NTU CE7454 Project 1: PSG Relation Classification

Siyue

Abstract

In addition to objects, images contain essential information such as relations between objects. This project aims to classify three salient relations in a given image. An exploratory data analysis was first conducted to understand the relation class distribution. A notable class imbalance was observed in the training dataset. To tackle the imbalance, images with frequent classes were under-sampled and images with infrequent classes were over-sampled. Three deep learning models have been used in the project: ResNet50, ResNet101, and Swin Transformer. The optimization and regularization techniques were applied for each model to improve the mean recall performance. The most effective measures included image augmentation, data whitening, learning rate adjustment, increasing binary cross-entropy loss weight for infrequent classes, and multiplying infrequent class prediction probabilities (i.e., reducing probability threshold). The model was re-trained in the union of training and validation datasets before being submitted to the testing dataset. Multi-task learning was implemented to leverage the semantic segmentation task to improve classification performance. Apart from the aforementioned approaches, zero-shot models were also proposed and explored to utilize the knowledge encapsulated in CLIP and Detectron2. Prompt engineering was performed for the relation classification. My optimized model is based on Swin Transformer at a test mean recall of 32.4%.

1. Introduction

This project is working on the problem of identifying 3 salient relations among objects in a given image. The problem originates from the Panoptic Scene Graph Generation (PSG) challenge proposed by [9].

A subset of the PSG dataset is provided for this project, which contains 4494 training images and 1001 validation images generated from the Microsoft COCO dataset. There are 50 classes of target relations/predicates to be predicted and 133 classes of objects existing in images. The pixel normalization statistics of training images are calculated as follows: mean $[0.495, 0.493, 0.491]$, standard deviation (std) $[0.320, 0.319, 0.320]$. Annotations of object detection and segmentation are also provided as supplemental data which can be used in the model training.

The relation classes are imbalanced in the training and validation datasets as shown in Fig. 1. There are 34 relation classes that have appeared less than 200 times in the training set. And only 5 classes have appeared more than 1000 times. A few relations do not exist in the training set but do in the validation set. This imbalance will make the model difficult to learn to predict infrequent classes. In order to obtain a high mean recall (mR) among classes, the detection of infrequent relations is of paramount importance. To balance the input dataset, images with frequent classes were randomly under-sampled and images with infrequent classes were manually over-sampled by repeating and image augmentation. The experiment result showed a significant mR improvement by data balancing.

2. Related Work

To solve this relation classification problem, a few related research works can provide good references and insights, e.g., Scene Graph Generation (SSG), Human-Object Interaction (HOI), and Visual Question Answering (VQA).

SSG aims to automatically map an image into a semantic structural scene graph, which requires the correct labeling of objects with bounding boxes and their relationships. SSG directly answers the need for relation classification by producing $\langle subject, verb, object\rangle$ triplets where the ”verb” is the relation to be predicted. State-of-the-art (SOTA) SSG models are based on CNNs and transformers [9].

HOI detects human and object instances and infers interactions between every pair of the detected instances, producing $\langle human, object, interaction\rangle$ triplets. As the most

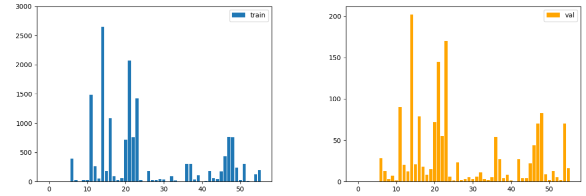


Figure 1. The imbalance of the frequency of 50 relation classes in training (blue) and validation (orange) datasets.

frequent object ”person” appears in more than 20% of images in this project, the knowledge of human-object relation is pivotal for the relation classification. SOTA HOI models include Graph Parsing Neural Network and Multi-task Human-Centric Network [1].

VQA systems take an image and a free-form, open-ended, natural-language question about the image as the input, and then produce a natural-language answer as the output [3]. Compared to SSG and HOI, VQA requires learning broader knowledge of image semantics such as location, time, quantity, and reasoning. We can formulate questions like ”Is this person driving a car?”, ”Where is the person?”, and ”What is the person doing?” for VQA systems to predict the salient relations in the image. SOTA VQA systems are such as BEiT-3 [7] and OFA [6].

3. Supervised Approach

The project begins by optimizing the baseline ResNet50 model with pre-trained weights in a supervised learning approach. Given the balanced training dataset, optimization is carried out to reduce training loss as much as possible. The same tuning is also performed for Swin Base Transformer (Swin B), which demonstrated better performance than ResNet50 and ResNet101.

The regularization is used to reduce the validation loss for better generalization capability. Fig. 2 shows the approximate mR improvement from applying optimization and regularization techniques sequentially from left to right.

Lastly, multi-task learning is explored, which aims to improve relation classification performance by learning to correctly detect objects or segment the image simultaneously.

3.1. Optimization: Lowering Train Loss

The initial image re-sizing dimension 1333 × 800 is first reduced to fit the computer's memory constraint. To leverage the transfer learning from the ImageNet1K dataset, the ResNet50 is customized with a 50-class classifier and then initialized with pre-trained weights. When pre-trained weights are used, input images are automatically re-sized to 232 × 232 and then cropped to 224 × 224.

The data whitening is carried out to speed up the convergence and improve the optimization. As the training set statistics (mean $[0.495, 0.493, 0.491]$, std $[0.320, 0.319, 0.320]$) is close to the ImageNet statistics (mean $[0.485, 0.456, 0.406]$, std $[0.229, 0.224, 0.225]$), channel-wise normalization based on both statistics improve the model recall significantly. Due to the pre-trained weights used in model initialization, the ImageNet statistics outperformed the training set statistics slightly. The mR was significantly improved from 7% to 16%.

Although a small batch size might lead to effective optimization, it fails to utilize parallel computing and produce terrible statistics for BN. The experiments on batch size 16 and 32 did not show apparent performance degradation when applying the larger value. To maximize the parallelism, batch size 32 was used in most experiments.

The learning rate (LR) is one of the most important parameters. The over-valued LR implies a large step toward the minimum loss, which could lead to jumping over the minimum and high training loss. The baseline LR 1e-3 was found over-valued because the decrease in training loss was too fast and the recall was not improved. The under-valued LR, such as 1e-5 as in Fig. 5a, leads to slow convergence and stagnates in the local minimum, which could also result in high training loss. The optimal value was found around 7e-5. A gradually decreasing LR makes the balance between the convergence speed and training loss. Besides the default cosine scheduler, the step scheduler, and cosine schedulers with warm-ups and restarts were tested. But no significant improvement was obtained, and the plain cosine scheduler was kept with the minimum rate tuned.

Two optimizers have been investigated in the project: SGD and Adam. Momentum is a key technique in SGD, which smooths out variations of the gradient and fastens convergence as found in Fig. 5b. In this case, a higher momentum is preferred because of the lower training loss and testing loss before the model becomes over-confident. Calibration such as label smoothing could be of use to reduce the testing loss rebound. As an excellent alternative to SGD optimizer, Adam adaptively scales the gradients and helps to reduce overshoots and stagnates. After experiments, Adma is more effective than SGD in the Swin B model while SGD performs better in the ResNet Models.

To enhance the ability to detect infrequent relations for higher recall value, apart from data balancing, the weights w_n of infrequent classes' binary cross-entropy (BCE) loss are increased to be higher than those of frequent classes' BCE loss in Eq. (1), and the predicted probability p_n of infrequent classes is enlarged by a multiplier m (e.g., 1.2) before the ranking for selecting top 3 classes. Thus, the chance of outputting infrequent classes is elevated, which benefits the mR performance.

$$\ell(\hat{y}, y) = \sum_{n=1}^{N} -w_n[y_n \cdot log\hat{y}_n + (1-y_n) \cdot log(1-\hat{y}_n)] \quad (1)$$

$$p_n = \begin{cases} m\hat{y}_n, & \text{if n is an infrequent relation class} \\ \hat{y}_n, & \text{otherwise} \end{cases} \quad (2)$$

3.2. Regularization: Lowering Test Loss

The model tends to be overfitted after optimization as it is specialized even capturing the noise in the training dataset and lacks generalization capability. The testing loss rebound is observed in most experiments. The regularization

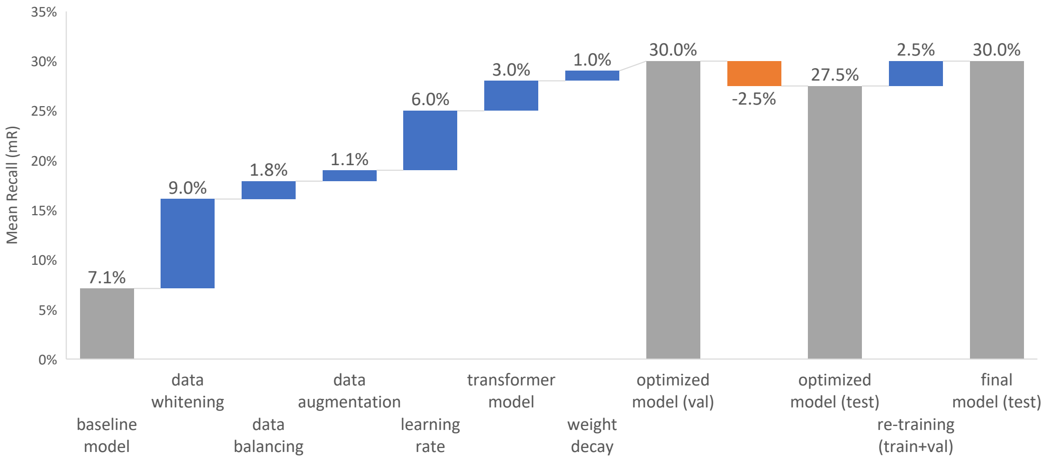


Figure 2. Performance improvement (mR) by sequentially applying optimization and regularization techniques.

techniques are introduced to discourage learning a more complex model, so as to avoid the risk of overfitting, including L1/L2 regularization, weight decay, batch normalization (BN), and data augmentation.

L1 regularization adds an L1 penalty (i.e., L1 norm of weights) to the BCE loss, which punishes the increase in model complexity. As observed in ResNet50 as Fig. 5c, the training loss is increased due to L1 regularization, and the testing loss is indeed reduced by introducing L1 norm. Similarly, L2 regularization adds an L2 norm of weights to the loss, which mathematically has the same effect as weight decay. A comparison between training with weight decay and without weight decay is demonstrated in Fig. 5d. Both training and testing loss reduction are observed as expected.

An over-valued l1 penalty or weight decay rate (e.g., 0.01) constrains the model complexity so that the model fails to learn complex features and has a high training error. The mR improvement from L1/L2 regularization is not significant based on experiment results.

BN has been widely used in convolutional layers and linear layers of models, which stabilizes the learning and improves the network's generalization properties. Image augmentation such as random flipping, color jetting, and cropping also helps to prevent the model from overfitting to the training dataset.

After obtaining the best combination of hyperparameters and before submitting to the test dataset, the model is retrained in the union of training and validation datasets after image balancing to leverage all labeled data available. As Fig. 2 illustrated, the model performance dropped by 2.5% when testing with unseen images, which was offset by retraining with images from both training and validation sets.

3.3. Multi-task Learning

Logically, people first identify objects in the image and then deduce relations based on objects' location, shape, size, intrinsic correlation, etc. Based on this intuition, the relation classification can be regarded as a high-level task. And low-level tasks such as classifying objects, detecting appearance and location, and learning pair-wise correlation are supposed to contribute to the high-level task. This explains the principle behind multi-task learning. As we inject more knowledge into the model by providing detection/segmentation annotations, better accuracy is expected from multi-task learning.

3.3.1 Relation classification and segmentation

Semantic segmentation contains fine-grained pixel information about object class, location, shape, size, and so on. To leverage this image information, I connected an auxiliary relation classifier to the latent features in the DeeplabV3 model as Fig. 3. Google DeeplabV3 is one of SOTA image segmentation models, which employs atrous convolution with upsampled filters to extract dense feature maps and to capture long-range context [2].

Fine-tuning multi-task learning models is more complicated because contributions from two tasks need to be well-balanced. The cross-entropy loss or dice loss for segmentation is naturally much larger than BCE loss for relation classification since the former is calculated for all pixels and all object classes while the latter is only computed for all relation classes. From experiments in Tab. 1, segmentation loss needs to be reduced to a similar magnitude with classification loss otherwise the classification performance deteriorates.

To verify the learning progress of the multi-task learning model, model predictions and testing metrics are visual-

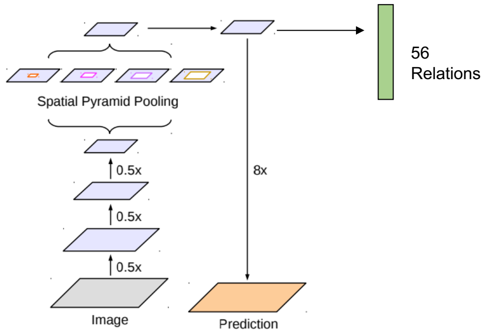


Figure 3. DeeplabV3 based multi-task learning architecture: relation classification and segmentation.

λ_{seg}	L_{seg}/L_{cls}	IOU	Accuracy	mR
4e-10	1.8	0.53	0.995	30.3%
1e-8	40.9	0.52	0.995	25.8%

Table 1. Experiment on the weight of segmentation loss. (L_{seg} segmentation loss, L_{cls} classification loss)

ized in Fig. 4. It was observed that the predicted segmentation mask was gradually becoming close to the ground truth mask during the training. Meanwhile, segmentation metrics, e.g., accuracy, recall, and IOU, and the mR of relation classification were improved altogether.

Although both classification and segmentation performance are improved during the training, the mR of the multi-task model has not outperformed the supervised model at around 30%. More time and effort will be needed for the architecture adjustment and subtle fine-tuning to further improve IOU and mR.

3.3.2 Relation classification, segmentation and object detection

The success of combining classification and segmentation motivates the further exploration of combining tasks of classification, segmentation, and object detection. Although object detection provides similar information to segmentation, the existence of object detection is expected to facilitate and ease task solving. The Mask R-CNN model is a suitable basis that already produces segmentation and classes with bounding boxes. A new head for relation classification can be added to the Mask R-CNN model as Fig. 6. Due to the limited data, the main feature extraction network could be fixed during the re-training. Seesaw loss could be beneficial to solve the long-tail issue of the class labels [5]. Due to the time limit, this idea has not been implemented.

4. Zero-shot Approach

Zero-shot learning can provide different solutions to solve this relation classification problem with the help of language without seeing the images provided.

4.1. CLIP

OpenAI CLIP, consisting of an image encoder and a text encoder, is trained on 400 million image-text pairs for predicting the level of alignment between the image and the text description [4]. To use CLIP for classification, prompt engineering is the key. The following strategies are tested:

(a) Relation-focused: Raw relations are kept and the template is ”relation of {relation}”. Validation mR is 30.5%.

(b) Action-focused: Raw relations are modified to emphasize the action, e.g., from ”playing with” to ”playing”. The template is ”action of {modified relation}”. While the validation mR is 34.2%, the test mR is 28.7%. The robustness of this kind of method is questionable considering such a big performance drop.

4.2. Detectron and CLIP

Experiments have been done to check if adding objects to the prompt can improve performance. Detectron2 [8] is used to detect the instances firstly. Candidate texts are generated by one or two instances detected and each relation from the PSG predicate list (Fig. 7). Using CLIP, candidate texts are ordered by the predicted alignment value with the image. However, the validation result shows that mR is 20.3% lower than previous results. As the instance detection by Detectron2 was verified as accurate, the error mainly comes from the CLIP.

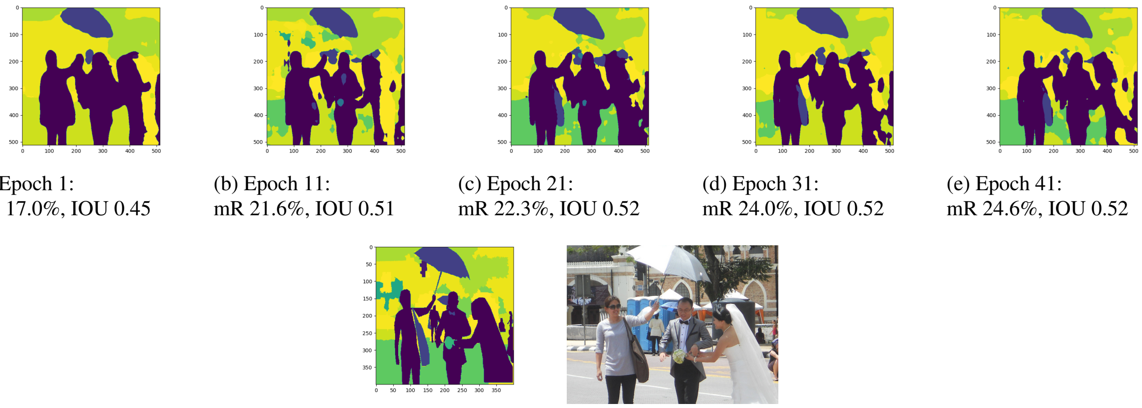
Supervised methods still outperform. Supplementary texts about relations and continual learning will be needed to enhance the CLIP model’s performance for this project.

5. Conclusion

Label imbalance is the most severe issue in the given datasets. Under-sampling and over-sampling are utilized to create a more balanced training dataset. ResNet50, ResNet101, and Swin B models have been optimized and regularized for the highest mR. Models are re-trained by training and validation images before testing. Multi-task learning models and zero-shot models have been proposed and explored preliminarily. The optimized Swin B model has the highest mR of 32.4% currently.

References

[1] Trevor Bergstrom and Humphrey Shi. Human-object interaction detection: A quick survey and examination of methods. *CoRR*, abs/2009.12950, 2020. 2



(a) Epoch 1:
mR 17.0%, IOU 0.45

(b) Epoch 11:
mR 21.6%, IOU 0.51

(c) Epoch 21:
mR 22.3%, IOU 0.52

(d) Epoch 31:
mR 24.0%, IOU 0.52

(e) Epoch 41:
mR 24.6%, IOU 0.52

(f) Ground truth mask

(g) Original figure

Figure 4. Evolution of predicted segmentation mask and relation classification mR in validation dataset for multi-task learning.

[2] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation, 2017. 3

[3] Sruthy Manmadhan and Binsu C Kovoor. Visual question answering: a state-of-the-art review. *Artificial Intelligence Review*, 53(8):5705–5745, 2020. 2

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 4

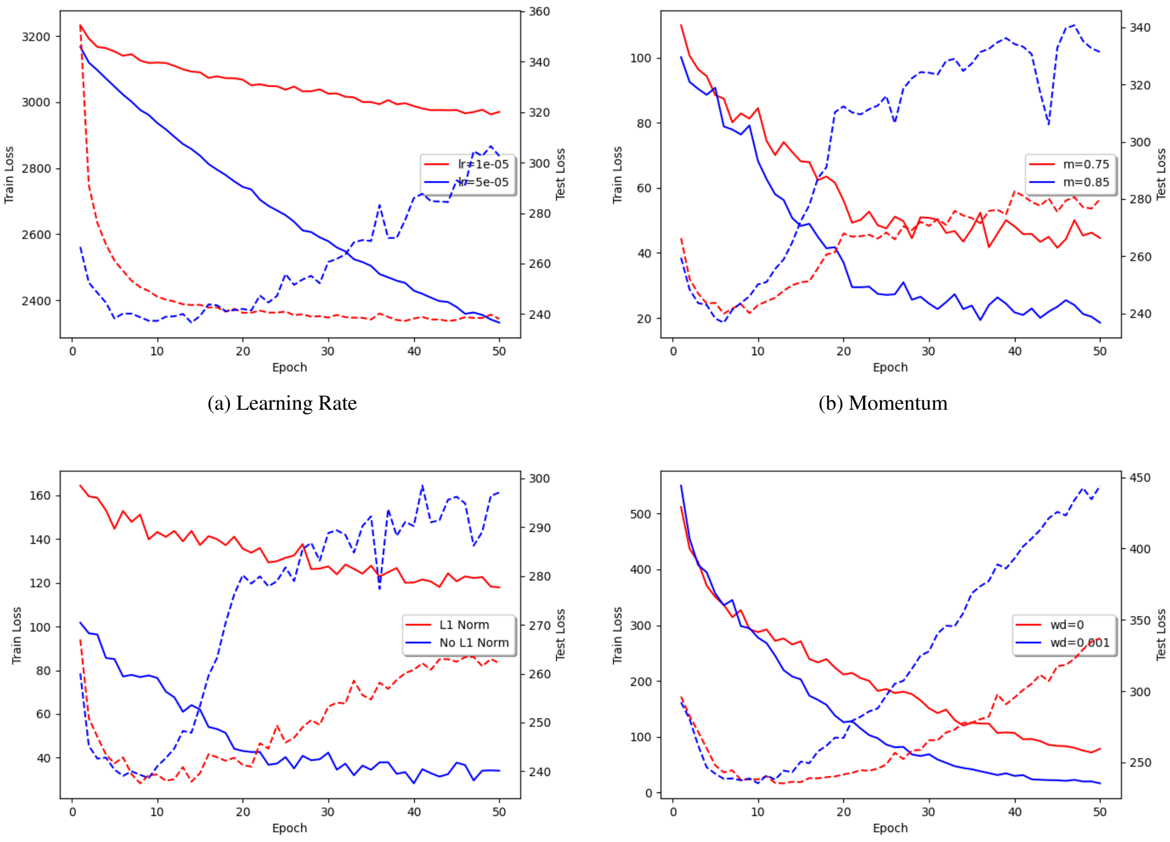
[5] Jiaqi Wang, Wenwei Zhang, Yuhang Zang, Yuhang Cao, Jiangmiao Pang, Tao Gong, Kai Chen, Ziwei Liu, Chen Change Loy, and Dahua Lin. Seesaw loss for long-tailed instance segmentation, 2020. 4

[6] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework, 2022. 2

[7] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: Beit pretraining for all vision and vision-language tasks, 2022. 2

[8] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 4

[9] Jingkang Yang, Yi Zhe Ang, Zujin Guo, Kaiyang Zhou, Wayne Zhang, and Ziwei Liu. Panoptic scene graph generation, 2022. 1



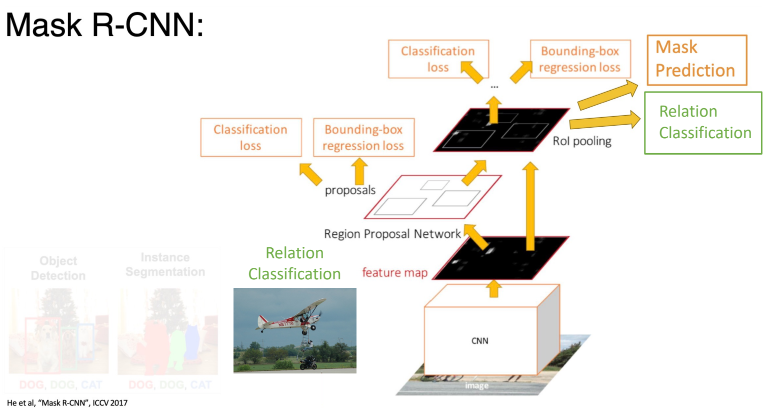
(a) Learning Rate

(b) Momentum

(c) L1 Regularization

(d) Weight Decay

Figure 5. Optimization and regularization training logs. (Training loss: solid lines, Validation loss: dashed lines)



Mask R-CNN:

He et al, "Mask R-CNN", ICCV 2017

Figure 6. Mask-RCNN based multi-task learning architecture: relation classification, instance segmentation, and object detection.

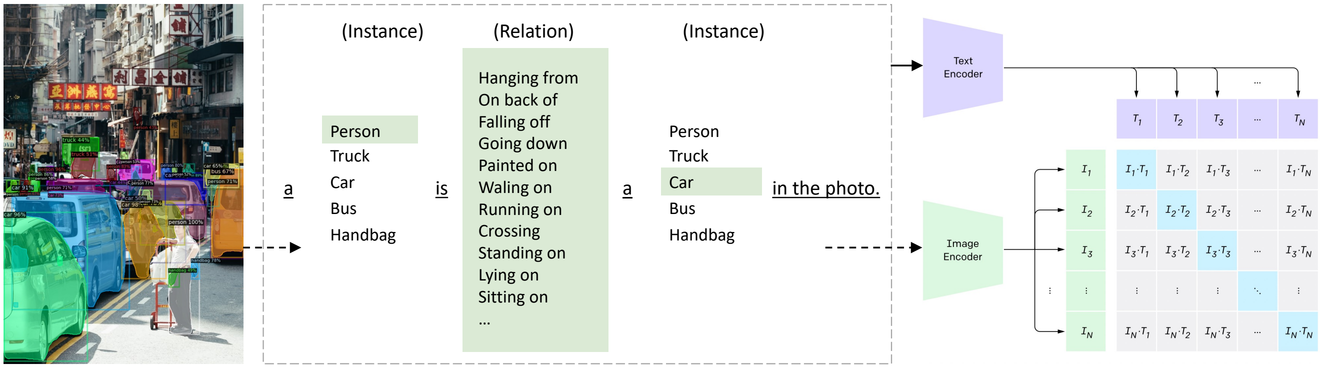

(Instance)
(Relation)
(Instance)
a
Person
Truck
Car
Bus
Handbag
is
Hanging from
On back of
Falling off
Going down
Painted on
Waling on
Running on
Crossing
Standing on
Lying on
Sitting on
...
a
Person
Truck
Car
Bus
Handbag
in the photo.
Text Encoder
Image Encoder
T_1 T_2 T_3 ... T_N
I_1 I_2 I_3 ⋮ I_N
$I_1{\cdot}T_1$ $I_1{\cdot}T_2$ $I_1{\cdot}T_3$ ... $I_1{\cdot}T_N$
$I_2{\cdot}T_1$ $I_2{\cdot}T_2$ $I_2{\cdot}T_3$ ... $I_2{\cdot}T_N$
$I_3{\cdot}T_1$ $I_3{\cdot}T_2$ $I_3{\cdot}T_3$ ... $I_3{\cdot}T_N$
$I_N{\cdot}T_1$ $I_N{\cdot}T_2$ $I_N{\cdot}T_3$ ... $I_N{\cdot}T_N$

Figure 7. Creating description texts using relation classes and Detectron2 instances.