MOST: Multiple Object localization with Self-supervised Transformers for object discovery

Sai Saketh Rambhatla[1], Ishan Misra[1], Rama Chellappa[2,3] and Abhinav Shrivastava[2]

[1]Meta AI [2]University of Maryland, College Park [3]Johns Hopkins University

Code Project page

Goal

- **Localize** objects in real world images **without** any human supervision and training

Main Idea

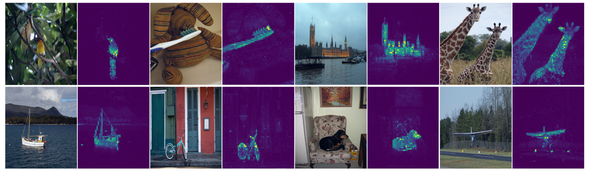
- Leverage similarity maps of features from large self-supervised transformers to localize objects without supervision
- Fractal analysis to identify similarity maps that belong to foreground objects

Results

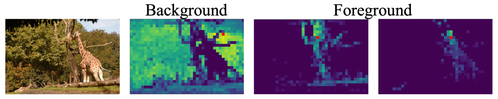
- State-of-the-art localization and discovery performance.
- Effective proposal generator for unsupervised pre-training of object detectors.

Motivation

- Self attention of features from last layer of Self-supervised transformers demonstrates interesting properties.



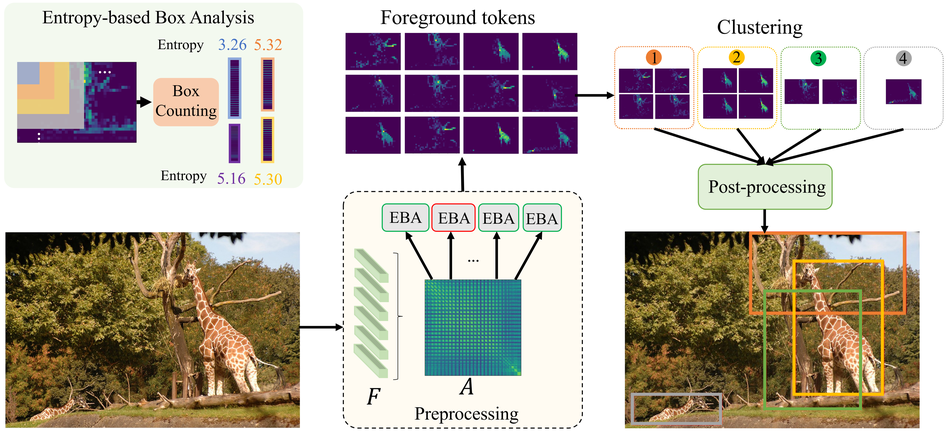
- Tokens within foreground patches exhibit higher similarity with other foreground tokens than background tokens and their similarity maps are "less" spatially random.



Drawbacks of prior work

- Can only localize the most salient object in an image
- Need training to localize multiple instances per image

Approach Overview

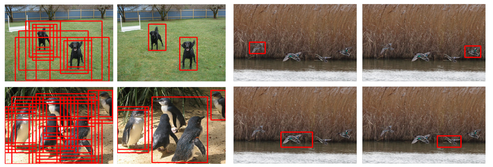


- Features F from an image are extracted using DINO and the outer product $A = FF^T$ is computed.
- The Entropy-based Box Analysis (EBA) module identifies similarity maps of foreground tokens from A.
- Highly redundant neighboring tokens are grouped using spatial clustering to obtain *pools*.
- Similarity maps of tokens from each *pool* are processed to obtain a single bounding box.
- MOST can automatically identify **multiple** objects in an image and can do so **without** any training.

Entropy-based Box Analysis

- We observe that similarity maps of tokens on foreground objects are "less" random spatially.
- EBA uses box counting to segregate similarity maps of tokens on foreground patches from those of background.
- On each similarity map, we perform a raster scan with increasing box sizes.
- Each map is then average pooled and flattened to compute entropy since we are interested in randomness of maps.
- A similarity map belongs to a token on foreground patch if its entropy $\leq a + b\log(n^2)$ where $a = 1$ and $b = 0.5$.

Effect of Clustering



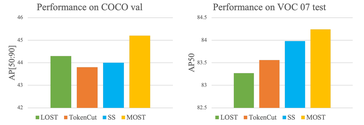
- EBA module identifies redundant tokens.
- Without clustering localization is very noisy.
- Redundancy can be reduced by spatial clustering.
- Each spatial cluster, aka *pool* focuses on a different object/instance.

Results

Proposal generation for Self-supervised pre-training

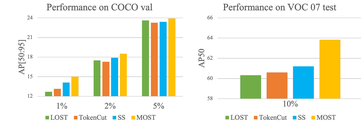
- DETReg pre-trained on COCO dataset using different localization methods.
- Pre-trained model is fine-tuned on full data for fully supervised setting and on $k\,\%$ of the data for semi-supervised setting.

Fully supervised setting



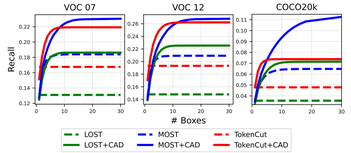
MOST can output multiple high quality boxes per image improving fully supervised training.

Semi-supervised setting



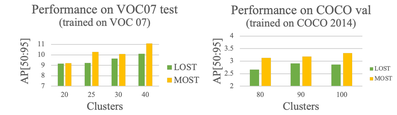
MOST is label efficient with higher gaps in performance at lower amounts of data.

Recall Analysis



MOST achieves the best recall with and without additional training across three standard datasets.

Object Discovery



- Cluster features of regions using K-Means and train object detectors on cluster labels.
- MOST outperforms LOST across multiple clusters and datasets on object discovery.

Qualitative results

Unsupervised Object Localization

MOST can localize multiple objects in cluttered real-world images.

Unsupervised Saliency Detection

MOST can also be used to perform unsupervised saliency detection by using the mask obtained from the largest *pool*.

Acknowledgements

This project was partially supported by DARPA SemaFor (HR001119S0085) and DARPA SAIL-ON (W911NF2020009), and Amazon Research Award to Abhinav Shrivastava. Rama Chellappa was sup- ported by an ONR MURI (N00014-20-1-2787)