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

<table>
<thead>
<tr>
<th>Image</th>
<th>Text</th>
<th>Score1</th>
<th>Score2</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.jpg" alt="Image" /></td>
<td>“a man with a red helmet on a small moped on a dirt road.”</td>
<td>0.85</td>
<td>0.25</td>
</tr>
<tr>
<td><img src="image2.jpg" alt="Image" /></td>
<td>“a man in a police uniform riding a motorcycle.”</td>
<td>0.30</td>
<td>0.70</td>
</tr>
</tbody>
</table>

University of Central Florida
Existing Solution

RCNN

Embedding

Bi-RNN Embedding

“a man with a red helmet on a small moped on a dirt road.”

“a man in a police uniform riding a motorcycle.”

\[
\mathcal{L}_{rank}(I, S) = \max[0, \delta - Sim_{\Phi,\theta}(I, S) + Sim_{\Phi,\theta}(I, \hat{S})] \\
+ \max[0, \delta - Sim_{\Phi,\theta}(I, S) + Sim_{\Phi,\theta}(\hat{I}, S)],
\]

University of Central Florida
Existing Solution

Cross-Stack Attention Network (SCAN)


Bidirectional Focal Attention Network (BFAN)


Visual Semantic Reasoning Network (VSRN)

Our Proposed Method:

Domain p

*Two zebras are grazing in a fenced enclosure.*

$f_p(x)$

Domain q

*Two zebras in a zoo lying down under a tree.*

$f_q(x)$

Training Sample p

Training Sample q
Our Framework:

Visual object features

\[ I_p = \{v_p^i\}_{i=1}^n \]

Logistic Regression domain discriminator

\[ h(x) = S_{\Phi,\theta}(I_p, S_p) \]

\[ L_{\text{rank}}, L_{\text{adv}}, L_p(f_p) \]

Set of textual features

\[ S_p = \{w_p^i\}_{i=1}^m \]

\[ f_p(x) \]

Constructive sample

\[ I_q = \{v_q^i\}_{i=1}^n \]

\[ S_{\Phi,\theta}(I_q, S_q) \]

\[ h(x) = 1 \]

\[ L_{\text{rank}}, L_{\text{adv}}, L_q(f_q) \]

Sample \( t \)

\[ g_\theta(\cdot) \]

\[ a \text{ man with a red helmet on a small moped on a dirt road.} \]

\[ f_q(x) \]

\[ h(x) = 0 \]

\[ L_{\text{reg}} \]

University of Central Florida
Adversarial Training

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

(2)

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

(3)
Discriminative Domain Regularization

Domain $p$

$\iff$

$f_p(x)$

$R_{Dp}(f_q) > R_{Dp}(f_p)$

$ADDR$

$R_{Dq}(f_q) < R_{Dq}(f_p)$

$\iff$

$f_q(x)$

Domain $q$

"Two zebras are grazing in a fenced enclosure."

"Two zebras in a zoo lying down under a tree."

University of Central Florida
Discriminative Domain Regularization

\[ R_{D_p}(f_p) \leq R_{D_p}(f_q) + \alpha \]  \hspace{1cm} (4)

\[ R_{D_p}(f_p) \leq R_{D_p}(f_r) + \alpha \]  \hspace{1cm} (5)

\[ R_{D_q}(f_q) \leq R_{D_q}(f_p) + \alpha \]  \hspace{1cm} (6)

\[ R_{D_r}(f_r) \leq R_{D_r}(f_p) + \alpha \]  \hspace{1cm} (7)

\[ L_{reg}^1(I_p, S_p, I_q, S_q) = \max[0, \alpha + L_p(f_p) - L_p(f_q),] \]
\[ + \max[0, \alpha + L_q(f_q) - L_q(f_p)] \]  \hspace{1cm} (8)

\[ L_{reg}^2(I_p, S_p, I_r, S_r) = \max[0, \alpha + L_q(f_q) - L_q(f_p)] \]
\[ + \max[0, \alpha + L_p(f_p) - L_p(f_r)], \]  \hspace{1cm} (9)

\[ L_{reg}(p, q, r) = L_{reg}^1(I_p, S_p, I_q, S_q) + L_{reg}^2(I_p, S_p, I_r, S_r) \]
The combined objectives

\[
\mathcal{L}^\vartheta_{ADD\text{R}} = \frac{1}{N} \sum_{p=1}^{N} \left\{ \mathcal{L}_{\text{adv}}(I_p, S_p) + \gamma \mathcal{L}_{\text{reg}}(p, q, r) \right\} \quad (10)
\]

\[
\mathcal{L}^g_{ADD\text{R}} = \frac{1}{N} \sum_{p=1}^{N} \left\{ \mathcal{L}_{\text{rank}}(I_p, S_p) - \beta \mathcal{L}_{\text{adv}}(I_p, S_p) \right\} \quad (11)
\]

**Algorithm 1** Adversarial Discriminative Domain Regularization (ADDR)

1: **Input:** Training Set \( Q = (I_p, S_p)_{p=1}^{N} \) with raw image features and sentence terms. Hyperparameters \( \delta, \alpha, \beta, \gamma \).
2: **Output:** Learned Parameters \( \theta, \Phi \)
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: \quad **for** each mini-batch of size \( k \) **do**
7: \quad \quad Select data \( I = (I_p)_{p=1}^{k}, S = (S_p)_{p=1}^{k} \)
8: \quad \quad Select Parameters \( W = \{W_p\}_{p=1}^{k} \) and \( b = \{b_p\}_{p=1}^{k} \)
9: \quad \quad Embedding \( \{v^p\}_{p=1}^{k} \leftarrow g_\theta(I), \{w^p\}_{p=1}^{k} \leftarrow g_\theta(S) \)
10: \quad \quad Calculate Metric Scores \( S = \{Sim_\Phi(I_p, S_p)\}_{p=1}^{k} \)
11: \quad \quad Select hard negative samples \( (I_q, S_q) \) and \( (I_r, S_r) \)
12: \quad \quad Calculate \( \Delta W, \Delta b \leftarrow \frac{\partial \mathcal{L}_{\text{adv}}}{\partial W, b} + \gamma \left( \frac{\partial \mathcal{L}_{\text{eq}}^1}{\partial W, b} + \frac{\partial \mathcal{L}_{\text{eq}}^2}{\partial W, b} \right) \)
13: \quad \quad Update \( W, b \leftarrow W, b - \text{Adam}(\Delta W, \Delta b) \)
14: \/* **Generator Training Phase** begin */
15: \quad **for** each mini-batch of size \( k \) **do**
16: \quad \quad Repeat L7 - L10
17: \quad \quad Calculate \( \Delta \Phi \leftarrow \frac{\partial \mathcal{L}_{\text{rank}}}{\Phi} \)
18: \quad \quad Calculate \( \Delta \theta \leftarrow \frac{\partial \mathcal{L}_{\text{rank}}}{\theta} - \beta \frac{\partial \mathcal{L}_{\text{adv}}}{\theta} \)
19: \quad \quad Update \( \Phi \leftarrow \Phi - \text{Adam}(\Delta \Phi) \)
20: \quad \quad Update \( \theta \leftarrow \theta - \text{Adam}(\Delta \theta) \)
Our Cross Modal Retrieval Result

<table>
<thead>
<tr>
<th>Method</th>
<th>Sentence Retrieval</th>
<th>Image Retrieval</th>
<th>Sum (ALL)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@1</td>
<td>R@5</td>
<td>R@10</td>
</tr>
<tr>
<td></td>
<td>R@1</td>
<td>R@5</td>
<td>R@10</td>
</tr>
<tr>
<td>1k Test Set (5-fold)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCAN [13] (2018)</td>
<td>72.7</td>
<td>94.8</td>
<td>98.4</td>
</tr>
<tr>
<td>MTFN [25] (2019)</td>
<td>74.3</td>
<td>94.9</td>
<td>97.9</td>
</tr>
<tr>
<td>BFAN [15] (2019)</td>
<td>74.9</td>
<td>95.2</td>
<td>98.3</td>
</tr>
<tr>
<td>VSRN [14] (2019)</td>
<td>76.2</td>
<td>94.8</td>
<td>98.2</td>
</tr>
<tr>
<td>DPRNN [3] (2020)</td>
<td>75.3</td>
<td>95.8</td>
<td>98.6</td>
</tr>
<tr>
<td>ADAPT [27] (2020)</td>
<td>76.5</td>
<td>95.6</td>
<td>98.9</td>
</tr>
<tr>
<td>ADDR-SCAN (Ours)</td>
<td>76.1</td>
<td>95.5</td>
<td>98.4</td>
</tr>
<tr>
<td>ADDR-BFAN (Ours)</td>
<td>76.4</td>
<td>95.8</td>
<td>98.3</td>
</tr>
<tr>
<td>ADDR-VSRN (Ours)</td>
<td><strong>77.4</strong></td>
<td><strong>96.1</strong></td>
<td><strong>98.9</strong></td>
</tr>
<tr>
<td>5K Test Set</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCAN [13] (2018)</td>
<td>50.4</td>
<td>82.2</td>
<td>90.0</td>
</tr>
<tr>
<td>MTFN [25] (2019)</td>
<td>48.3</td>
<td>77.6</td>
<td>87.3</td>
</tr>
<tr>
<td>BFAN [15] (2019)</td>
<td>52.9</td>
<td>82.8</td>
<td>90.6</td>
</tr>
<tr>
<td>VSRN [14] (2019)</td>
<td>53.0</td>
<td>81.1</td>
<td>89.4</td>
</tr>
<tr>
<td>ADDR-SCAN (Ours)</td>
<td><strong>57.3</strong></td>
<td><strong>86.0</strong></td>
<td><strong>92.7</strong></td>
</tr>
<tr>
<td>ADDR-BFAN (Ours)</td>
<td>54.3</td>
<td>84.0</td>
<td>91.5</td>
</tr>
<tr>
<td>ADDR-VSRN (Ours)</td>
<td>56.6</td>
<td>85.3</td>
<td>90.4</td>
</tr>
</tbody>
</table>

MSCOCO

<table>
<thead>
<tr>
<th>Method</th>
<th>Sentence Retrieval</th>
<th>Image Retrieval</th>
<th>Sum (ALL)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@1</td>
<td>R@5</td>
<td>R@10</td>
</tr>
<tr>
<td></td>
<td>R@1</td>
<td>R@5</td>
<td>R@10</td>
</tr>
<tr>
<td>RDAN [7] (2019)</td>
<td>68.1</td>
<td>91.0</td>
<td>95.9</td>
</tr>
<tr>
<td>ADDR-SCAN (Ours)</td>
<td>72.1</td>
<td><strong>93.1</strong></td>
<td>96.1</td>
</tr>
<tr>
<td>ADDR-BFAN (Ours)</td>
<td>71.3</td>
<td>91.5</td>
<td>96.4</td>
</tr>
<tr>
<td>ADDR-VSRN (Ours)</td>
<td><strong>73.0</strong></td>
<td><strong>92.5</strong></td>
<td><strong>96.6</strong></td>
</tr>
</tbody>
</table>
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 based 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