Exploiting Image-Text Synergy for Contextual Image Captioning

Sreyasi Nag Chowdhury
Rajarshi Bhowmik
Hareesh Ravi
Gerard de Melo
Simon Razniewski
Gerhard Weikum

Problem Statement

Design your own Spaghetti Arrabiata at our Pasta bar.

Multimodal documents on the web contain images placed at meaningful locations within the textual narrative.

The image captions are commonly conditioned on the surrounding text. We study the generation of such *Contextual Captions,* distinct from conventional image captioning.

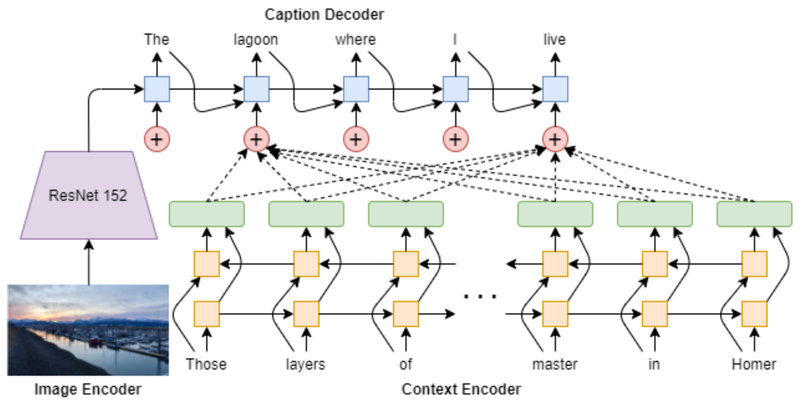
Dataset

Novel Contextual Captioning dataset:
- Data scrapping from subreddit /r/pics
- Domain-agnostic posts
- 250,000 posts spanning one year
- Post: 1 image, caption, 1-10 comments
- Captions contain 10.6 words on avg.
- Concatenated comments serve as image context or associated paragraph
- Paragraphs contain 59.2 words on avg.

Data splits based on Named Entities in image captions:
- 137,732 samples with NE
- 104,653 samples without NE
- Additional splits ensuring overlap between context and caption

SpaCy is leveraged to detect 14 types of NE in image captions.

Model Formulation



Text Encoder: BiLSTM
Image Encoder: ResNet152

Decoder: LSTM, attention-weighted sum of encoder states concatenated to current state to incorporate contextual information.

Objective: $$\mathcal{L}(\theta) = \sum_{t=1}^{N} -\log p(w_t^c \mid w_1^c, \ldots, w_{t-1}^c, I, P; \theta)$$

Output of BiLSTM: $$\tilde{\mathbf{G}}_\mathbf{t} = \sum_{i=i}^{M} \alpha_i^t \mathbf{g_i}$$

Attention weights: $$\alpha_i^t = \frac{\mathbf{v}^\intercal(\mathbf{W_g g_i} + \mathbf{W_h h_t} + \mathbf{b})}{\sum_{i'=1}^{M} \mathbf{v}^\intercal(\mathbf{W_g g_{i'}} + \mathbf{W_h h_t} + \mathbf{b})}$$

Results

	BLEU-1	*ROUGE-L*	*CIDEr*	*SPICE*	*SemSim*
Image-only	7.80	7.50	0.38	0.16	0.76
Text-only	6.87	6.54	0.61	0.36	0.72
Contextual	**9.30**	**9.68**	**0.78**	**0.50**	**0.77**

Table: Quantitative evaluation of baselines and Contextual Captioning on standard text similarity measures.

- Contextual Captions capture information from both visual and textual modalities.
- They are linguistically rich compared to text-only and image-only captions.

Context: I recently moved to Buffalo, NY... ...every day I am discovering how beautiful this town is. I wanted to share the pallet of colors the sunset had that evening...

Generated Image-only Caption:
- *A picture I took of a mossy branch through the shadows of a cloud.*

Generated Contextual Captions:
- *A beautiful sunset path to heaven.*
- *A sunset...unknown artist.*

References:
[1] Nag Chowdhury et al. "Illustrate Your Story: Enriching Text...", WSDM 2020
[2] Nag Chowdhury et al. "SANDI: Story-and-Images Alignment", EACL 2021

https://github.com/Sreyasi/contextual_captions

max planck institut informatik
HPI Hasso Plattner Institut Digital Engineering · Universität Potsdam
RUTGERS