DangerCLIP: Open Domain Weapon Detection with Using CLIP

Seungjun Lee, Desmond Tan, Thomas Liang

National University of Singapore

Introduction

Overfitting to the training dataset is a common problem in the fully supervised setting. Specifically in weapon detection task, this overfitting issue might manifest itself in two ways: 1) Model lacks the ability to detect the unseen weapons from the source dataset. 2) Model is unable to identify danger if the images come from different domain than the source dataset. To mitigate this issue, we propose **DangerCLIP**, the novel Open-Domain Weapon Detection method with leveraging the generality and robustness of foundation model like CLIP [1]. Our contribution is three-fold:

1. We propose novel open-domain weapon detection method with leveraging CLIP [1] that can detect weapon or other forms of danger within the images from open-domain.
2. We show that giving dense supervision to the model can improve the classification performance of the model. By giving pseudo-mask generated from pretrained segmentation model as supervision, model can understand image more locally with being able to identify more smaller weapons.
3. Our method surpasses the baseline models with large margin in precision, recall and f1-score metrics, and also shows better robustness to the open-domain dataset.

Methodology

Baseline: We construct our baseline with using pretrained model as backbone and attaching simple MLP head as classifier. Classifier is followed by sigmoid function with outputting the probability that image contains the weapon. The model is given supervison with Binary Cross Entropy (BCE) loss which is the standard in binary classification task. However, this simple approach suffers from overfitting to the training dataset since the whole network is trained by the downstream dataset.

DangerCLIP: To mitigate the aforementioned overfitting issue, we try to leverage CLIP's powerful generalization capability in weapon detection task. More specifically, motivated by the observation of recent work [2] that the text embedding can be implicitly matchable to patch-level image embeddings, we add light-weight transformer on the fixed CLIP's model to exploit the matching capability of CLIP between text features and patch-level image features.

Let's denote $\mathbf{T}$ as text features extracted from CLIP's text encoder. We generate $\mathbf{T} \in \mathbb{R}^{1\times d}$ by feeding word "weapon" to the encoder. The n patches tokens of an image is denoted as $\mathbf{H} = [h_1, h_2, ..., h_n] \in \mathbb{R}^{n\times d}$, extracted from CLIP's image encoder. We concatenate text features with patches tokens which can be denoted as $C \in \mathbb{R}^{(n+1)\times d}$ and feed C to the transformer layer. Transformer consists of self-attention module and feed forward network with treating C as query, key and value. Owing to the global locality of self-attention, T is internally matched with every patches of H within the transformer, aggregating the information about whether image contains the weapon or not. Finally, we extract the first token of aggregated tokens C' which is output from the transformer layer (Fig. 1) and feed it to the classification head followed by the sigmoid function, outputting the probability of danger. This probability is supervised by BCE loss λ_{BCE} as well as our baseline. While the aforementioned method already achieves competitive performance, we add two more techniques to further improve the model:

1. **Relational Descriptor** : Referring to the ZegCLIP [3], we decide to adopt Relational Descriptor (RD) firstly proposed in ZegCLIP. RD leverages matching capability learned from the original CLIP as guidance for the transformer layer, enhancing the generality of the model. Refer to the original paper for the details.
2. **Dense supervision with pseudo mask** : While only using BCE loss for the training can allow model to understand the semantic meaning of the weapon, model would lack the ability to understand image densely since BCE loss is the image-level supervision, not the pixel-level supervision. It motivates us to give model the dense labels as supervision with generating pseudo mask from pretrained segmentation model. We extract the attention map $\mathcal{A} \in \mathbb{R}^{(n+1)\times(n+1)}$ from every self-attention modules in transformer layer and only regard $\mathcal{A}[0, 1:] \in \mathbb{R}^{1\times n}$ with discarding the remaining parts. $\mathcal{A}[0, 1:]$ indicates matching scores between "weapon" and the image patches which can serve as mask. Therefore, We give

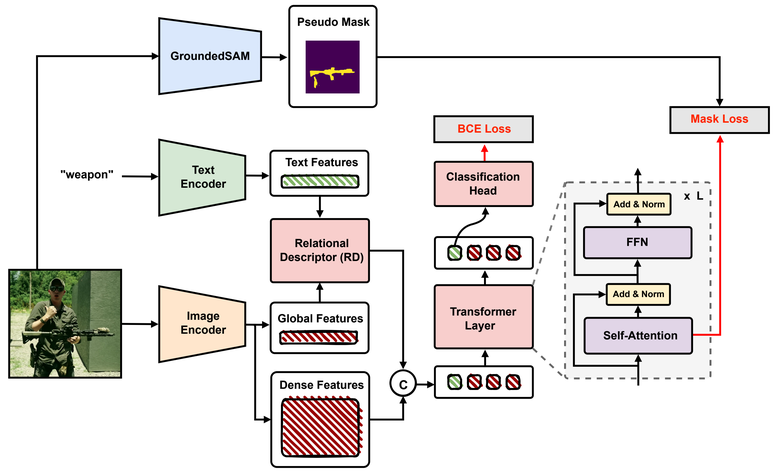

Figure 1. Overall architecture of DangerCLIP

pseudo mask as supervision to the $\mathcal{A}[0, 1:]$ with using focal loss [4] and dice loss [5]. The total objective can be formulated as : $\lambda_{total} = \alpha\lambda_{BCE} + \beta\lambda_{focal} + \gamma\lambda_{dice}$.

Implementation Details

DangerCLIP is based on Detectron2 [6] with using ViT-B/16 as the CLIP model. We only train transformer layer and classification head with using batch size of 32 for 2K iterations while CLIP's text encoder and image encoder are fixed. We use GroundedSAM [7, 8] as pretrained segmentation model to generate pseudo-mask for weapons within the dataset. The Number of transformer layers is set to 6. For the baseline, we adopt pretrained ResNet50 [9] and VGG16 [10] as backbone with fine-tuning it in the training phase. Both baselines are trained for 20 epochs with 0.005 learning rate. `RandomFlip` and `ColorAugSSDTransform` from Detectron2 are used for the data augmentation in all of the experiments. For the loss, we set $\alpha = 1$, $\beta = 2$, and $\gamma = 1$.

Experiments

Model	Precision	Recall	F1-score
VGG-16	41.85	60.33	49.42
Resnet50	56	58	57
DangerCLIP	**85.48**	**87.68**	**86.57**

Table 1. Quantitative result on source-domain dataset

Evaluation in source domain: We evaluate DangerCLIP on the test dataset provided by the course and compare the performance with our baselines. Precision, recall and f1-score are used to precisely evaluate the performance. As you can see in the Tab. 1, DangerCLIP surpasses the baselines in large margin with reporting **86.57** f1-score.

Model	F1-score
Resnet50	71.11
DangerCLIP	**89.23**

Table 2. Quantitative result on open-domain dataset

Evaluation in open domain: We evaluate DangerCLIP in the open-domain dataset to measure the generality of our model. To construct dataset that imposes different domain with the source dataset, we manually collect 50 historical or fictional images from the internet. As you can see in the Tab. 2, DangerCLIP shows higher f1-score than the baseline with proving the effectiveness of our method.

Ablation Study

Model	Precision	Recall	F1-score
DangerCLIP w/o dense supervision	75.69	**88.32**	81.52
DangerCLIP	**85.48**	87.68	**86.57**

Table 3. Comparison between *w dense supervision* and *w/o dense supervision*

We conduct ablation study to explore effectiveness of dense label supervision. As you can see in the Tab. 3, DangerCLIP supervised with pseudo mask shows higher f1-score than DangerCLIP trained without pseudo mask, reporting more balanced precision and recall score.

Supplementary Materials

Figure 2. Quality of pseudo mask from GroundedSAM (**Zoom in for details**).

Figure 3. Samples from collected open-domain dataset (**Zoom in for details**)

Challenges and Limitations

The current dataset is rather limited and only accounts for a small fraction of scenarios, which might pose a challenge when real world variability and environmental conditions that affect image quality are factored in. Additionally, multiple images in the dataset are labeled incorrectly.

There is not much precedent for measuring a model's generality across classification of all objects of one type, weapons in this case. More exploration must be done to standardize a metric for this.

Conclusion

In summary, the open domain model demonstrates promising results in detecting danger from static images over the baseline models VGG-16 as well as Resnet50. It is also effective in detecting danger across a wide variety of scenarios. The integration of the model into a system for alerting relevant authorities would definitely enhance public safety measures.

DangerCLIP: Open Domain Weapon Detection with Using CLIP

Seungjun Lee, Desmond Tan, Thomas Liang

National University of Singapore

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al.
Learning transferable visual models from natural language supervision.
In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[2] Chong Zhou, Chen Change Loy, and Bo Dai.
Extract free dense labels from clip.
In *European Conference on Computer Vision*, pages 696–712. Springer, 2022.

[3] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu.
Zegclip: Towards adapting clip for zero-shot semantic segmentation.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11175–11185, 2023.

[4] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi.
V-net: Fully convolutional neural networks for volumetric medical image segmentation.
In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016.

[5] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár.
Focal loss for dense object detection.
In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[6] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick.
Detectron2.
`https://github.com/facebookresearch/detectron2`, 2019.

[7] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick.
Segment anything.
arXiv:2304.02643, 2023.

[8] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al.
Grounding dino: Marrying dino with grounded pre-training for open-set object detection.
arXiv preprint arXiv:2303.05499, 2023.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun.
Deep residual learning for image recognition.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[10] Karen Simonyan and Andrew Zisserman.
Very deep convolutional networks for large-scale image recognition.
arXiv preprint arXiv:1409.1556, 2014.