Beyond the Deep Metric Learning: Enhance the Cross-Modal Matching with Adversarial Discriminative Domain Regularization

Li Ren, Kai Li, LiQiang Wang, Kien Hua

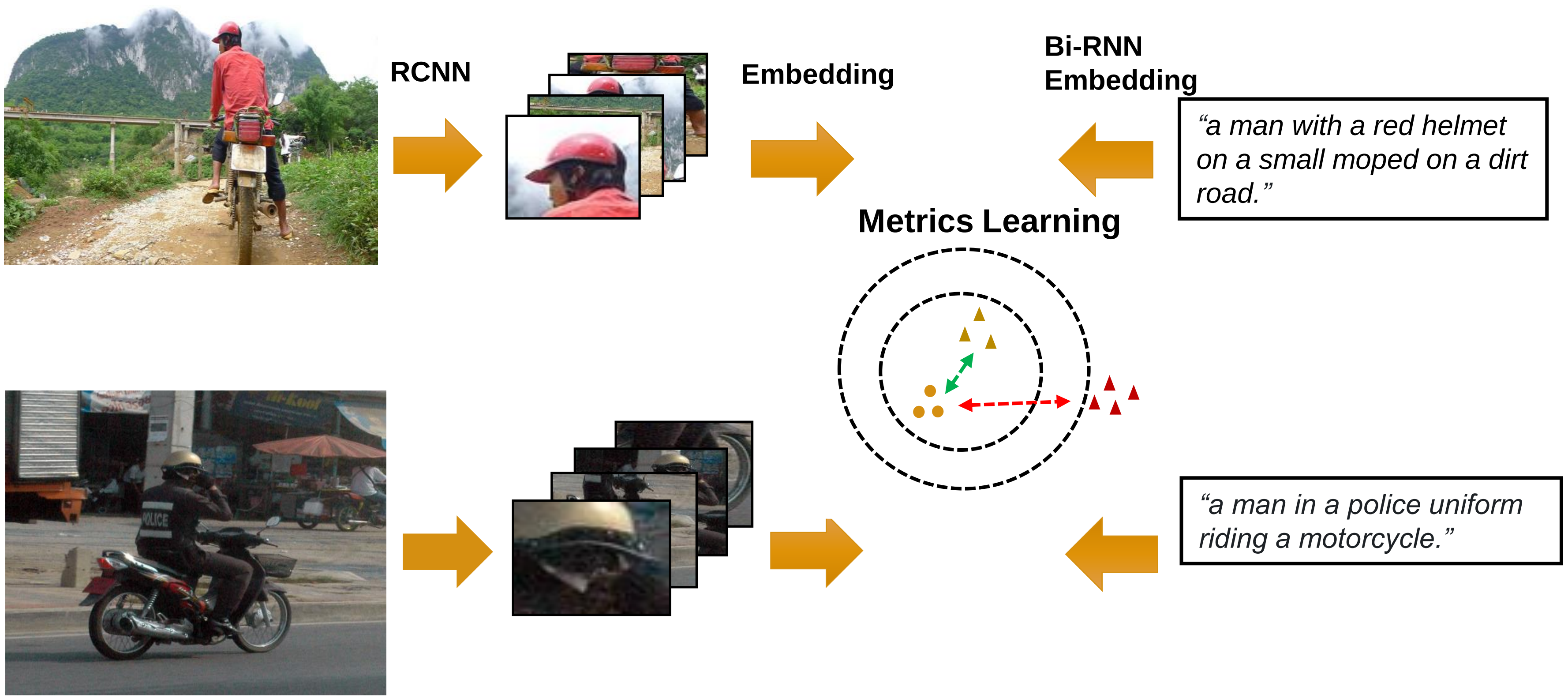
University of Central Florida, Orlando, FL

Cross-Modal Metric Learning and Matching

- **Input**: Image and sentence
- **Output**: Similarity scores of any pair of image and text
- **Application**: Image retrieval, Text retrieval

	"a man with a red helmet on a small moped on a dirt road."	*"a man in a police uniform riding a motorcycle."*
	0.85	0.25
	0.30	0.70

Existing Solution



$$\mathcal{L}_{rank}(I,S) = \max[0, \delta - \underset{\Phi,\theta}{Sim}(I,S) + \underset{\Phi,\theta}{Sim}(I,\hat{S})] + \max[0, \delta - \underset{\Phi,\theta}{Sim}(I,S) + \underset{\Phi,\theta}{Sim}(\hat{I},S)], \quad (1)$$

Existing Solution

Cross-Stack Attention Network (SCAN)

K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, “Stacked cross attention for image-text matching,” in ECCV, 2018.

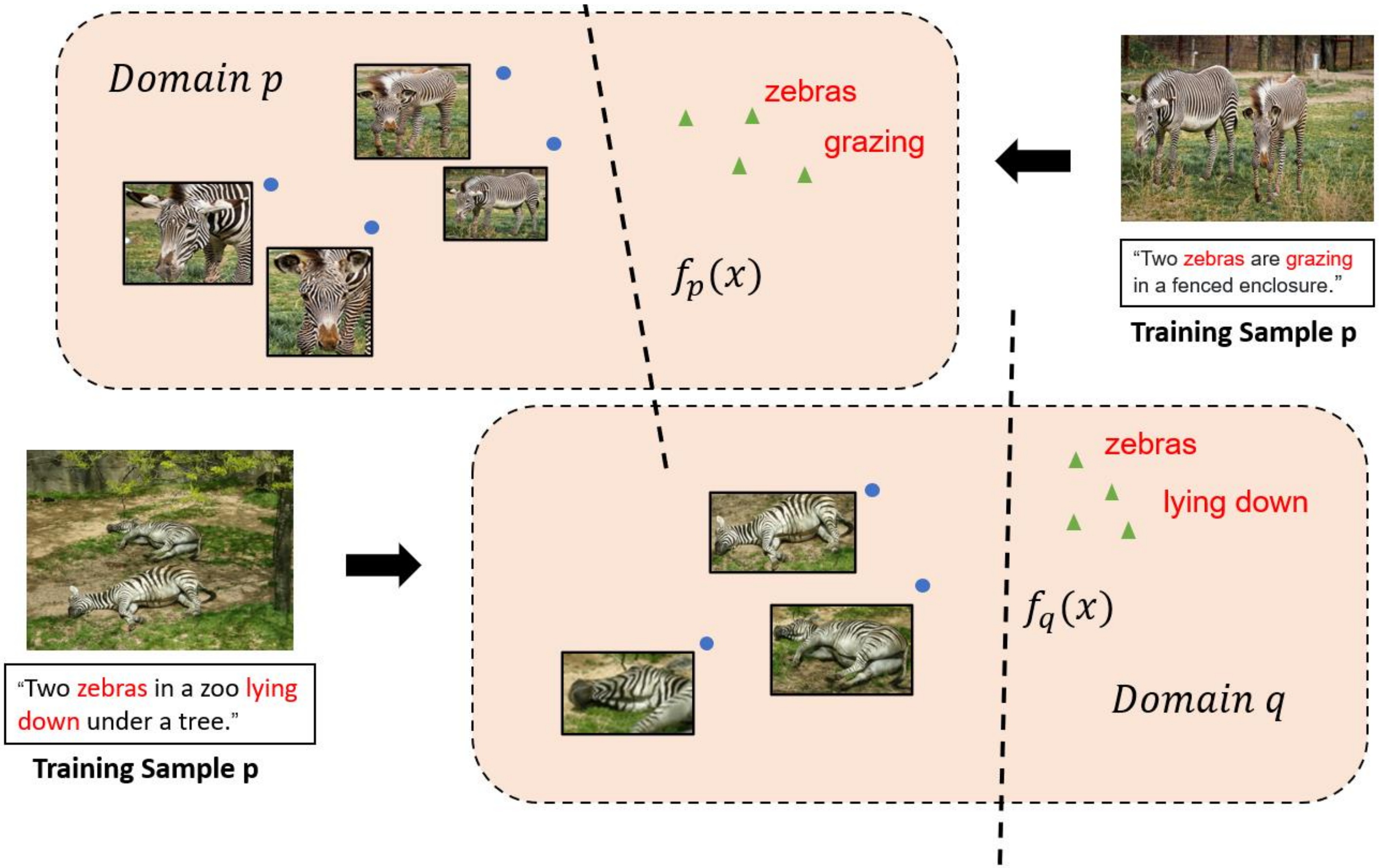
Bidirectional Focal Attention Network (BFAN)

C. Liu, Z. Mao, A.-A. Liu, T. Zhang, B. Wang, and Y. Zhang, “Focus your attention: A bidirectional focal attention network for image-text matching,” in ACM International Conference on Multimedia, 2019.

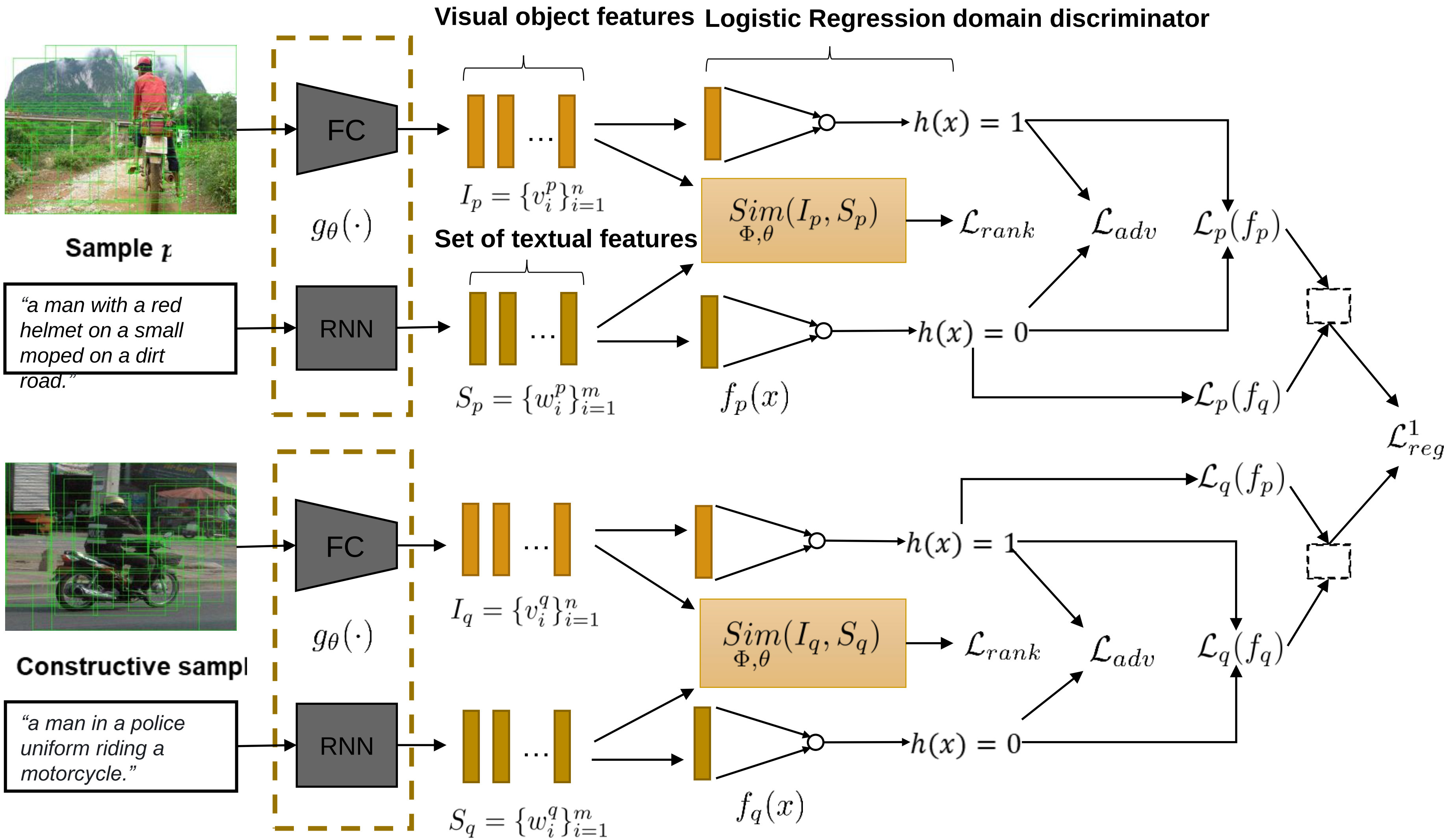
Visual Semantic Reasoning Network (VSRN)

K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, “Visual semantic reasoning for image-text matching,” in ICCV, 2019.

Our Proposed Method:



Our Framework:

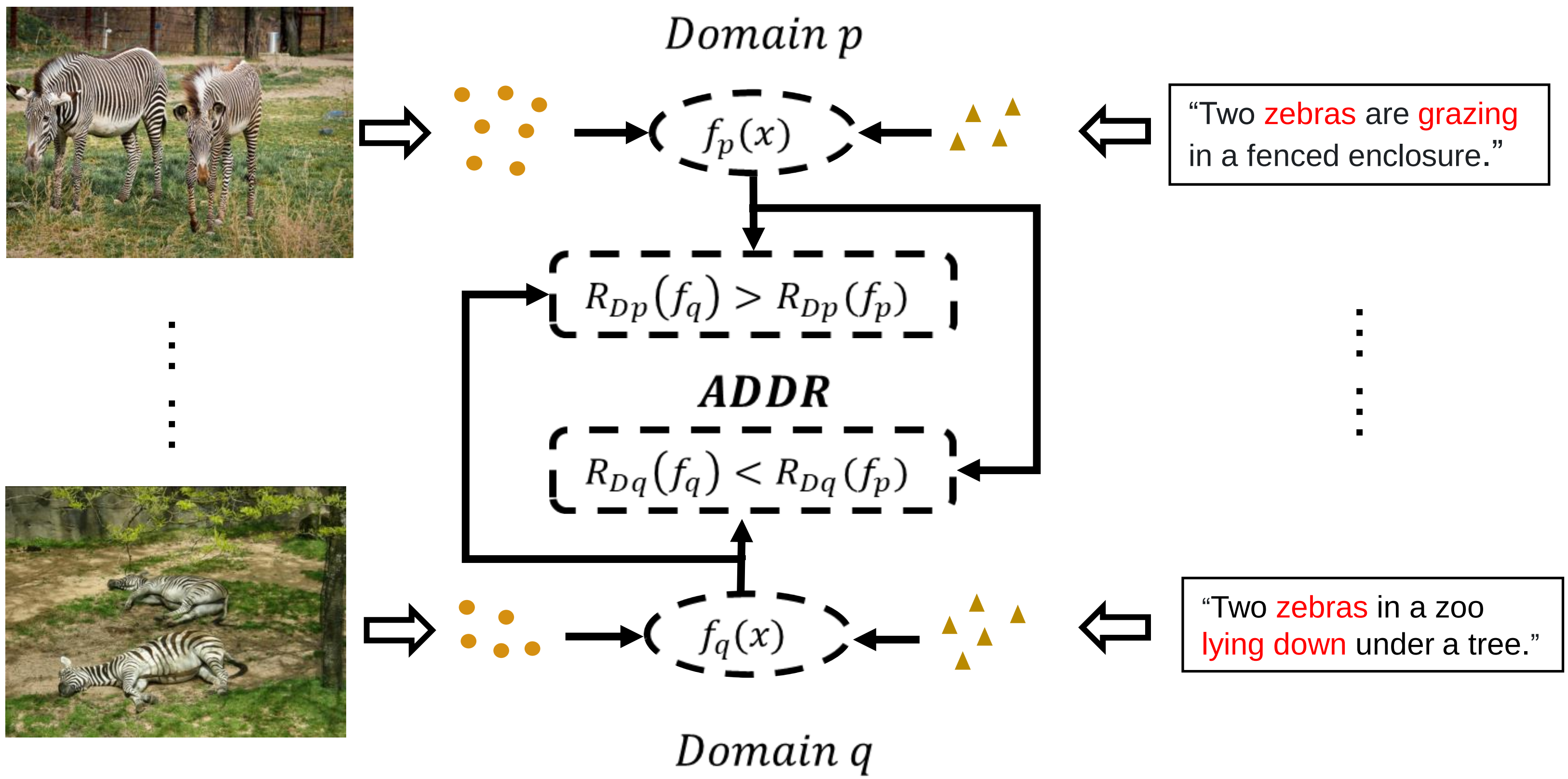


Adversarial Training

$$\min_{W_p, b_p} \mathcal{L}_{adv}(I_p, S_p) = \sum_{i=1}^{n} log(\sigma(W_p^T w_i^p + b_p)) + \sum_{i=1}^{m} log(1 - \sigma(W_p^T v_i^p + b_p)), \tag{2}$$

$$\min_{\Phi, \theta} \frac{1}{N} \sum_{p=1}^{N} [\mathcal{L}_{rank}(I_p, S_p) - \beta \mathcal{L}_{adv}(I_p, S_p)], \tag{3}$$

Discriminative Domain Regularization



Discriminative Domain Regularization

$$\mathcal{R}_{D_p}(f_p) \leq \mathcal{R}_{D_p}(f_q) + \alpha \tag{4}$$
$$\mathcal{R}_{D_p}(f_p) \leq \mathcal{R}_{D_p}(f_r) + \alpha \tag{5}$$
$$\mathcal{R}_{D_q}(f_q) \leq \mathcal{R}_{D_q}(f_p) + \alpha \tag{6}$$
$$\mathcal{R}_{D_r}(f_r) \leq \mathcal{R}_{D_r}(f_p) + \alpha \tag{7}$$

$$\begin{aligned} \mathcal{L}_{reg}^{1}(I_p, S_p, I_q, S_q) = &\max[0, \alpha + \mathcal{L}_p(f_p) - \mathcal{L}_p(f_q),] \\ &+ \max[0, \alpha + \mathcal{L}_q(f_q) - \mathcal{L}_q(f_p)] \end{aligned} \tag{8}$$

$$\begin{aligned} \mathcal{L}_{reg}^{2}(I_p, S_p, I_r, S_r) = &\max[0, \alpha + \mathcal{L}_q(f_q) - \mathcal{L}_q(f_p)] \\ &+ \max[0, \alpha + \mathcal{L}_p(f_p) - \mathcal{L}_p(f_r)], \end{aligned} \tag{9}$$

$$\mathcal{L}_{reg}(p, q, r) = \mathcal{L}_{reg}^{1}(I_p, S_p, I_q, S_q) + \mathcal{L}_{reg}^{2}(I_p, S_p, I_r, S_r)$$

The combined objectives

$$\mathcal{L}^{d}_{ADDR} = \frac{1}{N}\sum_{p=1}^{N}\{\mathcal{L}_{adv}(I_p, S_p) + \gamma\mathcal{L}_{reg}(p, q, r)\} \quad (10)$$

$$\mathcal{L}^{g}_{ADDR} = \frac{1}{N}\sum_{p=1}^{N}\{\mathcal{L}_{rank}(I_p, S_p) - \beta\mathcal{L}_{adv}(I_p, S_p)\} \quad (11)$$

Algorithm 1 Adversarial Discriminative Domain Regularization (ADDR)

1: **Input:** Training Set $\mathcal{Q} = (I_p, S_p)_{p=1}^{N}$ with raw image features and sentence terms. Hyperparameters $\delta, \alpha, \beta, \gamma$.
2: **Output:** Learned Parameters θ, Φ
3: **Initial:** $\theta, \Phi, \mathcal{W} = \{w_p, b_p\}_{p=1}^{N}$
4: **while** stop criteria is not satisfied **do**
5:     /* *Discriminator Training Phase* begin */
6:     **for** each *mini-batch* of size k **do**
7:         Select data $I = (I_p)_{p=1}^{k}, S = (S_p)_{p=1}^{k}$
8:         Select Parameters $W = \{W_p\}_{p=1}^{k}$ and $b = \{b_p\}_{p=1}^{k}$
9:         Embedding $\{v^p\}_{p=1}^{k} \leftarrow g_\theta(I)$, $\{w^p\}_{p=1}^{k} \leftarrow g_\theta(S)$
10:         Calculate Metric Scores $\mathcal{S} = \{Sim_\Phi(I_p, S_p)\}_{p=1}^{p=k}$
11:         Select *hard negative* samples (I_q, S_q) and (I_r, S_r)
12:         Calculate $\Delta W, \Delta b \leftarrow \frac{\partial \mathcal{L}_{adv}}{W,b} + \gamma(\frac{\partial \mathcal{L}^{1}_{reg}}{W,b} + \frac{\partial \mathcal{L}^{2}_{reg}}{W,b})$
13:         Update $W, b \leftarrow W, b - Adam(\Delta W, \Delta b)$
14:     /* *Generator Training Phase* begin */
15:     **for** each *mini-batch* of size k **do**
16:         Repeat L7 - L10
17:         Calculate $\Delta\Phi \leftarrow \frac{\partial \mathcal{L}_{rank}}{\Phi}$
18:         Calculate $\Delta\theta \leftarrow \frac{\partial \mathcal{L}_{rank}}{\theta} - \beta\frac{\partial \mathcal{L}_{adv}}{\theta}$
19:         Update $\Phi \leftarrow \Phi - Adam(\Delta\Phi)$
20:         Update $\theta \leftarrow \theta - Adam(\Delta\theta)$

Our Cross Modal Retrieval Result

MSCOCO

Method	Sentence Retrieval R@1	R@5	R@10	Image Retrieval R@1	R@5	R@10	Sum (ALL)
1k Test Set (5-fold)							
SCAN [13] (2018)	72.7	94.8	98.4	58.8	88.4	94.8	507.9
MTFN [25] (2019)	74.3	94.9	97.9	60.1	89.1	95.0	511.3
BFAN [15] (2019)	74.9	95.2	98.3	59.4	88.4	94.5	510.7
VSRN [14] (2019)	76.2	94.8	98.2	62.8	89.7	95.1	516.8
DPRNN [3] (2020)	75.3	95.8	98.6	62.5	89.7	95.1	517.0
ADAPT [27] (2020)	76.5	95.6	98.9	62.2	90.5	96.0	519.7
ADDR-SCAN (Ours)	76.1	95.5	98.4	61.2	88.9	94.8	514.9
ADDR-BFAN (Ours)	76.4	95.8	98.3	62.3	89.4	96.2	518.4
ADDR-VSRN (Ours)	**77.4**	**96.1**	**98.9**	**63.5**	**90.7**	**96.7**	**523.3**
5K Test Set							
SCAN [13] (2018)	50.4	82.2	90.0	38.6	69.3	80.4	410.9
MTFN [25] (2019)	48.3	77.6	87.3	35.9	66.1	76.1	391.3
BFAN [15] (2019)	52.9	82.8	90.6	38.3	67.8	79.3	411.7
VSRN [14] (2019)	53.0	81.1	89.4	40.5	70.6	81.1	415.7
ADDR-SCAN (Ours)	**57.3**	**86.0**	**92.7**	41.8	**72.0**	81.3	**431.1**
ADDR-BFAN (Ours)	54.3	84.0	91.5	40.1	69.2	80.6	419.7
ADDR-VSRN (Ours)	56.6	85.3	90.4	**42.5**	71.9	**82.0**	428.7

Flickr30k

Method	Sentence Retrieval R@1	R@5	R@10	Image Retrieval R@1	R@5	R@10	Sum (ALL)
SCAN [13] (2018)	67.4	90.3	95.8	48.6	77.7	85.2	465.0
MTFN [25] (2019)	65.3	88.3	93.3	52.0	80.1	86.1	465.1
BFAN [15] (2019)	68.1	91.4	95.9	50.8	78.4	85.8	470.4
VSRN [14] (2019)	71.3	90.6	96.0	54.7	81.8	88.2	482.6
RDAN [7](2019)	68.1	91.0	95.9	54.1	80.9	87.2	477.2
ADDR-SCAN (Ours)	72.1	**93.1**	96.1	53.5	80.4	87.4	482.6
ADDR-BFAN (Ours)	71.3	91.5	96.4	54.0	80.0	87.6	480.8
ADDR-VSRN (Ours)	**73.0**	92.5	**96.6**	**55.6**	**82.0**	**88.9**	**488.6**

Conclusion

- We propose a novel framework Adversarial Discriminative Domain Regularization (ADDR) that generally enhances the cross-modal metric learning networks. It is achieved by learning a group of discriminative domains regularized with a constructive learning term that explicitly aligned to each image-text pair.
- Our ADDR is compatible with existing metric learning networks. It is used as an add-on regularizer to their primary tasks to help match between a group of visual objects and the corresponding sentence.
- Our quantitative experiments show the effectiveness of our approach base on the recent popular metric learning frameworks: SCAN, VSRN and BFAN on the popular MS-COCO and Flickr30k datasets.

Thanks for watching

For more question you are pleased to email me at:

renli@knights.ucf.edu