Keeping an Eye on Context: Attention Allocation over Input Partitions in Referring Expression Generation

Simeon Schüz and **Sina Zarrieß**
Bielefeld University
{simeon.schuez,sina.zarriess}@uni-bielefeld.de

Abstract

In Referring Expression Generation, model inputs are often composed of different representations, including the visual properties of the intended referent, its relative position and size, and the visual context. Yet, the extent to which this information influences the generation process of black-box neural models is largely unclear. We investigate the relative weighting of target, location, and context information in the attention components of a Transformer-based generation model.[1] Our results show a general target bias, which, however, depends on the content of the generated expressions, pointing to interesting directions for future research.

1 Introduction

Context is crucial in multimodal language generation tasks such as Referring Expression Generation (REG), as descriptions for visible entities not only depend on their own appearance but also on their surroundings (e.g. Yu et al. 2016). For REG, this is especially evident, as the same expression can unambiguously describe an object in one context but be misleading in others (Schüz et al., 2023).

To this end, it has become a common practice to provide neural generation models for multimodal REG not only with visual representations for the target itself but also with information about its location and size and the visual context it appears in (see Figure 1). However, due to their black-box nature, it is not entirely clear to which extend state-of-the-art neural REG models take all of these representations into consideration. While ablation studies show how context information contribute to model performance (Yu et al., 2016; Zarrieß and Schlangen, 2018), they provide limited insight into how it is processed and to what extend it is relevant for e.g. lexical decisions.

[1]Code and models for this project are available at: https://github.com/clause-bielefeld/REG-Input-Partitions

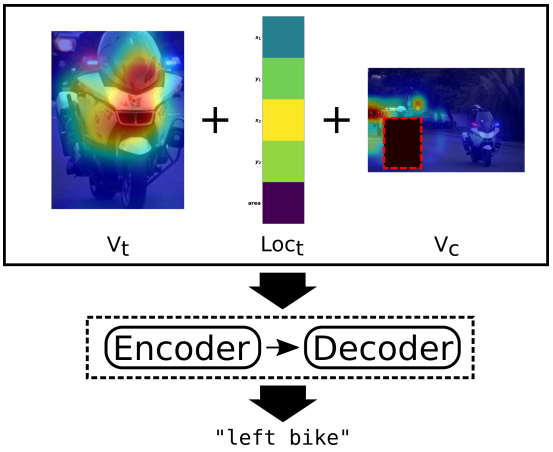


Figure 1: Input for our REG model. Input vectors are concatenations of visual (V_t) and location (Loc_t) features for targets and visual context features (V_c). We examine the relative attention weights of each partition.

Similar questions arise in other vision & language (V&L) tasks such as image captioning, where recent work has looked into analyzing the internal attention mechanisms of generation models (e.g. Ilinykh and Dobnik 2021, 2022). However, as the respective models usually take global images as input, analyses are mostly concerned with attention distribution *within* those representations rather than *across* different parts of the input. For REG, some authors use attention heatmaps as a method of model introspection (Tanaka et al., 2019; Liu et al., 2020; Sun et al., 2022), but without analyzing the patterns in detail.

As a first step for deeper investigations of how contextual information is processed in REG models, we quantify how attention is allocated over the partitioned inputs in a simple REG model. In more detail, we examine the model's relative focus on different parts of the input during inference, i.e. representations of the visual appearance of the referential target, its location in the image and the visual context it appears in. We analyze the attention weights on these input partitions both globally

Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023), pages 20–27

and for a subset of generated tokens to see if the weighting is affected by the expression content. To the best of our knowledge, no dedicated studies have yet been conducted on how attention is allocated across input partitions in REG models (but see Tanaka et al. 2019, who discuss this for a single example). Our results indicate that contextual information is utilized by the model in linguistically meaningful ways, highlighting promising directions for further research.

2 Background

Referring Expression Generation In REG, the goal is to generate descriptions for entities, which allow their identification in a given context (Reiter and Dale, 2000). Based on symbolic representations of objects in a given domain, classic work focused on rule-based approaches for determining *distinguishing* sets of attribute-value pairs, which identify a target by ruling out all other objects, cf. Krahmer and van Deemter 2012 for a survey.

In recent years, advances in neural modeling and vision and language corpora such as RefCOCO (Kazemzadeh et al., 2014) enabled REG set-ups based on real-world images. Neural REG models generally resemble architectures from e.g. image captioning, but are adapted in different ways to increase the discriminativeness of generated expressions (Schüz et al., 2023). This includes simulations of listener behaviour embedded in training objectives (Mao et al., 2016), comprehension modules (Luo and Shakhnarovich, 2017), reinforcement agents (Yu et al., 2017) or decoding strategies (Schüz and Zarrieß, 2021), but also supplementing model inputs with additional information.

For this, some works propose *visual comparisons* to encode differences in appearance between targets and context objects (Yu et al., 2016, 2017; Tanaka et al., 2019; Kim et al., 2020; Liu et al., 2020), whereas others directly use representations of the global image as context (Mao et al., 2016; Luo and Shakhnarovich, 2017; Zarrieß and Schlangen, 2018; Panagiaris et al., 2020, 2021). In addition to visual context, many approaches provide their models with the relative position and size of the target in the image (Mao et al., 2016; Yu et al., 2016, 2017; Liu et al., 2017; Luo and Shakhnarovich, 2017; Li and Jiang, 2018; Tanaka et al., 2019; Kim et al., 2020; Panagiaris et al., 2020; Liu et al., 2020). To be used as model inputs, different representations are usually concatenated, i.e. the inputs are composed of partitions of visual target and context features as well as location information.

Attention Analysis in V&L In recent years, attention mechanisms (Bahdanau et al., 2015; Luong et al., 2015) have become a cornerstone in generative V&L tasks like image captioning (Xu et al. 2015; Lu et al. 2016; Anderson et al. 2018; Herdade et al. 2019; Huang et al. 2019; Cornia et al. 2020; Pan et al. 2020, among many others, cf. Zohourianshahzadi and Kalita 2021). Despite some cautious remarks (Jain and Wallace, 2019), attention is used as a method for model introspection (e.g. Clark et al. 2019; Voita et al. 2019; Vig 2019 for text and Cao et al. 2020, 2022; Ilinykh and Dobnik 2021, 2022 for V&L settings). While recent REG approaches build on Transformer (Vaswani et al., 2017) architectures with attention as the key component (Panagiaris et al., 2021; Sun et al., 2022), the inner workings of the attention modules have only been studied in qualitative terms (Tanaka et al. 2019; Liu et al. 2020; Sun et al. 2022). Here, we perform a quantitative analysis of attention allocation in a simple Transformer-based REG model.

3 Experiments

3.1 Model

We implement a simple REG model which is based on an existing implementation for image captioning.[2] Following the general architecture in Vaswani et al. (2017), our model consists of transformer encoder and decoder and is largely comparable to the REG model in Panagiaris et al. (2021), but without self-critical sequence training and layer-wise connections between encoder and decoder. Unlike e.g. Mao et al. (2016), who enforce informativeness during training, we train our model with Cross Entropy Loss (cf. Limitations Section).

The model takes as input a concatenated feature vector $[V_t; Loc_t; V_c]$ where V_t is the visual representation of the target region, Loc_t is a vector of length 5 with the corner coordinates of the target bounding box and its area relative to the whole image, and V_c is the visual representation of the image context, i.e. the global image with the target bounding box masked out (cf. Figure 1). For both V_t and V_c the respective parts of the image are scaled to 224×224 resolution (keeping the original ratio and masking out the padding) and

[2] https://github.com/saahiluppal/catr

encoded with ResNet-101 (He et al., 2015), resulting in representations with 196 features (14×14) and embedding size 2048 for both target and context. Before being passed to the model encoder, V_t, Loc_t and V_c are concatenated into a sequence of 397 features ($196 + 5 + 196$). When generating an expression like *left bike* in Figure 1, we store one set of attention weights per input from the encoder self-attention component and one set per generated token from the decoder cross-attention component. In line with the input structure, the weights can be decomposed into partitions applying to V_t, Loc_t and V_c, with 196, 5 and 196 values, respectively.

3.2 Data

We use RefCOCO and RefCOCO+ (Kazemzadeh et al., 2014) for training and evaluation. Both are based on MSCOCO images (Lin et al., 2014) and contain references to the same objects, but the location attributes *left* and *right*, which are ubiquitous in RefCOCO, have been prohibited in RefCOCO+. The original test splits are separated for references to humans (*TestA*) and other objects (*TestB*). We combine both splits in our attention analysis but provide detailed results in the appendix.

3.3 Evaluation

Generation Quality To estimate the general generation capabilities of our models we rely on BLEU (Papineni et al., 2002), CIDEr (Vedantam et al., 2014) and METEOR (Banerjee and Lavie, 2005) as established metrics for automatic evaluation. However, we note that these metrics have been designed for other tasks and have limited utility for evaluating the overarching task objective in REG, i.e. identifying the intended referent.

Attention Allocation To examine attention allocation, we compare the summed attention weights directed to the target and its location and context for both the encoder (self-attention) and decoder (cross-attention) multi-head attention components. Importantly, we have different sample sizes for encoder and decoder, as the attention weights are calculated once per input for the encoder and once per inference step for the autoregressive decoder.

First, we compute α_t, α_l and α_c as the cumulative attention weights directed to V_t, Loc_t and V_c, respectively. For this, we calculate the sum of the attention weights assigned to each input partition, normalized such that $\alpha_t + \alpha_l + \alpha_c = 1$. As the dimensionality of V_l is considerably lower, we also report normalized scores α_{norm-t}, α_{norm-l} and α_{norm-c}, where we first divide the raw cumulative scores by the number of features in each partition.

Second, to investigate the respective influence of visual target and context features more closely, we quantify the attention difference between α_t and α_c as $\Delta_{t,c}$. As we exclude α_l here, we normalize the target and context scores such that $\alpha_t + \alpha_c = 1$. We then calculate $\Delta_{t,c} = \alpha_t - \alpha_c$, i.e. $0 < \Delta_{t,c} \leq 1$ if there is relative focus on the target, $-1 \leq \Delta_{t,c} < 0$ if there is relative focus on the context, and $\Delta_{t,c} = 0$ when both parts of the input are equally weighted. We also test whether target and context are weighted differently by the decoder when head or subordinate nouns are generated. In the referential noun phrases, head nouns generally represent the class label of the target and are therefore hypothesized to relate to V_t more clearly. To select nouns and determine NP heads in generated expressions, we rely on the POS tagger and dependency parser from the spaCy library.[3]

	$BLEU_1$	$BLEU_2$	CIDEr	METEOR
TestA	49.7	30.7	92.0	19.5
TestB	51.9	30.0	127.7	22.1
TestA+	24.4	13.6	80.3	12.3
TestB+	20.6	8.9	61.0	10.0

Table 1: Automatic Quality Metrics for RefCOCO TestA / TestB and RefCOCO+ TestA+ / TestB+.

	RefCOCO			RefCOCO+		
	α_t	α_l	α_c	α_t	α_l	α_c
Encoder	.55	.04	.40	.57	.01	.42
$\alpha_{norm-t/l/c}$	.21	.63	.16	.39	.32	.29
Decoder	.58	.03	.39	.64	.02	.34
$\alpha_{norm-t/l/c}$	.32	.49	.19	.42	.38	.20

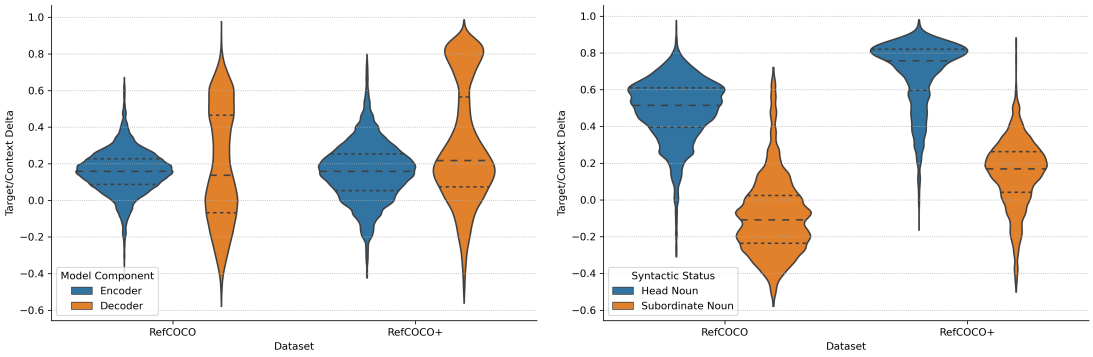
Table 2: Relative cumulative attention to the target (α_t), location (α_l) and context partitions (α_c) of inputs. $\alpha_{norm-t/l/c}$ scores are normalized by the respective dimensionality of V_t, Loc_t and V_c.

4 Results and Discussion

BLEU, CIDEr and METEOR are reported in Table 1. Although our scores fall below e.g. Panagiaris et al. (2021),[4] we note that our models are able to generate well-formed expressions with reasonable

[3] https://spacy.io/

[4] The discrepancies between RefCOCO and RefCOCO+ as well as the respective TestA and TestB splits can be also observed in other work on multimodal REG, cf. Section 2.



(a) $\Delta_{t,c}$ for Encoder and Decoder (all tokens)

(b) Decoder $\Delta_{t,c}$ for head and subordinate nouns

Figure 2: Target-Context Deltas ($\Delta_{t,c}$), inner lines denote quartiles. Left: Global distribution, averaged over all inputs (encoder) and generated tokens (decoder). Right: Decoder $\Delta_{t,c}$ for head and subordinate nouns.

	RefCOCO	RefCOCO+
Encoder	.16	.16
Decoder	.18	.30
Decoder / N_{head}	.49	.69
Decoder / N_{sub}	-.09	.15

Table 3: Mean target-context deltas ($\Delta_{t,c}$) for RefCOCO and RefCOCO+. *Decoder* / N_{head} and N_{sub} refer to deltas for generated NP-head or subordinated nouns.

similarity to the ground-truth annotations.

Table 2 shows the mean α_t, α_l and α_c scores for RefCOCO and RefCOCO+, as well as their dimension-normalized counterparts. Without normalizing for dimensionality, target features receive most attention in all cases, in line with the intuition that visual features of the referential target are the primary source of information for generating referring expressions. Loc_t features are consistently ranked last by a large margin. However, this changes if scores are normalized by their dimensionality: In this case, α_{norm-l} surpasses α_{norm-c} for both datasets and even α_{norm-t} for RefCOCO. The large differences between both datasets with respect to α_{norm-l} could be partly due to the location attributes *left* and *right*, which are highly common in RefCOCO but excluded in RefCOCO+.

Regarding the relative focus on target and context partitions, the mean $\Delta_{t,c}$ scores in Table 3 again show a general bias toward the target for both datasets, especially for decoder attention. Between datasets, the encoder scores are nearly identical, but the RefCOCO+ decoder is more biased to the target. Figure 2 (a) provides a more detailed view of the delta distribution: For both datasets, the median values of encoder and decoder are similar, but the decoder distributions extend more toward both extrema. This indicates that decoder attention shifts between target and context features during generation, possibly depending on which parts of the input are relevant for the generated tokens.

To investigate this in more detail, we report decoder deltas for generated nouns in Table 3 and Figure 2 (b), broken down by their syntactic status (head or subordinate noun in the referential noun phrase). For both datasets, mean scores and illustrations clearly show that there is a strong bias toward the target partition of the input when generating head nouns, supporting the intuition that information from this part of the inputs is particularly relevant for generating the class label for the referential target. However, the plots also reveal cases where context is given notable attention weight.

Overall, while generally biased towards the visual target features, our models allocate substantial attention weights to all input partitions. The increased target bias for head nouns and the high α_{norm-l} scores for RefCOCO indicate that attention allocation is sensitive to the respective relevance of different input partitions during the generation process. Importantly, as encoder and decoder are connected serially in our architecture, attention allocation in the decoder might be affected by attention biases in the encoder. Against this background, questions arise about the role of residual connections as well as layer-wise connections

between encoder and decoder in the style of Panagiaris et al. (2021) for processing contextual information, which we leave for future work.

5 Conclusion and Future Directions

In this paper, we investigated attention allocation across input partitions for a simple REG model. Our results show that the model attends to all sources of information, albeit with a general bias towards the target. In addition, our models show systematic differences between encoder and decoder attention across datasets, as well as sensitivity to the meaning of the generated tokens.

Importantly, this study only represents a small step toward a more thorough understanding of the significance of different types of information in REG. One limitation of this work is that our models are not explicitly optimized for the general objective of the REG task, i.e., unambiguously identifying the referential target (see Limitations Section). Consequently, as regarding for possible distractors is crucial for this, we see great potential in investigating how different approaches to increase the pragmatic informativeness of generated expressions (cf. Section 2) affect the relative weighting of input partitions. Along with this, given the multifaceted role of situational context for REG (Schüz et al., 2023), future work should take a closer look at attention allocation over semantic units in the visual context, e.g. to see whether objects with certain classes or relations to the target are weighted more or less during generation.

Limitations

We identify three main limitations in our study:

First, as our models are trained using Cross Entropy Loss as a target function, they have not been optimized for the general objective in the REG task and therefore may not be able to reflect certain pragmatic influences on attention allocation. We plan to address this weakness in future studies.

Second, the spaCy POS taggers and dependency parsers are prone to errors on RefCOCO and RefCOCO+ annotations, in part because they often do not consist of fully formulated sentences. This is a potential problem for selecting nouns from generated expressions and identifying the heads of noun phrases (although samples have shown that this still works reasonably well in practice).

Finally, as previous studies have found that large V&L models encode different kinds of information in different attention layers (e.g. Ilinykh and Dobnik 2021), a comprehensive investigation of attention allocation in Transformer-based generation models would require a comparison between different layers and heads of the multi-head attention modules. Likewise, further state-of-the-art explanation techniques for Transformer-based models (e.g. Abnar and Zuidema 2020; Mohebbi et al. 2023) could be instructive with respect to our research question. Due to time and space constraints, we leave this for future studies.

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online. Association for Computational Linguistics.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Feiqi Cao, Soyeon Caren Han, Siqu Long, Changwei Xu, and Josiah Poon. 2022. Understanding attention for vision-and-language tasks. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3438–3453, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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. In *Computer Vision – ECCV 2020*, pages 565–580. Springer International Publishing.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*,

pages 276–286, Florence, Italy. Association for Computational Linguistics.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition.

Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2019. *Image Captioning: Transforming Objects into Words*. Curran Associates Inc., Red Hook, NY, USA.

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE.

Nikolai Ilinykh and Simon Dobnik. 2021. What does a language-and-vision transformer see: The impact of semantic information on visual representations. *Frontiers in Artificial Intelligence*, 4.

Nikolai Ilinykh and Simon Dobnik. 2022. Attention as grounding: Exploring textual and cross-modal attention on entities and relations in language-and-vision transformer. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4062–4073, Dublin, Ireland. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar. Association for Computational Linguistics.

Jungjun Kim, Hanbin Ko, and Jialin Wu. 2020. CoNAN: A complementary neighboring-based attention network for referring expression generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1952–1962, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Emiel Krahmer and Kees van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Xiangyang Li and Shuqiang Jiang. 2018. Bundled object context for referring expressions. *IEEE Transactions on Multimedia*, 20(10):2749–2760.

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.

Jingyu Liu, Liang Wang, and Ming-Hsuan Yang. 2017. Referring expression generation and comprehension via attributes. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE.

Jingyu Liu, Wei Wang, Liang Wang, and Ming-Hsuan Yang. 2020. Attribute-guided attention for referring expression generation and comprehension. *IEEE Transactions on Image Processing*, 29:5244–5258.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2016. Knowing when to look: Adaptive attention via a visual sentinel for image captioning.

R. Luo and Gregory Shakhnarovich. 2017. Comprehension-guided referring expressions. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3125–3134.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Junhua Mao, J. Huang, A. Toshev, Oana-Maria Camburu, A. Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11–20.

Hosein Mohebbi, Willem Zuidema, Grzegorz Chrupała, and Afra Alishahi. 2023. Quantifying context mixing in transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3378–3400, Dubrovnik, Croatia. Association for Computational Linguistics.

Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. 2020. X-linear attention networks for image captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Nikolaos Panagiaris, Emma Hart, and Dimitra Gkatzia. 2020. Improving the naturalness and diversity of referring expression generation models using minimum risk training. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 41–51, Dublin, Ireland. Association for Computational Linguistics.

Nikolaos Panagiaris, Emma Hart, and Dimitra Gkatzia. 2021. Generating unambiguous and diverse referring expressions. *Computer Speech & Language*, 68:101184.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ehud Reiter and Robert Dale. 2000. *Building natural language generation systems*. Cambridge University Press, Cambridge, U.K. New York.

Simeon Schüz and Sina Zarrieß. 2021. Decoupling pragmatics: Discriminative decoding for referring expression generation. In *Proceedings of the Reasoning and Interaction Conference (ReInAct 2021)*, pages 47–52, Gothenburg, Sweden. Association for Computational Linguistics.

Simeon Schüz, Albert Gatt, and Sina Zarrieß. 2023. Rethinking symbolic and visual context in referring expression generation. *Frontiers in Artificial Intelligence*, 6.

Mengyang Sun, Wei Suo, Peng Wang, Yanning Zhang, and Qi Wu. 2022. A proposal-free one-stage framework for referring expression comprehension and generation via dense cross-attention. *IEEE Transactions on Multimedia*, pages 1–1.

M. Tanaka, Takayuki Itamochi, K. Narioka, Ikuro Sato, Y. Ushiku, and T. Harada. 2019. Generating easy-to-understand referring expressions for target identifications. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5793–5802.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2014. Cider: Consensus-based image description evaluation.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy. Association for Computational Linguistics.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. 37:2048–2057.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling context in referring expressions. In *Computer Vision – ECCV 2016*, pages 69–85, Cham. Springer International Publishing.

Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. 2017. A joint speaker-listener-reinforcer model for referring expressions. In *Computer Vision and Pattern Recognition (CVPR)*, volume 2.

Sina Zarrieß and David Schlangen. 2018. Decoding strategies for neural referring expression generation. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 503–512, Tilburg University, The Netherlands. Association for Computational Linguistics.

Zanyar Zohourianshahzadi and Jugal K. Kalita. 2021. Neural attention for image captioning: review of outstanding methods. *Artificial Intelligence Review*, 55(5):3833–3862.

A Appendix

	TestA	TestB	TestA+	TestB+
Encoder	.19	.12	.22	.09
Decoder	.16	.21	.34	.26
Decoder / N_{head}	.50	.48	.69	.69
Decoder / N_{sub}	-.07	-.10	.22	.07

Table 4: Mean target-context deltas ($\Delta_{t,c}$) for RefCOCO and RefCOCO+. *Decoder* / N_{head} and N_{sub} refer to the deltas when generating NP-head or subordinated nouns.

A.1 Data, Implementation and Training Details

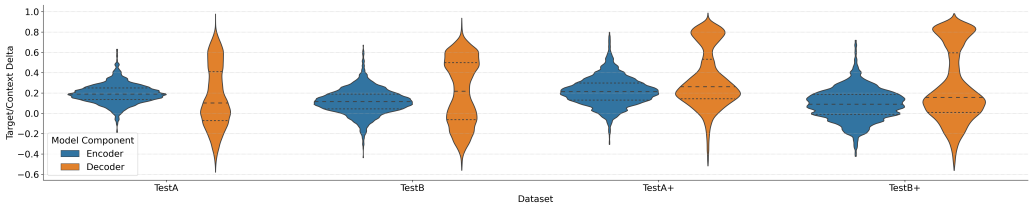
As described in Section 3, we trained our models on RefCOCO and RefCOCO+.[5] Both datasets consist of about 140k expressions for 50k objects in 20k images, out of which 42k objects in 17k images are assigned to the train splits.

Our model configurations for RefCOCO and RefCOCO+ are identical: In both cases, the model has 6 encoder and 6 decoder layers with 8 attention heads, a hidden dimension of 256, a feedforward dimension of 2048, and a total of $\sim$ 84,000,000 parameters. Our initial learning rate is set to 0.0001 for the transformer encoder and decoder, and 0.00001 for the pre-trained ResNet-101 backbone. We trained our models on an Nvidia RTX A40. The RefCOCO model was trained for 5 and the RefCOCO+ model for 7 epochs, with each epoch lasting approximately 2 hours.

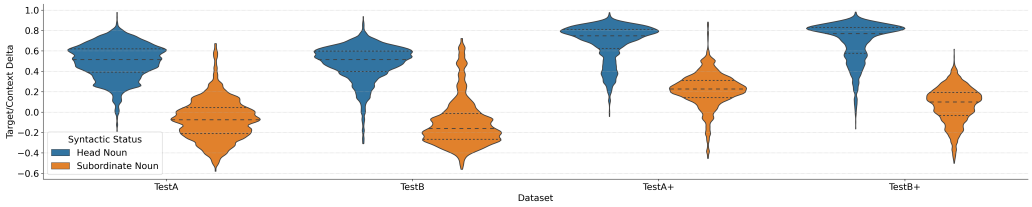
[5] https://github.com/lichengunc/refer

	TestA			TestB			TestA+			TestB+		
Encoder	.57	.05	.38	.54	.04	.42	.60	.02	.38	.54	.01	.45
$\alpha_{norm-t/l/c}$	.21	.65	.14	.22	.60	.18	.37	.40	.24	.42	.22	.35
Decoder	.56	.04	.40	.59	.02	.38	.65	.02	.32	.62	.02	.36
$\alpha_{norm-t/l/c}$	.28	.54	.18	.36	.44	.20	.40	.42	.18	.44	.34	.23

Table 5: Relative cumulative attention to the target (α_t), location (α_l) and context partitions (α_c) of inputs. $\alpha_{norm-t/l/c}$ scores are normalized by the respective dimensionality of V_t, Loc_t and V_c.



(a) $\Delta_{t,c}$ for Encoder and Decoder (all tokens)



(b) Decoder $\Delta_{t,c}$ for head and subordinate nouns

Figure 3: Target-Context Deltas ($\Delta_{t,c}$) for all splits in RefCOCO (TestA / TestB) and RefCOCO+ (TestA+, TestB+), inner lines denote quartiles. Top: Global distribution, averaged over all inputs (encoder) and generated tokens (decoder). Bottom: Decoder $\Delta_{t,c}$ for head and subordinate nouns.

A.2 Results for Individual Test Splits

In our attention analysis, we have combined the TestA and TestB splits of RefCOCO and RefCOCO+ for greater clarity. Table 5 shows the α_t, α_l and α_c scores as well as their normalized counterparts for each test split. In Table 4 we report and visualize target context deltas ($\Delta_{t,c}$) for the individual test splits (visualized in Figure 3).

These results are largely consistent with the findings in Section 4: Visual target features receive the most attention, but location features score high when normalized by partition size. With respect to target-context deltas, decoder attention exhibits greater variance, and there is strong target bias for generated head nouns.