Adaptive Gating Mechanism for Identifying Visually Grounded Paraphrases

Mayu Otani
CyberAgent, Inc.
mayu-ot@cyberagent.co.jp

Chenhui Chu
Osaka University
chu@ids.osaka-u.ac.jp

Yuta Nakashima
Osaka University
n-yuta@ids.osaka-u.ac.jp

1. Introduction

Visually grounded paraphrases (VGPs) [1] 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 Figure 1 are VGPs. Such textual representations can benefit various vision and language tasks such as captioning and visual question answering, where the same visual concept can be described in different ways.

We analyze VGPs on the Flickr30k entities dataset [2], and find that many of VGPs have high lexical similarity. Thus language clues often provide strong prior for VGP identification. For such VGPs, visual clues could lead to errors due to inaccurate visual grounding. On the other hand, many VGPs are difficult to be identified by language clues only.

Based on the observations, we propose a VGP identification model, which adaptively controls the weights for each modality based on input visual and language clues. Given a pair of phrases and a corresponding 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 could fail to detect image regions for input phrases, which means that the visual features can be completely spoiled. Our gating mechanism alleviates this issue by adaptively adjusting the weights for each modality. We expect the gating mechanism to use reliable modality more. The model predicts the probability that the input phrases are VGPs based on the features fused with their weights.

2. Our Model

We propose a VGP identification model with a gating mechanism (Figure 2). The model takes a pair of phrases and the associated image as input and predicts whether the input phrases are VGPs or not. The gating mechanism controls weights for each modality based on the input.

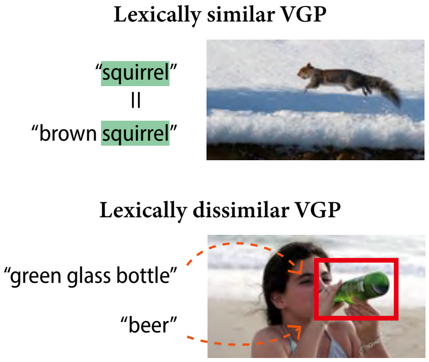

Figure 1. In the top example, input phrase pairs can be easily inferred that they are describing the same visual concept. On the other hand, for lexically dissimilar VGPs in the bottom example, visual clues can be helpful to estimate whether the two phrases describe the same visual concept in the image.

For phrase features, we use the 300-D word2vec word embeddings trained on the Google News corpus[1]. The word embeddings in a phrase are average-pooled to obtain a phrase embedding. Each phrase embedding goes through a two-layer MLP network with ReLU nonlinearity. We compute embeddings for each of an input phrase pair, and merge them by adding to produce phrase features $\mathbf{x}_\mathrm{l} \in \mathbb{R}^{1000}$.

To obtain visual features, our model localizes the input phrase in the image and extracts visual features from the detected image regions. We employ the phrase localization method [5] in our experiments. After obtaining a corresponding image region for each phrase, we extract an image region embedding with VGG16 [4] as in Faster R-CNN [3]. The image region embeddings are then fed into a two-layer MLP. The embeddings of the image regions are fused by adding to obtain output visual feature $\mathbf{x}_\mathrm{v} \in \mathbb{R}^{1000}$.

Our gating mechanism computes how much each modality should contribute for final output of VGP identification. Let $\mathbf{g}_l$ and $\mathbf{g}_v$ be the weights for language and visual features, respectively. They are computed by

$$\mathbf{g}_\mathrm{l} = \sigma(U_l[\mathbf{x}_\mathrm{v}, \mathbf{x}_\mathrm{l}] + \mathbf{s}_\mathrm{l}) \quad (1)$$

$$\mathbf{g}_\mathrm{v} = \sigma(U_v[\mathbf{x}_\mathrm{v}, \mathbf{x}_\mathrm{l}] + \mathbf{s}_\mathrm{v}), \quad (2)$$

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

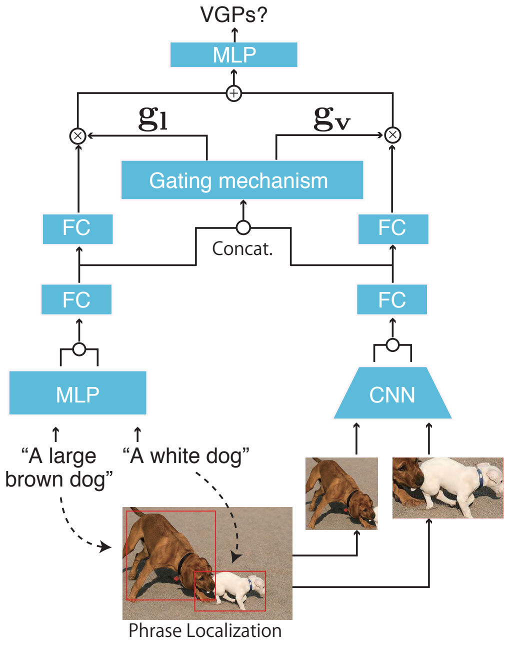

Figure 2. An overview of our VGP identification model.

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

$$\begin{aligned} \mathbf{y} \;=\; & \mathbf{g}_\mathrm{l} \odot \tanh(W_\mathrm{l}\mathbf{x}_\mathrm{l} + \mathbf{b}_\mathrm{l}) \\ & + \mathbf{g}_\mathrm{v} \odot \tanh(W_v\mathbf{x}_\mathrm{v} + \mathbf{b}_\mathrm{v}), \end{aligned} \tag{3}$$

where $\odot$ is the element-wise product and $\mathbf{y} \in \mathbb{R}^{300}$. Finally, we feed the gate network's output $\mathbf{y}$ to a two-layer MLP network to compute the probability of being VGPs.

3. Experiments

We evaluated our model on the Flicker30K entities dataset [2]. The performance of the following baseline models is also reported. **Word-overlap** predicts VGPs based on Jaccard similarity between phrases. **Phrase-only** model is a variant of our full model but uses only phrase features. **Visual-only** model, on the other hand, uses only visual-features. The first two are blind models, which do not use images, and the last one is without language clues.

Our model outperformed the baseline models. On the other hand, the word overlap and the phrase-only models demonstrate that the blind models are very efficient on this dataset. The main reason for the high F1 score of the word overlap model is that the Flickr30k entities dataset contains many lexically similar VGPs. Comparison between the word overlap and phrase-only models suggests the efficiency of the learned language features.

We also investigated the effects of the performance of phrase localization on our full model. Figure 3 shows that our model got some gains from visual features when phrases are successfully localized, *i.e.*, the average IoU is more than 0.5; otherwise, the performance drops compared to phrase-only models.

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

Method	F1	Prec.	Rec.
[1]	84.16	82.71	85.67
Word-overlap	61.25	74.15	52.18
Phrase-only	85.66	84.72	86.61
Visual-only	66.36	60.92	72.87
Ours	**86.48**	**85.81**	**87.16**

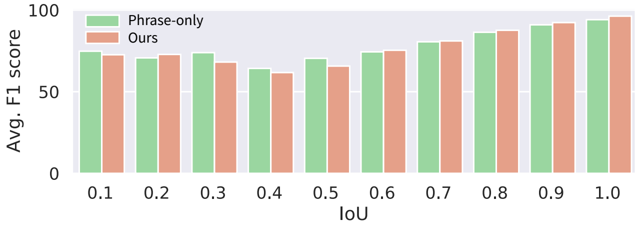

Figure 3. F1 scores computed for phrase pairs with different IoUs.

4. Conclusion

We proposed a gating mechanism for VGP identification. We observed that phrase features are often enough for lexically similar VGPs, but the visual features can improve the performance when phrase localization is accurate. Our gating mechanism learns to control the weights for each modality, and experimental results demonstrated the effectiveness of the proposed model.

References

[1] Chenhui Chu, Mayu Otani, and Yuta Nakashima. iParaphrasing: Extracting visually grounded paraphrases via an image. In *COLING*, pages 3479–3492, 2018.

[2] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, pages 2641–2649, 2015.

[3] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[4] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recoginition. In *ICLR*, pages 1–14, 2015.

[5] Zhou Yu, Jun Yu, Chenchao Xiang, Zhou Zhao, Qi Tian, and Dacheng Tao. Rethinking diversified and discriminative proposal generation for visual grounding. In *IJCAI*, pages 1114–1120, 2018.