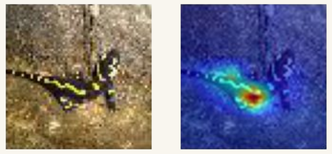
MapCap: Natural Language Descriptions of Saliency Maps

Kalen Frieberg Michael Lepori Matthew Murakami

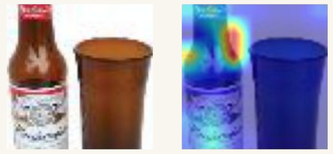
BROWN

Motivation

Saliency maps are great but…
What about full classes?



And what is the model “seeing” in each patch?



Problem

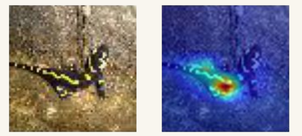
CLIP
Clip can perform zero-shot image classification, but can we explain its decisions?

We used an attention-based attribution method, as CLIP uses a ViT image encoder

Data
25 classes, 100 images from TinyImagenet

Goal

1. Generate saliency maps for CLIP images
2. Generate image representations for salient regions
3. Caption salient regions
4. Summarize captions

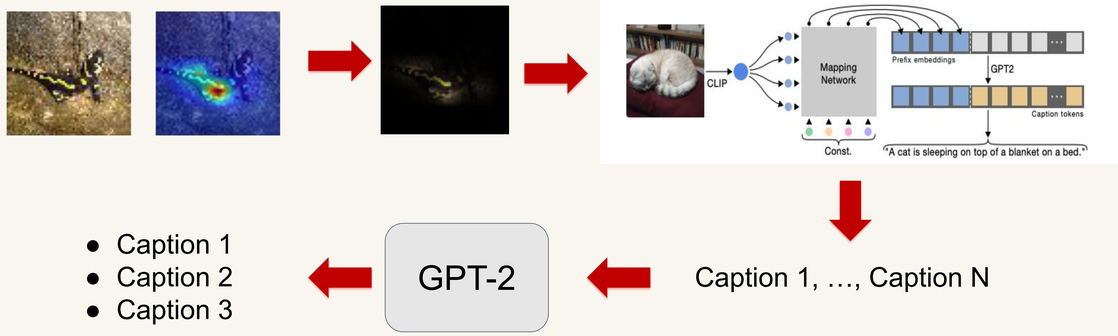


1. Bright yellow against black
2. Black torso against dirt

Insights

- Advances in LLMs have enabled caption generation from image representations
- LLMs can be used to perform unsupervised extractive summarization
- Combine these techniques to explain saliency maps for entire classes

Pipeline Diagram



Results (Zero Shot Classification)

- Can Clip perform zero shot classification on our data?

European Fire Salamander	1.0
Police Van	.99
Egyptian Cat	.98
Beer Bottle	.97
Monarch	.96

Tabby	.06
Plunger	.29
Tarantula	.43
Apron	.53
Banana	.63

Results (Summaries)

European Fire Salamander

- golden and red christmas lights on a black background.
- golden coins falling on a black background.
- 4k light leaks footage for different events and projects!.
- diamonds falling into the water on a black background.
- golden christmas tree on a black background.

Plunger

- a glass of water on a black background.
- a drop of water flying from a pink balloon on a black background.
- a burning candle on a black background.
- red flower in super slow motion being blown by the wind against a black background.
- a dolly shot of a heart shaped object moving slowly on a black background.

References

[1] Chefer, H., Gur, S., & Wolf, L. (2021). Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 397-406).
[2] Mokady, R., Hertz, A., & Bermano, A. H. (2021). Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.
[3] Padmakumar, V., & He, H. (2021). Unsupervised extractive summarization using pointwise mutual information. *arXiv preprint arXiv:2102.06272*.

Acknowledgements

We would like to thank the CS1430 staff for their support of this project