Knowledge Aided Consistency for Weakly Supervised Phrase Grounding (Supplementary Material)

Kan Chen Jiyang Gao Ram Nevatia

1. Hyperparameters Analysis

We evaluate KAC Net for different sets of hyperparameters. To evaluate one hyperparameter, we fix other hyperparameters to default values in Sec. 4.2.

Flickr30K Entities. In Table 1, we evaluate KAC Net's performance with different visual reconstruction loss weight λ. We observe the performance of KAC Net increases when λ is less than 10.0 and decreases when λ further increases.

Weight λ	0.5	1.0	5.0	10.0	20.0
Accuracy (%)	32.09	33.17	37.13	38.71	36.11

Table 1. KAC Net's performances on Flickr30K Entities for different weights λ of $\mathcal{L}_{lc}^{k}$.

We then evaluate KAC Net's performance for different dimensions m for multimodal features in Eq. 3. In Table 2, we observe KAC Net's performance is low when m is small ($m = 32$). When m is in the range from 64 to 128, KAC Net achieves small gain in performance. When m is larger than 128, the performance drops a lot.

Dimension m	32	64	128	256
Accuracy (%)	32.07	37.23	38.71	33.78

Table 2. KAC Net's performances on Flickr30K Entities for different dimensions m of $\mathbf{v}_i^q$.

Weight λ	0.5	1.0	5.0	10.0	20.0
Accuracy (%)	12.21	13.37	15.81	15.83	15.75

Table 3. KAC Net's performances on Referit Game for different weights λ of $\mathcal{L}_{lc}^{k}$.

Referit Game. We first evaluate KAC Net's performance for different weights λ for visual reconstruction loss $\mathcal{L}_{lc}^{k}$. As shown in Table 3, when λ is small, the performance is close to pure language consistency branch's performance (GroundeR [1] model). When λ increases, the effectiveness of visual consistency branch brings increase in grounding accuracy. When λ becomes too large, the performance decreases a little.

Dimension m	32	64	128	256
Accuracy (%)	11.74	14.95	15.83	12.68

Table 4. KAC Net's performances on Referit Game for different dimensions m of $\mathbf{v}_i^q$.

Threshold c	0.1	0.3	0.5	0.7	0.9
# queries	1987	5173	9285	12778	14870
GroundeR	25.57	26.10	27.03	28.51	29.05
G + KBP	26.32	27.55	29.84	31.29	32.06
KAC Net	28.15	29.32	31.35	34.01	36.98

Table 5. Different methods on Flickr30K Entities [31] for Type B queries under different thresholds. Accuracy is in %.

We then evaluate the KAC Net's performance for different multimodal dimensions m for $\mathbf{v}_i^q$ in Eq.3. In Table 4, we observe performance keeps increasing when $m < 128$, and decreases a lot when $m > 128$.

2. Discussion of Generalizability

Generalizability on open-world phrase grounding problem can be evaluated by calculating similarities between queries and names of fixed detection categories. Using the same notations in Sec. 4.5, we focus on Type B queries whose maximum similarity with 80 MSCOCO categories' names is less than a threshold, c. We vary c in the range of [0.1, 0.9], and evaluate different methods on these subsets of Type-B queries. The performance comparison is provided in Table 5. We find that when queries become more different to detection categories, visual features (VGG_{det}) used by GroundeR [1], which are pre-trained on PASCAL, also do not provide more useful information, and have a performance drop. However, KAC Net achieves consistently better performance compared with other methods, which shows the generalizability of visual consistency part for relatively different queries compared to fixed detection categories. We observe the improvement in performance from G+KBP to KAC Net becomes higher as c increases.

References

[1] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele. Grounding of textual phrases in images by reconstruction. In *ECCV*, 2016. 1