Neurocomputing 404 (2020) 165–172

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Visually grounded paraphrase identification via gating and phrase localization

Mayu Otani[a,*], Chenhui Chu[b], Yuta Nakashima[b]

[a] CyberAgent, Inc., Japan
[b] Institute for Datability Science, Osaka University, Japan

ARTICLE INFO

Article history:
Received 30 August 2019
Revised 31 January 2020
Accepted 11 April 2020
Available online 12 May 2020

Communicated by Dr. T. Mu

Keywords:
Visual grounded paraphrases
Gating
Phrase localization
Vision and language

ABSTRACT

Visually grounded paraphrases (VGPs) describe the same visual concept but in different wording. Previous studies have developed models to identify VGPs from language and visual features. In these existing methods, language and visual features are simply fused. However, our detailed analysis indicates that VGPs with different lexical similarities require different weights on language and visual features to maximize identification performance. This motivates us to propose a gated neural network model to adaptively control the weights. In addition, because VGP identification is closely related to phrase localization, we also propose a way to explicitly incorporate phrase-object correspondences. From our evaluation in detail, we confirmed our model outperforms the state-of-the-art model.

1. Introduction

Paraphrases are textual representations that describe the same semantics in different ways [4]. Paraphrase identification can facilitate flexible matching of text, which has a wide range of applications, including machine translation [7], summarization [49], question answering [37], text normalization [26], textual entailment recognition [1], and semantic parsing [3].

Text is often accompanied by images: for instance, many news articles come with symbolic images that well represent their content. Such images can convey more concrete ideas on described content and help smoother communication. These images also allow us to extend the idea of paraphrases to count for visual concepts (concepts appear in an image). One interesting work is *visually grounded paraphrases* (VGPs) coined by [8], which are paraphrase-like expressions that refer to the same visual concept in an image. For example, “a squirrel” and “a brown squirrel,” “a green glass bottle” and “a beer” for the images in Fig. 1 are VGPs. Because of the inherent nature of images, VGPs should refer to a concrete visual concept.

We argue that such textual representations can benefit various vision and language tasks. Take image captioning [41] as an example, one way of applying VGPs can be the improvement of the evaluation.[1] Other applications can be visual question answering [44] and visual dialogue [11], where the same visual concept can be described in different ways by different users or even the same user through the dialogue.

The work [8] is the pioneering and only study that we are aware of for VGP identification, which proposes a supervised similarity model using neural networks. They formulate the VGP identification task as a binary classification problem and use an attention mechanism over an image to incorporate visual features. They evaluate their model with the Flickr30k entities dataset [32] and show that visual features are helpful but give only a small gain.

We analyze VGPs on the Flickr30k entities dataset in Section 3, and show that on one hand, many pairs of VGPs have a high lexical similarity. Thus language features are often enough for VGP identification. Visual features for them could lead to errors due to inaccurate visual grounding. On the other hand, there are also many VGPs that are difficult to be identified by language features only.

We also argue that the VGP identification is highly related to phrase localization, which localizes an image region corresponding to a phrase [33,38]. If we have a perfect phrase localization algorithm, VGP identification turns out to be a trivial task. All we need is to apply phrase localization to VGP candidates and check if the corresponding image regions largely overlap or not. Unfortunately,

* Corresponding author.
E-mail addresses: otani_mayu@cyberagent.co.jp (M. Otani), chu@ids.osaka-u.ac.jp (C. Chu), n-yuta@ids.osaka-u.ac.jp (Y. Nakashima).

https://doi.org/10.1016/j.neucom.2020.04.066
0925-2312/

[1] Current evaluation metrics such as BLEU scores cannot evaluate generated captions with different wording from references. With VGP identification, the evaluation of phrases describing the same visual concept in different wording will be possible.

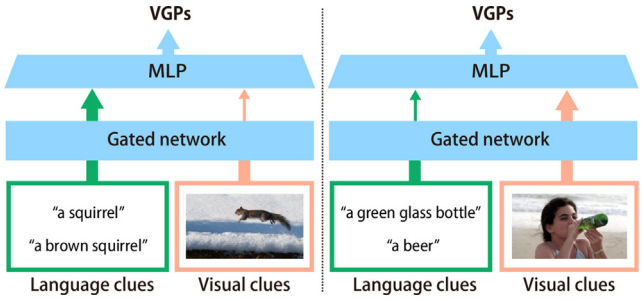


Fig. 1. Our gated network adaptively uses language and visual clues for VGP identification. In the left example, input phrase pairs may be easily inferred that they are describing the same visual concept in the image without looking at it. On the other hand, for lexically dissimilar VGPs in the right example, visual clues from a corresponding image can help predict whether the two phrases correspond to similar visual concepts.

it is not the case, at least currently, and dedicated methods for VGP identification are necessary.

Based on above observations, we propose a VGP identification model with a gated network. To facilitate the power of visual grounding, our model is built upon a phrase localization model, such as [31,48] (Fig. 3). Given a pair of phrases and an associated image, our model first applies phrase localization to get an image region for each phrase. Language features are then extracted from the phrases, and visual features are extracted from localized image regions. Phrase localization is a challenging task, and even the state-of-the-art model can fail to detect image regions for input phrases, which means that the visual features can be completely spoiled. Our gated network handles this issue by adaptively adjusting the weights for each modality. The model predicts the probability that the input phrases are VGPs based on the features fused with the weights. Our contributions are follows:

- Our VGP analysis points out the importance of the respective use of language and visual features for identifying different types of VGPs, which deepens the understanding of VGPs.
- We propose a novel model with explicit phrase localization and a gate network to balance the use of visual and language features adaptively.
- Experimental results on the Flickr30k entities dataset indicate that our model outperforms the state-of-the-art model [8] and that language and visual clues are complementary.

2. Related work

2.1. Phrase localization

Phrase localization is a task to find an image region that corresponds to a given phrase in a caption [13,31,32,38,42,45,48], which is closely related to the VGP identification task. Our model utilizes a phrase localization model, *e.g.*, [31]. Referring expression comprehension [9,30,46,47] is also closely related to phrase localization and so VGP identification, which addresses the problem of connecting a natural language query to an image region. Unlike phrase localization, referring expressions accompanied by properties, such as attributes of objects and relations to other objects, so that they can identify only one object in the image. Both phrase localization and referring expression comprehension involve computing vision-to-language similarity. On the other hand, VGP identification tries to compute the semantic similarity between phrasal expressions themselves conditioned on an image.

2.2. Paraphrase identification

Paraphrase identification has been studied in the NLP community in the last few decades. Previous works identify paraphrases from either monolingual corpora [5,24,27] or bilingual parallel corpora [2,6,14]. The former is based on distributional similarity, stating that paraphrases appear in similar context in monolingual corpora [5,24,27]. The latter uses bilingual pivoting, which assumes that paraphrases in one language are translated into a same phrase in another language [2,6,14]. These works only use languages for identifying paraphrases. VGP identification in contrast identifies VGPs from multimodal datasets consisting of images and their captions. Moreover, VGPs in the Flickr30k entities dataset are very different from the original definition of paraphrases because the dataset contains noun phrases and two phrases are counted as VGP when they refer to the same visual concept in a given image, so that they maybe not paraphrases linguistically. We follow the same definition of VGPs as [8] in this paper.

Paraphrases also have been studied in the multimodal context. Regneri et al. [35] collected sentence level paraphrases by aligning video scripts with the same time frame; these sentence level paraphrases are essentially similar to captions of an image. Lin et al. [25] introduced the visual paraphrasing task, namely identifying whether two descriptions consisting of several sentences are talking about the same visual scene or not. Han et al. [15] calculated the semantic relations such as paraphrases of a phrase pair with their image sets, which are retrieved by search engines given the phrase pair.

2.3. Coreference resolution

Coreference resolution is a task to find the expressions that refer to the same entity in text [40]. VGP identification differs from conventional coreference resolution in the way that it requires visual grounding. In addition, the targets of coreference resolution are the entities in a sentence or a document, while our targets are the entities in the captions of an image that are quasi-paraphrases but are not related to each other in the discourse level like sentences in a document. Conventional coreference resolution methods are based on a pipeline that first parses sentences to extract head-word features and then manually designs rules for entity proposals [10,12,34,43]. An end-to-end coreference resolution model also proposed, which jointly identifies entities and clusters them [23].

Coreference resolution also has been studied in multimodal tasks. Kong et al. [20] improved visual scene parsing by finding nouns and pronouns referring to the same object in an image. Kottur et al. [21] applied coreference resolution in visual dialogue, to identify what a pronoun is referring in a dialogue. Huang et al. [17,18] studied coreference resolution in instructional videos, which temporally links an entity to an action in a video.

3. Analysis of Flick30k VGP dataset

We introduce the dataset construction in [8] and provide lexical analysis on phrase pairs to justify our model design.

3.1. Dataset

VGP identification is first proposed by Chu et al. [8]. They use Flickr30k entities dataset [32], which is a large-scale image captioning dataset containing multiple captions for a single image. Captions for an image can be regarded as a quasi-sentence-level paraphrase corpus, which often contains VGPs. Each entity (*i.e.*, noun phrase) of the captions is manually aligned with an image region. This dataset contains approximately 30k images and each

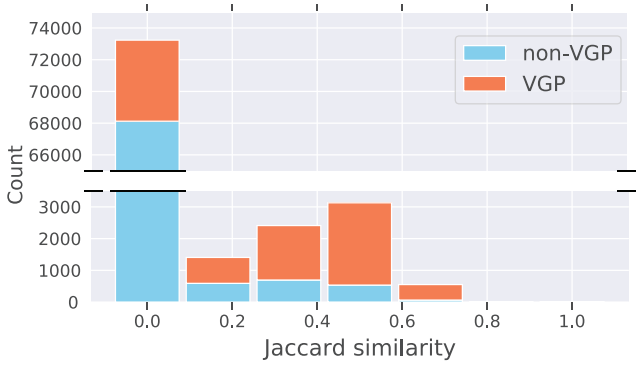


Fig. 2. Distribution of VGPs and non-VGPs for different word overlaps. Most phrases are VGPs when two phrases have at least one word in common.

image has 5 captions. Other image captioning dataset with multiple descriptions, such as Visual Genome [22], are also possible sources as VGPs. However, to apply such datasets for our task, expensive annotations to align paraphrases and corresponding visual concepts are needed, which unfortunately are unavailable besides Flickr30k.

We used the same splits as in [31], which have 29,769, 1000, and 1000 images for training, validation, and testing, respectively. We extracted all possible noun phrase pairs. Each pair is from two different captions associated with one image. We removed trivial pairs that are lexically identical. Consequently, we obtained roughly 2,383k, 80k, and 81k noun phrase pairs for training, validation, and testing, respectively. We treated phrase pairs associated with the same image region as VGP and all other pairs as non-VGP. The ratio of VGPs out of all phrase pairs is approximately 13% for all splits.

3.2. Word overlap

To understand the dataset, we explore lexical similarity between phrase pairs. As a metric of lexical similarity, we use Jaccard similarity between a phrase pair.[2] Fig. 2 shows the Jaccard similarity distribution on the validation set. We can see that most phrase pairs are non-VGPS when there are no word overlap. On the other hand, phrases with common words are more likely to be VGPs. Therefore, for identifying VGPs, a simple similarity based on word overlap can be a strong clue. However, there still are a large number of VGPs that share no words among them. For such VGPs, a model should exploit visual features to get additional cues.

4. Our model

Motivated by the analysis in Section 3, we propose a gated network for VGP identification. Fig. 3 shows an overview of our model. The model takes a pair of noun phrases and the associated image as input and predicts whether the input phrases are VGPs or not. The phrases are represented by phrase embeddings. To extract visual features from the image, we first apply phrase localization that aligns each phrase to their corresponding image regions. From the image regions, we extract visual features with a pre-trained deep model. Two distinct multi-layer perceptrons (MLPs) transform the features of each modality. Other architectures which take a pair of features and transform them into an output feature can be plugged into the model. As the model architecture of feature transformation is not the main contribution of this paper, we do not further explore architectures of this module. After transforming both phrase and visual features, our model predicts weights for each modality and fuse the features with the weights. The output feature is fed into an MLP to predict the probability of being VGPs.

4.1. Language features

For phrase features, we use the 300-D word2vec word embeddings trained on the Google News corpus[3] [28]. The embedding of each word are computed and average-pooled over all words in the phrase to obtain a phrase embedding. We remove stop words when computing the embedding. Chu et al. [8] tested their model with other types of phrase embeddings, such as Fisher vectors and Fisher vectors with CCA that projects phrase and image region features into a common space, but we found that the average-pooled word embeddings worked best. This may mean that simpler phrase embeddings work well for phrases consisting of few words. Each phrase feature goes through an MLP network ϕ, given by

$$\phi(\mathbf{x}) = f(W_2 f(W_1 \mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2), \tag{1}$$

and both are merged together as

$$\mathbf{x}_\mathrm{l} = f(\phi(\mathbf{x}_\mathrm{p1}) + \phi(\mathbf{x}_\mathrm{p2})) \tag{2}$$

to get the language feature $\mathbf{x}_\mathrm{l} \in \mathbb{R}^{1000}$, where $\mathbf{x}_\mathrm{p1}$ and $\mathbf{x}_\mathrm{p2}$ are an average-pooled 300-D word embedding for each phrase, respectively; and f is the ReLU nonlinearity.

4.2. Visual features

Our model localizes an image region for each phrase and uses the additional visual clues to identify VGPs. This is one of key differences from [8], which employs top-down attention on an input image. Top-down attention mechanism enhances visual features corresponding to uniformly located grid regions. However, object-level regions are more natural input to represent visual concepts. Based on this idea, we retrieve an object-level region for each phrase. We expect that the visual similarity between the image regions can help find VGPs even when the phrases have low lexical similarity.

We employ the methods [31,48] in our experiments, as the authors released their implementation. Moreover, the models trained on the Flickr30k entities dataset are also publicly available. Note that other phrase localization methods can be used in our model without significant modification.

After obtaining a corresponding image region for each phrase, we extract an image region embedding. Our method uses VGG16 [39] as in Faster R-CNN [36]. The image region embeddings are then fed into an MLP network ψ, defined as

$$\psi(\mathbf{x}) = f(W_4 f(W_3 \mathbf{x} + \mathbf{b}_3) + \mathbf{b}_4), \tag{3}$$

and are fused with

$$\mathbf{x}_\mathrm{v} = f(\psi(\mathbf{x}_\mathrm{v1}) + \psi(\mathbf{x}_\mathrm{v2})), \tag{4}$$

where $\mathbf{x}_\mathrm{v1}$ and $\mathbf{x}_\mathrm{v2}$ are image region embeddings in $\mathbb{R}^{4096}$, and $\mathbf{x}_\mathrm{v} \in \mathbb{R}^{1000}$ is the output visual feature.

4.3. Gated network

VGP identification can be an easy task when lexical similarity is high; however, this is not always the case and a large number of VGPs do not share common words. The visual features are useful for such VGPs. This observation motivates us to develop a gated network, which computes how much each modality should contribute when predicting the probability of being VGPs.

[2] Stop words are discarded when calculating Jaccard similarity.

[3] https://github.com/mmihaltz/word2vec-GoogleNews-vectors

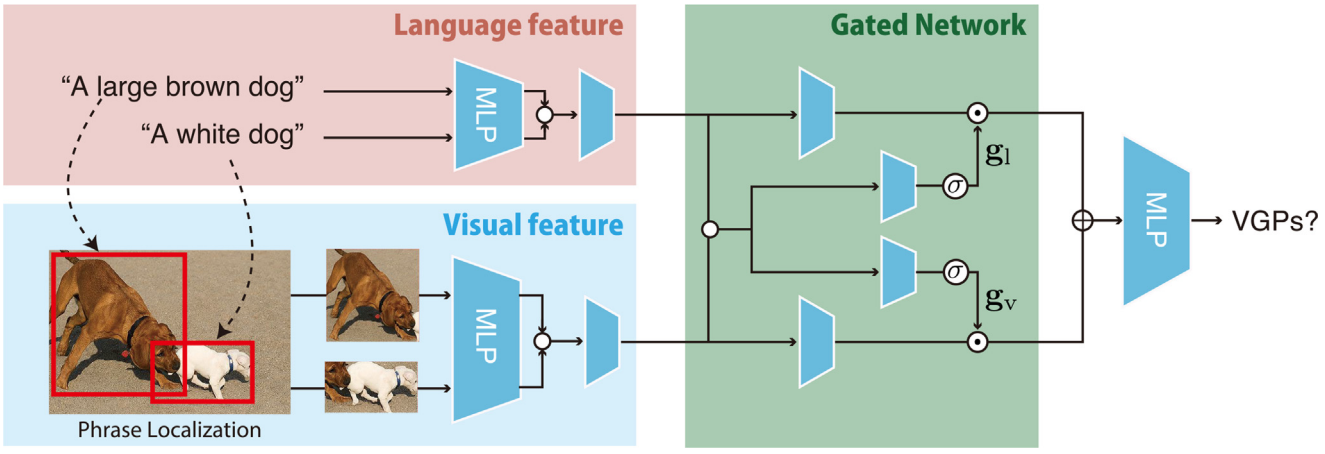


Fig. 3. An overview of our VGP identification model. It takes a pair of noun phrases in different captions and their associated image as input. Image regions corresponding to phrases are obtained by phrase localization models, *e.g.*, [31,48]. Visual features are fed into our model to compute the probability of being VGPs. Our model uses a gated network to adaptively weight language and visual features.

Our gated network computes the weights for language and visual features separately. The weights for language and visual features, $\mathbf{g}_l$ and $\mathbf{g}_v$, respectively, are computed by

$$\mathbf{g}_l = \sigma(U_l[\mathbf{x}_v, \mathbf{x}_l] + \mathbf{s}_l) \tag{5}$$

$$\mathbf{g}_v = \sigma(U_v[\mathbf{x}_v, \mathbf{x}_l] + \mathbf{s}_v), \tag{6}$$

where σ is the sigmoid nonlinearity and $[\,\cdot\,,\,\cdot\,]$ is the concatenation. After a fully-connected layer, the language and visual features are fused using the weights as

$$\mathbf{y} = \mathbf{g}_l \odot \tanh(W_l\mathbf{x}_l + \mathbf{b}_l) + \mathbf{g}_v \odot \tanh(W_v\mathbf{x}_v + \mathbf{b}_v), \tag{7}$$

where $\odot$ is the element-wise product and $\mathbf{y} \in \mathbb{R}^{300}$. The increase of the computational cost by the gated network is rather limited, compared to the whole computation. To be specific, the gate mechanism only increases computational time of the forward and backward processes by 2%.

4.4. VGP prediction

Finally, we feed the gate network's output $\mathbf{y}$ to a two-layer MLP network to compute the probability of being VGPs. The unit sizes for the two-layer MLP network are 128 and 1, respectively. That is,

$$\mathbf{h} = f(W_5\mathbf{y} + \mathbf{b}_5) \tag{8}$$

$$z = \sigma(W_6\mathbf{h} + \mathbf{b}_6). \tag{9}$$

We use dropout regularization before the second layer and batch normalization before every ReLU nonlinearity.

5. Experimental settings

5.1. Baselines

Chu et al. (2018) [8]**:** We report the best model in [8]. Their model simply fuses phrase and visual features. Instead of phrase localization, they use an attention mechanism over feature maps of images.

Word-overlap: We also tested a naive baseline model that predicts VGPs based solely on lexical similarity between phrases. We used the Jaccard similarity and classifies phrases into VGPs/non-VGPs when the similarity is larger than a threshold, which is optimized on the validation set.

BoundingBox-overlap: The BoundingBox overlap model classified VGPs based on the overlap of detected image regions. We trained a logistic regression model that takes IoU of two image regions and predicts the probability of being VGPs.

Phrase-only: The phrase-only model classifies VGPs without looking at images. This model is trained only with the phrase features; therefore, the phrase-only model computes $\mathbf{y}$ by

$$\mathbf{y} = \tanh(W_l\mathbf{x}_l + \mathbf{b}_l). \tag{10}$$

Visual-only: The visual-only model uses only visual features. Instead of fusing language and visual features, $\mathbf{y}$ is computed by

$$\mathbf{y} = \tanh(W_v\mathbf{x}_v + \mathbf{b}_v). \tag{11}$$

5.2. Gate mechanisms

We evaluated different configurations of gate mechanisms for comparison to ours.

Language gate: Instead of computing gates from both language and visual features, the gate weights are computed only from language features. Therefore, the gate weights are computed as

$$\mathbf{g}_l = \sigma(U_l\mathbf{x}_l + \mathbf{u}_l) \tag{12}$$

$$\mathbf{g}_v = \sigma(U_v\mathbf{x}_l + \mathbf{u}_v). \tag{13}$$

Visual gate: As in the language gate network, the visual gate network uses only visual features to control both gates; therefore, the gate weights are computed as:

$$\mathbf{g}_l = \sigma(U_l\mathbf{x}_v + \mathbf{u}_l), \tag{14}$$

$$\mathbf{g}_v = \sigma(U_v\mathbf{x}_v + \mathbf{u}_v). \tag{15}$$

5.3. Phrase localization methods

We used the following two phrase localization methods for the visual-only and our models.

PL-CLC: We tested an existing method in [31] (PL-CLC). Their method maps phrases and image regions using canonical correlation analysis (CCA) [16] and ranks image regions using the cosine CCA distance. They also combine other sorts of scores based on object detection results, region sizes, colors, and spatial relationships as clues for phrase localization. We used their pre-trained model, which is publicly available.[4] The resulting IoU distribution

[4] https://github.com/BryanPlummer/pl-clc

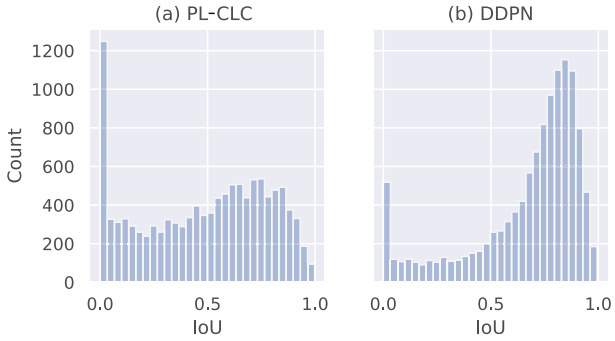


Fig. 4. Distribution of IoUs.

of this phrase localization method over the validation set is shown in Fig. 4 (a). The accuracy (IoU $\geq$ 0.5) is 52.53% and the average IoU is 0.48. We can see from this distribution that the most predictions are close to the gold standards in some extent, but many are completely off.

DDPN: We also tested our model with another recently published phrase localization method. For this, we used the implementation of DDPN [48],[5] which is the current state-of-the-art phrase localization method. This method increases both diversity and discriminative power of generated proposals, which improves the phrase localization accuracy. The accuracy (IoU $\geq$ 0.5) is 80.40% and the average IoU is 0.67. The IoU distribution is shown in Fig. 4 (b).

5.4. Training

We used the following settings for training the phrase-only, visual-only, and our models. Adam [19] was used for optimization with the mini-batch size of 500. We used the sigmoid cross-entropy as the loss function. During training, the learning rate was halved at every epoch. We terminated training after 5 epochs, at which the loss converged on the validation set. We tuned the hyperparameters of the initial learning rate, weight decay, and dropout, with Bayesian optimization [29] implemented in GPyOpt.[6] We initialized the parameters of VGG16 for visual feature extraction with the model trained on the PASCAL VOC2007 detection dataset provided in chainercv,[7] and fixed its parameters during training.

6. Results and discussion

6.1. Quantitative evaluation

We evaluated the performance of VGP identification in terms of the F1, precision, and recall scores. A phrase pair was predicted to be VGPs when its VGP identification probability is higher than a certain threshold. We report the performance with the threshold tuned on the validation set so that the F1 score is maximized. Table 1 shows the results.

Importance of language clues. The word overlap and the phrase-only models demonstrate that language clues are quite efficient to find VGPs on this dataset. The word overlap model achieved better performance than the visual-only models. This indicates that simple language similarity benefits more than visual features when a single modality is individually used as input. The main reason of this high F1 score of the word overlap model is the characteristic of the Flickr30k entities dataset that contains many VGPs with word overlap as described in Section 3, *e.g.*, "a hat" and "a pink hat." Language clues are often enough to find such VGPs. Comparison between the word overlap and phrase-only models shows a learning-based approach built upon word embedding boosts the performance.

Table 1
F1, precision, and recall scores of VGP identification on the Flickr 30k entities dataset.

	F1	Prec.	Rec.
Chu et al. (2018) [8]	84.16	82.71	85.67
Word-overlap	61.25	74.15	52.18
Phrase-only	85.66	84.72	86.61
Visual-only (PL-CLC)	57.73	51.86	65.09
Visual-only (DDPN)	66.36	60.92	72.87
BoundingBox-overlap (DDPN)	73.43	73.83	73.05
Ours (PL-CLC)	85.10	83.36	86.91
Ours (DDPN)	86.48	**85.81**	87.16
Ours+BBox (DDPN)	**86.50**	84.92	**88.15**

Table 2
Comparison of different gate mechanisms with DDPN for phrase localization.

Gate mechanism	F1	Prec.	Rec.
Without gate	86.43	84.54	**88.40**
Language gate	86.22	**86.05**	86.39
Visual gate	86.22	84.74	87.75
Multimodal gate	**86.48**	85.81	87.16

Table 3
L2 norm of gate weights.

Gate mechanism	$\|\mathbf{g}_l\|_2$		$\|\mathbf{g}_v\|_2$	
	Mean	Std.	Mean	Std.
Language gate	135.98	10.32	82.50	9.22
Visual gate	113.36	8.34	69.48	2.88
Multimodal gate	131.14	9.38	77.81	5.05

Comparison of different gate mechanisms. Table 2 compares different gate mechanisms. We can see that the language gate achieved a high precision, while the visual gate performed better in recall. One of the reasons for this can be that visual features are more noisy than language features due to erroneous phrase localization. Without gate mechanisms, the model got the high recall but the precision is lower than others. The model tends to depend too much on language clues. This may make it difficult to decline lexically similar non-VGPs. The multimodal gate showed the best performance, which is consistent with our assumption that both language and visual features have their advantage and disadvantage for VGP identification, and thus they should be used adaptively.

Table 3 shows the means and standard deviations of the L2 norm of the weights of each gate mechanism. A larger norm for gate weights indicates that the model leverages more features from that gate. Therefore the table implies that the language features contribute more than visual features among all gate mechanisms, which supports our analysis in Section 3. The large standard deviation of the language gate shows that the model aggressively changes gate weights according to language features. We assume that language features provide richer clues to control the flow from each modality. In contrast, the visual gate learned rather stable gate weights. The multimodal gate mechanism balances both characteristics and achieved the best performance.

How does phrase localization error affect VGP identification? Fig. 5 shows results with respect to the performance of phrase localization. For this, because the VGP identification takes a phrase

[5] https://github.com/XiangChenchao/DDPN
[6] https://github.com/SheffieldML/GPyOpt
[7] https://github.com/chainer/chainercv

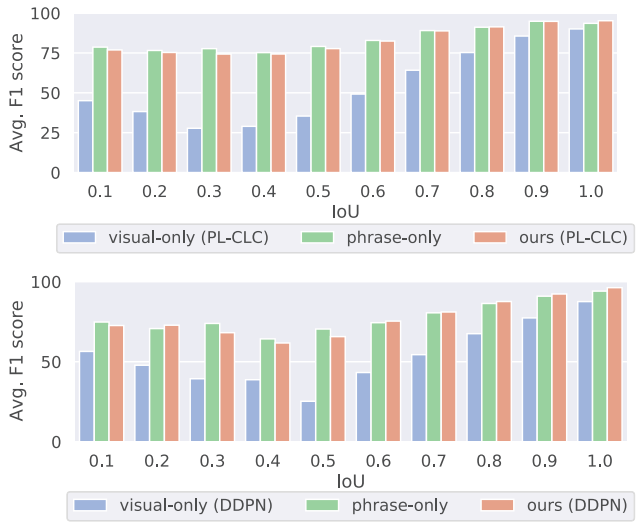


Fig. 5. F1 scores computed for phrase pairs with different IoUs. We split the test set into ten sections based on IoU of phrase localization. Top: F1 scores of models with PL-CLC. Bottom: Scores of models with DDPN.

Fig. 6. Hard VGP and non-VGP examples. "A stuffed animal" and "her teddy bear" are VGPs, but they have no common words. "A motorcycle" and "a motorcycle rider" depict different visual concepts, but they have a word "motorcycle" in common.

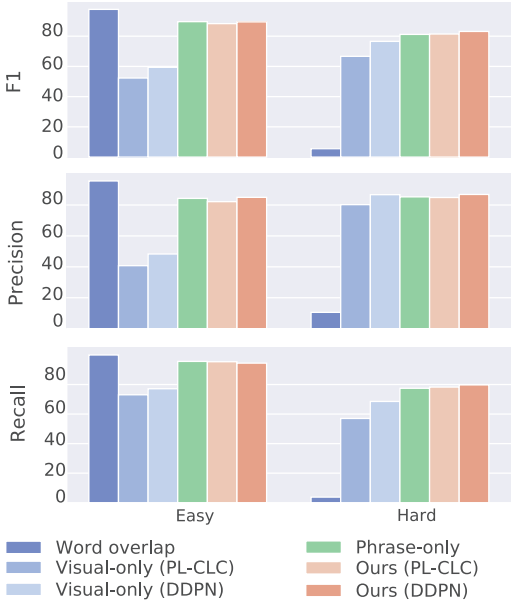


Fig. 7. Scores for easy and hard VGP identification.

pair as input, we computed the IoU of phrase localization results and took the average of them. We then split the test set based on the averaged IoUs.

With PL-CLC, the overall scores drop compared to the phrase-only model. We believe the reason is the erroneous phrase localization of PL-CLC; the extracted image regions can be irrelevant to the input phrases, which makes the training process hard to learn from visual features. This leads to a model that substantially ignores visual features even if the image region is correct (*i.e.*, with higher IoUs). Moreover, optimization of the model is harder than the phrase-only model because of the larger parameter size. On the other hand, ours with DDPN boosts the performance. Same as in ours with PL-CLC, the performance drops when phrase localization fails, *i.e.*, the average IoU is less than 0.5, but the proportion of errors in phrase localization is much smaller than PL-CLC (Fig. 4). This results in the improvement in the total F1 score.

The results suggest that visual features are helpful when phrase localization can find relevant image regions; otherwise, visual features are likely to work negatively.

Is phrase localization all you need? We observed that BoundingBox-overlap achieved F1 score of 73.43%, which is a strong baseline of VGP identification. We also tested Ours+BBox (DDPN) which inputs bounding boxes' IoU values to the last MLP network, in addition to the output of the gated network. However, the improvement by incorporating IoU values into our model is rather limited.

Visual features extracted from more accurate phrase localization boost the performance. However, the result by visual-only suggests that learning powerful visual features for the VGP task is still challenging. The visual-only model with DDPN got the F1 score of 66.36% (Table 1), while its phrase localization performs well.

Based on these observations, it is hard to further improve VGP identification using only phrase localization, on the other hand, incorporating language features drastically increases the F1 score. Therefore, language features are still important even when phrase localization performs well.

Fig. 5 provides insights in detail. As shown Fig. 5, the gain from language features is especially large when phrase localization fails. For example, there are significant differences between the performance of visual-only and ours for samples whose average IoU $\leq$ 0.5. On the other hand, we can still gain some boosts from language features when both phrases localized accurately.

When do visual features help? We also investigated how visual features affect VGP identification. We divided the test set into two categories, "easy" and "hard," in terms of word overlap measured by the Jaccard similarity. We computed the median of the Jaccard similarity of positive VGPs in the validation set and found that it was 0.25. Hard phrase pairs include hard positives and hard negatives, where hard positives are VGPs whose Jaccard similarity between phrases is less than the median, and hard negatives are non-VGPs whose Jaccard similarity is larger than 0.25. Others are grouped into easy phrase pairs. Fig. 6 shows some examples of hard phrase pairs.

Fig. 7 shows F1, precision, and recall scores computed for easy and hard phrase pairs. Note that the number of easy and hard phrase pairs are imbalanced. Most easy phrase pairs are non-VGPs; therefore, precision scores are likely to be low. For easy phrase pairs, the phrase-only model and ours, which use both language and visual features, do not show significant difference. For hard phrase pairs, compared to the phrase-only model, ours (DDPN) improved the performance by 1.56% and 2.30%, in precision and recall, respectively. Therefore, language features are highly efficient for lexically similar VGPs, but we can gain further improvement in finding lexically different VGPs by incorporating visual features.

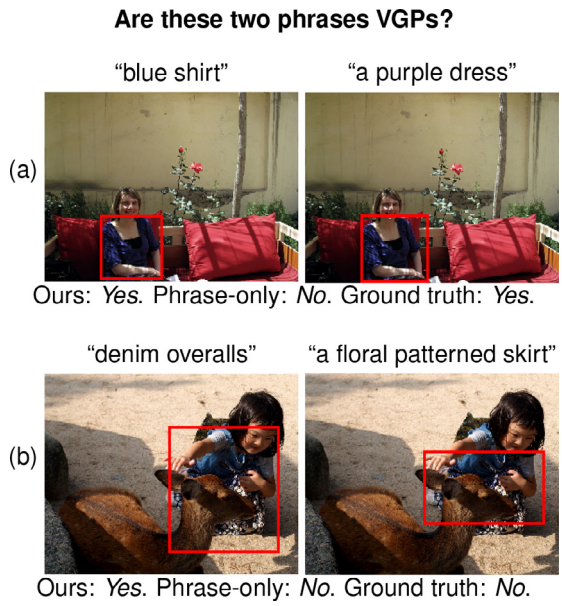


Fig. 8. Examples of VGP identification. The red boxes are phrase localization results.

6.2. Qualitative evaluation

We also conducted qualitative evaluation by comparing the results of the phrase-only model to our model. Fig. 8 (a) shows an example that VGP identification was improved by incorporating visual clues. In the example, DDPN successfully found image regions described by input phrases. On the other hand, the visual clues affected negatively in Fig. 8 (b). Localized image regions for "denim overalls" and "a floral patterned skirt" overlap. These image regions may result in similar visual features that can make our model incorrectly classify the phrases into VGPs. To avoid such errors, fine-grained image region extraction, *e.g.*, instance segmentation is required.

7. Conclusion

We proposed a gated network with phrase localization for VGP identification. Experimental results showed the effectiveness of the proposed model. We observed that visual features benefits VGP identification, especially for lexically different VGPs. However, the error in phrase localization propagates to VGP identification; therefore, phrase localization needs to be accurate.

Our future work includes building a novel VGP dataset. The Flickr30k entity dataset only has noun VGPs corresponding to single objects. We will extend the dataset by collecting more complex VGPs that describe relations between objects, *e.g.*, "a man riding a bike" and "a young man on a bicycle." Exploring fine-grained image region extraction such as instance segmentation, will be another interesting direction.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Mayu Otani: Methodology, Software, Formal analysis, Writing - original draft, Visualization. **Chenhui Chu:** Conceptualization, Software, Writing - original draft, Writing - review & editing, Project administration, Funding acquisition. **Yuta Nakashima:** Conceptualization, Writing - review & editing, Funding acquisition.

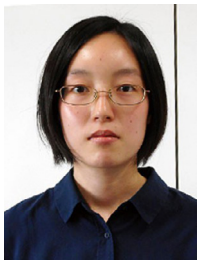
Acknowledgement

This paper is based on results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO). This work was supported by ACT-I, JST and JSPS KAKENHI No. 18H03264.

References

[1] I. Androutsopoulos, P. Malakasiotis, A survey of paraphrasing and textual entailment methods, J. Artif. Intell. Res. 38 (1) (2010) 135–187.
[2] C. Bannard, C. Callison-Burch, Paraphrasing with bilingual parallel corpora, in: Proceedings of the ACL, 2005, pp. 597–604, doi:10.3115/1219840.1219914.
[3] J. Berant, P. Liang, Semantic parsing via paraphrasing, in: Proceedings of the ACL, 2014, pp. 1415–1425. URL http://www.aclweb.org/anthology/P14-1133.
[4] R. Bhagat, E. Hovy, What is a paraphrase? Comput. Linguist. 39 (3) (2013) 463–472.
[5] R. Bhagat, D. Ravichandran, Large scale acquisition of paraphrases for learning surface patterns, in: Proceedings of the ACL, 2008, pp. 674–682. URL http://www.aclweb.org/anthology/P/P08/P08-1077.
[6] C. Callison-Burch, Syntactic constraints on paraphrases extracted from parallel corpora, in: Proceedings of the EMNLP, 2008, pp. 196–205. URL http://www.aclweb.org/anthology/D08-1021.
[7] C. Chu, S. Kurohashi, Paraphrasing out-of-vocabulary words with word embeddings and semantic lexicons for low resource statistical machine translation, in: Proceedings of the LREC, 2016, pp. 644–648.
[8] C. Chu, M. Otani, Y. Nakashima, iParaphrasing: extracting visually grounded paraphrases via an image, in: Proceedings of the COLING, 2018, pp. 3479–3492.
[9] V. Cirik, T. Berg-Kirkpatrick, L.-P. Morency, Using syntax to ground referring expressions in natural images, in: Proceedings of the AAAI, 2018, pp. 6756–6764.
[10] K. Clark, C.D. Manning, Improving coreference resolution by learning entity-level distributed representations, in: Proceedings of the ACL, 2016, pp. 643–653, doi:10.18653/v1/P16-1061.
[11] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J.M. Moura, D. Parikh, D. Batra, Visual Dialog, in: Proceedings of the CVPR, IEEE, 2017, pp. 326–335.
[12] G. Durrett, D. Klein, Easy victories and uphill battles in coreference resolution, in: Proceedings of the EMNLP, 2013, pp. 1971–1982. URL http://aclweb.org/anthology/D13-1203.
[13] A. Fukui, D.H. Park, D. Yang, A. Rohrbach, T. Darrell, M. Rohrbach, Multimodal compact bilinear pooling for visual question answering and visual grounding, in: Proceedings of the EMNLP, 2016, pp. 457–468. URL https://aclweb.org/anthology/D16-1044.
[14] J. Ganitkevitch, C. Callison-Burch, The multilingual paraphrase database, in: Proceedings of the LREC, ELRA, 2014, pp. 4276–4283. https://www.aclweb.org/anthology/L14-1520/.
[15] D. Han, P. Martínez-Gómez, K. Mineshima, Visual denotations for recognizing textual entailment, in: Proceedings of the EMNLP, 2017, pp. 2843–2849. URL https://www.aclweb.org/anthology/D17-1304.
[16] D.R. Hardoon, S.R. Szedmak, J.R. Shawe-taylor, Canonical correlation analysis: an overview with application to learning methods, Neural Comput. 16 (12) (2004) 2639–2664, doi:10.1162/0899766042321814.
[17] D. Huang, J.J. Lim, L. Fei-Fei, J.C. Niebles, Unsupervised visual-linguistic reference resolution in instructional videos, in: Proceedings of the CVPR, IEEE Computer Society, 2017, pp. 1032–1041.
[18] D.-A. Huang, S. Buch, L. Dery, A. Garg, L. Fei-Fei, J.C. Niebles, Finding "it": Weakly-supervised, reference-aware visual grounding in instructional videos, in: Proceedings of the CVPR, IEEE Computer Society, 2018, pp. 5948–5957.
[19] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, CoRR (2014). URL http://arxiv.org/abs/1412.6980.
[20] C. Kong, D. Lin, M. Bansal, R. Urtasun, S. Fidler, What are you talking about? text-to-image coreference, in: Proceedings of the IEEE CVPR, IEEE Computer Society, 2014, pp. 3558–3565.
[21] S. Kottur, J.M.F. Moura, D. Parikh, D. Batra, M. Rohrbach, Visual coreference resolution in visual dialog using neural module networks, in: Proceedings of the ECCV, Springer International Publishing, 2018.
[22] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D.A. Shamma, M.S. Bernstein, L. Fei-Fei, Visual genome: connecting language and vision using crowdsourced dense image annotations, Int. J. Comput. Vis. 123 (1) (2017) 32–73, doi:10.1007/s11263-016-0981-7.
[23] K. Lee, L. He, M. Lewis, L. Zettlemoyer, End-to-end neural coreference resolution, in: Proceedings of the EMNLP, 2017, pp. 188–197. URL https://www.aclweb.org/anthology/D17-1018.
[24] D. Lin, P. Pantel, DIRT–Discovery of inference rules from text, in: Proceedings of the ACM SIGKDD, 2001, pp. 323–328, doi:10.1145/502512.502559.
[25] X. Lin, D.P. Vinyals, Don't just listen, use your imagination: leveraging visual common sense for non-visual tasks, in: Proceedings of the CVPR, IEEE Computer Society, 2015, pp. 2984–2993.
[26] W. Ling, C. Dyer, A.W. Black, I. Trancoso, Paraphrasing 4 microblog normalization, in: Proceedings of the EMNLP, 2013, pp. 73–84. URL http://www.aclweb.org/anthology/D13-1008.

[27] Y. Marton, C. Callison-Burch, P. Resnik, Improved statistical machine translation using monolingually-derived paraphrases, in: Proceedings of the EMNLP, 2009, pp. 381–390. URL http://www.aclweb.org/anthology/D/D09/D09-1040.
[28] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, CoRR (2013). URL http://arxiv.org/abs/1301.3781.
[29] J. Mockus, Bayesian approach to global optimization: theory and applications, Mathematics and its applications: Soviet series, 1989. URL https://books.google.co.jp/books?id=FknvAAAAMAAJ.
[30] V.K. Nagaraja, V.I. Morariu, L.S. Davis, Modeling context between objects for referring expression understanding, in: Proceedings of the ECCV, Springer International Publishing, 2016, pp. 792–807.
[31] B.A. Plummer, A. Mallya, C.M. Cervantes, J. Hockenmaier, S. Lazebnik, Phrase localization and visual relationship detection with comprehensive image-language cues, in: Proceedings of the ICCV, IEEE Computer Society, 2017, pp. 1928–1937.
[32] B.A. Plummer, L. Wang, C.M. Cervantes, J.C. Caicedo, J. Hockenmaier, S. Lazebnik, Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models, in: Proceedings of the ICCV, IEEE Computer Society, 2015, pp. 2641–2649.
[33] B.A. Plummer, L. Wang, C.M. Cervantes, J.C. Caicedo, J. Hockenmaier, S. Lazebnik, Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models, in: Proceedings of the IJCV, 123, 2017, pp. 74–93, doi:10.1007/s11263-016-0965-7.
[34] K. Raghunathan, H. Lee, S. Rangarajan, N. Chambers, M. Surdeanu, D. Jurafsky, C. Manning, A multi-pass sieve for coreference resolution, in: Proceedings of the EMNLP, 2010, pp. 492–501. URL http://aclweb.org/anthology/D10-1048.
[35] M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, M. Pinkal, Grounding action descriptions in videos, Trans. Assoc. Comput. Linguist. 1 (2013) 25–36.
[36] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, in: Proceedings of the NIPS, Curran Associates, Inc., 2015, pp. 91–99.
[37] S. Riezler, A. Vasserman, I. Tsochantaridis, V. Mittal, Y. Liu, Statistical machine translation for query expansion in answer retrieval, in: Proceedings of the ACL, 2007, pp. 464–471. URL http://www.aclweb.org/anthology/P07-1059.
[38] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, B. Schiele, Grounding of textual phrases in images by reconstruction, in: Proceedings of the ECCV, Springer International Publishing, 2016, pp. 817–834.
[39] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recoginition, in: Proceedings of the ICLR, 2015, pp. 1–14.
[40] W.M. Soon, H.T. Ng, D.C.Y. Lim, A machine learning approach to coreference resolution of noun phrases, Comput. Linguist. 27 (4) (2001) 521–544.
[41] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: a neural image caption generator, in: Proceedings of the CVPR, IEEE Computer Society, 2015, pp. 3156–3164.
[42] L. Wang, Y. Li, S. Lazebnik, Learning deep structure-preserving image-text embeddings, in: Proceedings of the CVPR, IEEE Computer Society, 2016, pp. 5005–5013.
[43] S. Wiseman, A.M. Rush, S.M. Shieber, Learning global features for coreference resolution, in: Proceedings of the NAACL, 2016, pp. 994–1004, doi:10.18653/v1/N16-1114.
[44] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, A.v.d. Hengel, Visual question answering: a survey of methods and datasets, in: Proceedings of the CVIU, 2017, pp. 1–20.
[45] P. Young, A. Lai, M. Hodosh, H. Julia, From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions, Trans. Assoc. Comput. Linguist. 2 (1) (2014) 67–78.
[46] L. Yu, Z. Lin, X. Shen, J. Yang, X. Lu, M. Bansal, T.L. Berg, MAttNet: Modular attention network for referring expression comprehension, in: Proceedings of the CVPR, IEEE Computer Society, 2018, pp. 1307–1315.
[47] L. Yu, H. Tan, M. Bansal, T.L. Berg, A joint speaker-listener-reinforcer model for referring expressions, in: Proceedings of the CVPR, IEEE Computer Society, 2017, pp. 282–7290.
[48] Z. Yu, J. Yu, C. Xiang, Z. Zhao, Q. Tian, D. Tao, Rethinking diversified and discriminative proposal generation for visual grounding, in: Proceedings of the IJCAI, 2018, pp. 1114–1120, doi:10.24963/ijcai.2018/155.
[49] L. Zhou, C.-Y. Lin, D.S. Munteanu, E. Hovy, ParaEval: using paraphrases to evaluate summaries automatically, in: Proceedings of the ACL, Association for Computational Linguistics, 2006, pp. 447–454.



Mayu Otani received the B.S. degree from Kyoto University in 2013, and M.S, and Ph.D in engineering from Nara Institute of Science and Technology in 2015 and 2018. She is currently an Research Scientist at CyberAgent, Inc. Her research interests include video understanding and multimodal machine learning.

Chenhui Chu received his B.S. in Software Engineering from Chongqing University in 2008, and M.S., and Ph.D. in Informatics from Kyoto University in 2012 and 2015, respectively. He is currently a research assistant professor at Osaka University. His research interests center on natural language processing, particularly machine translation and multimodal machine learning.

Yuta Nakashima received the B.E. and M.E. degrees in communication engineering and the Ph.D. degree in engineering from Osaka University, Osaka, Japan, in 2006, 2008, and 2012, respectively. From 2012 to 2016, he was an Assistant Professor at the Nara Institute of Science and Technology. He is currently an Associate Professor at the Institute for Datability Science, Osaka University. His research interests include computer vision and machine learning and their applications. His main research includes video content analysis using machine learning approaches. He is a member of ACM, IEICE, and IPSJ.