Using Transformers in Image Captioning

Object Relation Transformer

Nikolai Ilinykh

Presentation for the course 'Frontiers in Language Modeling: The 'Transformer' And Its Applications'

Computational Linguistics
University of Gothenburg

13^{th} May 2020

Content of presentation

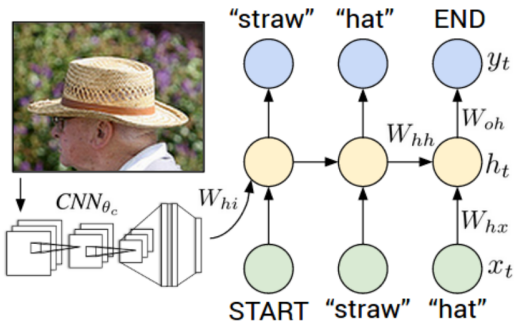
Paper

Herdade, S., Kappeler, A., Boakye, K., & Soares, J. (2019).
Image Captioning: Transforming Objects into Words.

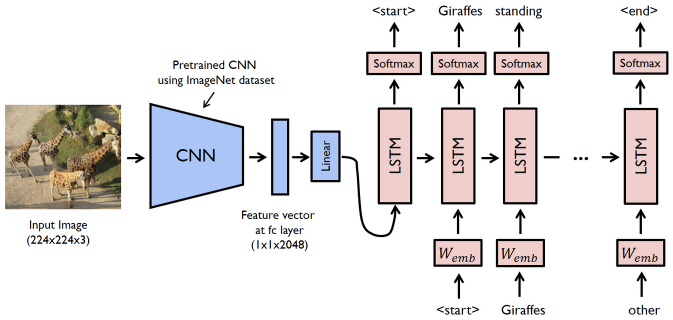
Outline

- the task of image captioning; incorporating spatial relations
- transformer encoder-decoder architecture (Object Relation Transformer)
- results, conclusion

Encoder-Decoder Architecture



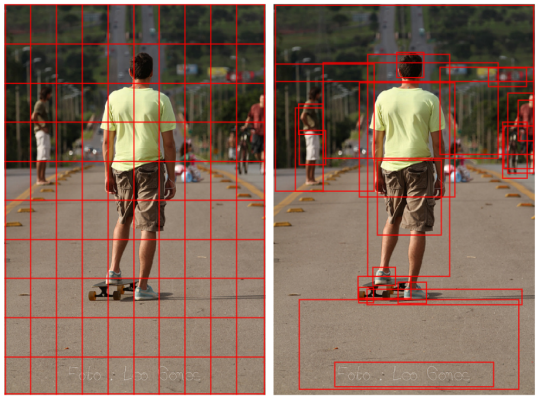
(a) Image Captioning I



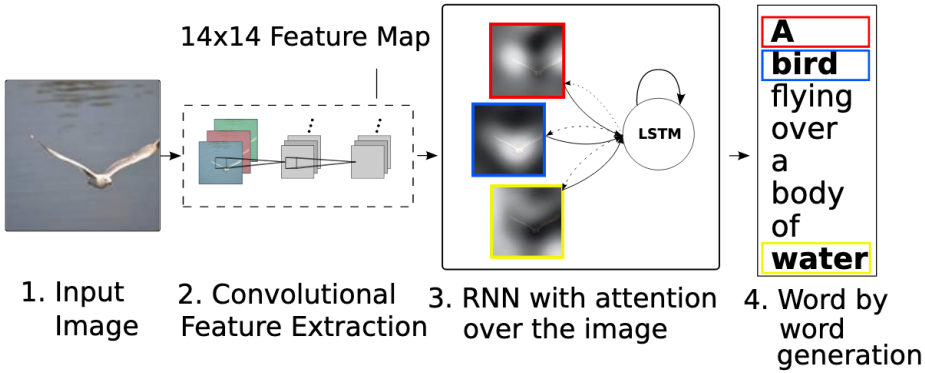
(b) Image Captioning II

- we want to build models that can input an image and output a caption in natural language [3, 6]
- the idea is to use convolutional neural network to obtain image representation and feed it to the recurrent neural network to generate a sentence (one-to-many)
- note that image can be used as initial hidden state of RNN **or** as its first input (need a fully-connected layer for that)

Region Features and Attention



(c) Region Features



(d) Attention

- image can be represented by collection of region features: from either uniform sampling (c-left) or object detector (c-right) [1]
- attention in RNN over image regions has shown to significantly improve image captioning [7]

Spatial Relations

There is a teddy bear partially under a go cart.

Figure: VisualGenome 2318741

- spatial information about image objects (size, shape, position, etc.) is crucial in understanding real-world images, which are typically rich with many similar objects [2].
- similarly, in machine translation tasks, positional information can be used to describe 'relations' between words. It is encoded in Transformers [5] as positional encoding.
- **That is, it should be beneficial to use spatial information in visual encoders as well.**

This Presentation

- object relation transformer, an encoder-decoder architecture, designed for image captioning, that incorporates information about the spatial relationships between detected objects through geometric attention
- experiments on one of the benchmark image captioning datasets, MSCOCO [4]
- qualitative analysis of generated captions (more geometry awareness)

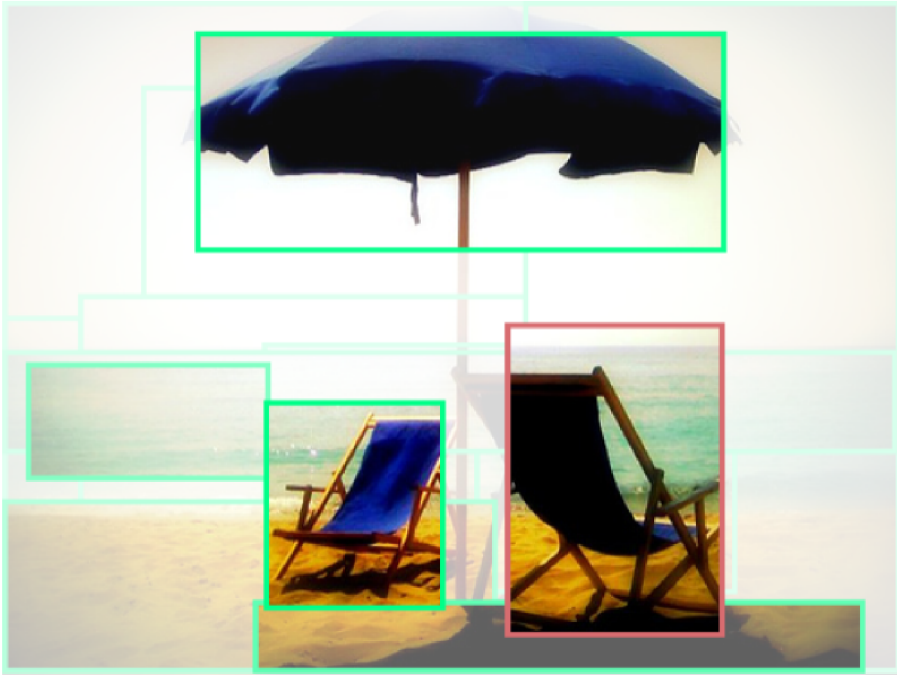
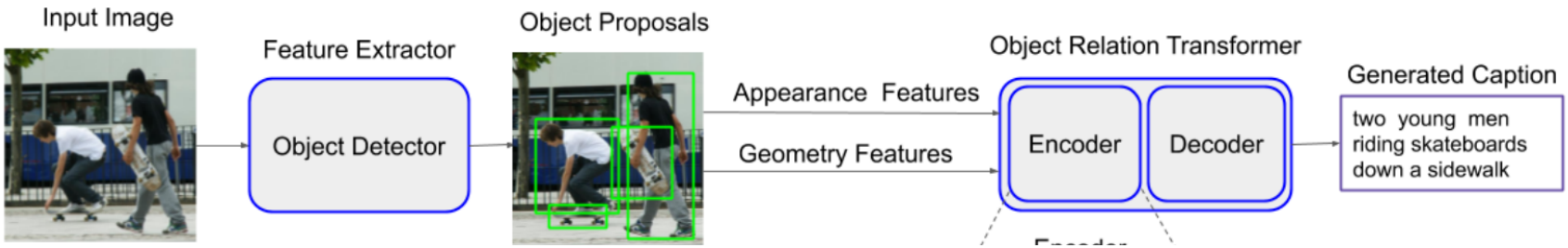


Figure: Example Caption generated by Object Relation Transformer

Model Overview



Model Overview

1. **Image**: use object detector to extract bottom-up region features; also, extract geometric features $\{x, y, w, h\}$ - center coordinates, width, height
2. Use original transformer architecture, modify **only** the encoder part (they treat left-side part of the original architecture as encoder)
3. **Language**: use encoder output as Keys and Values, and generate captions from masked ground-truth

Calculating geometry attention weights

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

Displacement Vector

- this is **modified bounding box regression method** ('modified', because the original formula has no logarithms applied to the first two elements, these 2 additional logarithms are supposed to help to model relations between distant objects)
- it is widely used in the task of object detection and its main purpose is to help mapping proposed bounding box to the ground truth box
- here, it is used to obtain some representation about relations between 2 objects

For more information on bounding box regression:
a) 'Relation Networks for Object Detection' by Hu et. al., 2017
b) 'Rich feature hierarchies for accurate object detection and semantic segmentation' by Girschick et. al., 2013

Calculating geometry attention weights

$$\omega_G^{mn} = ReLU\left(\text{Emb}(\lambda)W_G\right)$$

Positional Encoding of geometric features

- *Emb* follows the functions of positional encodings from original transformers
- it computes cosine and sine functions of fixed wavelengths for each value of displacement vector between bounding boxes
- then, the embedded feature is transformed by W_G to a scalar weight
- ReLU is applied for zero trimming, which (apparently) restricts relations only between objects of certain geometric relations

Code Piece for W_G and ReLU

```
self.WGs = clones(nn.Linear(geo_feature_dim, 1, bias=True),8)
relative_geometry_weights_per_head = [l(flatten_relative_geometry_embeddings).view(box_size_per_head) for l in self.WGs]
relative_geometry_weights = torch.cat((relative_geometry_weights_per_head),1)
relative_geometry_weights = F.relu(relative_geometry_weights)
```

Multi-Head and Feed-Forward Network

Multi-Head

Similar to the original Transformer architecture, the output of all 8 heads are concatenated and multiplied with a linear projection matrix W_O:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W_O$$

FFN

As the next step, the point-wise FFN with a ReLU activation is applied to each output of the attention layer:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

Note, that all layer-norm and skip-connections are included as well, following the original Transformer architecture.

Decoder

Citation from the paper

The decoder then uses the generated tokens from the last encoder layer as input to generate the caption text. Since the dimensions of the output tokens of the Transformer encoder are identical to the tokens used in the original Transformer implementation, we make no modifications on the decoder side.

Generation Results

Evaluation metrics are improved when using geometric attention with transformer

- trained on MSCOCO dataset (113K training images with 5 human annotated captions each, 5k in val and 5k in test sets)
- used direct optimising for CIDEr-D score (self-critical reinforcement learning)

Algorithm	CIDEr-D	SPICE	BLEU-1	BLEU-4	METEOR	ROUGE-L
Att2all [21]	114	-	-	34.2	26.7	55.7
Up-Down [2]	120.1	21.4	79.8	36.3	27.7	56.9
Visual-policy[15]	126.3	21.6	–	38.6	28.3	58.5
GCN-LSTM [29][1]	127.6	22.0	80.5	38.2	28.5	58.3
SGAE [27]	127.8	22.1	**80.8**	38.4	28.4	**58.6**
Ours	**128.3**	**22.6**	80.5	**38.6**	**28.7**	58.4

Positional Encoding Revisited

What are the replacements for original positional encoding?

- order of objects might be based on their size, location, saliency, importance, etc., several tests for positional encoding were presented:
- **box size**: calculate area of bounding boxes and order from largest to smallest
- **left-to-right** or **top-to-bottom**: x-coordinate or y-coordinate of the box centroids
- the results demonstrate that more sophisticated geometric attention gives a better CIDEr-D score

Positional Encoding	CIDEr-D
no encoding	111.0
positional encoding (ordered by box size)	108.7
positional encoding (ordered left-to-right)	110.2
positional encoding (ordered top-to-bottom)	109.1
geometric attention	112.6

Standard Transformer or Object Relation Transformer?

Algorithm	CIDEr-D	SPICE	BLEU-1	BLEU-4	METEOR	ROUGE-L
Standard Transformer	113.21	21.04	75.60	34.58	27.79	56.02
Ours	115.37	21.24	76.63	35.49	27.98	56.58
p-value	**0.01**	0.15	**<0.001**	0.051	0.24	**0.01**

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

Geometric Attention in Transformer helps to improve evaluation metrics

- for each evaluation metric, a two-tailed t-test with paired samples has been performed
- significant improvements in CIDEr-D and SPICE Relation score demonstrate that geometric attention helps model to be more precise in determining correct relations between objects
- interestingly, geometric attention seem to assist in defining the number of objects of the same type in an image

Generated Captions

			
Standard: a man on a motorcycle on the road	two chairs and an umbrella on a beach	a group of young men riding skateboards down a sidewalk	three children are sitting on a bunk bed
Ours: a man is working on a motorcycle in a parking lot	two beach chairs under an umbrella on the beach	two young men riding skateboards down a sidewalk	two young children are sitting on the bunk beds

Some of the current issues with generated captions:

- hard to identify unknown, unusual or rare objects
- generated captions tend to be less descriptive than the ground truth captions (probably, problem with decoding strategies?)


Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang.
Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering.
jul 2017.


Mehdi Ghanimifard and Simon Dobnik.
What goes into a word: generating image descriptions with top-down spatial knowledge.
In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 540–551, Tokyo, Japan, October–November 2019. Association for Computational Linguistics.


Andrej Karpathy and Li Fei-Fei.

Deep visual-semantic alignments for generating image descriptions, 2014.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollr.
Microsoft coco: Common objects in context, 2014.

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

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan.
Show and tell: A neural image caption generator, 2014.


Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio.
Show, attend and tell: Neural image caption generation with visual attention, 2015.