Supplementary Material for "Advancing weakly supervised cross-domain alignment with optimal transport"

1 IPOT algorithm

We use the inexact proximal point method optimal transport algorithm (IPOT) [8] to compute optimal transport. The algorithm and implementation details are shown in Algorithm 1.

Algorithm 1 IPOT($\mathbf{E}, \mathbf{V}, \beta$)

Input: $\mathbf{E} = \{\boldsymbol{e}_i\}_1^M, \mathbf{V} = \{\boldsymbol{v}_i\}_1^K$, hyper-parameter β
calcuate cost matrix $\mathbf{C} = C(\mathbf{V}, \mathbf{E})$
$\boldsymbol{b} \leftarrow \frac{1}{m}\mathbf{1}_m$, $\mathbf{T}^{(1)} \leftarrow \mathbf{1}\mathbf{1}^T$
$\mathbf{G}_{ij} \leftarrow \exp(-\frac{\boldsymbol{C}_{ij}}{\beta})$
for $t = 1, 2, 3, \ldots, N$ **do**
  $\boldsymbol{Q} \leftarrow \boldsymbol{G} \odot \mathbf{T}^{(t)}$
  for $l = 1, 2, 3, \ldots, L$ **do** // Usually set $L = 1$
    $\boldsymbol{a} \leftarrow \frac{1}{K\boldsymbol{Q}\boldsymbol{b}}$, $\boldsymbol{b} \leftarrow \frac{1}{M\boldsymbol{Q}^T\boldsymbol{a}}$
  end for
  $\mathbf{T}^{(t+1)} \leftarrow \mathrm{diag}(\boldsymbol{a})\boldsymbol{Q}\mathrm{diag}(\boldsymbol{b})$
end for

2 Qualitative results from image-text matching

2.1 Sample results

We provide samples of text to image retrieval results from Flickr30K test set in Figure 1. For each sentence query, we present top-3 images ranked by similarity score calculated by our model. For each image query, we present top-5 sentences ranked by similarity score. From the samples we can see that our model can match images and sentences with large correlations. Although the text and image pairs are not the exact pairs, therefore ruled incorrect in calculating recall rate, they are still highly correlated and share the same theme.

Figure 1: Examples of image-text matching results. The first row shows text to image retrieval results. For each sentence query, we present top-3 images ranked by similarity score calculated by our model. The mark in the right-bottom corner in each image shows if it is a ground truth image. Image-to-text retrieval results are shown in the second row, where the top-5 sentences given a image query are provided. The mark at the end of the sentence represents if it is a ground truth sentence. The green check mark represents ground truth while red cross mark represents otherwise. Although these results are ruled incorrect in calculating recall rate, they are still highly correlated and share the same theme.

2.2 Interpretable alignment

We present more illustrations of the interpretability property in 2. The darker shades implies stronger OT matching or attention weights. Specifically, we visualize $\mathbf{T}$ in comparison with the attention matrix $(1-\mathbf{C})$. We see that OT transport mapping is more interpretable, as the alignment is sparse and self-normalized.

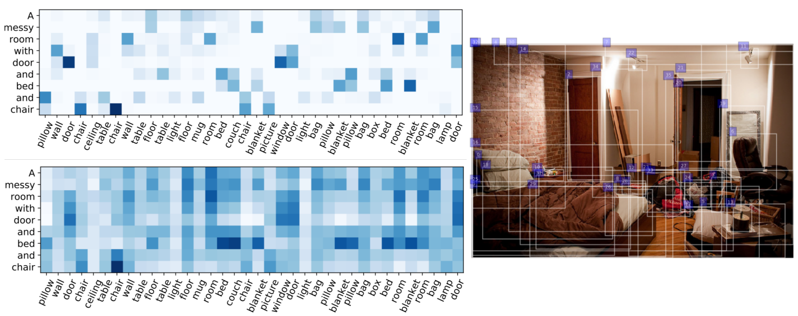

Figure 2: A comparison of OT transport mapping (top row) and attention matrix (bottom row). The horizontal axis represents image regions, and the vertical axis represents word tokens, the original image is on the right.

3 More Quantitative Results

3.1 VQA validation dataset results

We also tested our model on the VQA dataset[1], details can be found in Table 1.

Table 1: VQA validation dataset results

Method	VQA-score
BAN [2]	66.06
Ours	66.20

3.2 Ablation studies.

Adaptive number of regions. We use an adaptive number of regions as visual features. Specifically, we select all regions where any class detection probability exceeds a confidence threshold, set to 0.2. The alternative scheme is to fix the number of regions per image. We observed minor difference between these two schemes. Thus, we used features of top 36 objects ranked by *object* score to represent the image, $k = 36$. The results are shown in first two lines in Table 2, in which the hyper-parameters are selected by grid search on the validation set. In our experiment, using fixed number of objects outperforms the adaptive setting.

Effectiveness of the network architecture: We also consider using different network architectures to extract text sequence features, including Transformer [6] and basic GRU. The comparison results are shown in the last 3 lines of Table 2.

Add OT to a recent model. Considering adding more information usually improves performance, we add the OT alignment to the PFAN[7] model, which involves position information of bounding boxes into image features. By adding OT term, the model consistently improved on almost all metrics.

Table 2: Ablation study on Flickr30K. In the first section, we try to use adaptive features. In the second section, we compare the performance of different network architectures for text representations.

	Sentence Retrieval			Image Retrieval			
Method	R@1	R@5	R@10	R@1	R@5	R@10	Rsum
cos + OT, fixed	67.4	90.3	94.9	48.2	76.7	84.8	462.3
cos + OT, adaptive	64.8	88.3	94.5	45.9	74.5	83.5	451.5
cos+OT, bi-GRU	67.9	91.0	94.9	49.8	77.5	85.2	466.3
cos+OT, Transformer, 1 layer, 8 heads	57.8	86.8	93.1	46.4	74.4	82.9	441.4
cos+OT, Transformer, 2 layer, 8 heads	44.1	74.8	84.1	31.7	62.7	73.5	370.9
PFAN	66	89.6	94.3	49.6	77	84.2	460.7
PFAN + OT	67.1	89.2	94.3	50	78	85.7	464.3
SCAN	46.4	77.4	87.2	34.4	63.7	75.7	384.8
SCAN + OT	50.1	80.1	89.3	37.9	66.9	78.2	402.5

Table 3: Cross-domain matching results of ensembled model with Recall@K (R@K). Upper panel: Flickr30K, lower panel: MSCOCO.

	Sentence Retrieval			Image Retrieval			
Method	R@1	R@5	R@10	R@1	R@5	R@10	Rsum
SCAN (Faster R-CNN, ResNet) [3]	67.7	88.9	94.0	44.0	74.2	82.6	452.2
SCAN (Ensemble) [3]	67.4	90.3	95.8	48.6	77.7	85.2	465.0
BFAN (Ensemble)[5]	68.1	91.4	-	50.8	78.4	-	-
PFAN (Ensemble)[7]	70.0	91.8	95	50.4	78.7	86.1	472
VSRN (Ensemble, rerun) [4]	69.3	91.1	95.7	52.2	80	87.5	475.8
VSRN (Ensemble, paper) [4]	71.3	90.6	96.0	54.7	81.8	88.2	482.6
Ours (Faster R-CNN, ResNet):							
cos + OT (Ensemble)	70.3	**91.5**	**96.0**	52.2	79.4	87.1	476.4
SCAN (Faster R-CNN, ResNet)[3]	46.4	77.4	87.2	34.4	63.7	75.7	384.8
SCAN (Ensemble)[3]	50.4	82.2	90.0	38.6	69.3	80.4	410.9
VSRN (Ensemble, rerun)[4]	51.7	80.9	89.7	40.1	70.6	81.2	414.2
VSRN (Ensemble, paper)[4]	53.0	81.1	89.4	40.5	70.6	81.1	414.2
Ours (Faster R-CNN, ResNet):							
cos + OT (Ensemble)	52.1	**82.4**	**90.7**	39.1	68.2	79.2	411.7

3.3 Ensembled models

The performance of ensembled model are shown in Table 3. VSRN is the state-of-the-art model for image-text matching, which involves learning relationship between regions in the same image and using image captioning as assisting task during training, which is beyond the range of CDA models which we are discussing. For all CDA models using only matching loss, like SCAN, PFAN and BFAN, our model consistently outperforms CDA models.

4 More Visualization

4.1 Alignment visualization

We show the alignment visualization in Figure 3. This is a data from test set. The upper-left figure is the original image. And the attention mapping for each word is shown one by one. We view the matching transport matrix $\mathbf{T}$ as alignment. The regional brightness is determined by accumulated alignment strength. The region with highest alignment with respect to the word is rounded by blue boxes, with the corresponding word marked on top-left corner. The alignment in our model accurately aligns corresponding regions and words. For the words not directly representing certain area in image, like "a" in "a young person", "a" in "a bridge", "on" in "person on a motor bike", our model managed to align them with the region contains person, bridge, intersection of person and motor bike.

4.2 Visualization of transport matrix

We show a comparison of alignment between paired and unpaired data in Figure 4. The figure on upper left shows the alignment (matching transport matrix) of paired data. The figure on bottom left shows the alignment between the same sentence and a different image. The horizontal axis represents regions in image, marked by the region label. The vertical axis represents token in sequence, marked by the word. The figure on the right is the paired image, with location of regions marked by white bounding boxes.

Figure 3: Visualization of alignment. Showing attended image with respect to each word. The brightness represents the alignment strength. We view the matching transport matrix $\mathbf{T}$ as attention. The figure on upper-left corner is the original image. The bounding box represents the region with highest alignment score, with corresponding word marked.

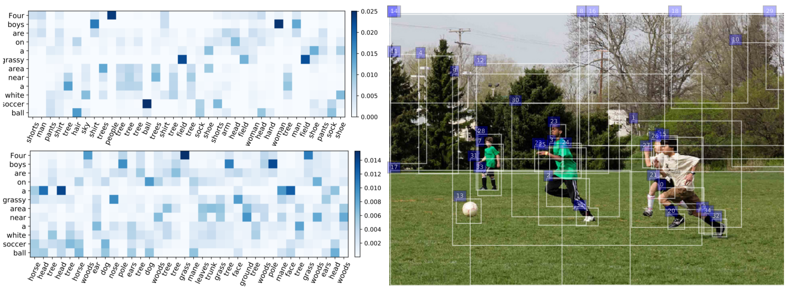

Figure 4: Visualization of alignment. The upper left figure shows the transport plan between paired data. The bottom left figure shows the transport plan between unpaired data, in which the sentence is the same, while image is different. The horizontal axis is marked by classification labels of the regions. The vertical axis is marked by words in sentence. It is clear that for paired data, the transport plan is more sparse. The figure in the right is the image in paired data with object location, marked by the rank of object score.

References

[1] Stanislaw Antol et al. Vqa: Visual question answering. In *ICCV*, pages 2425–2433, 2015.

[2] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Advances in Neural Information Processing Systems*, pages 1564–1574, 2018.

[3] Kuang-Huei Lee et al. Stacked cross attention for image-text matching. In *ECCV*, 2018.

[4] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4654–4662, 2019.

[5] Chunxiao Liu, Zhendong Mao, An-An Liu, Tianzhu Zhang, Bin Wang, and Yongdong Zhang. Focus your attention: A bidirectional focal attention network for image-text matching. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 3–11, 2019.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.

[7] Yaxiong Wang, Hao Yang, Xueming Qian, Lin Ma, Jing Lu, Biao Li, and Xin Fan. Position focused attention network for image-text matching. *arXiv preprint arXiv:1907.09748*, 2019.

[8] Yujia Xie, Xiangfeng Wang, Ruijia Wang, and Hongyuan Zha. A fast proximal point method for computing exact wasserstein distance. *arXiv preprint arXiv:1802.04307*, 2018.