.84	87.71	8.05	8.05	32.77	41.51
U + W + SF + L:S [18]	47.50	47.50	74.98	86.97	16.98	16.98	54.20	74.65
U + W + SF + L:T [18]	54.13	54.13	82.54	89.41	8.80	8.80	32.81	41.53
U + W + SF + L:T + S [18]	**55.16**	**55.16**	85.64	94.65	N/A	N/A	N/A	N/A
R [19]	51.16	51.16	N/A	N/A	12.98	12.98	N/A	N/A
S [19]	47.33	47.33	N/A	N/A	10.35	10.35	N/A	N/A
R-S [19]	51.50	51.50	N/A	N/A	14.60	14.60	N/A	N/A
L	44.09	44.09	75.48	86.69	10.86	10.86	50.55	69.71
LSV	48.19	48.19	78.31	88.40	15.82	15.82	55.09	74.85
SVUW	48.57	48.57	78.04	88.30	16.85	16.85	55.77	74.85
L + U	49.77	49.77	79.99	88.81	14.88	14.88	54.40	72.51
LSV + UW	53.05	53.05	85.12	93.17	20.78	20.78	64.67	81.35
LSV + SVU	53.37	53.37	85.61	93.74	21.21	21.21	65.78	82.37
LSV + SVUW	**55.16**	**55.16**	**88.88**	**95.18**	**21.38**	**21.38**	**64.49**	**83.49**

In [18], "U" denotes the union box that includes two objects, "SF" denotes the spatial vector in their study, "W" denotes the word-embedding-based semantic representation, "L" denotes the linguistic knowledge distillation, "S" denotes the student network, "T" denotes the teacher network, and "S + T" denotes the combination of two networks. In the study, "L" corresponds to a language module that uses word vectors, "SV" denotes the proposed spatial vector, "U" is identical to that in [18] and is termed as a visual module, "W" corresponds to a word vector in the visual module, and "+" implies that the two modules are placed before and after "+" are used together. Bold values mean the highest score for each metric.

TABLE 2 Predicate prediction on VG dataset

	R@50 $k = 1$	R@100 $k = 1$	R@50 $k = 70$	R@100 $k = 70$	R@50 $k = 1, z$	R@100 $k = 1, z$	R@50 $k = 70, z$	R@100 $k = 70, z$
U + W + SF + L:S	49.88	49.88	88.14	91.25	11.28	11.28	72.96	88.23
U + W + SF + L:T	55.02	55.02	91.47	94.92	3.94	3.94	47.62	62.99
U + W + SF + L:T + S	55.89	55.89	92.31	95.68	N/A	N/A	N/A	N/A
SVUW	65.59	65.73	96.37	98.90	16.82	16.82	86.33	95.02
LSV + SVUW	**70.99**	**71.12**	**97.98**	**99.37**	**19.68**	**19.68**	**89.00**	**95.72**

The notations are identical to those in Table 1. Bold values mean the highest score for each metric.

6.5 | Calibration Using The Spatial Module

As shown in Table 6, spatial module improves model performance. Three calibration methods were used in the experiment, namely calibrating all model results using element-wise multiplication, replacing the original result with the result from the spatial module when the original result is unreliable, and element-wise multiplication of the original result with the result from the spatial module when the original result is unreliable. A threshold is set to determine whether or not the result is unreliable. The result is unreliable if the maximum value among all the likelihoods is lower than the threshold. "*" denotes the first calibration. "T 0.1*" and "T 0.1 S" denote the second and third method, and the threshold corresponds to 0.1. The second or third method is used if the result is unreliable.

The zero-shot performance, $k = 1$, is improved via the first method. All the results from the spatial module are unreliable, and the performance decreases when the threshold increases. With respect to the last method, determination of the threshold is important in terms of performance. Based on Table 6, thresholds corresponding to 0.2, 0.3 or 0.4 are appropriate for the model for zero-shot, $k = 1$.

TABLE 3 Phrase detection result on VRD dataset

	R@50 $k = 1$	R@100 $k = 1$	R@50 $k = 10$	R@100 $k = 10$	R@50 $k = 70$	R@100 $k = 70$	R@50 $k = 1, z$	R@100 $k = 1, z$	R@50 $k = 10, z$	R@100 $k = 10, z$	R@50 $k = 70, z$	R@100 $k = 70, z$
VIP-CNN [24]	22.78	27.91	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
VRL [29]	21.37	22.60	N/A	N/A	N/A	N/A	9.17	10.31	N/A	N/A	N/A	N/A
Linguistic Cues [26]	N/A	N/A	16.89	20.70	N/A	N/A	N/A	N/A	10.86	15.23	N/A	N/A
U + W + SF + L:S [18]	19.15	19.98	22.95	25.16	22.59	25.54	10.44	10.89	13.01	17.24	12.96	17.24
U + W + SF + L:T [18]	22.46	23.57	25.96	29.14	25.86	29.09	6.54	6.71	9.45	11.27	7.86	9.84
U + W + SF + L:T + S [18]	23.14	24.04	26.47	29.76	26.32	29.43	N/A	N/A	N/A	N/A	N/A	N/A
LSV + SVU	**32.15**	**33.00**	**41.58**	**49.45**	**41.68**	**49.89**	**12.23**	**12.66**	**22.75**	**32.59**	**23.26**	**34.21**
VRD [15]	16.17	17.03	N/A	N/A	N/A	N/A	3.36	3.75	N/A	N/A	N/A	N/A
R [19]	15.84	18.13	N/A	N/A	N/A	N/A	2.24	2.57	N/A	N/A	N/A	N/A
S [19]	15.12	17.09	N/A	N/A	N/A	N/A	2.29	2.40	N/A	N/A	N/A	N/A
R-S [19]	16.94	18.89	N/A	N/A	N/A	N/A	3.24	3.52	N/A	N/A	N/A	N/A
LSV + SVU	**17.00**	**19.03**	**18.94**	**23.01**	**18.95**	**23.06**	**7.35**	**8.12**	**9.83**	**13.08**	**9.92**	**13.43**

The notations are identical to those in Table 1. The performance that is compared with the result obtained by [18] is the first LSV + SVU, and the performance that is compared with the result obtained by [15,19] is the second LSV + SVU. Bold values mean the highest score for each metric.

TABLE 4 Relationship detection on the VRD dataset

	R@50 $k = 1$	R@100 $k = 1$	R@50 $k = 10$	R@100 $k = 10$	R@50 $k = 70$	R@100 $k = 70$	R@50 $k = 1, z$	R@100 $k = 1, z$	R@50 $k = 10, z$	R@100 $k = 10, z$	R@50 $k = 70, z$	R@100 $k = 70, z$
VIP-CNN [24]	17.32	20.01	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
VTRANS [25]	14.07	15.20	N/A	N/A	N/A	N/A	1.71	2.14	N/A	N/A	N/A	N/A
VRL [29]	18.19	20.79	N/A	N/A	N/A	N/A	7.94	8.52	N/A	N/A	N/A	N/A
Linguistic Cues [26]	N/A	N/A	15.08	18.37	N/A	N/A	N/A	N/A	9.67	13.43	N/A	N/A
U + W + SF + L:S [18]	16.57	17.69	19.92	27.98	20.12	28.94	8.89	9.14	12.31	16.15	12.02	15.89
U + W + SF + L:T [18]	18.56	20.61	21.91	29.41	21.98	31.13	6.07	6.44	7.82	9.71	8.75	10.21
U + W + SF + L:T + S [18]	19.17	21.34	22.56	29.89	22.98	31.89	N/A	N/A	N/A	N/A	N/A	N/A
LSV + SVU	**30.25**	**31.06**	**39.49**	**47.38**	**39.60**	**47.80**	**11.97**	**12.40**	**22.07**	**31.73**	**22.58**	**33.36**
VRD [15]	13.86	14.70	N/A	N/A	N/A	N/A	3.13	3.52	N/A	N/A	N/A	N/A
R [19]	13.40	15.17	N/A	N/A	N/A	N/A	1.79	2.07	N/A	N/A	N/A	N/A
S [19]	13.01	14.52	N/A	N/A	N/A	N/A	1.85	1.96	N/A	N/A	N/A	N/A
R-S [19]	14.31	15.77	N/A	N/A	N/A	N/A	2.85	3.08	N/A	N/A	N/A	N/A
LSV + SVU	**15.05**	**16.73**	**16.82**	**20.49**	**16.83**	**20.54**	**6.75**	**7.35**	**8.98**	**11.80**	**9.06**	**12.14**

The notations are identical to those in Table 1. The performance that is compared with the result obtained by [18] is the first LSV + SVU, and the performance that is compared with the result obtained by [15,19] is the second LSV + SVU. Bold values mean the highest score for each metric.

TABLE 5 Predicate prediction on VRD dataset to verify proposed spatial vector

	R@50 $k = 1$	R@100 $k = 1$	R@50 $k = 70$	R@100 $k = 70$	R@50 $k = 1, z$	R@100 $k = 1, z$	R@50 $k = 70, z$	R@100 $k = 70, z$
(SF)UW	45.58	45.58	77.10	87.98	13.60	13.60	53.63	74.16
SVUW	**48.57**	**48.57**	**78.04**	**88.30**	**16.85**	**16.85**	**55.77**	**74.85**
L(SF)+(SF)UW	50.53	50.53	81.99	91.08	15.99	15.99	56.63	76.98
LSV + SVUW	**55.16**	**55.16**	**88.88**	**95.18**	**21.38**	**21.38**	**64.49**	**83.49**

The notations are identical to those in Table 1. "SF" denotes the spatial vector in [18]. Bold values mean the highest score for each metric.

TABLE 6 Predicate prediction on the VRD dataset to verify the spatial module

	R@50 $k = 1$	R@50 $k = 70$	R@50 $k = 1, z$	R@50 $k = 70, z$
LV + SVUW	55.16	**88.88**	21.38	64.50
*	53.99	87.26	22.75	**66.38**
T 0.1 S	55.15	88.84	21.85	64.49
T 0.2 S	55.10	88.63	22.07	64.67
T 0.3 S	53.94	86.61	21.12	63.81
T 0.4 S	50.95	83.59	20.61	62.27
T 0.1 *	55.16	88.84	21.38	64.49
T 0.2 *	**55.25**	88.09	22.24	62.87
T 0.3 *	55.19	84.69	22.49	60.51
T 0.4 *	54.77	82.42	**23.09**	58.51

The notations are identical to those in Table 1. Bold values mean the highest score for each metric.

7 | DETAIL STUDY

7.1 | Phrase and relation detection

The study uses separated object detection modules, and the experiment results are based on the object detection results of the test images. The result from [15] by RCNN [3] is poor due to many false positives, missing objects, and incorrect bounding box(es). The predicate classifier is unable to correct this. Therefore, the score is significantly lower than the predicate prediction.

7.2 | How language and visual modules work

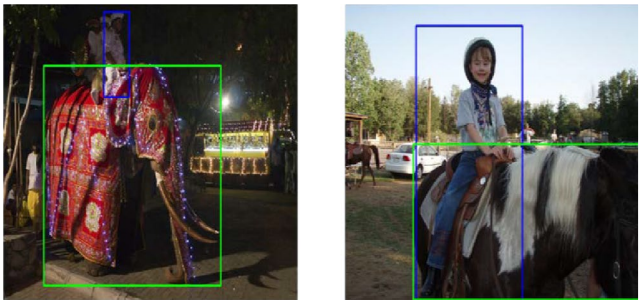
Following the VRD test dataset result, the language module determines most of the visual relationship and the visual module supports it. Therefore, several of the final results match results from the language module. This implies that the language aspect is more important than the visual on visual relationship detection. All filters in the convolution layers are black and only one of the biases in fully connected layers are used for classifying predicates in the visual module. This is because the intra-class variance affects the visual module due to the visual gap on the predicate.

7.3 | Effectiveness of the spatial vector

As shown in Table 5, the proposed spatial vector is better than the existing vector proposed by [18] in terms of predicate prediction. The zero-shot and general performance drop occur for the model that uses the existing spatial vector [18]. A high amount of relative information is useful in distinguishing between (unseen) visual relationships. The cflag aids in better understanding the property of predicate such as "on" or "in." The IOU provides the degree of overlapped area between 0 and 1 for two boxes, and the normalized location aids the IOU because it corresponds to a vector. When an unseen (zero-shot) visual relationship is observed in the model, the spatial vector is potentially similar to training data. The aforementioned parameters (ie, IOU and normalized location) can support zero-shot performance. Figure 8 shows two examples of visual relationships. If one of them corresponds to zero-shot visual relationship, then a possibility exists that it will be detected due to similarities with a person's pose, and two word vectors, namely "elephant," "horse" are close in the embedding space.

7.4 | Degradation from class-overlapping

Figure 9 details why performance gain is not significant although all components and a spatial module are applied to the model. In a predicate list, a few predicates exhibit an undistinguishable meaning. This leads to non-dominated classification (ambiguous) results. Although feature vectors are close, they are included in different classes such as "above" and "over."



<person-ride-elephant> <person-ride-horse>

FIGURE 8 Semantically close visual relationships

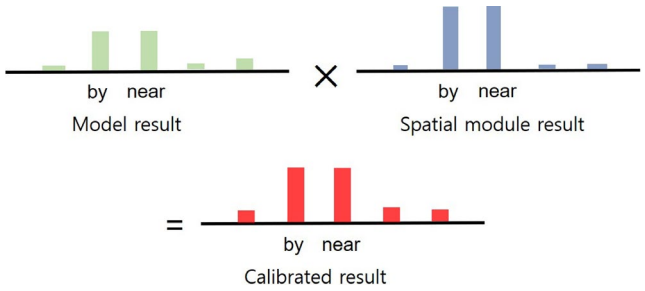


FIGURE 9 Performance degradation from class-overlapping

This classification can confuse the model. Additionally, performance does not exhibit further improvements although all resources are applied to the model as shown in Table 1.

The two results from the spatial module and model are under class-overlapping. The calibrated result also falls under the aforementioned problem. Therefore, performance improvement due to the spatial module is not significant.

8 | CONCLUSION

The results of the study exhibit significant improvements when compared with those of an extant study. However, improvement gain is effortless and cost-effective when compared with the use of linguistic knowledge distillation. The proposed spatial vector includes relative information on two objects in an image and exhibits its effectiveness in terms of general and zero-shot visual relationship detection. All zero-shot visual relationships can be detected using the spatial module via calibration and the model performance is increased. Class-overlapping is first described and its effect on the deterioration of the model's performance in detecting visual relationship is demonstrated in the study.

9 | SUBSEQUENCE WORK

A multi-visual relationship corresponds to the next research theme in this field. All categories can be simultaneously annotated between the two objects in an image. Therefore, the model can detect the visual relationship of several categories while avoiding class-overlapping. Existing ambiguities are removed between categories. Re-organizing the predicate and object list corresponds to an upcoming task. It is necessary to merge redundant meaning predicates and objects to improve result classification.

ORCID

Jaewon Jung https://orcid.org/0000-0001-8282-2873

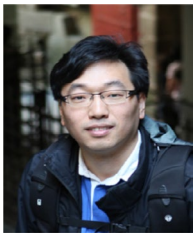
REFERENCES

1. K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, 2014, CoRR abs/1409.1556.
2. H. Kaiming et al., *Deep residual learning for image recognition*, in Proc. IEEE Conf. Comput. Vision Pattern Recogn., Las Vegas, NV, USA, June 2016, pp. 770–778.
3. G. Ross et al., *Region-based convolutional networks for accurate object detection and segmentation*, IEEE Trans. Pattern Anal. Mach. Intell. **38** (2016), no. 1, 142–158.
4. R. Girshick, *Fast R-CNN*, in Proc. IEEE Int. Conf. Comput. Vision, Santiago, Chile, Dec. 2015, pp. 1440–1448.
5. S. Ren et al., *Faster R-CNN: Towards real-time object detection with region proposal networks*, Adv. Neural Inf. Process. Syst., Montreal, Canada, Dec. 2015, pp. 91–99.
6. K. He et al., *Mask R-CNN*, in IEEE Int. Conf. Comput. Vision (ICCV), Venice, Italy, Oct. 2017, pp. 2980–2988.
7. Z. Ren et al., *Deep reinforcement learning-based image captioning with embedding reward*, in Proc. IEEE Conf. Comput. Vision Pattern Recogn. (CVPR), Honolulu, HI, USA, July 2017, pp. 1151–1159.
8. S. Li et al., *Person search with natural language description*, in Proc. IEEE Conf. Comput. Vision Pattern Recogn. (CVPR), Honolulu, HI, USA, July 2017, pp. 5187–5196.
9. J. Johnson et al., *Image retrieval using scene graphs*, in Proc. IEEE Conf. Comput. Vision Pattern Recogn., Boston, MA, USA, June 2015, pp. 3668–3678.
10. Y. Li et al., *Scene graph generation from objects, phrases and region captions*, in Proc. IEEE Conf. Comput. Vision Pattern Recogn., Venice, Italy, Oct. 2017, pp. 1261–1270.
11. X. Danfei et al., *Scene graph generation by iterative message passing*, in Proc. IEEE Conf. Comput. Vision Pattern Recogn., Honolulu, HI, USA, July 2017, pp. 3097–3106.
12. Y. Goyal et al., *Making the V in VQA matter: Elevating the role of image understanding in visual question answering*, in IEEE Conf. Comput. Vision Pattern Recogn. (CVPR), Honolulu, HI, USA, July 2017, pp. 6325–6334.
13. S. Hochreiter and J. Schmidhuber, *Long short-term memory*, Neural Comput **9** (1997), no. 8, 1735–1780.
14. M. A. Sadeghi and A. Farhadi, *Recognition using visual phrases*, in Proc. IEEE Conf. Comput. Vision Pattern Recogn. (CVPR), Providence, RI, USA, June 2011, pp. 1745–1752.
15. L. Cewu et al., *Visual relationship detection with language priors*, in Proc. Eur. Conf. Comput. Vision, Amsterdam, Netherlands, Oct. 2016, pp. 852–869.
16. R. Krishna et al., *Visual genome: Connecting language and vision using crowdsourced dense image annotations*, Int. J. Comput. Vision **123** (2017), no. 1, 32–73.
17. T. Mikolov et al., *Efficient estimation of word representations in vector space*, in Proc. Int. Conf. Learn. Representations (ICLR) Workshop, Scottsdale, AZ, USA, 2013, pp. 1–12.
18. Y. Ruichi et al., *Visual relationship detection with internal and external linguistic knowledge distillation*, in Proc. IEEE Int. Conf. Comput. Vision (ICCV), Venice, Italy, Oct. 2017, pp. 1068–1076.
19. Y. Zhu, S. Jiang, and X. Li, *Visual relationship detection with object spatial distribution*, in Proc. IEEE Int. Conf. Multimedia Expo (ICME), Venice, Italy, Oct. 2017, pp. 379–384.
20. Y. W. Chao et al., *Learning to detect human-object interactions*, in Proc. IEEE Winter Conf. Applicat. Comput. Vision, Lake Tahoe, NV, USA, Mar. 2018, pp. 381–389.
21. G. Gkioxari et al., *Detecting and recognizing human-object interactions*, in Proc. Conf. Vision Pattern Recong., Salt Lake City, UT, USA, June 2018, pp. 8359–8367.
22. T. Y. Lin et al., *Microsoft coco: Common objects in context*, in Proc. Comput. Vision – ECCV, Zurich, Switzerland, Sept. 2014, pp. 740–755.

23. B. Dai, Y. Zhang, and D. Lin, *Detecting visual relationships with deep relational networks*, in Proc. IEEE Conf. Comput. Vision Pattern Recogn. (CVPR), Honolulu, HI, USA, July 2017, pp. 3298–3308.
24. Y. Li et al., *ViP-CNN: Visual phrase guided convolutional neural network*, in Proc. IEEE Conf. Comput. Vision Pattern Recogn. (CVPR), Honolulu, HI, USA, July 2017, pp. 7244–7253.
25. H. Zhang et al., *Visual translation embedding network for visual relation detection*, in Proc. IEEE Conf. Comput. Vision Pattern Recogn. (CVPR), Honolulu, HI, USA, July 2017, pp. 3107–3115.
26. B. A. Plummer et al., *Phrase localization and visual relationship detection with comprehensive image-language cues*, in Proc. IEEE Conf. Comput. Vision Pattern Recogn., Venice, Italy, Oct. 2017, pp. 1928–1937.
27. X. Liang, L. Lee, and E. P. Xing, *Deep variation-structured reinforcement learning for visual relationship and attribute detection*, 2017, CoRR abs/1703.03054.
28. X. Shang et al., *Video visual relation detection*, in Proc. ACM Int. Conf. Multimedia, Mountain View, CA, USA, Oct. 2017, pp. 1300–1308.
29. X. Liang, L. Lee, and E. P. Xing, *Deep variation-structured reinforcement learning for visual relationship and attribute detection*, in Proc. IEEE Conf. Comput. Vision Pattern Recogn. (CVPR), Honolulu, HI, USA, July 2017, pp. 4408–4417.

AUTHOR BIOGRAPHIES

Jaewon Jung received his BS in computer science and engineering from the Soongsil University, Seoul, Rep. of Korea in 2016, and his MS in the Department of Computer and Software engineering at the University of Science and Technology Electronics and Telecommunications Research Institute School, Daejeon, Rep. of Korea in 2019. His research interests include generative adversarial networks, image and video understanding, object recognition, and deep learning.



Jongyoul Park received his BS in computer engineering from the Chunanam National University, Daejeon, Rep. of Korea, in 1996, and his MS and PhD in Information and Communication Engineering from the Gwangju Institute of Science and Technology, Rep. of Korea, in 1999 and 2004, respectively. From 2001 to 2002, he was a visiting researcher at the School of Computing, University of Utah, UT, USTA. Since 2004, he is working with the Visual Intelligence Research Section in the Artificial Intelligence Laboratory at the Electronics and Telecommunications Research Institute, Daejeon, Rep. of Korea. His research interests include large scale machine learning, object recognition, and behavior understanding.