

Supplementary Materials: Visual-linguistic Cross-domain Feature Learning with Group Attention and Gamma-correct Gated Fusion for Extracting Commonsense Knowledge

Anonymous Authors

1 MORE EXPERIMENTAL RESULTS

1.1 Effects of Gamma Correction

To investigate the best choice of γ for the proposed Gamma-corrected Gated Fusion, additional experiments have been carried out by varying the γ value from 1 to 5, while keeping all the other components unchanged. The ablation results are presented in Table 1. It can be witnessed that when the γ value increases from 1 to 3, the results on each evaluation metric consistently show performance gains, *i.e.*, 0.93%, 0.35%, 0.18%, 1.13%, 1.90%, and 0.56% improvement on AUC, F1, P@2%, mAUC, mF1, and mP@2%, respectively. These improvements indicate that enlarging the contrast between the informative instances and less informative ones indeed helps reduce the negative effects of less informative instances. However, when the value of γ increases from 3 to 5, the performance starts to decrease by 3.34%, 3.27%, 3.77%, 2.61%, 2.73%, and 0.94% on AUC, F1, P@2%, mAUC, mF1, and mP@2%, respectively. This decrease in performance may be attributed to an overly large contrast that suppresses some informative instances by only highlighting the most informative ones. Considering these results, $\gamma = 3$ is hence selected as the default value.

Table 1: Ablation study of γ in Gamma-corrected gated fusion.

Value	AUC	F1	P@2%	mAUC	mF1	mP@2%
1	48.23	50.95	48.05	27.96	42.50	24.14
2	48.51	51.01	48.00	28.59	43.69	24.38
3	49.16	51.30	48.23	29.09	44.40	24.70
4	48.34	50.84	47.47	27.32	42.10	24.33
5	45.82	48.03	44.46	26.48	41.67	23.76

1.2 Plots of Gated Ratios

To visualize the effects of the proposed Gamma-corrected Gated Fusion, we plot the gated ratios for a bag of instances without Gamma Correction in an ascending order, and the gated ratios for the corresponding instances with Gamma Correction in Figure 1. From Figure 1, it can be observed that when Gamma correction is not applied, the gated ratios do not change significantly, and hence the negative effects of non-informative instances will significantly influence the group-level decision on the predicted entity relations. However, when Gamma correction is applied, the gated ratios vary more significantly for different instances, where the weights become smaller for less informative ones and become larger for more informative ones. As a result, the model will focus more on the informative instances and reduce the negative effects of non-informative ones. This clearly demonstrates the benefits of the proposed Gamma-corrected Gated Fusion.

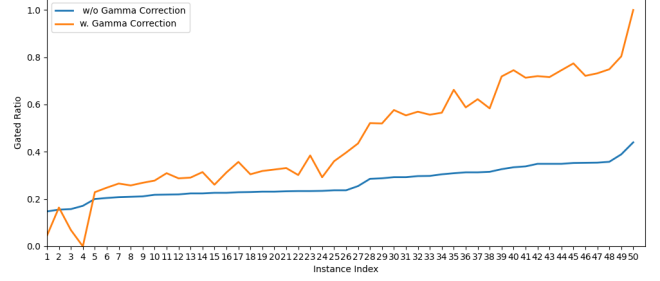


Figure 1: Plot of the gated ratios for a bag of instances with/without Gamma correction. When no Gamma correction is applied, the gated ratios for different instances do not vary significantly, while after applying Gamma correction, the gated ratios show significantly different values, where those non-informative ones have smaller values and the informative ones have larger values.

1.3 Per-class Comparisons to CLEVER

We have demonstrated the superiority of the proposed method over the previous best solution model, CLEVER [2] in the manuscript. We further compare the performance in precision@2% for each relation between CLEVER [2] and the proposed method, and list the top 20 relations that exhibit the most significant performance gain in Figure 2.

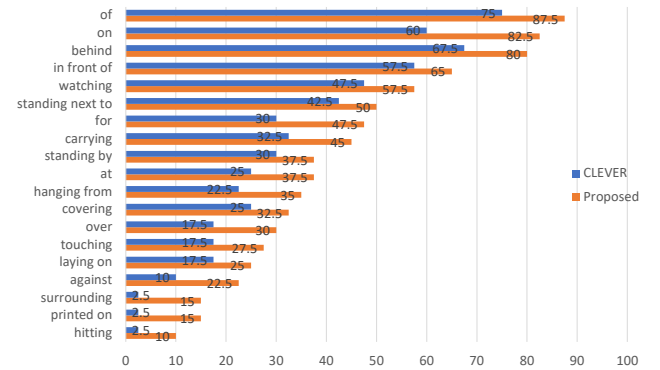


Figure 2: Compared with CLEVER [2] for the top 20 relations that exhibit the most significant performance gain in terms of precision@2%.

First of all, as observed from Figure 2, most of the performance gains are significant. The proposed method surpasses CLEVER [2] by a minimum of 10% across thirteen relation types, achieving a

notable performance gain of 22.5% and 17.5% for the relational categories ‘on’ and ‘for’, respectively. Secondly, even for some common relations such as ‘of’, ‘on’ and ‘behind’, where CLEVER already yields good prediction results, the proposed method still achieves significantly better performance. Lastly, it is indeed difficult to predict some less common relations, such as ‘printed on’ and ‘hitting’, for both methods, while the proposed method consistently outperforms CLEVER in these cases as well.

1.4 More Visual Comparison Results with CLEVER

We provide additional visual comparison results with CLEVER in Figure 3. Relation facts that are missed by CLEVER but accurately extracted by our method are highlighted in green, while relations that are incorrectly predicted are marked in red.

As depicted in the first three cases in Figure 3, the proposed model is capable of summarizing more precise relation triplets than CLEVER, such as (*dog*, ‘with’, *man*) in the first example, (*man*, ‘standing next to’, *cow*) in the second, and (*bottle*, ‘sitting on’, *bench*) in the third. These relational triplets, revealed in corresponding captions, demonstrate the advantages of leveraging general domain knowledge and the effectiveness of the proposed multi-modal cross-domain feature learning.

In addition, the proposed method correctly predicts relations that CLEVER [2] inaccurately extracts. As illustrated in the fourth row of Figure 3, the proposed method correctly identifies the relations between the ‘bottle’ and ‘bowl’ as ‘behind’, ‘near’ and ‘on’, while CLEVER [2] mistakenly identifies the additional relationship of ‘in’. Similar cases are evident in the last two rows, *i.e.*, CLEVER inaccurately summarizes (*bike*, ‘has’/‘on’/‘under’, *dog*) and (*woman*, ‘above’/‘on’/‘under’/‘with’, *giraffe*). The higher accuracy of the proposed model may be attributed to the proposed Group Attention module, which greatly enhances the representations of single-instance features, coupled with the Gamma-corrected Gated Fusion to mitigate the influence of noisy instances.

1.5 Failure Cases

In the manuscript, we conducted an analysis of failure cases and categorized them into two types. 1) The first type comprises triplets that are not labeled in the VG-CKE dataset but can be inferred from the images. 2) The second type consists of incorrectly recognized relational facts by the proposed method.

Besides the examples shown in the manuscript, we provide additional failure cases here in Figure 4, highlighting the first type of failure cases with red and underlines, while highlighting the second type of failure cases with red only. For instance, in the first case, the relational triplets (*giraffe*, ‘behind’/‘near’, ‘bird’) are reasonably summarized by the proposed method, yet not included in the ground-truth annotations. The relational facts in the benchmark are automatically created by aligning relational facts from knowledge bases to the Visual Genome dataset [1] through distant supervision, potentially resulting in missing labels. Despite these challenges, the proposed method correctly identifies these relations from images. Similar cases can be observed from rows 2 to 4 in Figure 4.

Samples of the second type of failure cases are also exemplified in Figure 4. For instance, in the last two rows, the relational facts of

‘behind’ and ‘under’ for the entity pair (*bag*, *dog*), ‘in front of’ and ‘on’ for the entity pair (*zebra*, *car*) are erroneously inferred by both the proposed model and CLEVER [2]. CLEVER [2] has identified more incorrect relations than the proposed method, as evident from the respective images.

2 MORE DETAILS OF VG-CKE DATASET

In this study, we adopt the large dataset developed by Yao *et al.* [2] for multi-modal commonsense knowledge extraction. As the images in this dataset are sourced from the Visual Genome dataset [1], it is denoted as VG-CKE. The VG-CKE dataset [2] focuses on the top 100 entities and relations. We list all these entities and potential relations of this benchmark in Table 2 and Table 3, respectively.

The VG-CKE dataset [2] formulates pairs of entities by selecting the subject and object from Table 2 and aligning relational facts in Table 3 to these pairs automatically through distant supervision. Some instances are mislabeled due to distant supervision [2]. The VG-CKE dataset comprises 6,443/1,964/678 pairs of entities and 13,780/3,496/1,166 associated commonsense facts for training, testing, and validation, respectively. Images are organized as image-bags for each entity pair. Ultimately, there are 55,911 images for training, 13,722 for testing, and 5,224 for validation.

REFERENCES

- [1] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV* 123 (2017), 32–73.
- [2] Yuan Yao, Tianyu Yu, Ao Zhang, Mengdi Li, Ruobing Xie, Cornelius Weber, Zhiyuan Liu, Hai-Tao Zheng, Stefan Wernter, Tat-Seng Chua, and Maosong Sun. 2023. Visually Grounded Commonsense Knowledge Acquisition. *AAAI* 37 (2023), 6583–6592.


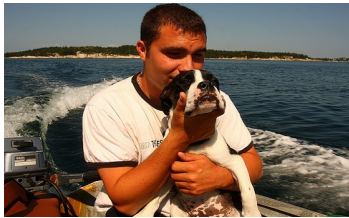










		Entity Pair: (dog, man) CLEVER: behind, in front of, near, on Proposed: behind, in front of, looking at , near, on, with Caption1: A man sitting on a couch with a dog Caption2: A man sitting on a boat with a dog
		Entity Pair: (man, cow) CLEVER: behind, holding, near, watching, with Proposed: behind, holding, looking at , near, watching, with, standing next to Caption1: A man standing next to a group of cows Caption2: A man standing next to a brown and white cow
		Entity Pair: (bottle, bench) CLEVER: above, in , near, on Proposed: above, near, on, sitting on , under Caption1: A bottle of water sitting on top of a wooden bench Caption2: A blue water bottle sitting on top of a wooden bench
		Entity Pair: (bottle, bowl) CLEVER: behind, in , near, on Proposed: behind, near, on Caption1: A table topped with a bowl of fruit and a bottle of wine Caption2: A bowl of food sitting on a counter next to a bottle of wine
		Entity Pair: (bike, dog) CLEVER: behind, has , on , under , with Proposed: behind, with Caption1: A brown dog standing next to a bike Caption2: A little girl riding a bike with a dog
		Entity Pair: (woman, giraffe) CLEVER: above , near, on , under , with Proposed: near Caption1: A woman holding a baby in front of a giraffe Caption2: A giraffe standing next to a woman on a horse

Figure 3: Visual comparison with CLEVER on the VG-CKE dataset [2]. The proposed method accurately extracts more relational facts from the specified entity pairs. Facts that are missed by CLEVER but correctly summarized by the proposed methods are highlighted in green while the relations that are wrongly predicted are highlighted in red. Compared to CLEVER, the proposed method predicts less incorrect relation labels and more correct relation labels.



Entity Pair: (giraffe, bird)
 CLEVER: behind, near, has, holding, with
 Proposed: near, behind, with

Caption1: A bird perched on top of a giraffe
 Caption2: A giraffe standing next to a bunch of birds



Entity Pair: (book, lamp)
 CLEVER: behind, near
 Proposed: behind, near

Caption1: A lamp sitting on top of a table next to a book
 Caption2: A bookshelf filled with books and a lamp



Entity Pair: (bowl, dog)
 CLEVER: behind, in front of, near, of, on
 Proposed: behind, for, in front of, near, on

Caption1: A dog is sniffing a bowl of food
 Caption2: A black dog sitting next to a bowl of food



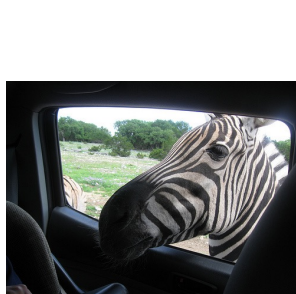
Entity Pair: (motorcycles, sidewalk)
 CLEVER: in, near, on, sitting on
 Proposed: at, in, near, on, parked on

Caption1: Motorcycles are parked on the sidewalk in front of a brick building
 Caption2: Motorcycles are lined up on a sidewalk



Entity Pair: (bag, dog)
 CLEVER: behind, in front of, near, of, on, under
 Proposed: behind, near, on, under

Caption1: A brown dog sitting on top of a luggage bag
 Caption2: A dog laying on the ground next to a bag of luggage



Entity Pair: (zebra, car)
 CLEVER: behind, in front of, near, in, on
 Proposed: in, in front of, near, on

Caption1: A zebra standing in the back seat of a car
 Caption2: A zebra is looking out the window of a car

Figure 4: Visualization of the failure cases. The predicted relation labels mismatched with the ground truth are highlighted in red. The relations not labeled in the VG-CKE dataset [2] but reasonably identified by the proposed method or CLEVER are underlined. Compared to CLEVER, the proposed method discovers more new reasonable relations while retains the error rates.

Table 2: List of entities in the VG-CKE dataset [2].

arm	bottle	chair	face	handle	leaf	nose	roof	snow	truck
bag	bowl	clock	fence	hat	leg	pant	seat	street	trunk
banana	box	coat	flower	head	letter	paper	sheep	surfboard	umbrella
beach	boy	cow	food	helmet	light	person	shelf	table	wave
bench	branch	cup	giraffe	hill	logo	pillow	shirt	tail	wheel
bike	building	dog	girl	horse	man	plant	shoe	tile	window
bird	bus	door	glass	house	motorcycle	plate	short	tire	wing
board	cabinet	ear	glove	jacket	mountain	pole	sidewalk	track	wire
boat	cap	elephant	hair	jean	mouth	post	sign	train	woman
book	car	eye	hand	lamp	neck	rock	skateboard	tree	zebra

Table 3: List of relations in the VG-CKE dataset [2].

above	between	driving on	hanging from	in middle of	near	parked on	shows	standing on	walking
across	built into	eating	hanging in	laying in	next	part of	sitting	supporting	walking down
adorning	carrying	filled with	has	laying on	of	playing	sitting at	surrounding	walking in
against	connected to	floating in	has on	leaning on	on	playing with	sitting on	swinging	walking on
along	contains	flying	held by	lining	on back of	printed on	sitting on top of	throwing	watching
at	covered in	for	hitting	looking at	on bottom of	pulling	standing	to	wearing
attached to	covered with	from	holding	lying on	outside	resting on	standing behind	touching	wears
behind	covering	full of	holding up	lying on top of	over	riding	standing by	under	with
belonging to	crossing	growing in	in	made of	painted on	riding on	standing near	underneath	worn by
beneath	cutting	growing on	in front of	mounted on	parked in	says	standing next to	using	written on