More Than Just Attention: Improving Cross-Modal Attentions with Contrastive Constraints for Image-Text Matching (Supplementary Material)

Yuxiao Chen[1], Jianbo Yuan[2], Long Zhao[1], Tianlang Chen[2], Rui Luo[2],
Larry Davis[2], Dimitris N. Metaxas[1]

[1]Rutgers University, [2]Amazon.com Services, Inc

1. Implementation Details

SCAN and PFAN. These two methods separately train the text-to-image attention models where words are query fragments, and the image-to-text attention models where image regions are query fragments. When training the text-to-image attention models, we randomly sample one word fragment from each matched image-text pair to apply the proposed constraints. The image-to-text attention models are trained in a similar way by sampling image fragments.

BFAN. The method jointly trains the text-to-image and image-to-text attention models. In order to jointly supervise both attention models and reduce computation cost, for each matched image-text pair, we apply our constraints to either a sampled word fragment for the text-to-image attention model or a sampled image region for the image-to-text attention model with a probability of 50%.

SGRAF. This approach has a text-to-image attention model to learn the alignment between words and regions. We randomly sample one word fragment from each matched image-text pair to apply the proposed constraints.

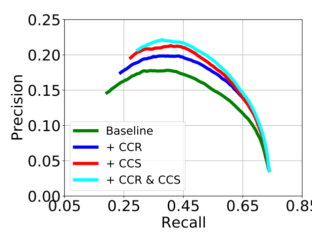
2. Additional Attention Evaluation Results

Quantitative Analysis. To demonstrate the influence of different T_{IoU} on Attention Precision, Attention Recall, and Attention F1-Score, we report the results when T_{IoU} is set to 0.6 on Table 1 and Figure 1. We can observe similar performance improvements as when T_{IoU} is set to 0.4 (shown in the main paper). It demonstrates that the proposed constraints achieve consistently better results than baseline methods when different T_{IoU} values are chosen.

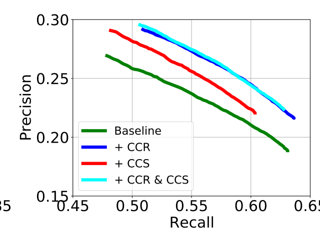
Qualitative Analysis. We report three cases of BFAN and SGRAF trained on the Flickr30K (see Figure 2 and 4) and MS-COCO dataset (see Figure 3 and 5). We find that the attention models trained with the proposed constraints can assign attention weights more accurately than the correspondent baselines across different datasets.

Method	Attention Precision	Attention Recall	Attention F1-Score
SCAN	16.88	47.40	22.88
+ CCR	18.87	**49.96**	25.05
+ CCS	20.30	48.58	26.22
+ CCR & CCS	**21.31**	47.15	**26.80**
BFAN	27.50	46.52	31.92
+ CCR	30.17	48.72	34.49
+ CCS	29.55	46.97	33.55
+ CCR & CCS	**30.24**	**49.20**	**34.69**
SGRAF	25.93	47.93	31.23
+ CCR	27.13	48.49	32.27
+ CCS	28.24	48.31	33.08
+ CCR & CCS	**28.94**	**49.20**	**33.94**

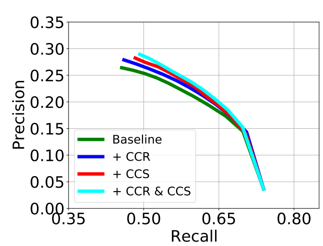
Table 1: Attention Precision, Attention Recall, and Attention F1-Score (%) of the SCAN, BFAN, and SGRAF models trained on the Flickr30K dataset when T_{IoU} is 0.6.



(a) PR curves of SCAN



(b) PR curves of BFAN



(c) PR curves of SGRAF

Figure 1: Attention PR curves of SCAN, BFAN, and SGRAF trained on the Flickr30K dataset when T_{IoU} is 0.6.

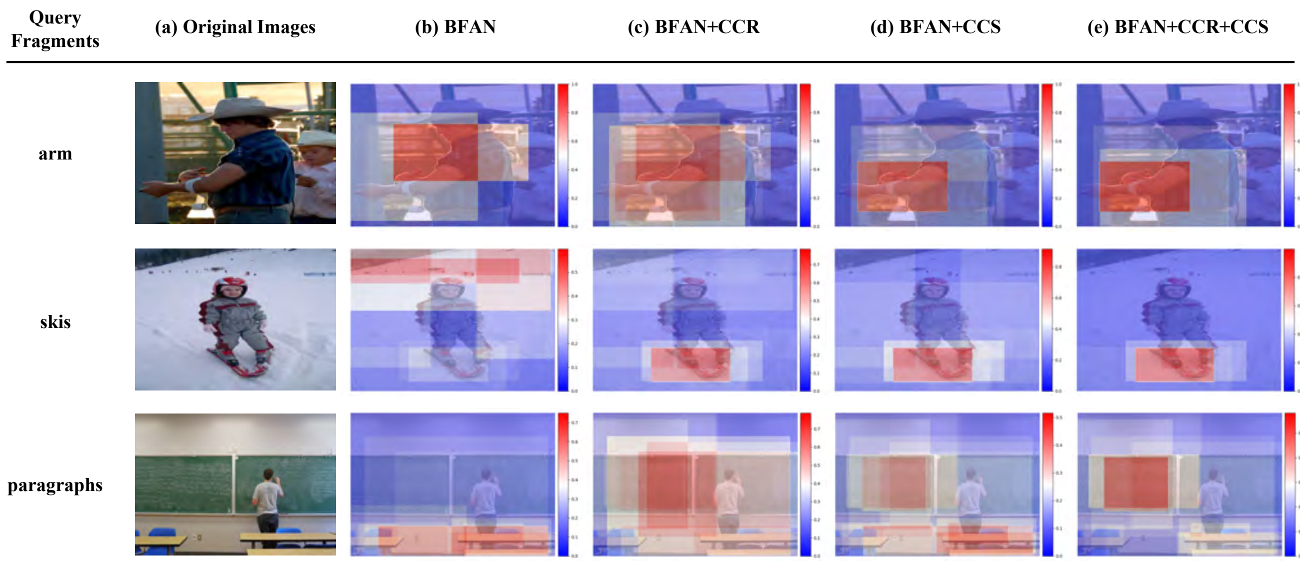


Figure 2: Examples of attended image regions with respect to the given words for the BFAN model on the Flickr30K dataset.

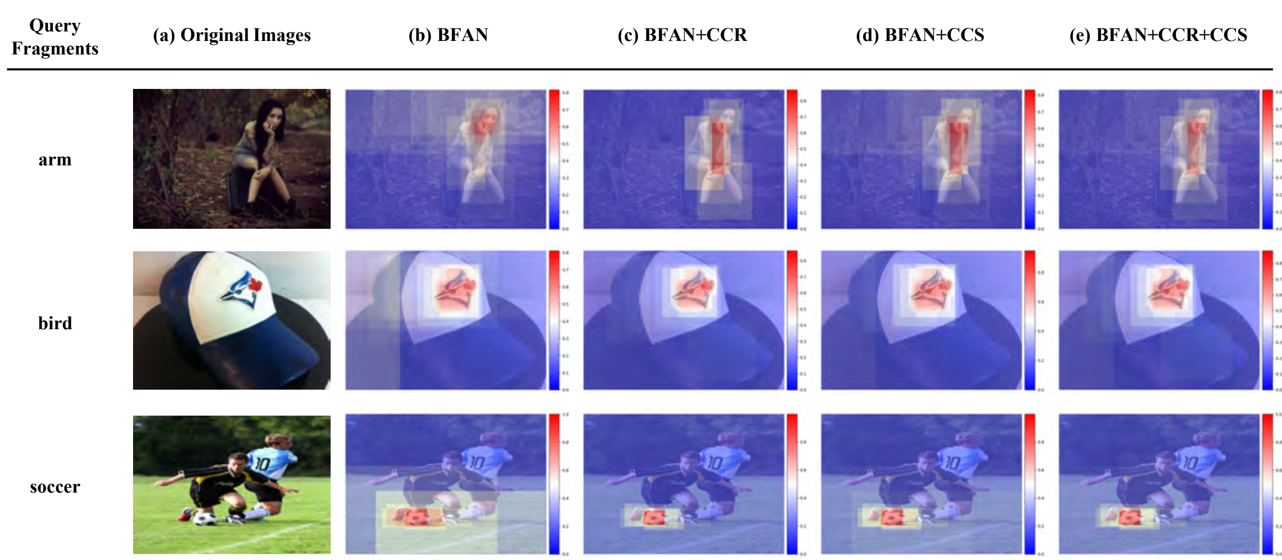


Figure 3: Examples of attended image regions with respect to the given words for the BFAN model on the MS-COCO dataset.

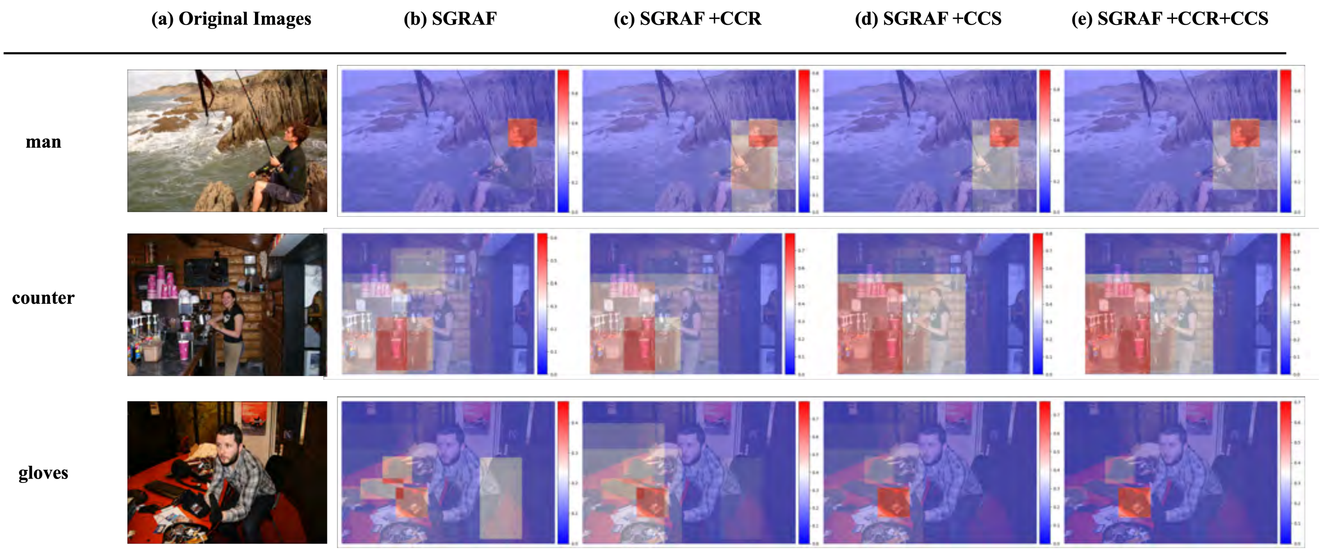


Figure 4: Examples of attended image regions with respect to the given words for the SGRAF model on the Flickr30K dataset.

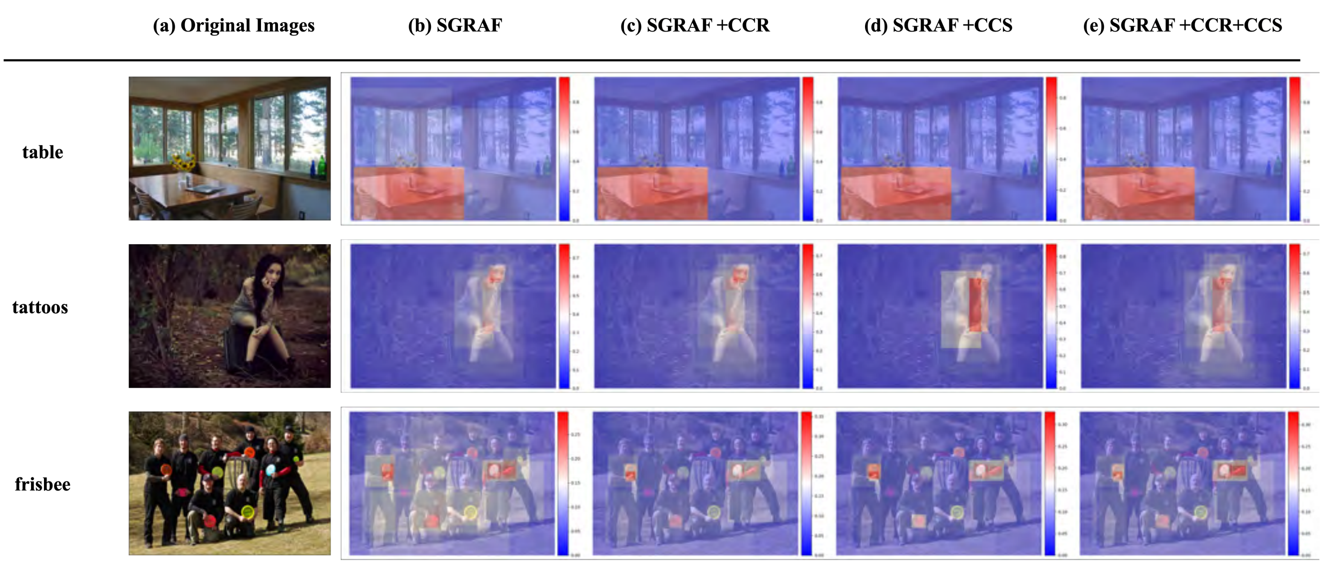


Figure 5: Examples of attended image regions with respect to the given words for the SGRAF model on the MS-COCO dataset.