2nd Place Solution to Open Images 2019 - Visual Relationship

Takuya Ito
Universal Knowledge Inc.
tito@universal-knowledge.jp

Abstract

This article describes the model that achieved 2nd place in the Open Images 2019 - Visual Relationship Detection Challenge on Kaggle.

1. Object Detection

I made cascade-rcnn model using mmdetection. Configuration is almost same as X-101-64x4d-FPN in mmdetection model zoo[1]. Main difference is img scale which was set to (1024, 768).

2. Visual Relationship

This part can be split up into two subproblems. The first subproblem involves relation ’is’ for example *chair is wooden*. And the second subproblem is based on triplet relationships, such as *chair at table*.

2.1. Relation ’is’

I made 3 models for this part, and then I made ensemble of them.

2.1.1 Relation ’is’ (2 Stage Model)

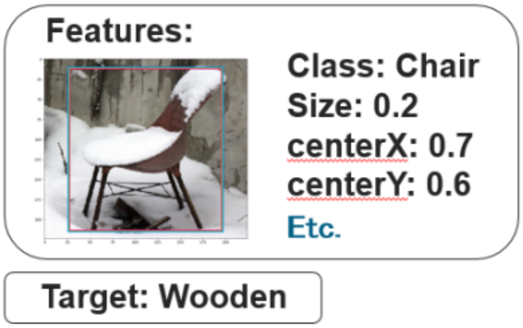

Fig. 1. Relation ’is ’ (2 Stage Model)

I made training data by mapping the ground truth BBOX(challenge-2019-train-vrd-bbox.csv) to ground truth relational data(challenge-2019-train-vrd.csv), giving a target.

- Target: Material (wooden, plastic, ..., None).
- Features: Cropped image, BBOX class, BB position / size etc.

Example:

bbox data

image1,bbox1,Man
image1,bbox2,Guiter
image1,bbox3,Chair

Relationship data

image1, bbox2, bbox2, Guitar, Wooden, is
image1, bbox3, bbox3, Chair, Wooden, is
image1, bbox1, bbox2, Man, Guitar, hold

Then, training data would be

image1,bbox1,bbox1, Man,None,is
image1,bbox2,bbox2, Guiter,Wooden,is
image1,bbox3,bbox3, Chair ,Wooden,is

Here, target is None if the data is not in Relationship data.

2.1.2 Relation ’is’ (1 Stage Model)

The distinct triplets of “ is ” relation has only 42 classes. I made cascade-rcnn model which detect 42 ’is-relation’ classes.

2.1.3 Relation ’is’ (1 Stage Model with Material Head)

I added ’material’ detection head to cascade-rcnn. This model predict Bounding Box and class and material at the same time.

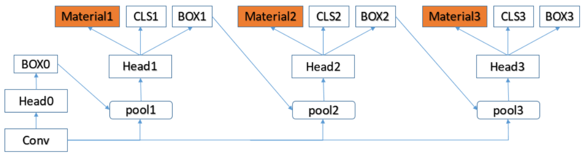

Fig. 2. Relation ’is ’ (1 Stage Model with Material Head)

2.1.4 Results of Relation ’is’

Table 1 is public and private LB scores for relation ’is’.

Table 1. Public And Private LB Score of Relation ’is’

model	public	private
2.1.1	0.07523	0.07264
2.1.2	0.08332	0.08075
2.1.3	0.08191	0.07948
ensemble	0.08514	0.08232

I expected model2-1-3 to have better score...

2.2. Triplet Relationships

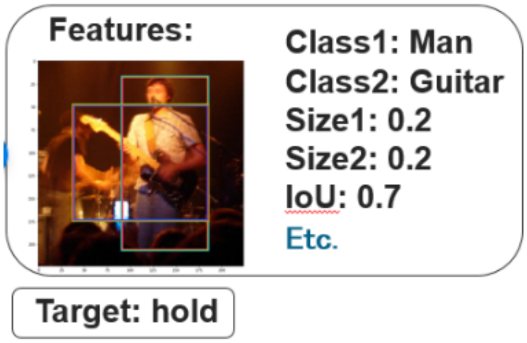

Fig. 3. Triplet Relationships

First, I created all pairs of BBOX in the same image with some filter. Then, I made training data by mapping the pairs to ground truth relational data, giving a target.

- Target: Relationship (at, on, ..., None).
- Features: Cropped image including two BBOX with box line, LabelName1, LabelName2, XCenter1, YCenter1, XCenter2, YCenter2, Size1, Size2, Aspect1, Aspect2, XCenterDiff, YCenterDiff, CenterDiff, XCenter, YCenter, IOU

Example:

bbox data

```
image1,bbox1,Man
image1,bbox2,Guiter
image1,bbox3,Chair
```

Relationship data

```
image1,bbox1,bbox2,Man,Guiter,hold
```

Then, training data would be

```
image1,bbox1,bbox2,Man,Guiter,hold
image1,bbox1,bbox3,Man,Chair,None
image1,bbox2,bbox1,Guiter,Man,None
image1,bbox2,bbox3,Guiter,Chair,None
image1,bbox3,bbox1,Chair,Man,None
image1,bbox3,bbox2,Chair,Guiter,None
```

here, target is None if the data is not in Relationship data.

I made 8 expert models which only in charge of small sample class and made ensemble of them with weighted average of their probability.

Table 2. Expart Models

model name	target classes
full model	at,on,holds,plays,interacts,inside,wears,hits,under
expert model1	on,holds,plays,interacts,inside,wears,hits,under
expert model2	holds,plays,interacts,inside,wears,hits,under
expert model3	plays,interacts,inside,wears,hits,under
expert model4	interacts,inside,wears,hits,under
expert model5	inside,wears,hits,under
expert model6	wears,hits,under
expert model7	hits,under
expert model8	under

Table3 is the result AP for validation data:

Table 3. Result AP and mAP

class	ground truth BB	predicted BB
at	93%	31%
on	92%	32%
holds	89%	54%
plays	94%	58%
interacts with	82%	45%
inside of	72%	37%
wears	94%	55%
hits	55%	57%
under	50%	20%
mAP without hits/under	88%	45%
mAP	80%	43%

For ground truth BB pairs, this relationships prediction model has very high accuracy. mAP without hits/under which have very small samples is 88%.

3. Final Score Prediction

I just used simple formula for model2-2.

$$Score = SubjectScore * ObjectScore * RelationsipScore \quad (1)$$

This year, my LB score improved to 0.38818 from last year score 0.23709. Most of this improvement comes from object detection improvement.

It seems that good object detection is the most important part of this competition.

4. Experiments

In order to see the importance of image feature to detect the relationship, I made a model without CNN and checked its performance by AP and LB scores.

Table 4 is the result of no CNN model AP for validation data, and Table 5 is the result LB Scores:

Table 4. Ablation Study: AP and mAP

class	ground truth BB	predicted BB
at	86%	29%
on	85%	27%
holds	82%	49%
plays	88%	53%
interacts_with	79%	41%
inside_of	68%	33%
wears	99%	56%
hits	26%	22%
under	100%	100%
mAP without hits/under	84%	41%
mAP	79%	46%

Table 5. Ablation Study: Result LB Score

LB (without 'is')	without CNN	with CNN
Public LB	0.38267	0.44079
Private LB	0.36283	0.38818

Private LB Score dropped about 0.025 from 0.38818 to 0.36283.

5. Hardware

I used local 1080ti x 2 and titan TRX. And in the very ending of this competition, I used V100 x 8 instance on GCP.

These resources are shared by 3 open image competitions.

6. Data and Pre-Trained Networks Used

I did not use external dataset. I used pre-trained weights for initialization of model2.1.1 and 2.1.2, 2.1.3, 2.2.

References

[1] mmdetection model zoo https://github.com/open-mmlab/mmdetection/blob/master/docs/MODEL_ZOO.md