Weakly Supervised Localization with CLIP

Ștefan Popa, Robert-Gabriel Trifan, Ștefan Smeu*, Elisabeta Oneață*
University of Bucharest, Romania
*Bitdefender, Romania

popashtefan10@gmail.com, trifangrobert@gmail.com, ssmeu@bitdefender.com, eoneata@bitdefender.com

1. Introduction

1. We propose to analyze the zero-shot capabilities of CLIP models for object localization tasks.
2. Raw outputs from CLIP can be used to accurately pinpoint the objects, but fail to cover the full shape and produce many separate localizations.

3. Dataset Description

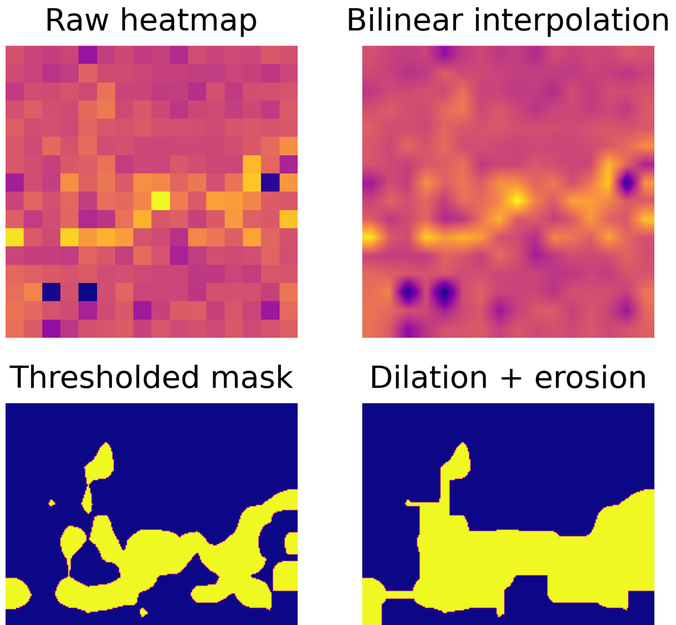
COCO is a large-scale dataset used for object detection and segmentation which contains around 330k images. We worked on a smaller subset, **CocoGlide**, of 512 images with captions and pixel-level masks for objects.

Dataset stats	
NUMBER OF IMAGES	512
NUMBER OF CLASSES	73
MIN NUMBER OF IMAGES PER CLASS	1
MAX NUMBER OF IMAGES PER CLASS	17
MEAN NUMBER OF IMAGES PER CLASS	7.014
STD OF NUMBER OF IMAGES PER CLASS	4.446

4. Data processing

Preprocessing:
We have tried another format for the input prompt: *There is a <label> in this image.* However, CLIP performed the best with the original format: *An image of a <label>.*
Postprocessing:

- According to each method, CLIP's activations are turned into heatmaps that are further used as image overlays.
- We focus our attention on handling the heatmaps, over which we apply **interpolations** (nearest, bilinear, bicubic), **thresholding** and **image transforms** (dilation, erosion).
- A gridsearch approach allowed us to find the best hyperparameters among all possible preprocessing combinations.



2. Models

We evaluate three CLIP-based methods, pretrained on different data sets. Each approach uses a different component for the visual encoder.

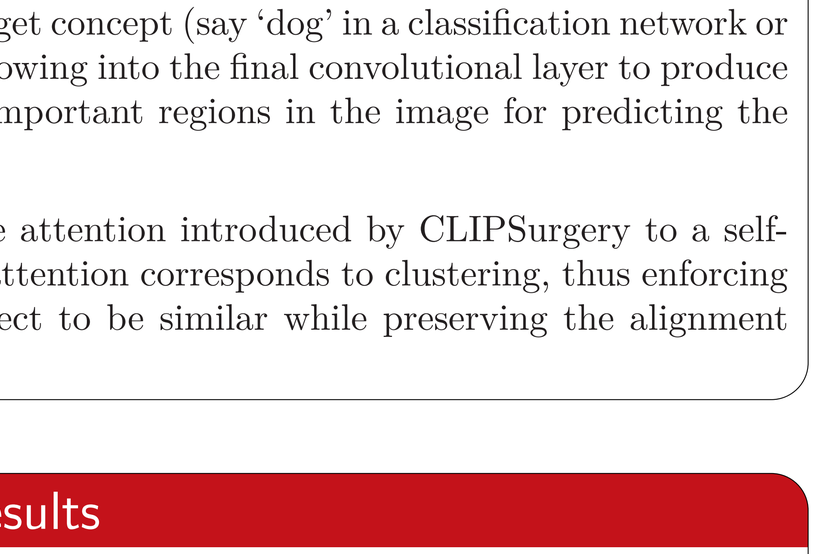
Method	Training Data	Visual Encoder
ClipTextSpan	LAION-2B	ViT-L/14
Grad-CAM	ImageNet	ResNet-50
GEM	WIT-400M	ViT-B/16

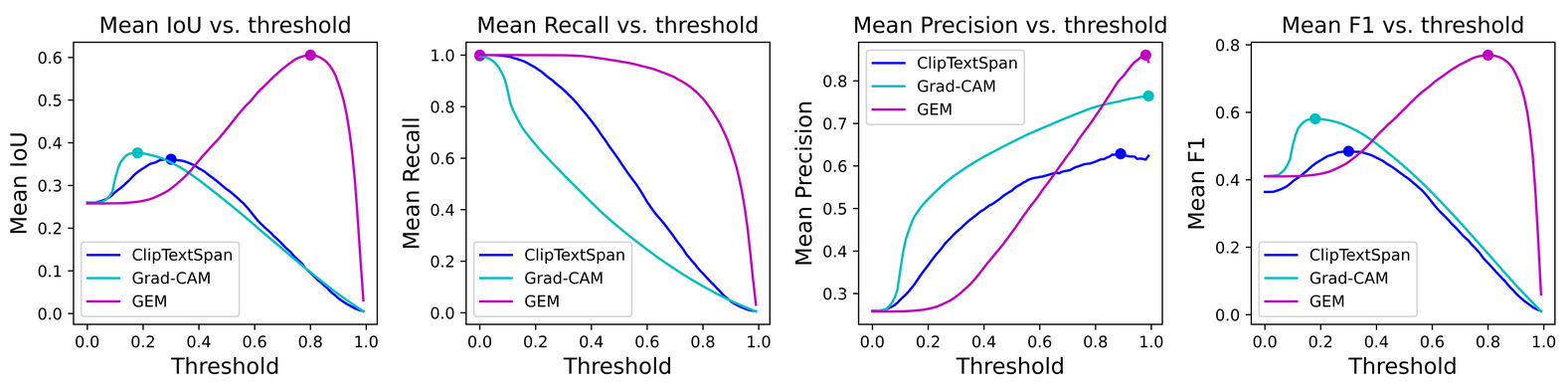
1. **ClipTextSpan** [2] decomposes the image representation as a sum across image patches, model layers, and attention heads, and uses CLIP's text representation to interpret the summands.
2. **Grad-CAM** [3] uses the gradients of any target concept (say 'dog' in a classification network or a sequence of words in captioning network) flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the concept.
3. **GEM** [1] generalizes the idea of value-value attention introduced by CLIPSurgery to a self-self attention path. The concept of self-self attention corresponds to clustering, thus enforcing groups of tokens arising from the same object to be similar while preserving the alignment with the language space

5. Results

- We obtained the best results after applying dilation followed by erosion over the heatmaps. Kernel sizes for each operation were between 41 and 61 depending on the method.
- With the exception of Grad-CAM, image transforms led to an increase in Mean IoU of 7.4-8.6%. This improvement is likely due to the fact that dilation covers holes in the masks and creates connections between neighboring localizations. Then, erosion restores the heatmap to its previous size.
- We compared the three methods by measuring 4 metrics across 100 thresholds uniformly arranged between 0 and 1.
- The GEM methods yields better results at higher thresholds due to a more precise localization.

Method	Mean IoU	AP
ClipTextSpan	0.361	0.535
Grad-CAM	0.376	0.555
GEM	**0.605**	**0.786**





6. Conclusion

- After experimenting with Grad-CAM, GEM, and ClipTextSpan, we can conclude that GEM is the best model for the weakly supervised segmentation task, achieving the highest mean IoU of 0.60539 after applying post-processing filters such as dilation and erosion.

References

[1] Walid Bousselham et al. "Grounding Everything: Emerging Localization Properties in Vision-Language Transformers" (2023).

[2] Yossi Gandelsman, Alexei A. Efros, and Jacob Steinhardt. "Interpreting CLIP's Image Representation via Text-Based Decomposition" (2023).

[3] Ramprasaath R. Selvaraju et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization" (2016).