Image Captioning - Object Relation Transformer

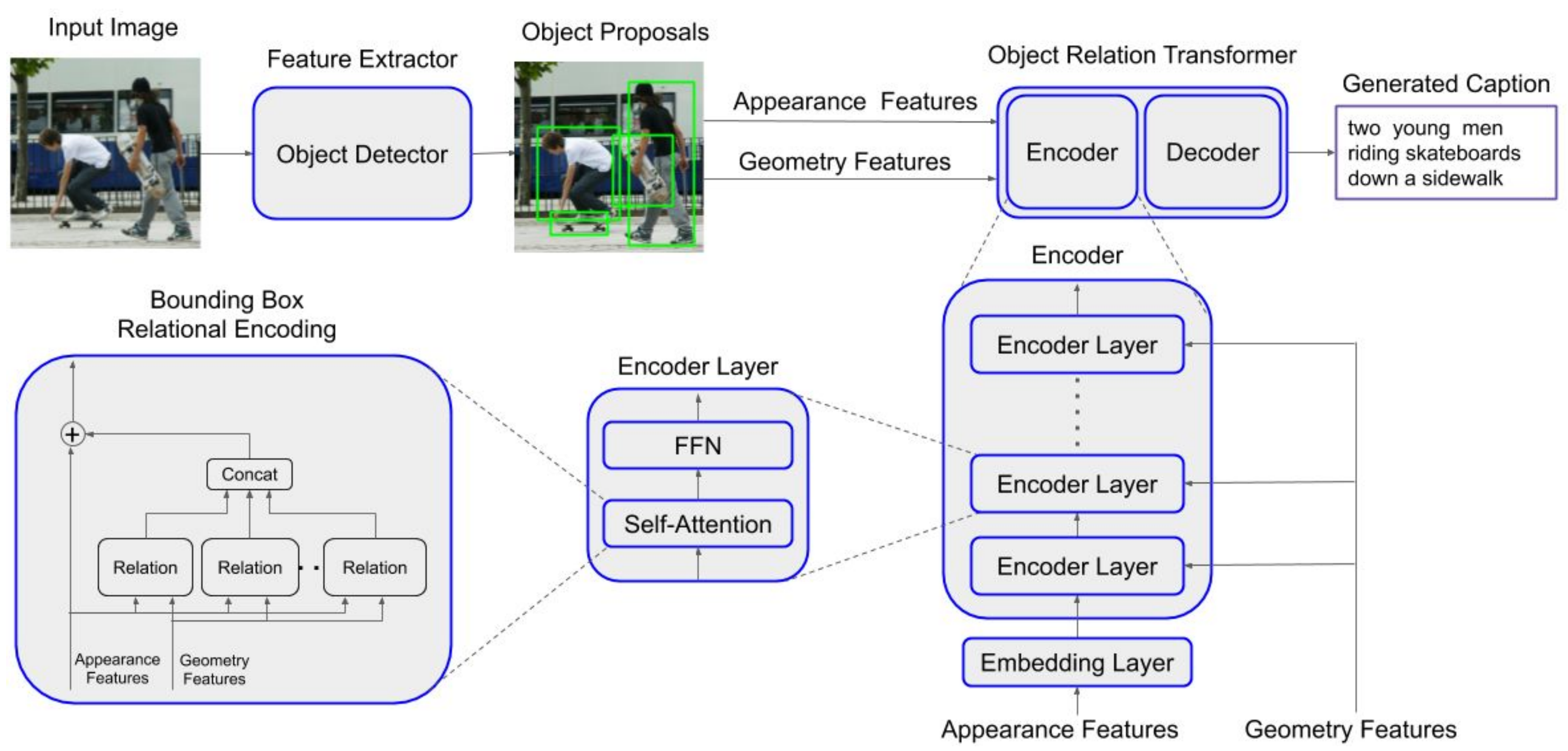
yahoo! research

Simão Herdade, Armin Kappeler, Kofi Boakye, João Soares

Motivation

- **Goal**
  - Utilize spatial relationships between objects to improve image captioning
- **Motivation**
  - The Transformer model has achieved great success in machine translation
  - Image captioning can be viewed as translating a collection of object image features into a sentence
- **Problem**
  - There is no natural order in a collection of object image features
- **Applications**: Improve accessibility of image content
  - 285 million people have some type of visual impairment
  - Most Yahoo! images lack descriptive alt-text
  - Automatic caption generation could greatly improve accessibility

Object Relation Transformer



- **Idea** - We propose a novel encoder-decoder architecture consisting of:
  - Faster R-CNN for object detection and feature extraction
  - Transformer Encoder, with a **geometric self-attention layer,** acting on the object crop image features
  - Standard Transformer Decoder, to generate sentences
- **Observations**
  - Extracting features for the object region proposals, rather than generic grid spatial features, yields better image captions [2]
  - Incorporating spatial relationships improves object detection [3]
- **Hypothesis**
  - **Explicitly encoding spatial relationships between detected objects should enable better image captions**

    Ex. 1: *"a girl riding a horse" vs. "a girl standing beside a horse"* (position)

    Ex. 2: *"a woman playing the guitar" vs. "a woman playing the ukulele"* (size)

Geometric Attention

- **Object Relation Transformer Encoder**
  - The encoder of the object relation transformer uses the feature vectors from the object detector
  - Each encoder layer consists of a multi-head self-attention layer followed by a small feed-forward neural network
- **Transformer Self-Attention**
  - Self-attention first takes a query Q, key K, and value V, where

$$Q = XW_Q, K = XW_K, V = XW_V,$$

  X contains all input vectors $x_1 \dots x_N$ stacked into a matrix.

  - The appearance-based attention weights are computed as

$$\omega_A = \frac{QK^T}{\sqrt{d_k}} \quad (d_k = 64 \text{ a scaling factor})$$

- **Geometric Attention**
  - The geometric attention weights are computed as

$$\omega_G^{mn} = ReLU(Emb(\lambda)W^G)$$

  where λ represents a displacement vector for two bounding boxes

$$\lambda(m,n) = \left(\log\left(\frac{|x_m - x_n|}{w_m}\right), \log\left(\frac{|y_m - y_n|}{y_m}\right), \log\left(\frac{w_n}{w_m}\right), \log\left(\frac{h_n}{h_m}\right)\right)$$

  and Emb($\cdot$) calculates a high-dimensional embedding following [3]

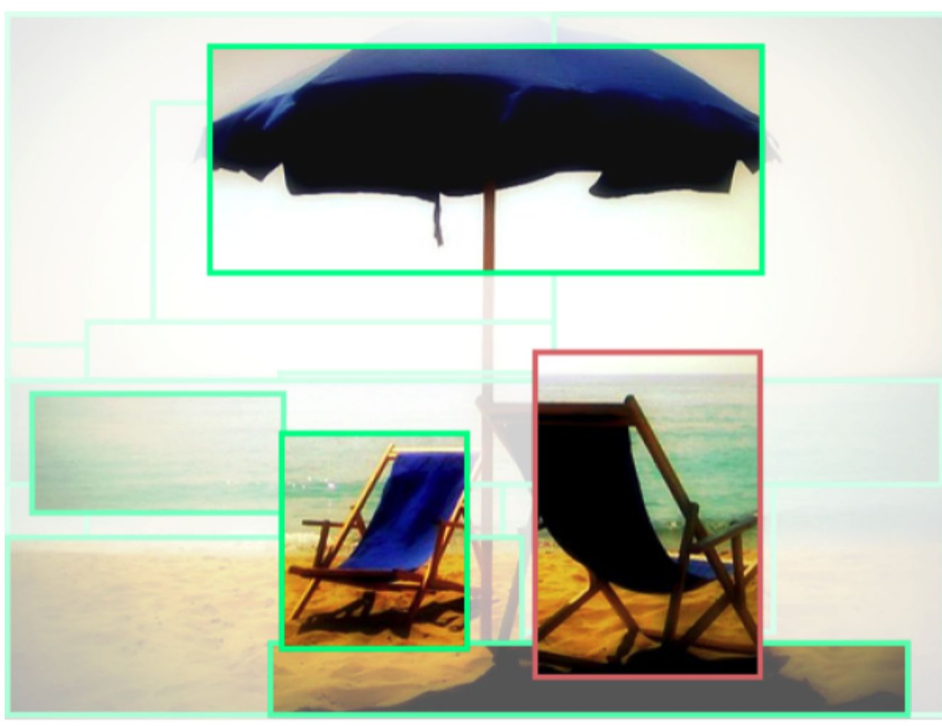
  - The final attention weights are given by

$$\omega^{mn} = \frac{\omega_G^{mn}\exp(\omega_A^{mn})}{\sum_{l=1}^{N}\omega_G^{ml}\exp(\omega_A^{ml})}$$

  - The output of each attention head can be calculated as

$$\text{head}(X) = \text{self-attention}(Q, K, V) = \Omega V$$

  where Ω is the NxN matrix whose elements are given by ω^{mn}



Generated Caption: *two beach chairs under an umbrella on the beach*

Visualization of self-attention in our Object Relation Transformer. Each detected object's transparency is proportional to its attention weight with respect to the chair outlined in red.

Results

- **Best performing model on the MSCOCO dataset**
  - Cross-entropy loss training (30 epochs), followed by self-critical reinforcement learning (30 epochs), optimizing for CIDEr-D

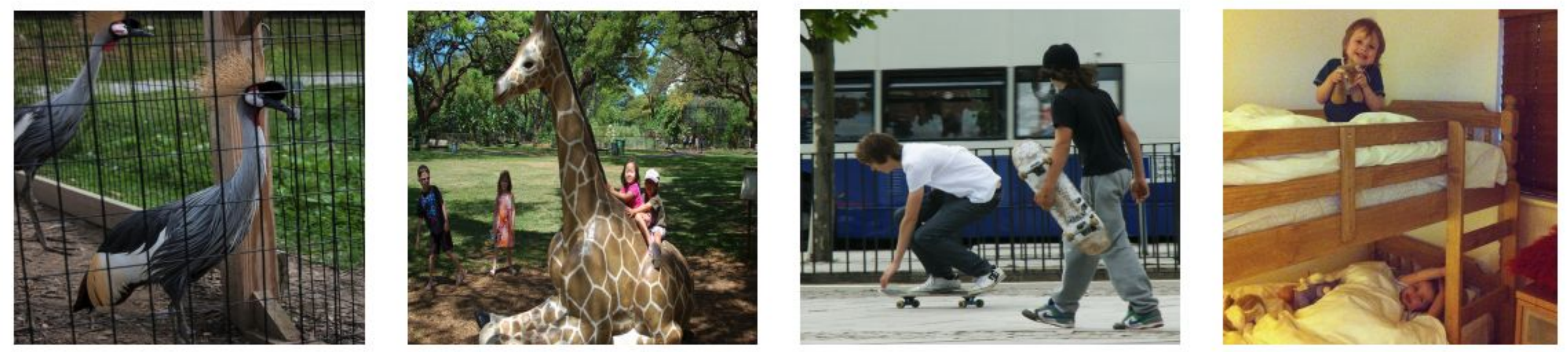
Algorithm	CIDEr-D	SPICE	BLEU-1	BLEU-4	METEOR	ROUGE-L
Att2all [20]	114	-	-	34.2	26.7	55.7
Up-Down [2]	120.1	21.4	79.8	36.3	27.7	56.9
Ours	**128.3**	**22.6**	**80.5**	**38.6**	**28.7**	**58.4**

- **Geometric Attention's ability to count and relate objects**
  - Quantitative comparison (SPICE metric, 30 epochs)

Algorithm	SPICE						
	All	Object	Relation	Attribute	Color	Count	Size
Standard Transformer	21.04	37.83	5.88	11.31	14.88	11.30	5.82
Ours	21.24	37.92	6.31	11.37	15.49	17.51	6.38
p-value	0.15	0.64	**0.01**	0.81	0.35	**<0.001**	0.34

  - Qualitative comparison (relations and count)

Standard: a man on a motorcycle on the road	a couple of bears standing on top of a rock	two chairs and an umbrella on a beach	a laptop computer sitting on top of a wooden desk
Ours: a man is working on a motorcycle in a parking lot	two brown bears standing next to each other on a rock	two beach chairs under an umbrella on the beach	a desk with a laptop and a keyboard



Standard: a large bird is standing in a cage	a little girl sitting on top of a giraffe	a group of young men riding skateboards down a sidewalk	three children are sitting on a bunk bed
Ours: two large birds standing in a fenced in area	a giraffe with two kids sitting on it	two young men riding skateboards down a sidewalk	two young children are sitting on the bunk beds

Future Work

- Extend geometric attention to the decoder's cross-attention layers
- Scale to larger datasets (e.g., Google's Conceptual Captions)

References and Acknowledgments

We thank Ruotian Luo for making his image captioning code available [1].

[1] R. Luo. An image captioning codebase in pytorch. https://github.com/ruotianluo/ImageCaptioning.pytorch, 2017

[2] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In CVPR, 2018.

[3] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei. Relation networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3588–3597, 2018.