Towards Context-aware Interaction Recognition for Visual Relationship Detection

Zhang, Liqiang

May 17,2018

1 Visual Relationship Detection

Recognizing how objects interact with each other is a crucial task in visual recognition. Then most current methods to define the context of the interaction suffer limitations, some scales poorly with the number of combinations and fails to generalize to unseen combinations, the other often leads to poor interaction recognition performance due to the difficulty of designing a interaction classifier [2]. To mitigate those drawbacks, the author proposes an alternative, context-aware interaction recognition framework. The key to the method is to explicitly construct an interaction classifier which combines the context, and the interaction. The context is encoded via word2vec into a semantic space, and is used to derive a classification result for the interaction. Fig. 2 is Comparison of two baseline interaction recognition methods and the proposed approach.

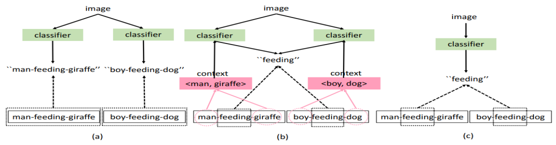

Figure 1: The two baseline methods take two extremes. (a) treats the combination of the interaction and its context as a single class. (c) classifies the interaction separately from its context. (b) lies somewhere between (a) and (c).

2 Context-aware interaction classification framework

In general, an interaction and its context can be expressed as a triplet $O1-P-O2$, where P denotes the interaction, and $O1$ and $O2$ denote its subject and object respectively. In this study, the author assume the interaction context $(O1, O2)$ has been detected by a detector. It is designed as the summation of the following two terms:

$$w_p(O1, O2) = \bar{w}_p + r_p(O11, O2), \qquad (1)$$

where the first term $\bar{w}_p$ is independent of the context; it plays a role which is similar to the traditional context-independent interaction classifier. The second term $r_p(O1, O2)$ can be viewed as an auxiliary classifier generated from the information of context $(O1, O2)$.

$$r_p(O1, O2) = V_P f(QE(O1, O2)), \qquad (2)$$

where $E(O1, O2) \in R^{2e}$ is the concatenation of the edimensional word2vec embeddings of $(O1, O2)$. Note that V_P and $\bar{w}_p$ in Eq. 2 are distinct per interaction type p while the projection matrix Q 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 1, AP+C+CAT again achieves the best performance. In comparison with the performance of [1], the 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] L Itti. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Trans Image Process*, 13(10):1304–1318, 2004.

[2] Tie Liu, Jian Sun, Nan Ning Zheng, Xiaoou Tang, and Heung Yeung Shum. Learning to detect a salient object. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, 2007.