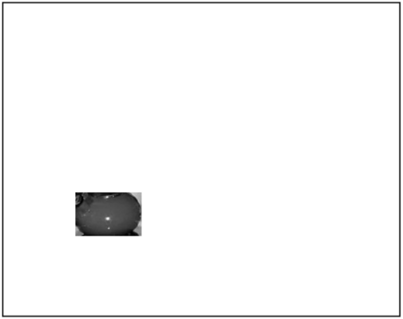
Resilience through Scene Context in Visual Referring Expression Generation

Simeon Junker and Sina Zarrieß

Bielefeld University

26 September 2024



(a)

(from Galleguillos and Belongie 2010)

(from Galleguillos and Belongie 2010)

(from Galleguillos and Belongie 2010)

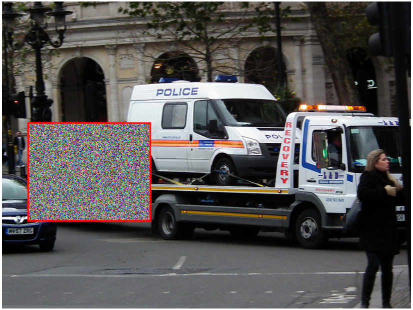
- Visual objects commonly appear in **typical surroundings** with other **related objects**
- **Scene context** helps us to process the visual world, e.g. recognize objects more quickly and reliably

- **V&L systems** also often process “real-world” scenes
  - **visual REG**: Objects in Photographs

Example from RefCOCO (Kazemzadeh et al., 2014)

- Often lots of relations between target and context!

- **V&L systems** also often process “real-world” scenes
  - **visual REG**: Objects in Photographs



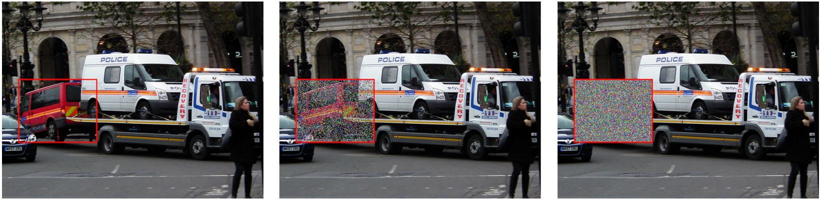
Example from RefCOCO (Kazemzadeh et al., 2014)

- Often lots of relations between target and context!

Do REG systems exploit Scene Context in similar ways?

Experimental Setup

- Question: Does scene context help REG systems to process target objects, if they are not clearly seen?
- Method: Train and test **REG systems with and without scene context** with target representations **obscured with varying degrees of random noise**



- Expectation:
  - Model performance degrades with increasing noise
  - **exploiting context mitigates the loss**

Experimental Setup / Models

Variants of two Transformer-based systems:

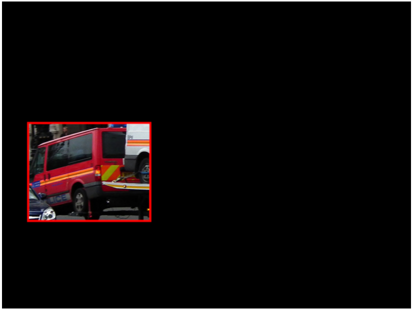
1. **TRF**: Standard Transformer (similar to Panagiaris et al. 2021)
   - ResNet as visual backbone
2. **CC**: *ClipCap* captioning model (Mokady et al., 2021) applied to the REG task
   - CLIP as visual backbone, with pre-trained GPT-2

   Here: **Only discuss TRF results**

Experimental Setup / Models

TRF_{tgt}: Target-only

- Target, but no context features
- Input: $[\mathbf{V}_t; \mathbf{Loc}_t]$
  - V_t: ResNet encodings of the target bounding box content
  - Loc_t: Target location / size relative to global image



Input for TRF_{tgt}

Experimental Setup / Models

TRF$_{vis}$: Visual context variant:

- Target + visual context features
- Input: [$\mathbf{V}_t$; $\mathbf{Loc}_t$; $\mathbf{V}_c$]
  - V_c: ResNet encoding of the global image (without target)

Input for TRF$_{vis}$

Experimental Setup / Models

TRF_{sym}: Symbolic context variant

- Target + symbolic context features
- Input: $[\mathbf{V}_t; \mathbf{Loc}_t; \mathbf{S}_c]$
  - S_c: Symbolic information about **what kind of objects and stuff the context is composed of**
    - e.g. 25 % street; 15 % vehicles; 15 % buildings; ...

$$+ \begin{bmatrix} 0.25 \\ 0.15 \\ 0.01 \\ 0.00 \\ 0.15 \\ 0.00 \\ \cdots \end{bmatrix}$$

Input for TRF_{sym}

S_c features based on dense 2D maps for Panoptic Segmentation (Kirillov et al., 2018)

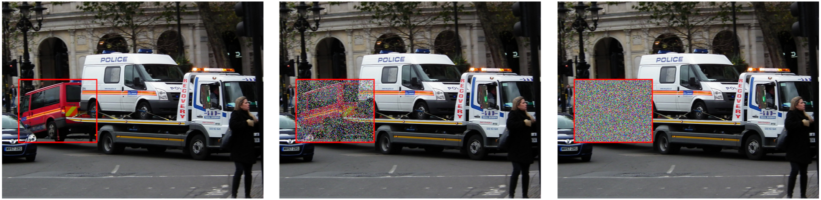
$\rightarrow$ Details in paper

Experimental Setup / Models

All variants are trained and tested for three noise settings:

- **0.0** $\rightarrow$ no noise
- **0.5** $\rightarrow$ 50 % of target bounding box replaced with noise
- **1.0** $\rightarrow$ full target bounding box replaced with noise (no visual target information)

We always use the same setting for training and evaluation.



Results

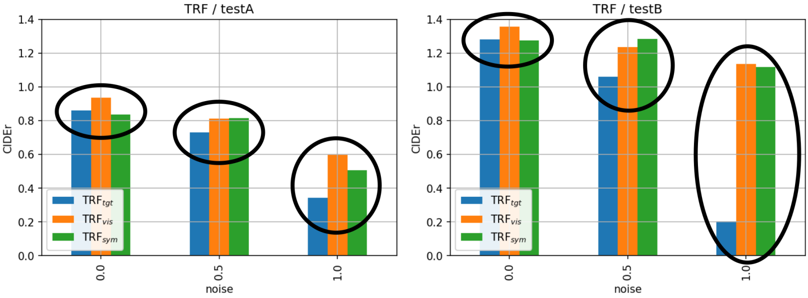
	TRF_{tgt}	cow (A)
noise 0.0	TRF_{vis}	left cow (A)
	TRF_{sym}	cow on left (A)

	TRF_{tgt}	cow (A)
noise 0.0	TRF_{vis}	left cow (A)
	TRF_{sym}	cow on left (A)
	TRF_{tgt}	white horse (F)
noise 0.5	TRF_{vis}	cow on left (A)
	TRF_{sym}	cow (A)

noise 0.0	TRF_{tgt}	cow (A)
	TRF_{vis}	left cow (A)
	TRF_{sym}	cow on left (A)
noise 0.5	TRF_{tgt}	white horse (F)
	TRF_{vis}	cow on left (A)
	TRF_{sym}	cow (A)
noise 1.0	TRF_{tgt}	man (F)
	TRF_{vis}	left cow (A)
	TRF_{sym}	cow on left (A)

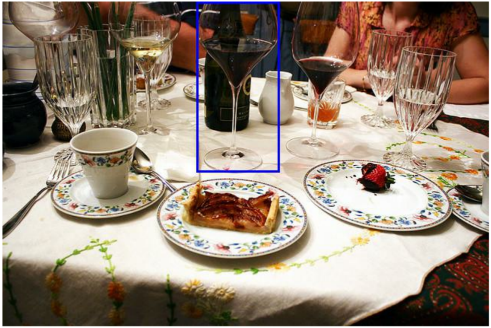
Results: CIDEr/BLEU

- context very effective for compensating noise
  - scores drop with increasing noise, but mitigated by context
  - visual context more effective than symbolic context
- differences between testA (humans) and testB (other objects)
  - target-only suffers less on testA
    - $\rightarrow$ human referents are very frequent
  - context is more helpful on testB
    - $\rightarrow$ other objects are more varied, but appear in more specific contexts



Human Evaluation

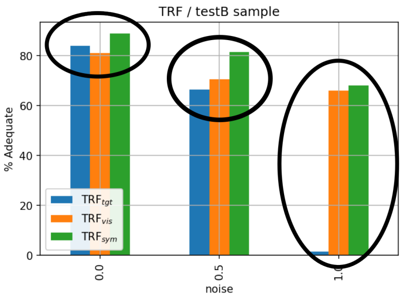
- 200 item sample from RefCOCO testB
- Instruction: Rate the expression parts which refer to the object type (e.g. “a black **dog**”)



Adequate:	wine glass
False:	fork
Misaligned:	bottle
Omission:	thing in center

Results: Human Evaluation

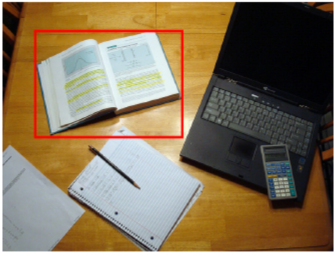
- context again very effective for compensating noise
  - Adequacy rates drop with increasing noise, but mitigated by context
  - symbolic context is more effective than visual context
- identification with only context works surprisingly well: 68 % for TRF_{sym} with full occlusion!



How exactly does context improve the predictions?

Copying Strategy

- Observation: Systems often predict referent types which are also present in the surrounding scene
- Often effective, as many objects tend to appear in groups



	TRF_{tgt}	top left (O)
noise 1.0	TRF_{vis}	left laptop (F)
	TRF_{sym}	laptop on left (F)

Copying Strategy: Statistical Analysis

- is exploiting context more effective if the target class is present in the scene?
- **correlation study**: adequacy of descriptions vs. context area covered by target class
- results: systems rather pick the correct target class, if objects of the same type are present in the context

	noise	corr.	p
TRF_{tgt}	0.0	0.128	–
TRF_{vis}		0.109	–
TRF_{sym}		0.154	< 0.05
TRF_{tgt}	0.5	0.071	–
TRF_{vis}		0.186	< 0.01
TRF_{sym}		0.157	< 0.05
TRF_{tgt}	1.0	0.046	–
TRF_{vis}		0.321	< 0.001
TRF_{sym}		0.277	< 0.001

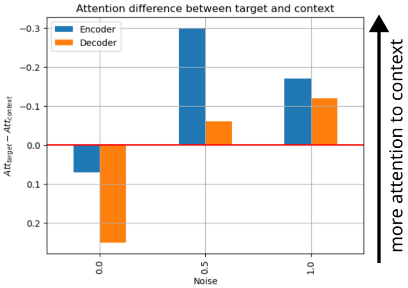
Attention Analysis (TRF$_{vis}$)

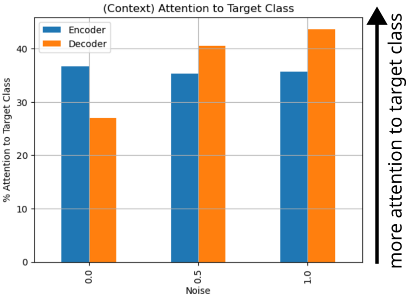
Encoder / Decoder attention to

1. target and context features
2. object types in context (target class vs. other classes)

Results:

- No clear picture for Encoder
- Decoder Attention: More attention to context and target class for higher noise

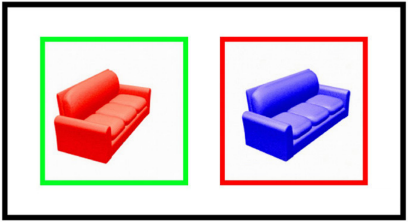




How does Scene Context fit into the REG task?

Scene Context in REG

- In classical works (Incremental Algorithm) and work on visual REG: **Distractors** taken as most relevant form of context
- considered during Content Determination: pick target properties that **do not** apply to distractor



(TUNA, van Deemter et al. 2006)

The red couch ~~facing right~~

Scene Context in REG

- **Scene context is different, but complimentary**: Which properties are **true** (not distinctive) for the target?
  - rather effects **semantic** than **pragmatic** aspects
    - (or other pragmatic aspects, e.g. Gricean Maxim of Quality instead of Quantity/Relevance)
- possibly important for subsequent pragmatic processing!

The **truck** being towed

Conclusion

Do REG systems exploit Scene Context?

- Scene Context makes models more resilient against perturbations in visual target representations
- Context affects reference generation at different levels: Can be exploited to generate distinguishing expressions **but also** to ensure that expressions are true in the first place
- Is reliance on copying strategy cognitively plausible? Perhaps not.
  - further research!

Citations I

Galleguillos, Carolina and Serge Belongie (June 2010). "Context based object categorization: A critical survey". In: *Computer Vision and Image Understanding* 114.6, pp. 712–722. doi: 10.1016/j.cviu.2010.02.004.

Kazemzadeh, Sahar et al. (Oct. 2014). "ReferItGame: Referring to Objects in Photographs of Natural Scenes". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 787–798. doi: 10.3115/v1/D14-1086. url: https://www.aclweb.org/anthology/D14-1086.

Kirillov, Alexander et al. (Jan. 2018). "Panoptic Segmentation". In: doi: 10.48550/ARXIV.1801.00868. arXiv: 1801.00868 [cs.CV].

Citations II

- Mokady, Ron, Amir Hertz, and Amit H. Bermano (Nov. 2021). "ClipCap: CLIP Prefix for Image Captioning". In: doi: 10.48550/ARXIV.2111.09734. arXiv: 2111.09734 [cs.CV].
- Panagiaris, Nikolaos, Emma Hart, and Dimitra Gkatzia (2021). "Generating unambiguous and diverse referring expressions". In: *Computer Speech & Language* 68, p. 101184. issn: 0885-2308. doi: 10.1016/j.csl.2020.101184. url: https://www.sciencedirect.com/science/article/pii/S0885230820301170.
- Van Deemter, Kees, Ielka van der Sluis, and Albert Gatt (July 2006). "Building a Semantically Transparent Corpus for the Generation of Referring Expressions.". In: *Proceedings of the Fourth International Natural Language Generation Conference*. Sydney, Australia: Association for Computational Linguistics, pp. 130–132. url: https://aclanthology.org/W06-1420.