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

Simeon Schüz and Sina Zarrieß

Bielefeld University

12 September 2023

Scope of this work

- Task: **Referring Expression Generation (REG)**
- We explore **attention allocation over input partitions**
  - Typically, model inputs are **concatenations of different representations**, e.g. visual properties of target and context, and location of the target
  - Question: To which extent does our model focus on each of those?

Outline

1. Background
2. Experimental Setup
3. Results and Discussion
4. Future Directions

Background

The REG task

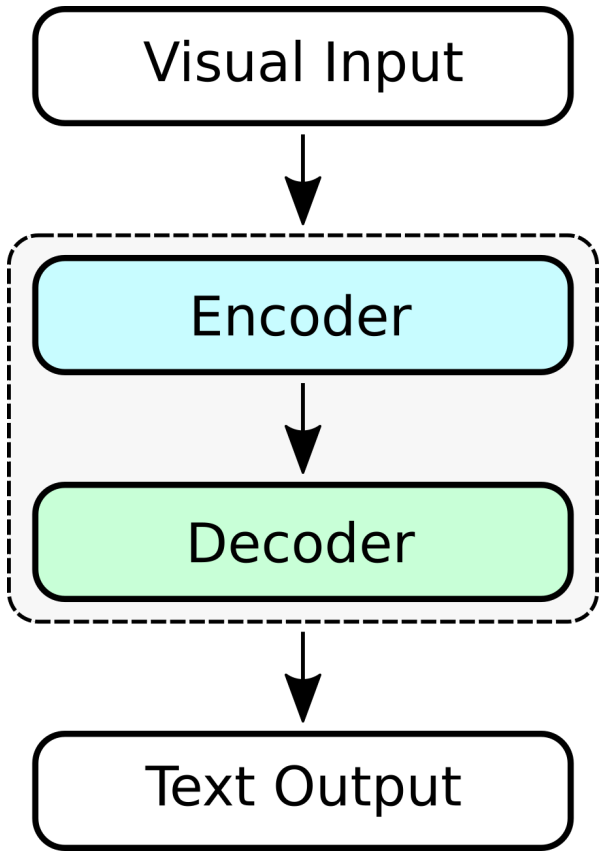
- **Referring Expression Generation (REG)**: Long-standing field of interest in computational linguistics
- Goal: Generating expressions for entities, which allow their identification in context
- Here: REG as a Vision & Language task (V&L) task: expressions for objects in photographs
- Context is crucial (to ensure discriminativeness, but also e.g. for relational descriptions)

Possible REs for target (green):

“left bike”, “cop bike closest to car”, . . .

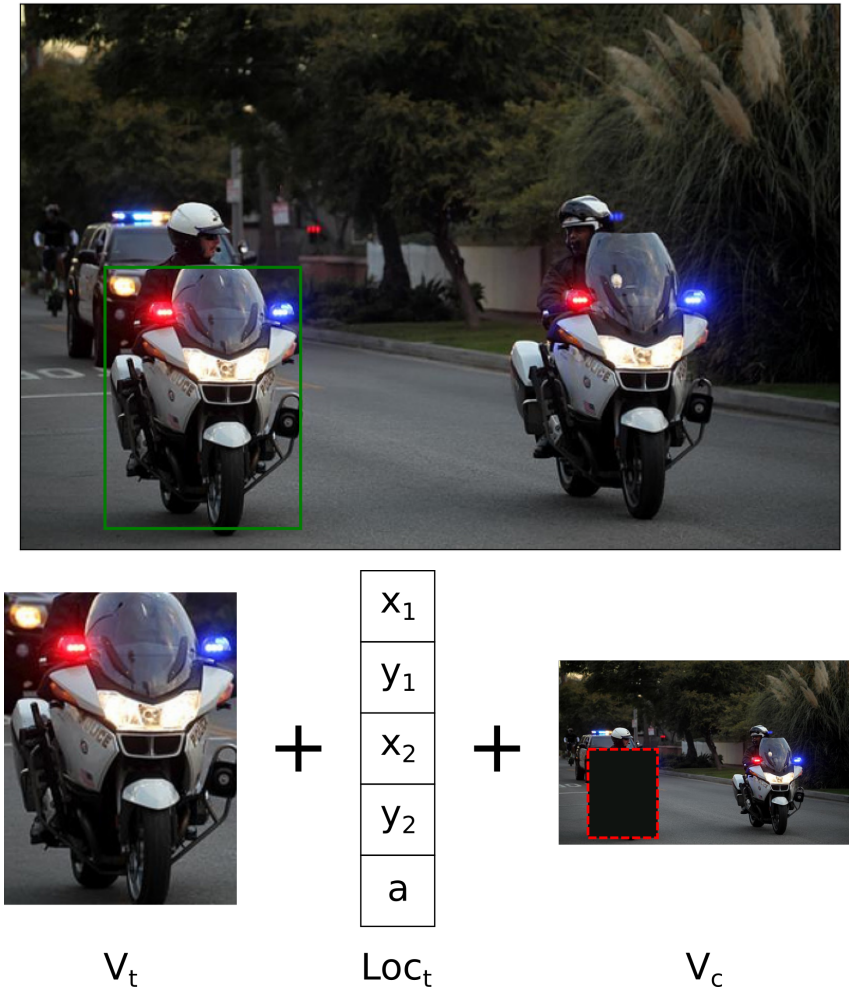
Visual REG Models

- At their core: Similar to image captioning
  - Encoder processes visual information, Decoder autoregressively generates descriptions
  - e.g. based on Transformer
- adaptations to pragmatic aspects of the task
  - $\rightarrow$ generate more discriminative descriptions
  - contrastive loss functions, internal listener modules, decoding methods, …
- more complex input representations
  - $\rightarrow$ account for referential targets **and** their context



Model Inputs

- combinations of (structurally) different representations:
  1. visual representations of the target bounding box (V_t)
  2. visual representations of the context or global image (V_c)
  3. location and size of the target (Loc_t)
- common: combination via concatenation
  - also more sophisticated techniques, e.g. Yu et al. 2016



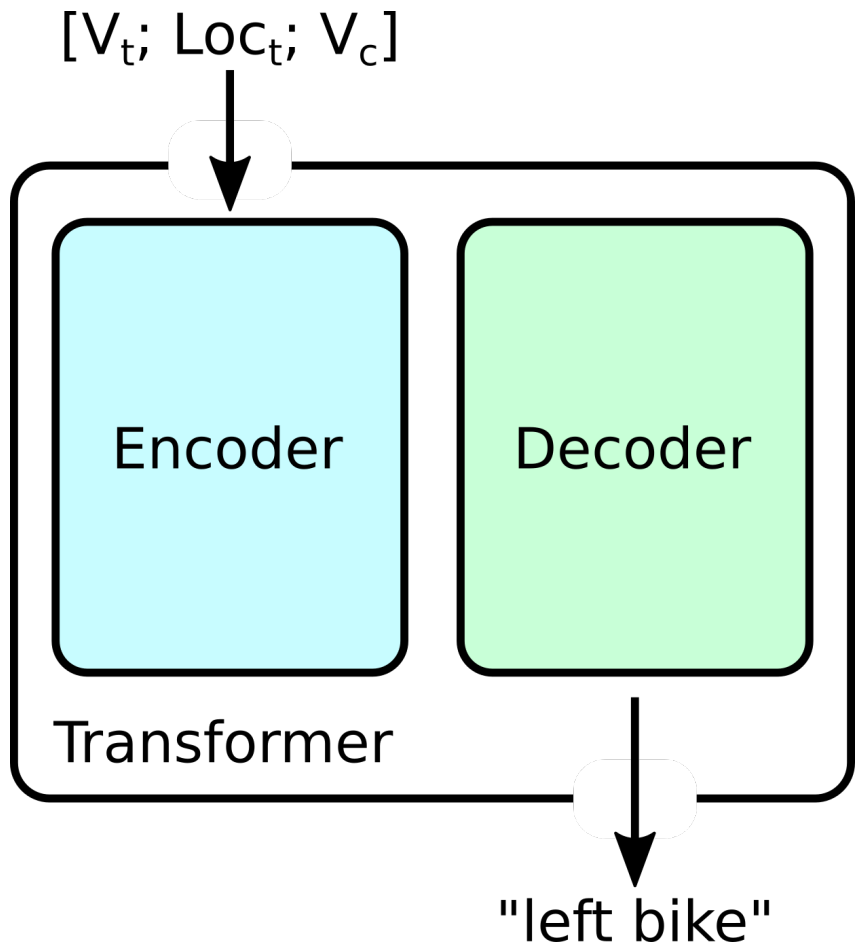
Model Inputs

- ablation studies: context and location features contribute to model performance (e.g. Yu et al. 2016; Zarrieß and Schlangen 2018)
- but still open questions:
  - are they equally relevant in all cases?
  - which function do they serve? (cf. Schüz et al. 2023)
  - how are they processed by the model?
- As a starting point: **Analyzing model attention over different input partitions** ($\rightarrow$ kinds of representations)

Experimental Setup

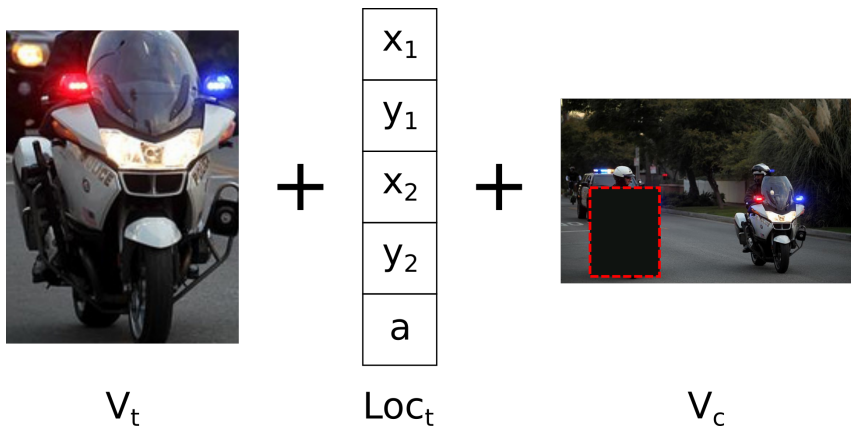
Our Model

- Standard Encoder-Decoder Transformer Model
  - Encoder processes the concatenated features and passes intermediate representation to Decoder (with same general structure, cf. residual connections)
- similar to the REG model in Panagiaris et al. 2021 (but without e.g. self-critical learning)
- Trained with Cross-Entropy Loss on RefCOCO and RefCOCO+



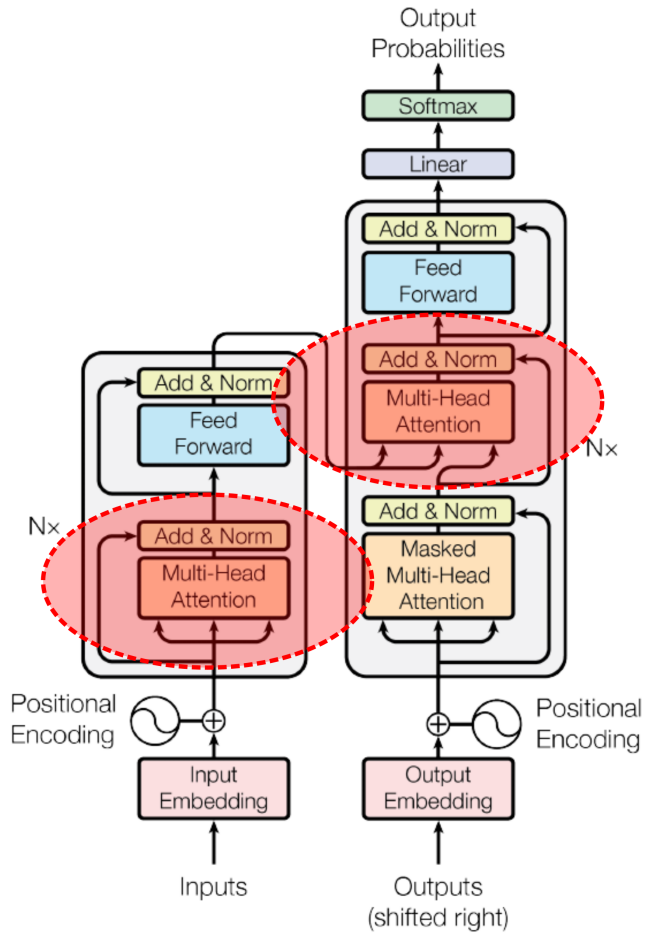
Input Features

- V_t: ResNet encodings / target bounding box content ($dim = 196$)
- V_c: ResNet encodings / full image (target masked out; $dim = 196$)
- Loc_t: location and size of target bounding box (relative to full image; $dim = 5$)
- concatenated as $[V_t; V_c; Loc_t]$ and passed to Transformer Encoder



Attention in our Model

- attention weights over concatenated feature vector:
  - Encoder Self-Attention (left; computed once per input)
  - Decoder Cross-Attention (right; computed once per generated token)



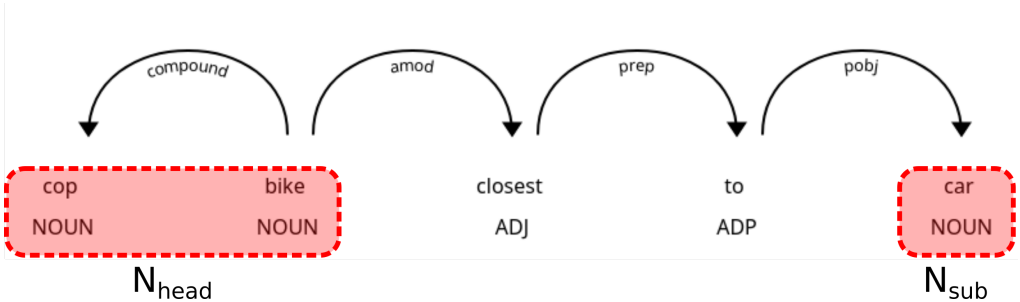
(Vaswani et al., 2017)

Metrics and Measurements

1. CIDEr, BLEU, METEOR
   - as sanity check
2. α_t, α_l, α_c: Cumulative attention weights to V_t, Loc_t, V_c
   - all layers / heads
   - normalized such that $\alpha_t + \alpha_l + \alpha_c = 1$
   - also normalized variants to account for low dimensionality of Loc_t

Metrics and Measurements

3. $\Delta_{t,c}$: Attention delta between target and context
   - $\Delta_{t,c} > 0$ target is focused; $\Delta_{t,c} < 0$ if context is focused
   - also decoder $\Delta_{t,c}$ restricted to nouns, syntactic status in referential NP as variable (based on the spaCy dependency parser)
     - Head Nouns: Relate to the target
     - Subordinate Nouns: Relate to other objects (in context)

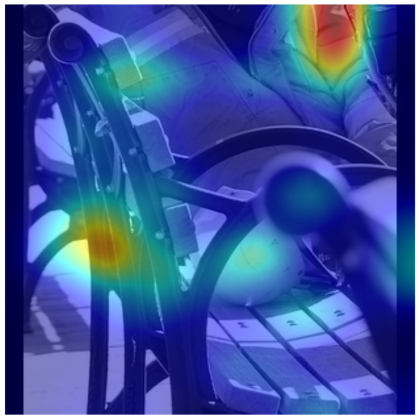


Results

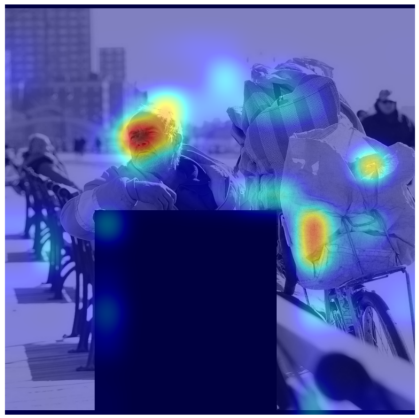
Example

"chair in front of man"

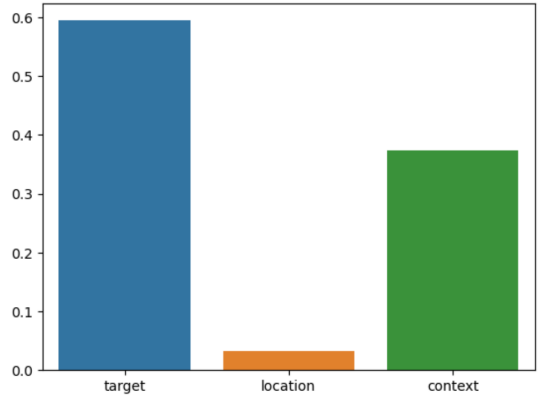
Example – Encoder Attention



V_t



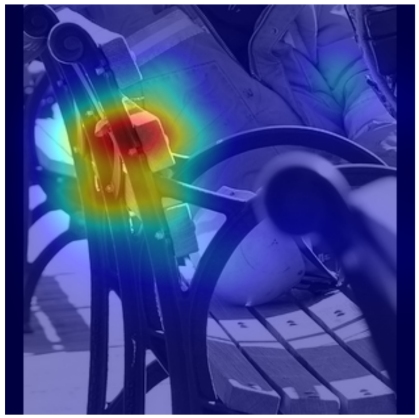
V_c



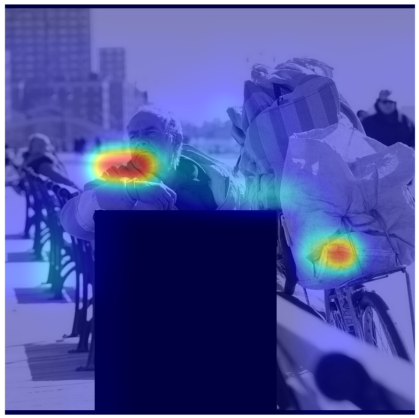
α scores

"chair in front of man"

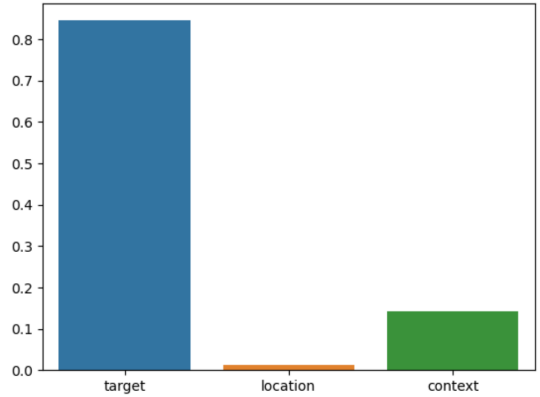
Example – Decoder Attention



V_t



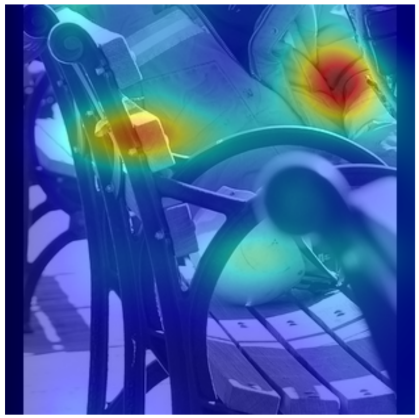
V_c



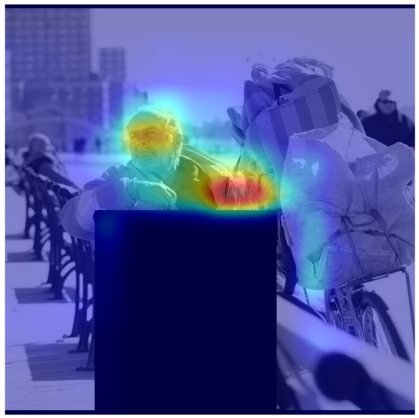
α scores

"**chair** in front of man"

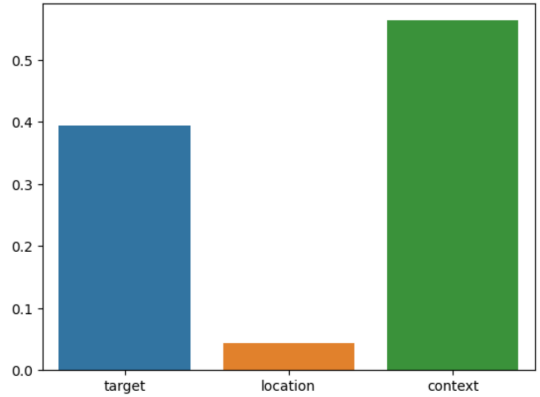
Example – Decoder Attention



V_t



V_c



α scores

"chair in front of **man**"

N-Gram Metrics

	$BLEU_1$	$BLEU_2$	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

- indicate decent generation capabilities

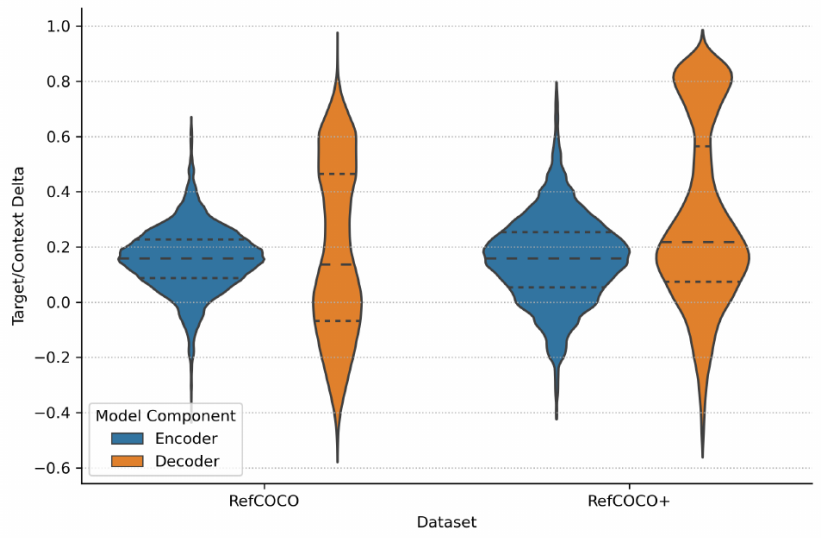
Cumulative Attention Weights

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

- for (combined) RefCOCO and RefCOCO+ test splits
- target bias: visual target receives most attention in all cases
  - arguably most important information for truthful descriptions
- α_c also considerably high, low α_l scores
- when normalized for dimensionality:
  - $\alpha_{norm-l} > \alpha_{norm-c}$ in all cases
  - $\alpha_{norm-l} > \alpha_{norm-t}$ for RefCOCO $\rightarrow$ high frequency of location attributes

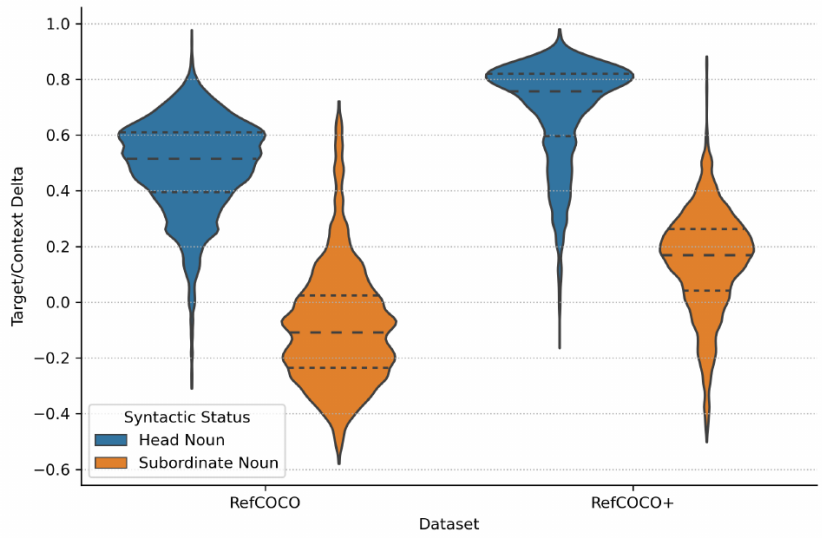
Attention Deltas

- for (combined) RefCOCO and RefCOCO+ test splits
- again: target bias
- differences between encoder and decoder
  - encoder deltas roughly normally distributed (with median ≈ 1.7)
  - higher variation in decoder deltas, including many cases with strong focus on target



Attention Deltas / Head and Subordinate Nouns

- attention for head nouns shifted towards target partition
  - in line with assumption: target features most important for determining type of referent
- but:
  - standard dependency parsers do not work very well on RefCOCO(+) expressions
  - head noun almost always at beginning of sentences $\rightarrow$ positional or semantic / syntactic effect?



Summary

- implemented a simple transformer-based REG model
- measured attention allocation over partitions of the model input
  - Visual target features; target location; visual context features
- general target bias, but also attention on context / location partitions
  - depending on properties of generated tokens and ground-truth annotations

Future Directions

- refining analysis methodology
- extending to more SOTA models / architectures with pragmatic components
- zoom in on visual context $\rightarrow$ which (kinds of) objects are attended by the model?

Citations

- Panagiaris, Nikolaos, Emma Hart, and Dimitra Gkatzia (2021). “Generating unambiguous and diverse referring expressions”. In: *Computer Speech & Language* 68, p. 101184.
- Schüz, Simeon, Albert Gatt, and Sina Zarrieß (2023). “Rethinking symbolic and visual context in Referring Expression Generation”. In: *Frontiers in Artificial Intelligence* 6.
- Vaswani, Ashish et al. (2017). “Attention is All you Need”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc.
- Yu, Licheng et al. (2016). “Modeling Context in Referring Expressions”. In: *Computer Vision – ECCV 2016*. Ed. by Bastian Leibe et al. Cham: Springer International Publishing, pp. 69–85.
- Zarrieß, Sina and David Schlangen (Nov. 2018). “Decoding Strategies for Neural Referring Expression Generation”. In: *Proceedings of the 11th International Conference on Natural Language Generation*. Tilburg University, The Netherlands: Association for Computational Linguistics, pp. 503–512.