Attending Self-Attention: A Case Study of Visually Grounded Supervision in Vision-and-Language Transformers (Supplementary Material)

Jules Samaran[1] **Noa Garcia**[2]
Mayu Otani[3] **Chenhui Chu**[4] **Yuta Nakashima**[2]
[1]PSL Research University [2]Osaka University [3]CyberAgent, Inc. [4]Kyoto University
jules.samaran@mines-paristech.fr
{noagarcia, n-yuta}@ids.osaka-u.ac.jp
otani_mayu@cyberagent.co.jp chu@i.kyoto-u.ac.jp

1 Visually Grounded Paraphrases

A VGP pair, by definition, always refers to something in a given image and thus corresponds to a certain concrete concept; however, different expressions may, e.g., explain some different information or emphasize certain aspects that the concept has, as we can see in the above examples. Treating such VGPs equally may spoil semantics enriched by different expressions. There are five classes of VGPs: *equivalence, forward entailment, reverse entailment, alternation, and independence* (see Figure 1).

The **equivalence** relationship is present when phrase X and Y entails each other in both directions. In particular, they are linguistic paraphrases, such as different expressions with the same meaning or abbreviation of the other phrase (e.g., Chevrolet/chevy).

The **entailment** relationship is categorized into forward entailment and reverse entailment. Forward entailment relationship is present when phrase X is a subtype of phrase Y (e.g., a rock climb wall/a rock). Reverse entailment is to be given when Y is a subtype of X (e.g., a stuffed animal/her teddy bear). We treat them as different classes to store the entailment direction.

The **alternation** class applies when phrase X and phrase Y are mutually exclusive: the same visual concept cannot have X and Y being true at the same time. For example, some VGPs in alternation may describe the same visual concept with gender difference (e.g., a man/a woman) or age-related difference (e.g., baby/child).

In the **independence** relationship, phrase X and phrase Y describe different attributes of the same visual concept that are true at the same time (e.g., competitors/a group of bicyclist).

2 Visualization

Figure 2 displays the text-to-region in one attention head for every method (without fine-tuning on downstream tasks)[1] on two sample images. The Indirect method's attention is uniformly distributed over all image regions for every text token and the attention head doesn't display any kind of visual grounding pattern. On the other hand, we can see that in both the Direct and Semi-direct approach's attention heads the visually grounded entities attend primarily to their associated image regions. This visual grounding ability is even more accentuated for the Semi-direct method.

Looking also at a bird's eye view of the attention heads on examples can provide insights on the impact of different attention supervision methods. We can see for instance in Figure 3 that supervising indirectly the attention doesn't modify greatly the patterns in the attention heads. On the other hand, the directly supervised method displays a much more different pattern which is pretty uniform over attention heads, which could be explained by the fact that the exact same classification problem is supervised on every attention head in every layer. The semi-directly supervised method displays a much sparser pattern which hints at the fact that with this freer constraint than the one enforced by the direct approach, only some attention heads will attempt to align phrases with their visual grounding.

[1]Our motivation for visualization without fine-tuning on downstream tasks is to visualize the impact our fine-tuning would have generally on the model, hoping that it would improve the performance on several tasks.

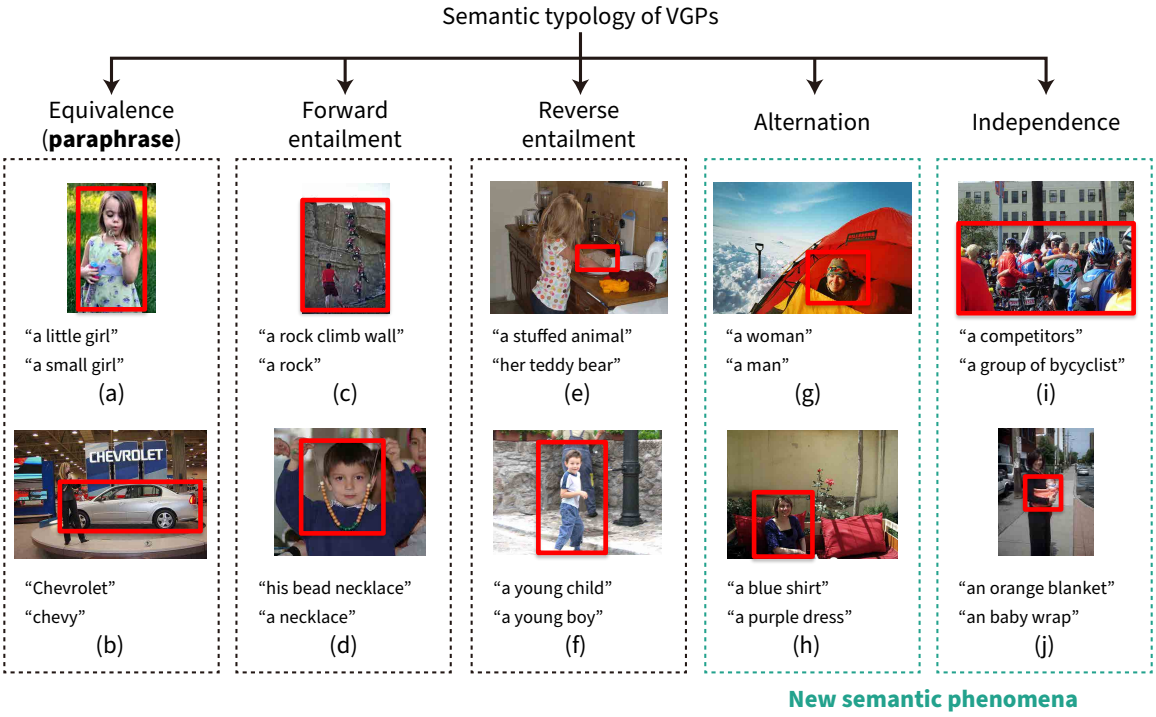
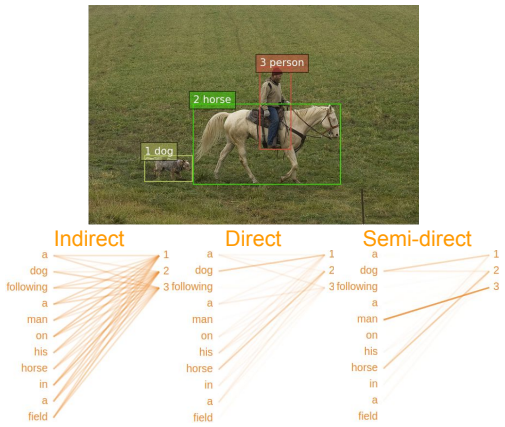
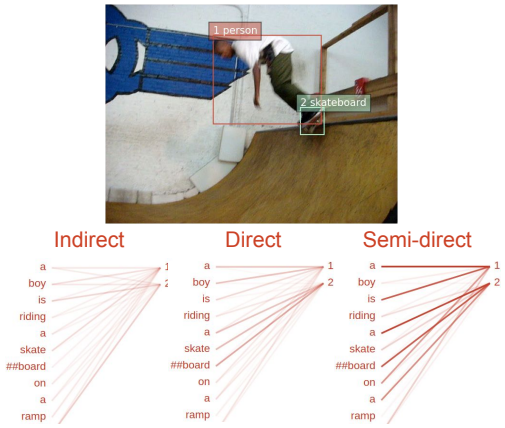


Figure 1: Semantic VGP typology examples. VGPs are categorized into 5 semantic relations. Equivalence is linguistic paraphrase. In addition, forward and reverse entailment, alternation and independence were introduced as relations for VGPs, which are not linguistic paraphrases but describing the same visual concepts.



(a) Layer 2, head 11.



(b) Layer 4, head 8.

Figure 2: Visualization of attention heads on sample images. Line intensity indicates the magnitude of attention probability.