公開

（事後評価・研究者用）

研 究 報 告 書

「視覚に基づく言い換えのセマンティック類型」

研究期間： 2019 年 4 月～2021 年 3 月
研究者番号： 50226
研 究 者： チョ シンキ

1． 研究のねらい

Visually grounded paraphrases (VGPs) are different phrasal expressions describing the same visual concept in an image. Our research until now treats VGP identification as a binary classification task, which ignores various phenomena behind VGPs. E.g., Figure 1 (a) "field hockey" and "lacrosse" are linguistic paraphrases; Figure 1 (g) "competitors" and "a group of bicyclist" describe the same visual concept from different aspects; however, these two pairs of VGPs have been treated equally, which is obvious undesirable. In the acceleration phase, we aim to **create typology of VGPs to elucidate the phenomena behind VGPs**. Studying typology for linguistic paraphrases is not new in natural language processing. The paraphrase typology focuses on lexicon, syntax and discourse phenomena in paraphrases. VGPs differ from paraphrases that they focus on the semantic aspects of images to describe the same visual concept. Therefore, we believe that the paraphrase typology is unsuitable for VGPs, and **propose to create VGP typology based on semantics**.

In **natural language inference**, seven basic semantic relations between two phrases have been defined, i.e., *equivalence, forward entailment, reverse entailment, negation, alternation, cover and independence*. **As VGPs describe the same concrete visual concept, *negation* and *cover* relations should not exist in VPGs.**

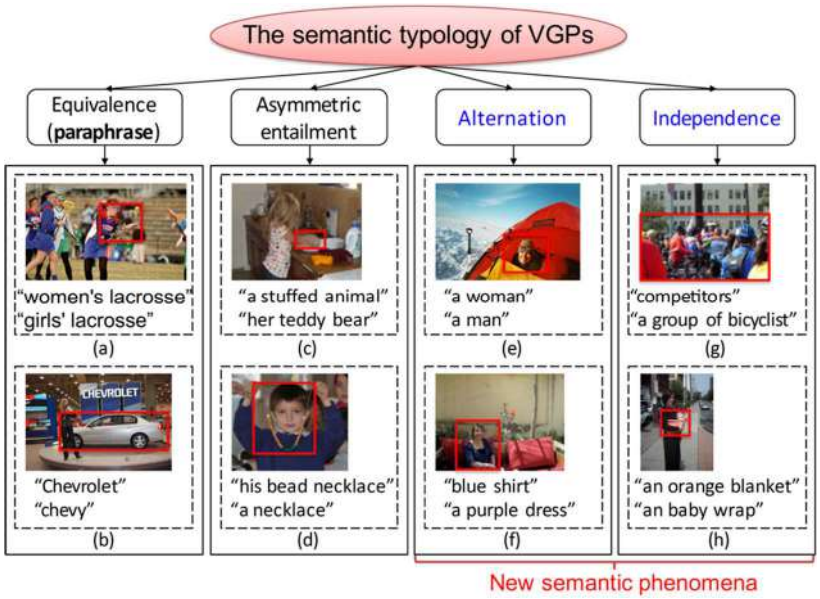

Figure 1: The semantic typology of VGPs.

Detailed analyses on the Flickr30k entities dataset indicate that the other five semantic relations cover VGPs. Figure 1 shows our semantic VGP typology (note that forward and reverse entailment are combined, but these can be further classified according to applications). Figure 1 (a) and (b) are linguistic paraphrases; Figure 1 (c) and (d) are entailment VGPs where one phrase

contains more fine object attribute description of the same concept, i.e., "teddy bear" and "bead." Knowledge such as "teddy bear *is a* stuffed animal" can be mined from this type. Figure 1 (e) and (f) use alternate phrases to describe the same visual concept, which may come from the difference in human recognition. The phrases in Figure 1 (g) and (h) are linguistically independent but become VGPs upon grounding. We believe that ***alternation* and *independence* VGPs are new semantic phenomena that cannot be explained only with language**, and visual inference is required to automatically identify them. More interestingly, when we look at a combination of two noun phrases, **the semantic relation can change from the sub-noun phrases**. E.g., "A woman speaking in front of a chevy" and "A woman promoting the Chevrolet" of Figure 1 (b), although both the sub-noun phrases are *equivalence*, the relation of the entire phrase pair is *independence*.

In the acceleration phase, **we plan to annotate the VGP typology on the Flickr30k entity dataset and design novel language and vision models to automatically classify VGPs**. The creation of the semantic VGP typology will not only elucidate the phenomena behind VGPs but also **open up novel ways of utilizing VPGs for various language and vision tasks, which require semantic understanding**. E.g., it can significantly boost the performance of textual entailment via visual grounding, which is a fundamental but very challenging natural language understanding problem. Understanding the semantic relation via visual grounding is also crucial towards machine reading, which is the main challenge in the Todai robot project. Therefore, we believe that this work will **significantly deepen and promote the research in both language and vision understanding**.

2. 研究成果

(1)概要

In summary, we mainly studied the following VGP topics in the acceleration phase.

1. **VGP identification improvement.** We improved the VGP identification model proposed in the ACT-I research period via a gated network and visual grounding. Our model can adaptively use visual and language features based on different VGPs.
2. **VGP typology annotation and classification.** Following the VGP typology explained in the previous section, we annotated a large-scale VGP typology dataset based on the original VGP dataset and designed a model for VGP classification.
3. **Cross-lingual visual grounding.** As the core technology of VGP, visual grounding has been studied for English only. We created a French visual grounding dataset and created a cross-lingual visual grounding model for French visual grounding via transferring knowledge from English.

Besides the core VGP research, we also studied its applications of visual question answering and human prediction. We published our work at international journals of Neurocomputing [1], IEEE Access [2], and international conferences of AAAI, WACV etc. [3-7]

（2）詳細

We detail the main VGP topics in the acceleration phase that we studied.

1. VGP identification improvement

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 (Figure 2) 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.

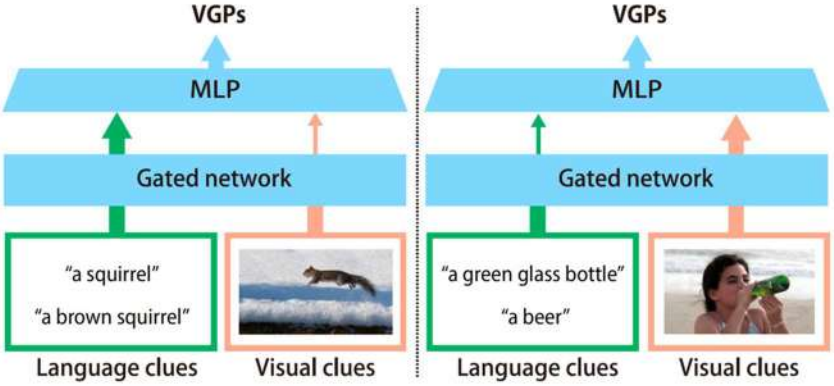

Figure 2: Our gated network for VGP identification.

2. VGP typology annotation and classification

Previous studies treat VGP identification as a binary classification task, which ignores various phenomena behind VGPs such as linguistic paraphrases and VGPs from different aspects. In this work, we propose semantic typology for VGPs, aiming to elucidate the VGP phenomena. We construct a large VGP dataset that annotates the class to which each VGP pair belongs according to our typology. In addition, we present a classification model that fuses language and visual features for VGP classification on our dataset (Figure 3). Experiments indicate that joint language and vision representation learning is important for VGP classification.

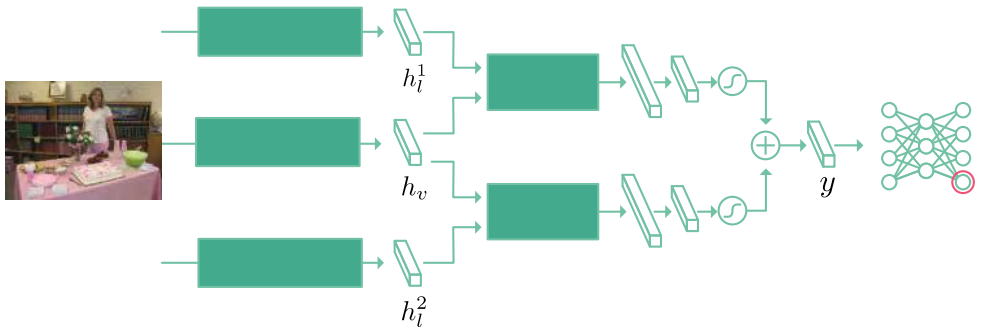

Figure 3: Our VGP classification model.

3. Cross-lingual visual grounding

Visual grounding is a vision and language understanding task aiming at locating a region in an image according to a specific query phrase. However, most previous studies only address this task for the English language. In this work, we present the first work on cross-lingual visual grounding to expand the task to different languages to study an effective yet efficient way for visual grounding on other languages. We construct a visual grounding dataset for French via crowdsourcing. Our dataset consists of 14k, 3k, and 3k query phrases with their corresponding image regions for 5k, 1k, and 1k training, validation and test images, respectively. In addition, we propose a cross-lingual visual grounding approach that transfers the knowledge from a learnt English model to a French model (Figure 4). Despite that the size of our French dataset is 1/6 of the English dataset, experiments indicate that our model achieves an accuracy of 65.17%, which is comparable to the accuracy 69.04% of the English model.

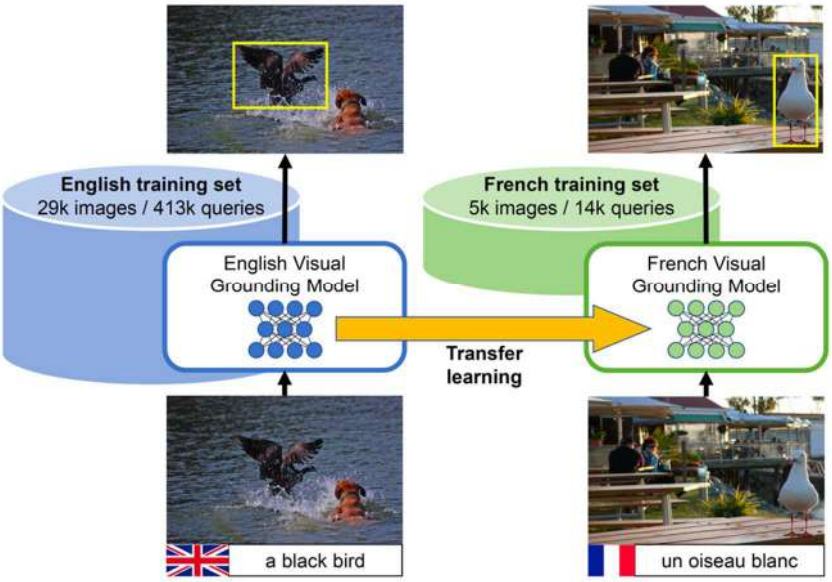

Figure 4: Cross-lingual visual grounding.

3. 今後の展開

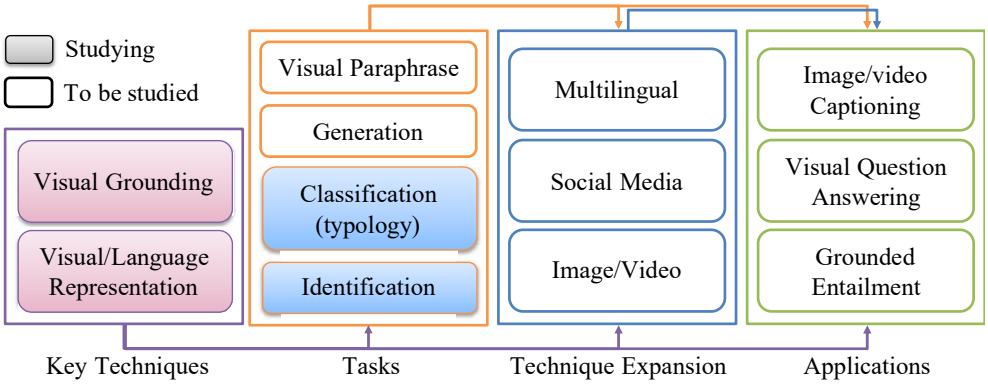

Figure 5: The VGP research flied that we aim to establish (arrows denote the technical hierarchy between each module category).

Although we originally created the concept of VGP and its typology in our JST ACT-I project, this is just **a start point**. There is a large amount of problems to be studied, by addressing which we can: **1) establish a research field of VGP; 2) establish vision and language processing techniques based on VGP understanding**; **3)** it has the potential to **elucidate how humans learn and process paraphrases in vision and language understanding**. Figure 5 summarizes these problems and shows the VGP research field that we aim to establish.

4. 自己評価

We have successfully achieved our research goal, which was creating typology for VGP. We further improved VGP identification and proposed cross-lingual visual grounding.

・研究の進め方（研究実施体制及び研究費執行状況）

Regarding to the research implementation system, because of the multimodality of the project, we closely collaborated with researchers in the computer vision (CV) field. We also employed RA students for creating VGP related datasets.

Regarding to the research funding execution status, we reasonably used it for purchasing GPU servers, dataset annotation, RA employment, and travel expenses. However, due to COVID-19, we have to apply for an extension of half a year for our project.

・研究成果の科学技術及び学術・産業・社会・文化への波及効果

VGP identification and classification that we have started in our ACT-I research period is a great start point for VGP research. VGP research is crucial for both language and vision understanding. We believe that our pioneering work on VGPs in the ACT-I research period has the potential to make VGPs a new language and vision research field in both NLP and CV fields. We are eager to promote the VGP research field with other researchers further.

・研究課題の独創性・挑戦性

We originally proposed the concept of VGPs, which is an achievement inspired by our study in

both NLP and CV. Due to the pioneer and multimodality characteristics, this work is very challenging. We overcame the problems and systematized VGP research via close collaboration between NLP and CV researchers.

5. 主な研究成果リスト

（1）論文（原著論文）発表

[1]. Wenjian Dong, Mayu Otani, Noa Garcia, Yuta Nakashima, Chenhui Chu. Cross-Lingual Visual Grounding. IEEE Access, (2021).

[2]. Mayu Otani, Chenhui Chu, Yuta Nakashima. Visually Grounded Paraphrase Identification via Gating and Phrase Localization. Neurocomputing, Volume 404, Pages 165-172, (2020).

（2）特許出願

研究期間累積件数：0 件

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

[3] Yuta Kayatani, Zekun Yang, Mayu Otani, Noa Garcia, Chenhui Chu, Yuta Nakashima, Haruo Takemura. The Laughing Machine: Predicting Humor in Video. In Proceedings of the IEEE 2021 Winter Conference on Applications of Computer Vision (WACV 2021), (2021).

[4] Zekun Yang, Noa Garcia, Chenhui Chu, Mayu Otani, Yuta Nakashima, Haruo Takemura. BERT Representations for Video Question Answering. In Proceedings of the IEEE 2020 Winter Conference on Applications of Computer Vision (WACV 2020), (2020).

[5] Noa Garcia, Mayu Otani, Chenhui Chu, Yuta Nakashima. KnowIT VQA: Answering Knowledge-Based Questions about Videos. In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2020), (2020).

[6] Mayu Otani, Chenhui Chu, Yuta Nakashima. Adaptive Gating Mechanism for Identifying Visually Grounded Paraphrases. ICCV 2019 MDALC Workshop, (2019).

[7] Chenhui Chu. Visually Grounded Paraphrase Identification. Third International Workshop on Symbolic-Neural Learning, (2019).