Supplementary Material for:
Adversarial Representation Learning for Text-to-Image Matching

Nikolaos Sarafianos Xiang Xu Ioannis A. Kakadiaris
Computational Biomedicine Lab
University of Houston
{nsarafianos, xxu21}@uh.edu, ikakadia@central.uh.edu

Discussion on Novelty

What is novel in TIMAM? What has been done and what is new in the proposed approach? Since all these are valid questions that the interested reader might have, we aim to provide the reader with an in-depth understanding of the contributions of our work and how these result in advantages over previous work.

Our method is novel in that we take a different approach in learning the embeddings from both modalities. Specifically, we leverage an adversarial discriminator, which helps TIMAM to learn modality-invariant discriminative representations. It is a very effective addition that can easily be applied to other cross-modal matching applications (*e.g.*, audio-visual retrieval). Our ablation studies, in two very challenging datasets, showed improvements of ~3% on the CUHK-PEDES dataset and ~1.8% on the Flickr30K dataset, when the adversarial discriminator is used. These results demonstrate that adversarial learning is well-suited for cross-modal matching.

The second contribution of this work is that we demonstrated that a pre-trained language model can successfully be applied (with some fine-tuning) to computer vision applications such as text-to-image matching. Our results demonstrated that we can improve our feature representations when better word embeddings are learned in this manner. In summary, the advantages of TIMAM over prior work can be described as follows:

- TIMAM improves upon CMPM [26] (previous best performing method on the CUHK-PEDES, Flowers, and Birds datasets) by employing a domain discriminator, which results into more discriminative representations. Unlike CMPM, we do not perform cross-modal projections in the classification loss since we observed that their contribution is insignificant. Instead we employ identification losses for the visual and textual features and present their impact in the first ablation study.
- TIMAM improves upon the previous best performing method on the Flickr30K dataset by learning better textual embeddings using the fine-tuning capabilities of BERT (as well as employing the adversarial representation learning framework).
- TIMAM is very easy to reproduce, which is not the case with prior work that requires complex attention mechanisms at both modalities [20] or text reconstruction objectives [2]. To obtain our results one can simply fetch an image backbone and a deep language model and then follow the steps described in Alg. 1.

Algorithm 1: Training Procedure of TIMAM

```
Input : Batch ($B$) of image-text pairs $(V_i, T_i)$ with
        their label $Y_i$, pre-trained ResNet-101
        weights, pre-trained BERT weights
1 $\phi(V_i) \leftarrow$ extract visual embedding by feeding $V_i$ to
    image backbone
2 $\tau(T_i) \leftarrow$ extract textual embedding by feeding $T_i$ to
    text backbone and then to the LSTM
3 $L_I^V \leftarrow$ compute identification loss for the images
    using $(V_i, Y_i)$. Similarly compute $L_I^T$ for the text.
4 $p_{i,j} \leftarrow$ compute the probability of matching $\phi(V_i)$
    to $\bar{\tau}(T_j)$
5 $q_{i,j} \leftarrow$ compute the true matching probability using
    $Y_i$ as well as the rest of true labels in $B$)
6 $L_M^V \leftarrow$ compute cross-modal projection matching
    loss as the KL divergence from $q_i$ to $p_i$
7 Repeat steps 4-6 for the text modality to compute
    $L_M^T$ by normalizing $\phi(V_i)$ instead of $\tau(T_i)$
8 $L_D \leftarrow$ compute adversarial loss by passing $\phi(V_i)$
    and $\tau(T_i)$ through the discriminator
9 Update network parameters using:
    $L = L_D + L_I^V + L_I^T + L_M^V + L_M^T$
Output: Network weights
```

Implementation Details

Datasets: The first dataset used was the CUHK-PEDES [14] which consists of 40,206 images of in-

Table 1: Notation used throughout our paper

Notation Sign	Description
V	The visual (*i.e.*, image) modality
T	The textual modality
V_i	A sample from the visual modality
T_i	A sample from the textual modality
Y_i	The ID/Category label of the pair
$\phi(\cdot)$	The feature extractor at the image modality (*i.e.*, ResNet-101)
$\tau(\cdot)$	The feature extractor at the textual modality (*i.e.*, BERT, the LSTM and the FC- layer)
$\bar{\tau}(\cdot)$	Normalized textual features
G^V, G^T	Generators from the visual and textual modality (*i.e.*, $\phi()$ and $\tau()$)
D	Cross-modal discriminator
(W_i, b_i)	Weights and bias of the last FC-layer that produces the embedding
B	Batch size
$p_{i,j}$	Probability of matching each visual embedding to each normalized textual embedding in the batch
$q_{i,j}$	True matching probability for each pair in the batch
$s_{i,j}$	Cosine similarity between i^{th} probe and j^{th} gallery sample
L_I^V	Norm-softmax cross entropy loss used for identification for the visual embedding
L_I^T	Norm-softmax cross entropy loss used for identification for the textual embedding
L_I	Summation of the two identification losses from both modalities
L_M^V	KL-divergence loss used for cross-modal $(V- > T)$ projection matching
L_M^T	KL-divergence loss used for cross-modal $(T- > V)$ projection matching
L_M	Summation of the two cross-modal projection losses from both modalities
L_D	Adversarial loss of the discriminator
L	Loss used to train our network: summation of individual sub-losses

dividuals of 13,003 identities, and each image is described by two textual descriptions. The dataset is split into 11,003/1,000/1,000 identities for the training/validation/testing sets with 34,054, 3,078 and 3,074 images respectively, in each subset. The second dataset was the Flickr30K [21] which contains 31,783 images with five text descriptions each. The data split introduced in the work of Karpathy and Fei-Fei [10] is adopted which results in 29,783/1,000/1,000 images for training validation and testing respectively. The third dataset was the Caltech-UCSD Birds (CUB) [22], which comprises 11,788 bird images from 200 different categories. Each image is labeled with 10 descriptions and the dataset is split into 100 training, 50 validation and 50 test categories. Finally, the Oxford102 Flowers (Flowers) [22] dataset was used, which consists of 8,189 flower images of 102 different categories. Each image is accompanied by 10 descriptions and the dataset is split into 62 training, 20 validation, and 20 test categories.

Data Pre-processing: For the CUHK-PEDES dataset all images were resized to 128 × 256 since pedestrians walking are usually rectangular. For the rest of the datasets, all images were resized to 224 × 224. For the textual input, basic word tokenization was performed by mapping each word to the vocabulary accompanying the base BERT model pre-trained on the uncased book corpus and English Wikipedia datasets.[1] For the CUHK-PEDES dataset the maximum length of the sentences was set to 50 words (following the pre-processing steps of Li *et al.* [14]) whereas for the rest of the datasets it was set to 30 words (following the pre-processing steps of Zhang and Lu [26] for Flickr30K and Reed *et al.* [22] for the CUB and Flowers datasets). Thus, sentences shorter than the maximum length were zero-padded, whereas those longer than the threshold were trimmed.

Data Augmentation: During data augmentation images were upscaled to $\times 1.25$ the original size in both dimensions and random crops of the original dimensions were extracted and fed to the model. In addition, data shuffling, random horizontal flips with 50% probability and color jittering were employed.

Architecture Details: We used the pre-trained models of ResNet-101 and BERT available online for the backbone architectures of the two modalities while the rest of the layers were initialized with Xavier initialization.

- **Image domain**: Our backbone architecture on the vi-

[1] The pre-trained model we used is available at the Gluon-NLP website: https://gluon-nlp.mxnet.io/model_zoo/bert/index.html

sual domain is a ResNet-101 that extracts feature representations of dimensionality $7 \times 7 \times 2,048$ (for an input image with dimensions $224 \times 224 \times 3$). These representations are then fed to a fully-connected layer after performing global-average pooling to extract the image embedding of size equal to 512.

- **Text domain**: For the textual domain, each tokenized input sentence of length is fed to the deep language model which extracts a 768-D vector for each word. The sequence of word embeddings is then fed to a bidirectional LSTM with 512 hidden dimensions and its output is then projected to a fully-connected layer which outputs the text embedding of size equal to 512.
- **Discriminator**: We opted for a simple discriminator comprising two fully-connected layers [FC(256)-BN-LReLU(0.2)-FC(1)] that reduce the embedding size to a scalar value which is used to predict the input domain.

Training Details: We present all the notation used throughout our work in Table 1 for easier reference. We used MXNet/Gluon as our deep learning framework and a single NVIDIA GeForce GTX 1080 Ti GPU. We used stochastic gradient descent (SGD) with momentum equal to 0.9 to train the image and discriminator networks and the Adam optimizer [11] for the textual networks. The learning rate was set to 2×10^{-4} and was divided by ten when the loss plateaued at the validation set until 2×10^{-6}. The batch-size was set to 64 and the weight decay to 4×10^{-4}. The deep language model was initially frozen and the rest of the parameters were updated until convergence. After this step, we unfroze its weights and the whole network was fine-tuned with a learning rate equal to 2×10^{-6} for 30 epochs. Successfully, training the discriminator required maintaining an adequate balance between the two feature generators and the discriminator. To accomplish that, we relied on several of the tricks presented by Chintala *et al.* [4] on how to train a GAN: (i) different mini-batches were constructed for the features of each domain, (ii) labels were smoothed by replacing each positive label (visual domain) with a random number in [0.8, 1.2], and each label equal to zero (textual domain) with a random number in [0, 0.3], and (iii) labels were flipped with 20% probability to introduce some noise.

Extended Quantitative Results

Due to space constraints in the main paper, we provided quantitative results that contained the 8 best-performing methods in each dataset. In Tables 2 and 3, we present complete results against all approaches test in the CUHK-PEDES and Flickr30K datasets. TIMAM surpasses all methods in text-to-image matching but demonstrates inferior performance compared to GXN [6] in image-to-text matching.

Table 2: Text-to-image results on the CUHK-PEDES dataset. Results are ordered based on the rank-1 accuracy.

Method	Rank-1	Rank-5	Rank-10
iBOWIMG [28]	8.00	-	30.56
Word CNN-RNN [22]	10.48	-	36.66
Neural Talk [23]	13.66	-	41.72
GMM+HGLMM [12]	15.03	-	42.47
deeper LSTM Q+norm I [1]	17.19	-	57.82
GNA-RNN [14]	19.05	-	53.64
IATV [13]	25.94	-	60.48
PWM-ATH [3]	27.14	49.45	61.02
GLA [2]	43.58	66.93	76.26
Dual Path [27]	44.40	66.26	75.07
CAN [9]	45.52	67.12	76.98
CMPM + CMPC [26]	49.37	-	79.27
TIMAM	**54.51**	**77.56**	**84.78**

Extended Qualitative Results

In Figure 1 we present additional qualitative text-to-image matching results on the CUHK-PEDES and Flickr30K datasets. We observe that TIMAM is capable of learning visual attributes related to soft-biometrics (*e.g.*, sex of the individual), clothing (gray t-shirts on second row to the left or teal t-shirts on the third row) as well as objects such as hats (first row to the right) and backpacks (third and fourth rows to the left). To our surprise, we can effectively learn to match descriptions and images of scenes/actions (biking in the woods, or people on a subway) while having only a handful of such examples in the whole dataset.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *Proc. Conference on Computer Vision and Pattern Recognition*, Boston, MA, June 8-10 2015. 3

[2] Dapeng Chen, Hongsheng Li, Xihui Liu, Yantao Shen, Jing Shao, Zejian Yuan, and Xiaogang Wang. Improving deep visual representation for person re-identification by global and local image-language association. In *Proc. European Conference on Computer Vision*, Munich, Germany, Sept. 8-14 2018. 1, 3

[3] Tianlang Chen, Chenliang Xu, and Jiebo Luo. Improving text-based person search by spatial matching and adaptive threshold. In *Proc. Winter Conference on Applications of Computer Vision*, Lake Tahoe, NV, Mar. 12-15 2018. 3

[4] Soumith Chintala, Emily Denton, Martin Arjovsky, and Michael Mathieu. How to train a GAN? Tips and tricks to make GANs work. github.com/soumith/ganhacks, 2016. 3

[5] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: Improved visual-semantic embeddings. In

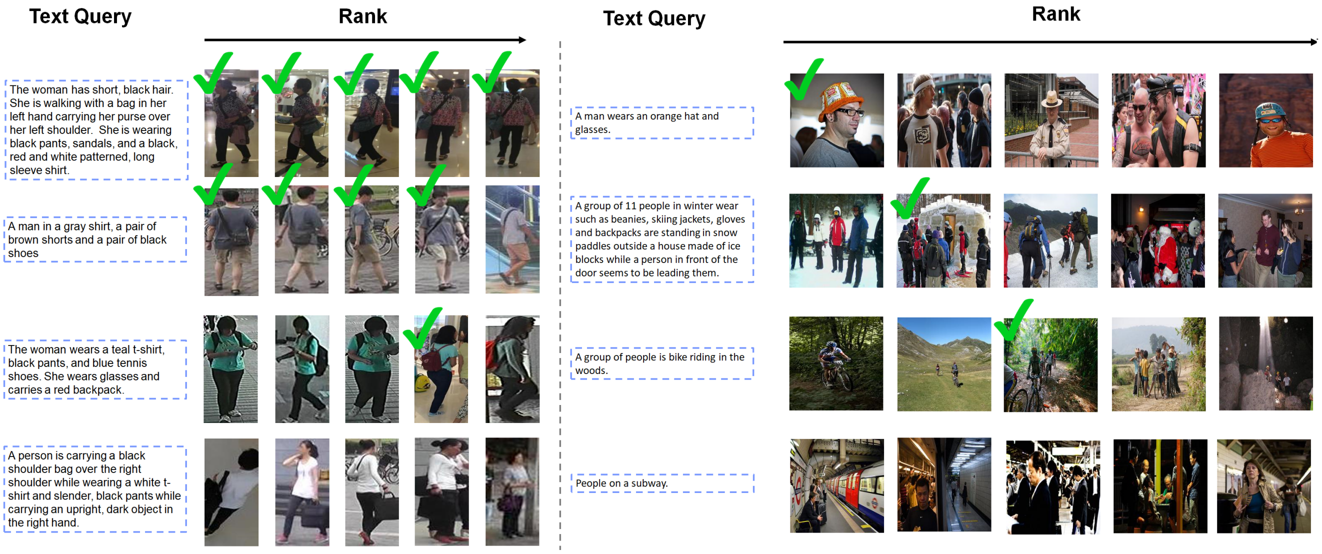
Text Query
Rank
The woman has short, black hair. She is walking with a bag in her left hand carrying her purse over her left shoulder. She is wearing black pants, sandals, and a black, red and white patterned, long sleeve shirt.
A man in a gray shirt, a pair of brown shorts and a pair of black shoes
The woman wears a teal t-shirt, black pants, and blue tennis shoes. She wears glasses and carries a red backpack.
A person is carrying a black shoulder bag over the right shoulder while wearing a white t-shirt and slender, black pants while carrying an upright, dark object in the right hand.
Text Query
Rank
A man wears an orange hat and glasses.
A group of 11 people in winter wear such as beanies, skiing jackets, gloves and backpacks are standing in snow paddles outside a house made of ice blocks while a person in front of the door seems to be leading them.
A group of people is bike riding in the woods.
People on a subway.

Figure 1: Additional qualitative text-to-image retrieval results on the CUHK-PEDES (left) and Flickr30K (right) datasets.

Table 3: Matching results on the Flickr30K dataset. The results are ordered based on their text-to-image rank-1 accuracy.

Method	Image Backbone	Image-to-Text			Text-to-Image		
		Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10
DVSA [10]	RCNN	22.2	48.2	61.4	15.2	37.7	50.5
m-RNN-VGG [19]	VGG-19	35.4	63.8	73.7	22.8	50.7	63.1
HGLMM FV [21]	VGG-19	36.5	62.2	73.3	24.7	53.4	66.8
VQA-A [15]	VGG-19	33.9	62.5	74.5	24.9	52.6	64.8
GMM+HGLMM [12]	VGG-19	35.0	62.0	73.8	25.0	52.7	66.0
m-CNN [18]	VGG-19	33.6	64.1	74.9	26.2	56.3	69.6
DCCA [25]	AlexNet	27.9	56.9	68.2	26.8	52.9	66.9
DSPE [24]	VGG-19	40.3	68.9	79.9	29.7	60.1	72.1
sm-LSTM [7]	VGG-19	42.5	71.9	81.5	30.2	60.4	72.3
DAN [20]	VGG-19	41.4	73.5	82.5	31.8	61.7	72.5
RRF-Net [17]	ResNet-152	47.6	77.4	87.1	35.4	68.3	79.9
CMPM +CMPC [26]	ResNet-152	49.6	76.8	86.1	37.3	65.7	75.5
DAN [20]	ResNet-152	55.0	81.8	89.0	39.4	69.2	79.1
NAR [16]	ResNet-152	55.1	80.3	89.6	39.4	68.8	79.9
VSE++ [5]	ResNet-152	52.9	80.5	87.2	39.6	70.1	79.5
SCO [8]	ResNet-152	55.5	82.0	89.3	41.1	70.5	80.1
GXN [6]	ResNet-152	56.8	-	89.6	41.5	-	80.1
TIMAM	ResNet-152	53.1	78.8	87.6	**42.6**	**71.6**	**81.9**

Proc. British Machine Vision Conference, Newcastle, UK, Sep. 3-6 2018. 4

[6] Jiuxiang Gu, Jianfei Cai, Shafiq Joty, Li Niu, and Gang Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *Proc. Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, June 18-22 2018. 3, 4

[7] Yan Huang, Wei Wang, and Liang Wang. Instance-aware image and sentence matching with selective multimodal LSTM. In *Proc. Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, July 21-26 2017. 4

[8] Yan Huang, Qi Wu, Chunfeng Song, and Liang Wang. Learning semantic concepts and order for image and sentence matching. In *Proc. Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, June 18-22 2018. 4

[9] Ya Jing, Chenyang Si, Junbo Wang, Wei Wang, Liang Wang, and Tieniu Tan. Cascade attention network for person search: Both image and text-image similarity selection. *arXiv preprint arXiv:1809.08440*, 2018. 3

[10] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proc. Confer-*

ence on Computer Vision and Pattern Recognition, Boston, MA, June 8-10 2015. 2, 4

[11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[12] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *Proc. Conference on Computer Vision and Pattern Recognition*, Boston, MA, June 8-10 2015. 3, 4

[13] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang. Identity-aware textual-visual matching with latent co-attention. In *Proc. International Conference on Computer Vision*, Venice, Italy, Oct. 22-29 2017. 3

[14] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *Proc. Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, July 21-26 2017. 1, 2, 3

[15] Xiao Lin and Devi Parikh. Leveraging visual question answering for image-caption ranking. In *Proc. European Conference on Computer Vision*, Amsterdam, The Netherlands, Oct. 8-16 2016. 4

[16] Chunxiao Liu, Zhendong Mao, Wenyu Zang, and Bin Wang. A neighbor-aware approach for image-text matching. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, Brighton, United Kingdom, May 12-17 2019. 4

[17] Yu Liu, Yanming Guo, Erwin M Bakker, and Michael S Lew. Learning a recurrent residual fusion network for multimodal matching. In *Proc. International Conference on Computer Vision*, Venice, Italy, Oct. 22-29 2017. 4

[18] Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li. Multimodal convolutional neural networks for matching image and sentence. In *Proc. International Conference on Computer Vision*, Santiago, Chile, Dec. 13-16 2015. 4

[19] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-RNN). In *Proc. International Conference on Learning Representations*, San Diego, CA, May 7-9 2015. 4

[20] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *Proc. Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, July 21-26 2017. 1, 4

[21] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proc. International Conference on Computer Vision*, Santiago, Chile, Dec. 13-16 2015. 2, 4

[22] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *Proc. Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, June 26 - July 1 2016. 2, 3

[23] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proc. Conference on Computer Vision and Pattern Recognition*, Boston, MA, June 8-10 2015. 3

[24] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proc. Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, June 26 - July 1 2016. 4

[25] Fei Yan and Krystian Mikolajczyk. Deep correlation for matching images and text. In *Proc. Conference on Computer Vision and Pattern Recognition*, Boston, MA, June 8-10 2015. 4

[26] Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *Proc. European Conference on Computer Vision*, Munich, Germany, Sept. 8-14 2018. 1, 2, 3, 4

[27] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, and Yi-Dong Shen. Dual-path convolutional image-text embedding with instance loss. *arXiv preprint arXiv:1711.05535*, 2017. 3

[28] Bolei Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. Simple baseline for visual question answering. *arXiv preprint arXiv:1512.02167*, 2015. 3