Towards Context-aware Interaction Recognition for Visual Relationship Detection

Zhang, Liqiang

May 13,2018

1 Object Interaction Recognition

Object interaction recognition is a fundamental problem in computer vision and it can serve as a critical component for solving many visual recognition problems such as action recognition, visual phrase recognition, sentence to image retrieval and visual question answering. There are two ways to model the interaction and its context. The first one treats the combination of interaction and its context as a single class, but it is very inefficient to collect training images for each combination. Another way is to model the interaction and the context separately, but which can lead to poor recognition performance due to the difficulty of associating the interaction with certain visual appearance in the absence of context information.

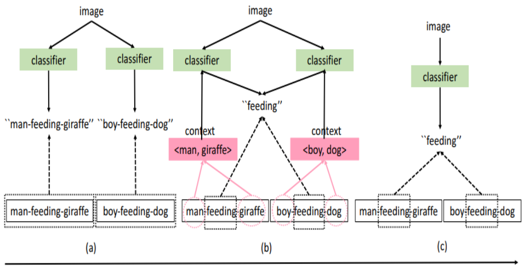

Figure 1: (Comparison of two baseline interaction recognition methods and the proposed approach.

The two baseline methods take two extremes. For one extreme, Fig. 1(a) treats the combination of the interaction and its context as a single class. For another extreme, Fig. 1(c) classifies the interaction separately from its context. Our method Fig. 1(b) lies somewhere between Fig. 1(a) and Fig. 1(c). We still build one classifier for each interaction but the classifier parameter is also adaptive to the context of the interaction, as shown in the example in Fig. 1(b).

2 Methods

The author assume that interaction classifier takes a linear classifier form $y_p = \mathbf{w}_p^T \phi(I), \mathbf{w}_p \in R^d$, where$y_p$ is the classification score for the p-th interaction and is the feature representation extracted from the input image. The classifier parameters for the p-th interaction $\mathbf{w}_p$ are a function of $(O1, O2)$, that is, the context of the p-th h interaction. It is designed as the summation of the following two terms:

$$\mathbf{w}_p(O1, O2) = \bar{\mathbf{w}} - p + r_p(O1, O2), \qquad (1)$$

where the first term ¯ wp is independent of the context; it plays a role which is similar to the traditional context-independent interaction classifier. The second term rp$(O1, O2)$ can be viewed as an auxiliary classifier generated from the information of context $(O1, O2)$.
Eq. 1 are distinct per interaction type p while

the projection matrix Q is shared across all interactions. All of these parameters are learnt at training time.

3 Evaluation on the visual phrase dataset

As seen from Table 4, AP+C+CAT again achieves the best performance. In comparison with the performance of [1], our method improves most in the zero-shot learning setting.

Method	Phrase	Detection	Zero-Shot
Visual Phrase	52.7	49.3	-
Language Priors	82.7	78.1	23.9
Baseline1-app	70.1	65.6	12.4
Baseline1-spatial	68.3	63.6	10.3
Baseline2-app	77.5	72.3	11.0
Baseline2-spatial	15.7	10.4	1.1
Spatial+C	84.9	80.8	27.6
AP+C	85.9	81.6	28.5
AP+C+AT	86.2	82.1	28.8
AP+C+CAT	**86.8**	**82.9**	**30.2**

Table 1: Comparison of performance on the Visual Phrase dataset.

References

[1] Khurram Soomro and Amir Roshan Zamir. Ucf101: A dataset of 101 human actions classes from videos in the wild. *Computer Science*, 2012.