Quantifying CLIP's ability to Perform Cross-Modal Grounding Using Attention-Model Explainability

Interpretability & Explainability in AI, 2022

Piyush Bagad, Danilo de Goede, and Paul Hilders

University of Amsterdam

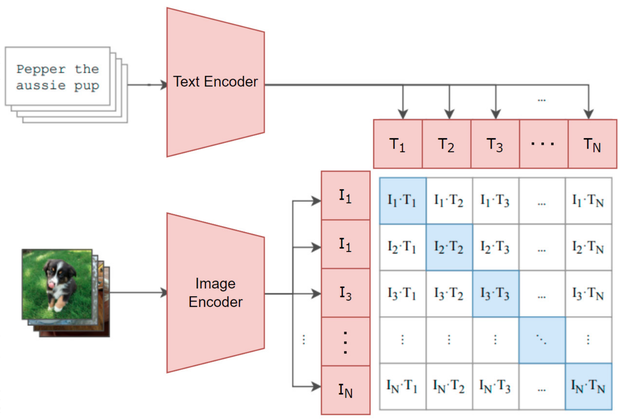
Motivation

Multimodal models such as CLIP [1] are expected to combine Natural Language and Visual concepts to find matching text-image pairs. However, it is unclear whether CLIP attends to the correct signals when looking for Cross-modal correspondences. Can we use a State-of-the-Art Transformer attribution method [2] to verify and quantify CLIP's behavior?

CLIP

Connecting Text and Images [1]

- Effective for finding correspondences between images and captions.
- Is not restricted to a fixed set of class labels --> CLIP accepts basically any word on sentence in the English language.

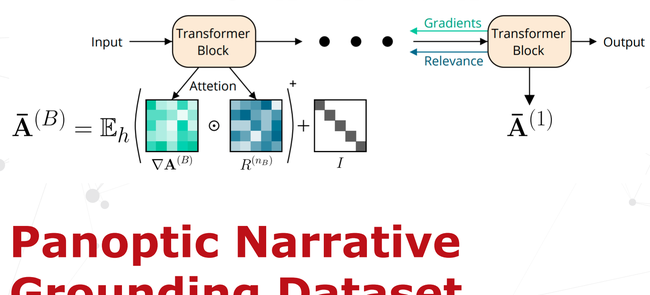


RQ: If CLIP is not enforced to explain **why** a caption and image correspond, how can we verify if CLIP actually looks at the relevant signals?

Transformer Explainability

- We use the Transformer Explainability method by Chefer et al. (CVPR'21) [2] to visualize CLIP's attention
- The generated explanations have been shown to be useful for generating CLIP explanations, improving accuracy for image classification, and to mitigate biases.

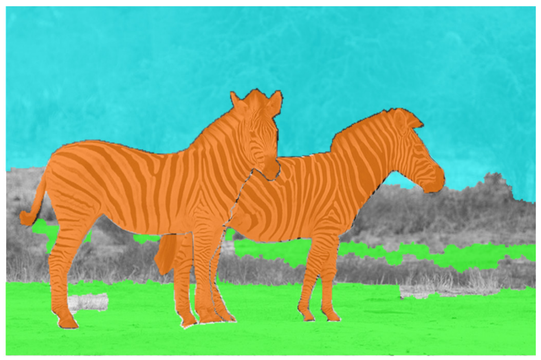
$$\mathbf{C} = \bar{\mathbf{A}}^{(1)} \cdot \bar{\mathbf{A}}^{(2)} \cdot \ldots \cdot \bar{\mathbf{A}}^{(B)}$$



$$\bar{\mathbf{A}}^{(B)} = \mathbb{E}_h\left(\nabla\mathbf{A}^{(B)} \odot R^{(n_B)}\right)^{+} + I \qquad \bar{\mathbf{A}}^{(1)}$$

Panoptic Narrative Grounding Dataset

- *[3] provides fine-grained grounding (segmentations) of parts of a sentence.*
- *Can be used to quantify how well components of text and image "align"*

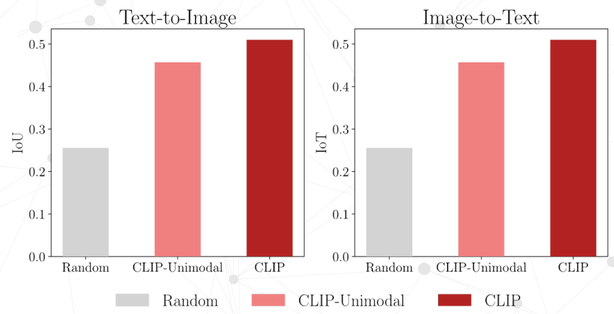


In this image i can see **two zebras** which are in black white color. these are standing on the **ground**. in the back i can see **many trees**.

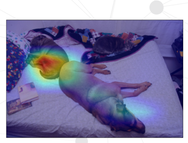
Quantitative Results

- **Random Baseline:** Sample attribution from $U(0,1)$.
- **Unimodal Baseline:** CLIP when provided with only one modality.

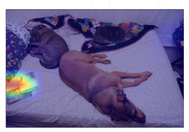
CLIP **does** focus on the relevant signals to find correspondences between texts and images.



Qualitative Analysis



In this image a bed is visible on which two dogs and cat are sleeping, cushions and blankets are visible and book visible. Background walls are white in color and a curtain visible and a table visible. This image is taken inside a room

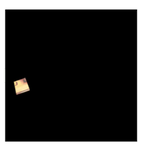


In this image a bed is visible on which two dogs and cat are sleeping, cushions and blankets are visible and book visible. Background walls are white in color and a curtain visible and a table visible. This image is taken inside a room

Predicted Relevance

In this image a bed is visible on which two dogs and cat are sleeping, cushions and blankets are visible and book visible. Background walls are white in color and a curtain visible and a table visible. This image is taken inside a room

In this image a bed is visible on which two dogs and cat are sleeping, cushions and blankets are visible and book visible. Background walls are white in color and a curtain visible and a table visible. This image is taken inside a room



Input text

Predicted Relevance

Input text

Discussion & Limitations

- We evaluate CLIP's ability to align vision and language at fine-grained level using transformer-explainability
- CLIP indeed is capable, to an extent, of cross-modal grounding
- We rely on the assumption that the attribution method perfectly represents CLIP's behaviour.

Future work

Study how to incorporate explainability maps into the pre-training of CLIP in order to improve its cross-modal grounding abilities.

References:

[1] Radford et al, Learning Transferrable Visual Models from Natural Language Supervision, Arxiv, 2021

[2] Chefer et al, Generic Attention-model Explainability for Interpreting Bi-modal and Encoder-Decoder architectures, ICCV 2021

[3] Gonazlez et al, Panoptic Narrative Grounding, ICCV 2021