GROOViST: A Metric for Grounding Objects in Visual Storytelling

Aditya K Surikuchi[1] | Sandro Pezzelle✦[1] | Raquel Fernández✦[1]
✦Institute for Logic, Language and Computation [1]University of Amsterdam

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

UvA

EMNLP 2023

Visual Storytelling

Input: sequence of images

Task: to generate a textual story consistent with the input

Human-annotated story: *It's parade day, and the whole town turns out to watch. There are those who serve our country, and the crowds cheer. There are the bands, and the music is loud but thankfully well performed. The flags are always fun to watch. And of course you get the old cars and their owners traveling through.*

Challenge: plausibility of several creative stories for a single given image sequence, makes reference-based NLG metrics (e.g., METEOR) inappropriate for the task.

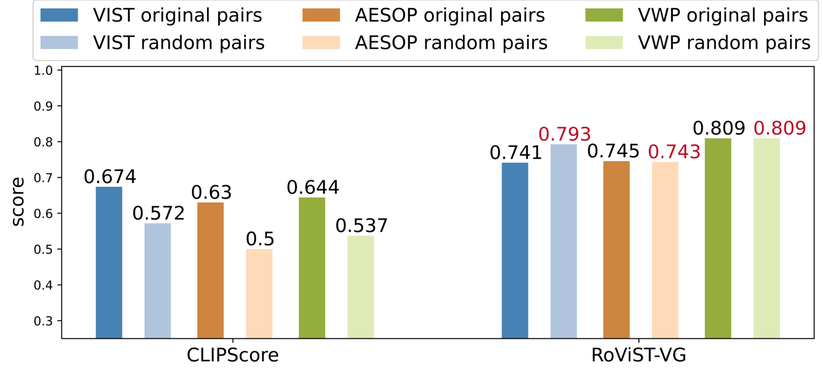
Previous approaches: Reference-free metrics

RoViST-VG:

1. Trains a model to contrastively learn relationships between nouns and image regions.
2. Evaluates of visual grounding of a story as the max cos similarity between extracted regions of the sequence (10 per image) and all the nouns in the story.

CLIPScore: average cos similarity between CLIP representations of each image of the sequence and its corresponding sentence in the story

RoViST-VG fails to distinguish original pairs from random ones on 3 different visual storytelling datasets!

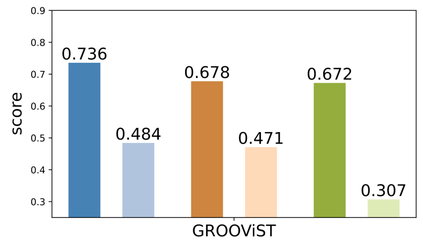


GROOViST

Informed by insights from both RoViST-VG and CLIPScore, we propose different components to address different aspects of visual grounding:

- **Noun phrases (NPs)** based on the prevalence of compound words (e.g., *parking lot*) in stories.
- **Vision-language alignment** between CLIP representations of extracted image regions and NPs.
- **Penalization of poorly grounded NPs** to accurately account for different degrees of grounding: negative values are attributed to NPs with scores below a pre-determined threshold (θ).
- **Concreteness weighting** to differentiate abstract words from concrete ones.

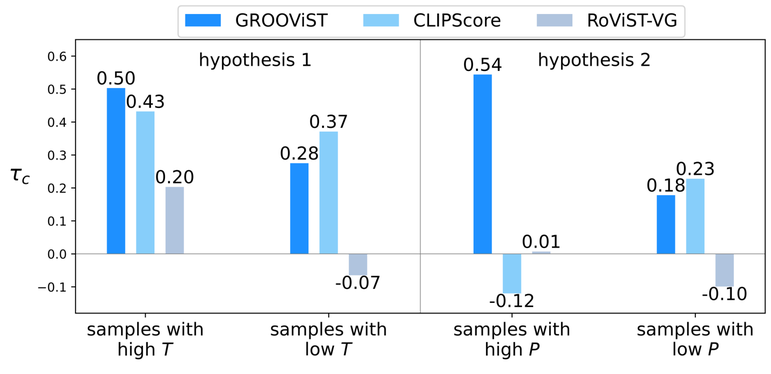
Overall grounding score is the sum of the final NP scores **normalized** by the total number of NPs.



1) it 's parade day , and the whole town turns out to watch . 2) there are those who serve our country , and the crowds cheer . 3) there are the bands , and the music is loud but thankfully well performed . 4) the flags are always fun to watch . 5) and of course you get the old cars and their owners traveling through .

GROOViST clearly distinguishes original *<image sequence, story>* pairs from random ones.

GROOViST correlates better with human scores on samples with high *temporal misalignment, T* (hypothesis 1) and on samples with high *proportion of NPs, P* (hypothesis 2):



$$t(sent_i) = \frac{\#(NPs\ matching\ img_{j!=i})}{\#(NPs\ in\ sent_i)}, \quad T(story) = \frac{1}{n}\sum_{i=1}^{n} t(sent_i)$$

$$P(story) = \frac{\#(NPs\ in\ story)}{\#(all\ words\ in\ story)}$$

Takeaways

1. Evaluating visual grounding is essential for the visual storytelling task and existing reference-free metrics have shortcomings.
2. We propose GROOViST — a modular and interpretable metric — for evaluating grounding in visual storytelling.
3. We show that GROOViST is well-aligned with human intuitions on grounding.