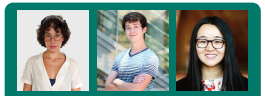
Impressions: Understanding Visual Semiotics and Aesthetic Impact

SALT



Julia Kruk
Georgia Institute of Technology

Caleb Ziems
Stanford University

Diyi Yang
Stanford University

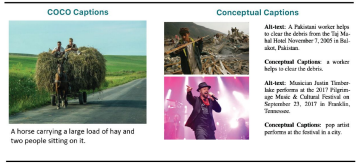
arXiv GitHub

Research Questions

- **Does visual salience extend past formal definitions of beauty, to an image's capacity for effective communication?**
- **How can we design datasets that capture the connotation of visual features and aesthetic elements?**
- **Can such a resource improve multimodal architectures' ability to model human impressions of images?**

Motivations

- Popular image captioning datasets contain terse and reductive captions that describe their visual counterparts.
- Annotators are encouraged to focus solely on the most concrete elements of an image: objects, entities, colors, etc.
- Such datasets inhibit models' ability to reason about the ***semiotics of images***, or the ***connotation*** of visual elements.



Each visualization is taken directly from the COCO (Chen et al., 2015) and Conceptual Captions (Sharma et al., 2018) publications.

Impressions Dataset

The ***Impressions*** dataset, a multimodal benchmark that consists of 4,320 unique annotations over 1,440 image-caption pairs from the photography domain. Each annotation explores:

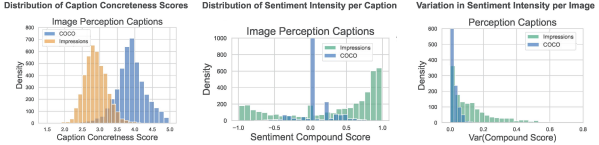
1. The aesthetic impactfulness of a photograph.
2. Image descriptions in which pragmatic inferences are welcome.
3. Emotions/thoughts/beliefs that the photograph may inspire.
4. The aesthetic elements that elicited the expressed impression.

Images are collected via the New York Times official API and the Google Search API to encompass a wide distribution of styles and visual elements.

Dataset Qualities

Impressions captures rich, diverse, and expressive commentary on image features and aesthetic elements. This demonstrated by:

- Increased variance in the distributions of sentiment intensity.
- Increased subjectivity.
- Lower concreteness scoring of linguistic data.



Improved Image Impression Modeling

	Description	Impression	Aesthetic Evaluation	All Captions
GIT†	0.750	0.920	0.780	0.815
BLIP†	0.690	0.840	0.880	0.805
OpenFlamingo-16*	0.610	0.960	1.000	0.857
LLaVA-7b-v0*	0.560	0.530	0.590	0.560

- We fine-tune / few-shot adapt four multimodal architectures on the three different caption categories present in ***Impressions***: GIT, BLIP, OpenFlamingo, and LLaVA.
- In a human evaluation task, annotators preferred captions generated by the tuned / adapted models 76% of the time on average, across all architectures.

Persona-Specific Generation

To investigate the variation in human perceptions captured by ***Impressions***, we leverage personality and demographic information provided by annotators to explore the unique generation qualities that emerge when training on annotations created by distinct groups.

- We find that certain opposing personality groups, such as ***introvert vs extrovert*** and ***3+ years art experience vs no art experience***, yielded distinct generation qualities with statistical significance on caption length and mean concreteness, respectively.