Self-Supervised Object Detection with Multimodal Image Captioning

UNIVERSITY OF MICHIGAN

Max Hamilton, Madhav Kumar, Kemmannu Vineet Rao, Kshama Nitin Shah, Wen Jay Lim

Motivation

- **Most SOTA object detection methods use supervised learning**
  - High reliance on large amounts of labeled data
  - Heavy reliance on representation of training data
  - Can have poor generalization to real world examples

Background

- **Multimodal Representation Learning**
  - Joint representation of images and text
  - Image summarization gives context to the image
  - Can be used to "label" data without expensive human supervision
- **Redcaps Dataset**
  - Used image captioning on Reddit data as a pre-training task to learn joint representations of images and text
  - The curated captions in Redcaps are created by the users,hence it contains rich semantic information compared to other generic images
- **VirTex-v2 Model**
  - *Desai et al.* proposed VirTex model that jointly learns visual representations from semantically rich captions
  - Then discards the textual backbone to finetune on several downstream tasks

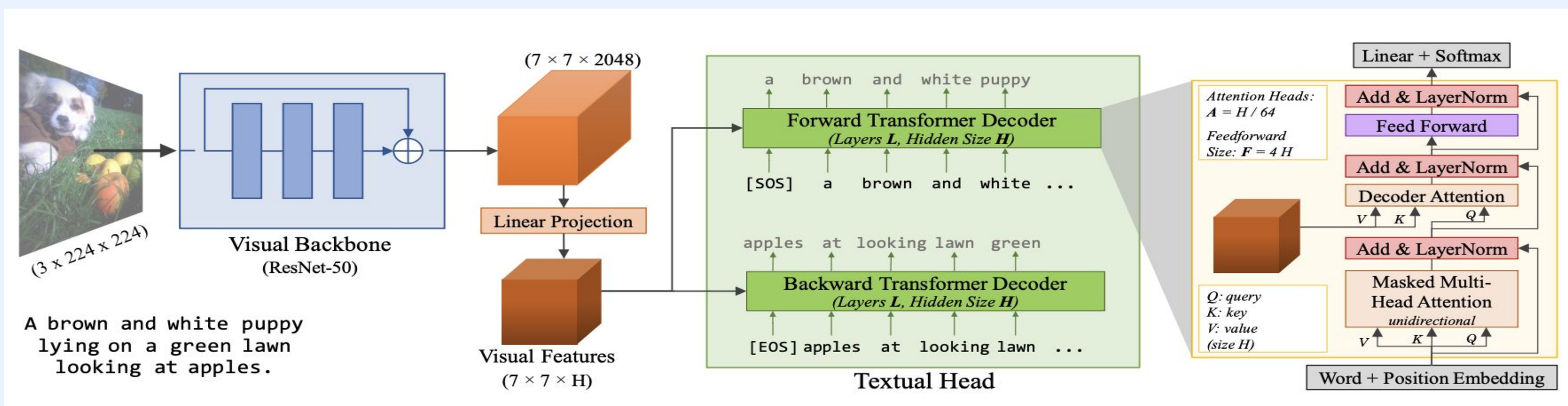


Fig:1 *Desai et al* ,"VirTex:Learning Visual Representations from Textual Annotations ",CVPR 2020

- **Target Extraction using Wordnet Database**
  - Used to connect nouns from generated captions to dataset labels
  - Lexical database linking words by their semantic similarity
  - Can be used with parts-of-speech tagging to calculate similarity code to generate words of interest

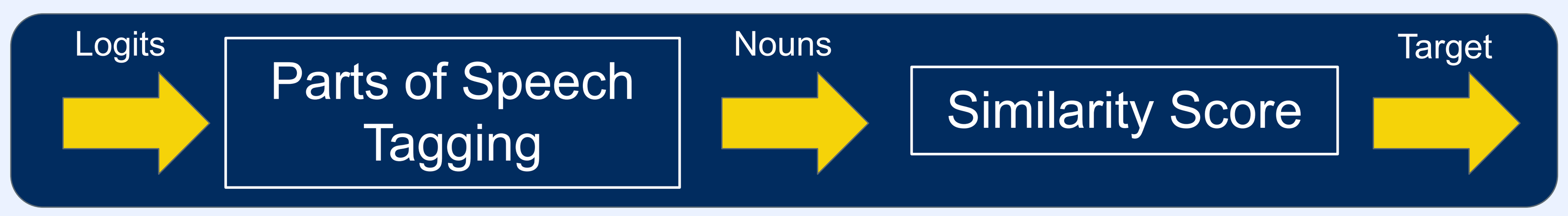


Fig:2 Workflow of Target Extraction Module

- **GradCam**
  - Computes gradient of input image with respect to words of interest in caption
  - Generates a heatmap localizing objects in image corresponding to a particular word in the generated captions
- **FCOS Object Detection Module**
  - FCOS is a fully convolutional one stage object detector which is anchor-free
  - We use a novel modified version of GIoU loss to account for the noise in the pseudo bounding boxes

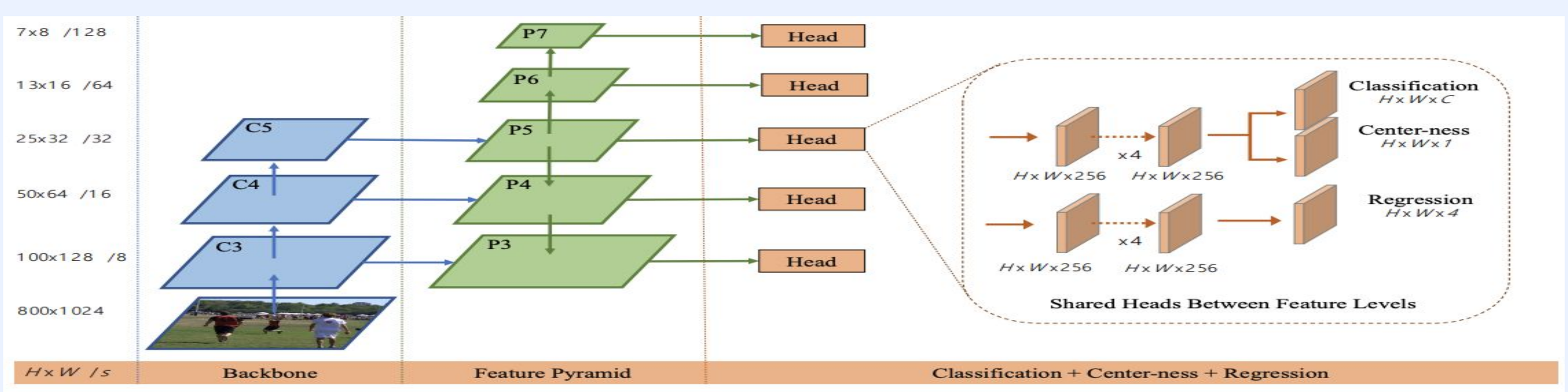


Fig 3: Tian et al, "FCOS: Fully Convolutional One-Stage Object Detection", ICCV 2019

Methodology

Our novel method consists of the following steps:

1. **Generating Captions :** We modified the existing VirTex-v2 to get output logits as well as logit to word mapping for the generated caption.
2. **Prompt Engineering :** Performed zero-shot transfer using the sub-prompt "I took a picture of and prompt "Itap" to generate pseudo class labels.
3. **Generating Heat Maps :**The pseudo class labels are used to generate class activation maps for that specific query to spatially locate the corresponding object.
4. **Object Detector :** Finally, we used the generated location information to train an object detector using novel modified box loss without any ground truth annotations.

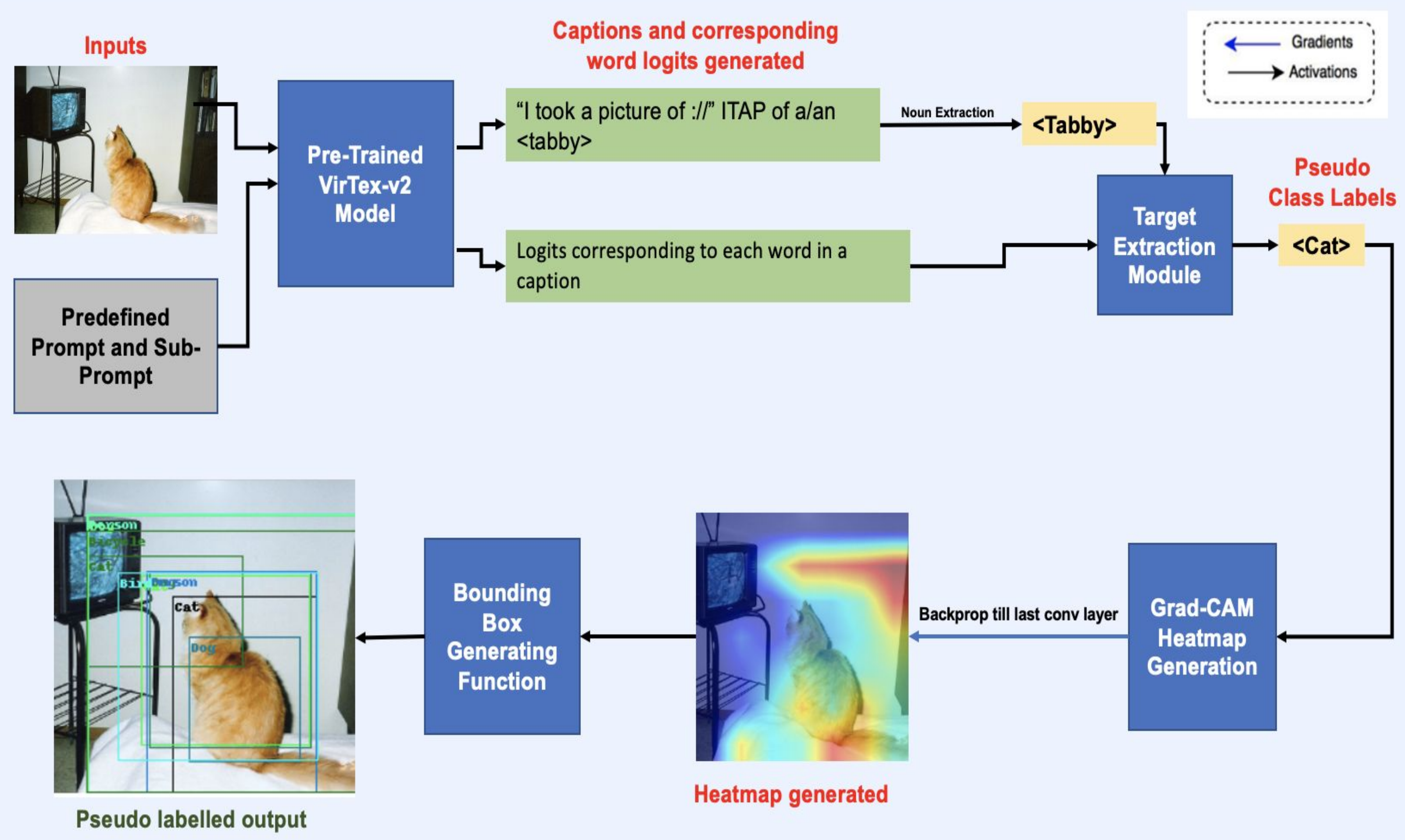


Fig:4 Illustration of our novel approach of object detection

Experiments and Results

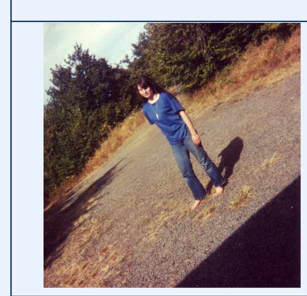
Input Image	Thresh	Logits	Nouns	Target(n)
	Logit: -7.13 Sim: 0.62	'picture', 'dog', 'person', 'man', 'girl', 'friend', 'beautiful', 'hand', 'boy', 'guy', 'blue', 'baby', 'beach', 'street', 'woman', …	'dog', 'person', 'man', 'girl', 'friend', 'boy', 'guy', 'baby', 'bird', 'woman', 'friends', 'bike', 'boat', …	person(40), dog(2), bicycle(2), sheep(2), bird(1), boat(1), car(1), cat(1), chair(1), cow(1), horse(1)
	Logit: -7.13 Sim: 0.62	'dog', 'beach', 'street', 'pup', 'bull', 'husky', 'bulldog', 'horse', 'cow', 'doggo', 'goat', 'sheep', 'stray', …	'dog', 'beach', 'street', 'pup', 'bull', 'husky', 'bulldog', 'horse', 'cow', 'goat', 'sheep', 'stray', …	dog(14), sheep(7), cat(4), bird(3), horse(1)

Fig:5 Two Representative cases showing Target Extraction using WordNet
Sim: similarity score between classes in the dataset and words in the generated captions
Logit: Probability of token appearing in the caption

Fig:6 Experimental Heat Maps Generated

Experiments and Results

- **Dataset Used for Evaluation and Fine-Tuning:** Pascal VOC 2007
- **Evaluation Metric Used:** mAP
- **Hyperparameters used for experimentation:** Logit Thresholds, GradCAM method, Duplicate Predictions Removal Function, usage of Eigen smoothing

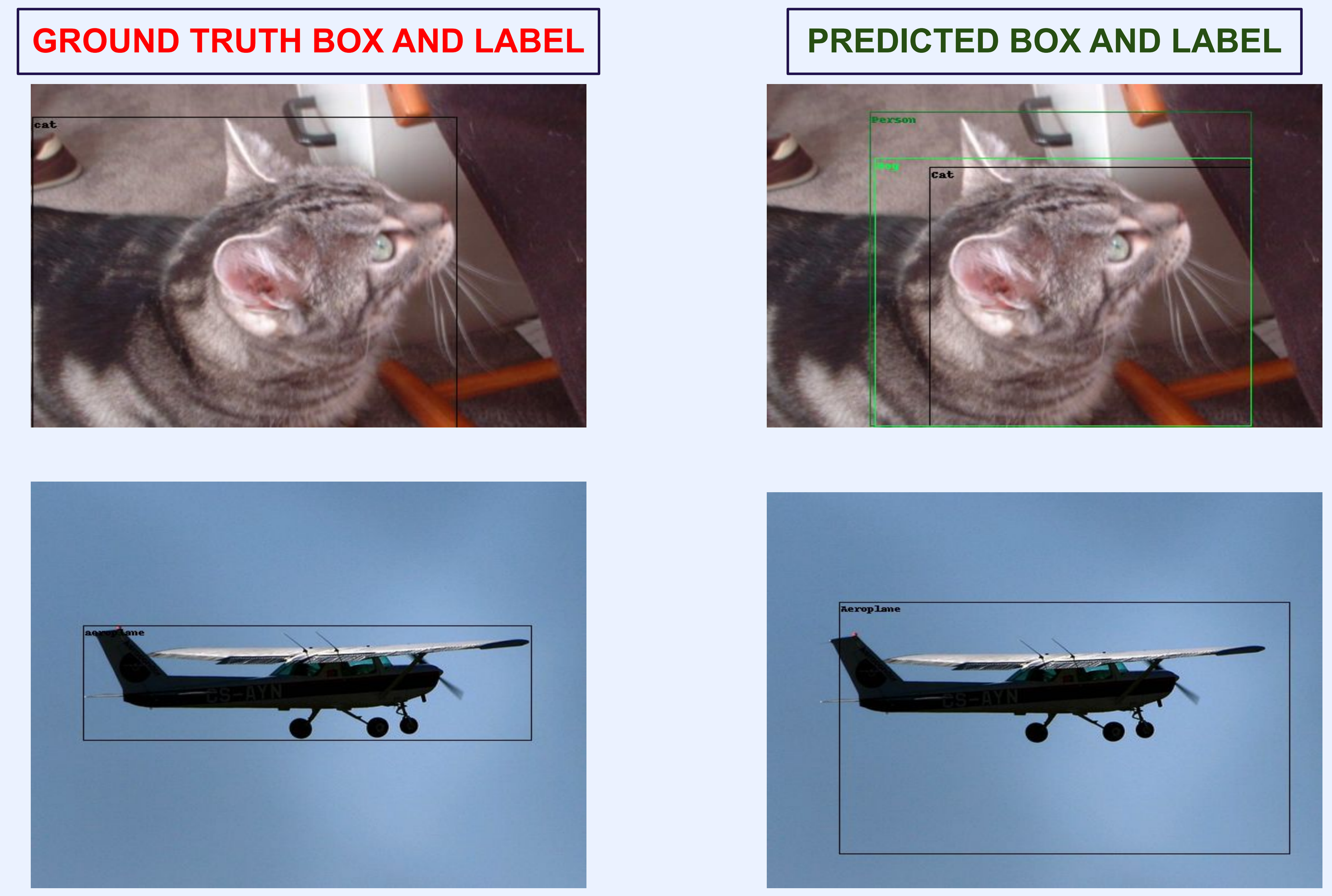


Fig 7: Illustration of Ground truth bounding boxes and Pseudo bounding boxes

mAP	GradCAM Method Used	Duplicate Removal Function	Eigen Smoothing
10.64%	GradCAM	Mean	No
12.4%	GradCAM++	Median	Yes
17.43%	**GradCAM++**	**Mean**	**No**

Fig 8: mAP results after testing with various hyperparameters

Conclusions

- With a pipeline combining image captioning, gradient heatmap generation, and hand crafted methods to fit each part together, we were able to create a system to generate labelled object detection data autonomously
- We were able to achieve the best mAP score of 17.43% when compared to the original datasets
- Although there is further work to be done, we have shown that our method of autonomous data generation is feasible and yields fruitful results

Future Work

- **Research improvements for bounding box noise**
  - Many bounding boxes are generated for each class and merging them into a single accurate bounding box is an ongoing challenge
  - It is difficult to differentiate between multiple instances of a single class in an image and multiple bounding boxes generated for the same instance - we currently threshold IoU of boxes to determine instance separation
- **Improvements to label assignment**
  - Some words, like names, are hard to assign to a particular class from text alone
  - Slang words or abbreviations are tough to deal with

Selected References

- K. Desai, G. Kaul, Z. Aysola, and J. Johnson. RedCaps: Web-curated image-text data created by the people, for the people. In NeurIPS Datasets and Benchmarks, 2021.
- R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision, 128(2):336–359, Oct 2019.