Zero-Shot Visual Grounding of Referring Utterances in Dialogue

Anonymous ACL submission

Abstract

This work explores whether current pretrained multimodal models, which are optimized to align images and captions, can be applied to the rather different domain of referring expressions. In particular, we test whether one such model, CLIP, is effective in capturing two main trends observed for referential chains uttered within a multimodal dialogue, i.e., that utterances become less descriptive over time while their discriminativeness remains unchanged. We show that CLIP captures both, which opens up the possibility to use these models for reference resolution and generation. Moreover, our analysis indicates a possible role for these architectures toward discovering the mechanisms employed by humans when referring to visual entities.

1 Introduction

During a conversation, speakers can refer to an entity (e.g., the girl in Fig. 1) multiple times within different contexts. This has been shown to lead to subsequent referring expressions that are usually shorter and based on the most communicatively effective words from the previous mentions (Krauss and Weinheimer, 1967; Brennan and Clark, 1996). This well known trend has been confirmed in recent vision-and-language (V&L) work (Shore and Skantze, 2018; Haber et al., 2019; Takmaz et al., 2020; Hawkins et al., 2020): referring utterances become more compact (i.e., less descriptive), and yet participants are able to identify the intended referent (i.e., they remain pragmatically informative).

Several approaches have tackled the generation of image captions from the perspective of pragmatic informativity (Mao et al., 2016; Luo et al., 2018; Cohn-Gordon et al., 2018; Schüz et al., 2021, i.a.) and Coppock et al. (2020) have compared the informativity of image captions and of referring expressions. However, no work to date has investigated how these two dimensions, *descriptiveness* and *discriminativeness* or pragmatic informativity, interact in referring expressions uttered in dialogue.

Figure 1: Referring utterance chain from PhotoBook (Haber et al., 2019). The chain has 4 ranks (4 references to the target image, in red outline). For simplicity, only the 5 distractor images from rank 1 are shown.

In this work, we use a transformer-based pretrained multimodal model to study the interplay between descriptiveness and discriminativeness in human referring utterances produced in dialogue. Due to their unprecedented success in numerous tasks, pretrained V&L models—such as LXMERT (Tan and Bansal, 2019), VisualBERT (Li et al., 2019), UNITER (Chen et al., 2020) and ALIGN (Jia et al., 2021)—have recently attracted a lot of interest aimed at understanding the properties and potential of their learned representations. This includes probing them in a zero-shot manner, i.e., without any specific fine-tuning, on some diagnostic tasks, e.g., image-text alignment or counting (Hendricks and Nematzadeh, 2021; Parcalabescu et al., 2021); quantifying, via input ablations, the role of each modality on the resulting multimodal representations and model performance (Frank et al., 2021); inspecting models' attention patterns when performing a specific task, i.e., visual coreference resolution (Cao et al., 2020); devising a unified experimental framework to compare various architectures fairly (Bugliarello et al., 2021).

Here, we focus on one model: Contrastive Language-Image Pre-training (CLIP, Radford et al., 2021), which has been shown to outperform several V&L models on zero-shot image-sentence alignment for object- and scene-level descriptions (Cafagna et al., 2021) and been proposed as a reference-free image caption evaluator (Hessel et al., 2021). However, CLIP's ability to encode discriminativeness in dialogue and to capture referring utterances in a zero-shot fashion has not yet been demonstrated. Here, we evaluate it on this capability for the first time, obtaining very promising results. This allows us to gain insight into both the strategies used by humans in sequential reference settings and CLIP's potential for reference resolution and generation.

2 Data

We focus on PhotoBook (PB; Haber et al., 2019), a dataset of multimodal task-oriented dialogues where players aim to pick the images they have in common without seeing each other's visual contexts (which consist of 6 images coming from the same domain). The game is played over several rounds in which the previously seen images reappear in different visual contexts, giving the players an opportunity to refer to such images again. As a result, *chains* of utterances referring to a single image are formed over the rounds as the players build common ground. See Fig. 1 for a simplified representation of a chain.[1] In total, PB consists of 2,500 games, 165K utterances, and 360 unique images from COCO (Lin et al., 2014).

All our experiments are conducted on a subset of 50 PB games with manually annotated referring utterances, which contains 364 referential chains about 205 unique target images. We refer to this subset as PB-GOLD.[2] Although a dataset of automatically-extracted chains using all PB data was recently made available (Takmaz et al., 2020), as reported by the authors these chains may contain errors. We therefore opt for using the relatively small but high-quality PB-GOLD subset since, as described in Sec. 3, we evaluate a pre-trained model without fine-tuning and hence do not need large amounts of data.

PB-GOLD's chains contain 1,078 utterances, i.e., 2.96 utterances per chain on average (min 1, max 4). We henceforth use the term 'rank' to refer to the position of an utterance in a chain. The average token length of utterances is 13.34, 11.03, 9.23, and 7.82, respectively, for ranks 1, 2, 3, and 4.[3] This decreasing trend, which is statistically significant at $p < 0.01$ with respect to independent samples t-tests between the ranks, is in line with the trend observed in the whole dataset (Haber et al., 2019). PB-GOLD's vocabulary consists of 926 tokens.

[1] Only 1 player's perspective for 1 context is represented.

[2] We use the gold set of the utterance-based chains v2 available at `https://dmg-photobook.github.io/`.

[3] We use TweetTokenizer: `https://www.nltk.org/api/nltk.tokenize.html`

3 Model

We use CLIP (Radford et al., 2021), a model pre-trained on a dataset of 400 million image-text pairs collected from the internet using a contrastive objective to learn strong transferable vision representations with natural language supervision.[4] In particular, we employ the ViT-B/32 version of CLIP, which utilizes separate transformers to encode vision and language (Vaswani et al., 2017; Dosovitskiy et al., 2021; Radford et al., 2019, 2021).

As the model learns to align images and texts, this enables zero-shot transfer to various V&L tasks such as image-text retrieval and image classification and even certain non-traditional tasks in a simple and efficient manner (Radford et al., 2019; Agarwal et al., 2021; Shen et al., 2021; Cafagna et al., 2021; Hessel et al., 2021). In this work, we freeze CLIP's weights and do not fine-tune the model or perform prompt engineering, since we aim to evaluate the model on referring utterances taken out of dialogue in a zero-shot setting.

4 Descriptiveness

In our first experiment, we investigate whether CLIP is effective in capturing the degree of descriptiveness exhibited by referring utterances in the PhotoBook game, i.e., the amount of information they provide about the image out of context. We consider each target image and corresponding referential utterance at a give rank *in isolation*, i.e., without taking into account the other competing images. We quantify descriptiveness as the alignment between an utterance and its image referent using `CLIPScore` (Hessel et al., 2021). For all the target image-utterance pairs in the chains of PB-GOLD, we use CLIP to obtain a vector t representing the utterance and a vector v representing the image. `CLIPScore` is then computed as the scaled cosine similarity between these two vectors, with range $[0, 2.5]$:[5]
$\texttt{CLIPScore}(t, v) = 2.5 * max(cos(t, v), 0)$.

We compute the average `CLIPScore` per rank over the whole PB-GOLD dataset.

[4] `https://github.com/openai/CLIP`

[5] The scaled factor was introduced by Hessel et al. (2021) to account for the relatively low observed cosine values.

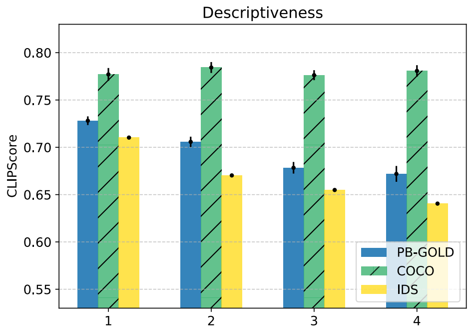

Figure 2: `CLIPScore` values for PB-GOLD, COCO and IDS. We only plot the first 4 'ranks' (x-axis) for COCO and IDS for comparability with PB-GOLD. The error bars illustrate the standard error.

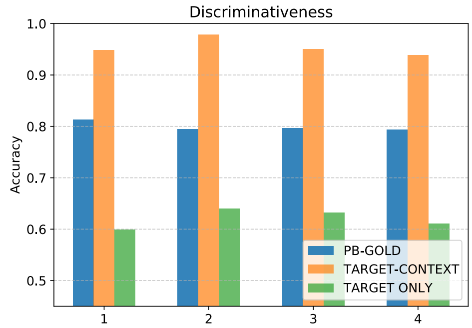

Figure 3: Reference resolution accuracy per rank with PB-GOLD utterances, word retrieved in context (TARGET−CONTEXT) and word retrieved from the image in isolation (TARGET ONLY).

Results. We find that earlier utterances are better aligned with the target image features and that there is a monotonically decreasing trend over the 4 ranks (Fig. 2, blue bars). The differences between all pairs of ranks are statistically significant (according to independent samples t-tests, $p < 0.01$), except for the comparison between the last 2 ranks ($p > 0.05$). Since earlier referring utterances tend to be longer (see Sec. 2), we check to what extent length may be a confounding factor. We find that there is only a weak correlation between token length and `CLIPScore` (Spearman's $\rho = 0.29, p < 0.001$).

Analysis. Our results indicate that CLIP is able to capture that earlier referring utterances contain more descriptive information about the target image than later referring utterances, and that this is only weakly related to length. We compare these results on PhotoBook with text-to-image alignment computed with the same method on two other datasets: (1) COCO (Lin et al., 2014),[6] which includes 5 captions per image provided independently by different annotators; here we do not expect to find significant differences in the level of descriptiveness across the captions, and (2) Image Description Sequences (IDS, Ilinykh et al., 2019)[7] where one participant describes an image incrementally, by progressively adding sentences with further details; here we do expect a similar pattern to PhotoBook, albeit for different reasons (because participants mention the most salient information at the beginning; Ilinykh et al., 2019).

Fig. 2 shows that these expectations are confirmed. According to CLIP, COCO captions (green bars) are more descriptive than IDS descriptions and PB referring utterances, and are equally aligned with the image across 'ranks' (the order is arbitrary in this case). In contrast, IDS incremental descriptions (yellow bars) are intrinsically ordered and show a significant decreasing trend.

Overall, these findings show that CLIP is effective as a *reference-free image caption evaluator*, as claimed by Hessel et al. (2021), as well as being able to capture the trends in sequential settings such as IDS and PB. However, this does not shed light on whether (and how) the model accounts for the degree of discriminativeness of a referring utterance in a given context, which is critical in PB. We explore this issue in the next section.

5 Discriminativeness

In order for a listener to select the target image among distractor images, a referring utterance should be discriminative in its visual context. Our results in the previous section show that descriptiveness decreases over time—what is the trend in discriminativeness when we encode the utterances in CLIP? To address this question, in our second experiment we investigate the use of CLIP from the perspective of reference resolution and generation.

We focus on local text-to-image alignment, ignoring the previous dialogue history. To this end, we feed CLIP a single referring utterance together with the visual context of the speaker who produced that utterance. CLIP yields softmax probabilities for each image contrasted with the single text. As a metric, we use accuracy: 1 if the target image gets the highest probability; 0 otherwise.

Results. The overall accuracy is 80.15%, which shows that CLIP performs well above the random baseline of 16.67%. In Fig. 3, we break down the

[6] We use the set of COCO images in PB-GOLD (N=205).

[7] The images are from ADE20k corpus (Zhou et al., 2017)

results per rank (blue bars). A 4×2 chi-square test (4 ranks vs. correct/incorrect) did not yield significant differences in accuracy between the ranks, $p > 0.05$. Thus, although descriptiveness decreases over time, discriminativeness is not significantly affected. Interestingly, an analysis of the entropy of the softmax distributions reveals that entropy increases monotonically over the ranks (this difference is statistically significant according to an independent samples t-test between ranks 1 and 4, $p < 0.01$). That is, the model is more uncertain when trying to resolve less descriptive utterances,[8] yet still performs remarkably well at this task.

Analysis. Our results show that CLIP is very effective in resolving referring utterances, even for later ranks where their form is more likely to rely on common ground established over the previous dialogue history, which we do not exploit in our setup. To better understand CLIP's abilities, we explore to what extent the model can *extract* what is discriminative in the images, which would provide a basis for using CLIP not only for resolution but also for referring expression generation.

We encode all the words in the vocabulary of PB-GOLD using CLIP. For each target image, we retrieve two words: the word whose representation is the closest to the features of the target image in isolation (TARGET ONLY); the word whose representation is the closest to the discriminative features of the target image in context (TARGET−CONTEXT). For the latter, we compute the discriminative features by average-pooling the visual representations of distractor images to end up with the mean context vector and then subtracting this vector from the visual representation of the target image.

To check whether these retrieved one-word utterances would be enough to identify the target image in context, we plug them into the CLIP-based reference resolution mechanism described earlier.[9] We observe very high resolution performance with the discriminative words (TARGET−CONTEXT; orange bars in Fig. 3), with all ranks achieving above 94% accuracy. The accuracies obtained from TARGET ONLY (green bars), however, are lower than those of the original utterances.

We also check whether at least one of the top-10 retrieved words are mentioned in the original human utterance: words retrieved in context are less frequently (59.83%) mentioned than the words retrieved for the image on its own (77.09%). As an illustration, the TARGET ONLY word retrieved for the example in Fig. 1 is *umbrella*, which is present in all the human utterances in this chain, although not discriminative. The TARGET−CONTEXT words retrieved are *beach*, *teal*, *blue*, and *beach* for ranks 1, 2, 3, and 4, respectively. As can be seen, the word is either present in the human utterance (*blue* in rank 3) or similar to other words mentioned (*teal* instead of *blue*, *beach* instead of *water*). Reference resolution succeeds with both the human utterances and the generated TARGET−CONTEXT words, but fails with the TARGET ONLY word.

6 Conclusion

We explored whether a pretrained multimodal model claimed to be a reference-free caption evaluator, CLIP (Radford et al., 2021), is effective in capturing two main trends observed for referential chains uttered within a multimodal dialogue, i.e., that (1) the utterances become less descriptive over time while (2) their discriminativeness remains unchanged. We showed that CLIP captures both, which sheds new light on the abilities of this model to deal with referential utterances besides standard image descriptions.

At the same time, the findings that CLIP can identify the correct referent without exploiting any dialogue history and that the retrieved TARGET ONLY words are more often used by the participants than the retrieved TARGET−CONTEXT words are intriguing, and suggest that participants playing the PhotoBook game (Haber et al., 2019) seek a trade-off between relying on contrastive and non-contrastive information. This could be due to perceptual salience, previously established conceptual pacts (Brennan and Clark, 1996), or to control referential entropy even though the discriminative utility of such information is not necessarily high (Rehrig et al., 2021; Tourtouri et al., 2018; Gatt et al., 2013). Interestingly, this opens up the possibility, parallel to the present work, to use CLIP to identify the mechanisms employed by humans when referring to visual entities. Moreover, future work could explore novel ways to incorporate the CLIP model or its representations into a reference resolution model embedding dialogue history and visual context.

[8] There is indeed a negative correlation between entropy and CLIPScore (Spearman's $\rho = -0.5, p < 0.001$).

[9] Note that since for resolution CLIP compares the word to the images one by one, this mechanism is independent from the subtraction method used to generate the TARGET−CONTEXT words.

References

Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. 2021. Evaluating clip: Towards characterization of broader capabilities and downstream implications.

Susan E. Brennan and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22:1482–1493.

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. 2021. Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language BERTs. *Transactions of the Association for Computational Linguistics*.

Michele Cafagna, Kees van Deemter, and Albert Gatt. 2021. What vision-language models 'see' when they see scenes. *ArXiv*, abs/2109.07301.

Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. 2020. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. *ECCV Spotlight*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer.

Reuben Cohn-Gordon, Noah Goodman, and Christopher Potts. 2018. Pragmatically informative image captioning with character-level inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 439–443, New Orleans, Louisiana. Association for Computational Linguistics.

Elizabeth Coppock, Danielle Dionne, Nathanial Graham, Elias Ganem, Shijie Zhao, Shawn Lin, Wenxing Liu, and Derry Wijaya. 2020. Informativity in image captions vs. referring expressions. In *Proceedings of the Probability and Meaning Conference (PaM 2020)*, pages 104–108, Gothenburg. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale.

Stella Frank, Emanuele Bugliarello, and Desmond Elliott. 2021. Vision-and-language or vision-for-language? On cross-modal influence in multimodal transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*. Association for Computational Linguistics.

Albert Gatt, Roger P.G. van Gompel, Kees van Deemter, and Emiel Krahmer. 2013. Are we Bayesian referring expression generators? In *Proceedings of the CogSci workshop on the production of referring expressions*, pages 1–6.

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. The PhotoBook dataset: Building common ground through visually-grounded dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910, Florence, Italy. Association for Computational Linguistics.

Robert Hawkins, Minae Kwon, Dorsa Sadigh, and Noah Goodman. 2020. Continual adaptation for efficient machine communication. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 408–419, Online. Association for Computational Linguistics.

Lisa Anne Hendricks and Aida Nematzadeh. 2021. Probing image-language transformers for verb understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3635–3644, Online. Association for Computational Linguistics.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. Tell me more: A dataset of visual scene description sequences. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 152–157, Tokyo, Japan. Association for Computational Linguistics.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*.

Robert M. Krauss and Sidney Weinheimer. 1967. Effect of referent similarity and communication mode on verbal encoding. *Journal of Verbal Learning & Verbal Behavior*, 6(3):359–363.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Ruotian Luo, Brian L. Price, Scott D. Cohen, and Gregory Shakhnarovich. 2018. Discriminability objective for training descriptive captions. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6964–6974.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.

Letitia Parcalabescu, Albert Gatt, Anette Frank, and Iacer Calixto. 2021. Seeing Past Words: Testing the Cross-Modal Capabilities of Pretrained V&L Models. In *Proceedings of the First Workshop on Multimodal Semantic Representations (MMSR)*, Groningen. To appear.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.

Gwendolyn Rehrig, Reese A. Cullimore, John M. Henderson, and Fernanda Ferreira. 2021. When more is more: redundant modifiers can facilitate visual search. *Cognitive Research: Principles and Implications*, 6:10.

Simeon Schüz, Ting Han, and Sina Zarrieß. 2021. Diversity as a by-product: Goal-oriented language generation leads to linguistic variation. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 411–422, Singapore and Online. Association for Computational Linguistics.

Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. 2021. How Much Can CLIP Benefit Vision-and-Language Tasks? *arXiv*, abs/2107.06383.

Todd Shore and Gabriel Skantze. 2018. Using lexical alignment and referring ability to address data sparsity in situated dialog reference resolution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2288–2297, Brussels, Belgium. Association for Computational Linguistics.

Ece Takmaz, Mario Giulianelli, Sandro Pezzelle, Arabella Sinclair, and Raquel Fernández. 2020. Refer, Reuse, Reduce: Generating Subsequent References in Visual and Conversational Contexts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4350–4368, Online. Association for Computational Linguistics.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Elli Tourtouri, Francesca Delogu, and Matthew W. Crocker. 2018. Specificity and entropy reduction in situated referential processing. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, pages 3356–3361.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene parsing through ade20k dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5122–5130.