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

Abstract

For many computer vision applications such as image captioning, visual question answering, and person search, learning discriminative feature representations at both image and text level is an essential yet challenging problem. Its challenges originate from the large word variance in the text domain as well as the difficulty of accurately measuring the distance between the features of the two modalities. Most prior work focuses on the latter challenge, by introducing loss functions that help the network learn better feature representations but fail to account for the complexity of the textual input. With that in mind, we introduce TIMAM: a Text-Image Modality Adversarial Matching approach that learns modality-invariant feature representations using adversarial and cross-modal matching objectives. In addition, we demonstrate that BERT, a publicly-available language model that extracts word embeddings, can successfully be applied in the text-to-image matching domain. The proposed approach achieves state-of-the-art cross-modal matching performance on four widely-used publicly-available datasets resulting in absolute improvements ranging from 2% to 5% in terms of rank-1 accuracy.

1. Introduction

We set out to develop a cross-modal matching method that given a textual description, identifies and retrieves the most relevant images. For example, given the sentence “A woman with a white shirt carrying a black purse on her hand” we aspire to obtain images of individuals with such visual attributes. The first challenge of matching images and text is the large word variability in the textual descriptions even when they are describing the same image. What is considered as important information for one is not necessarily the same for another annotator. At the same time, textual descriptions might contain mistakes, the descriptions can be too long or the annotator might describe additional information that is available on the image but is not related to the primary point of interest (e.g., human, object). All of these factors make text-to-image matching a difficult problem since learning good feature representations from such descriptions is not straightforward.

A second major challenge of text-to-image matching is how to accurately measure the distance between the text and image features. During deployment, the distance between the probe text features and all the gallery image features is computed and the results are ranked based on this criterion. Most existing methods introduce loss functions to tackle this challenge. For example, Li et al. [25] proposed to “bring close-together” cross-modal features originating from the same identity and “push away” features from different identities. Although such methods have continuously outperformed previous state-of-the-art, their performance remains unsatisfactory. For example, the best performing text-to-image matching method on the CUHK-PEDES dataset [26] is below 50% in terms of rank-1 accuracy. Finally, most methods usually rely on some assumptions under which they perform matching. For example, in the work of Chen et al. [6], part-of-speech (POS) tagging is performed to extract local phrases (e.g., nouns with adjectives). However, when we performed POS tagging on the same textual inputs, we observed that important information is lost since the same word can be tagged differently depending on the context or its location within the sentence.
(e.g., the word “t-shirt” was frequently identified as an adjective although it was used as a noun in the description).

In this paper, our objectives are: to (i) learn discriminative representations from both the visual and the textual inputs; and (ii) improve upon previous text-to-image matching methods in terms of how the word embeddings are learned. To accomplish these tasks, we introduce TIMAM: a Text-Image Modality Adversarial Matching method that performs matching between the two modalities and achieves state-of-the-art results without requiring any additional supervision.

The first contribution of this work is an adversarial representation learning (ARL) framework that brings the features from both modalities “close-to-each-other”. The textual and visual feature representations are fed to a discriminator that aims to identify whether the input is originating from the visual or textual modality. By learning to fool the discriminator, we can learn modality-invariant feature representations that are capable of successfully performing text-to-image matching. The adversarial loss of the discriminator along with an identification loss and a cross-modal projection matching loss are used to jointly train the whole network end-to-end. We demonstrate that adversarial learning is well-suited for cross-modal matching and that it results in improved rank-1 accuracy. Our second contribution stems from improving upon previous text-to-image matching methods in terms of how the word embeddings are learned. We borrow from the NLP community a recent language representation model named BERT [10], which stands for Bidirectional Encoder Representations from Transformers. We demonstrate that such a model can successfully be applied to text-to-image matching and can significantly improve the performance of existing approaches. Each description is fed to the language model which extracts word representations that are then fed to an LSTM and mapped to a final sentence embedding.

Thus, TIMAM results in more discriminative feature representations learned from the proposed objective functions, while using the learning capabilities of the backbones of the two modalities. Through experiments, ablation studies and qualitative results, we demonstrate that:

- Adversarial learning is well-suited for cross-modal matching and that it results in more discriminative embeddings from both modalities. Using our proposed learning approach, we observe improvements ranging from 2% to 5% in terms of rank-1 accuracy over the previous best-performing techniques.
- Pre-trained language models can successfully be applied to cross-modal matching. By leveraging the fine-tuning capabilities of BERT, we learn better word embeddings. Our experimental results indicate rank-1 accuracy improvements ranging from 3% to 5% over previous work when features are learned in this manner.

2. Related Work

Text-Image Matching: Learning cross-modal embeddings has numerous applications [55, 63] ranging from PINs using facial and voice information [32], to generative feature learning [15] and domain adaptation [57, 59]. Nagrani et al. [32] demonstrated that a joint representation can be learned from facial and voice information and introduced a curriculum learning strategy [3, 40, 41] to perform hard negative mining during training. Text-to-image matching is a well-studied problem in computer vision [4, 19, 24, 33, 44, 46, 53, 54, 60, 62] facilitated by datasets describing objects, birds, or flowers [27, 39, 61]. A relatively new application of text-to-image matching is person search the task of which is to retrieve the most relevant frames of an individual given a textual description as an input. Most methods [21, 25, 26] rely on a relatively similar procedure: (i) extract discriminative image features using a deep neural network, (ii) extract text features using an LSTM, and (iii) propose a loss function that measures as accurately as possible the distance between the two embeddings. To improve the performance, some interesting ideas include jointly learning the 2D pose along with attention masks [21] or associating image features from sub-regions with the corresponding phrases in the text [6]. While such approaches have shown significant improvements, they: (i) ignore the large variability of the textual input by solely relying on an LSTM to model the input sentences, and (ii) demonstrate unsatisfactory results as the best text-to-image rank-1 accuracy results in the literature are below 50% and 40% in the CUHK-PEDES [26] and Flickr30K [36] datasets, respectively. Finally, there have been a few works recently [15, 16, 34, 51, 67] that have applied adversarial learning in cross-modal matching applications. Zhu et al. [67] introduced $R^2$GAN: a text-to-image matching method accompanied by a large-scale dataset to perform retrieval of food recipes. Unlike the rest of the datasets where the textual input is a description of the image, in this case the input consists of a title, a set of ingredients and a list of instructions for the recipe, which introduce additional challenges on how to properly handle and learn discriminative textual representations.

Overview of BERT: BERT [10] is a deep language model capable of extracting discriminative embeddings. Unlike previous approaches [35, 38] that processed the input sequences either from left to right or combined left-to-right and right-to-left training, BERT relies on applying the bidirectional training of Transformer [50] to language modeling. By leveraging the Transformer architecture (which is an attention mechanism) BERT learns the contextual relations between the words in a textual description. In addition, it introduces a word masking mechanism, which masks 15% of the words with a token and then a model is trained to predict the original value of the masked words based on the context. In that way, BERT learns robust word embeddings.
3. Methodology

In this section, we introduce TIMAM: a cross-modal matching approach that learns to match the feature representations from the two modalities in order to perform both text-to-image and image-to-text retrieval.

3.1. Joint Feature Learning

During training, our objective is to learn discriminative visual and textual feature representations capable of accurately retrieving the ID (or the category) of the input from another modality. The training procedure is depicted in Figure 2 and is described in detail below. Specifically, our input at training-time consists of triplets \((V_i, T_i, Y_i)\) where \(V_i\) is the image input from the visual domain, \(T_i\) a textual description from the textual domain describing that image, and \(Y_i\) is the identity/category of the input. To learn the visual representations denoted by \(\phi(V_i)\), any image classification model can be used as a backbone network (a ResNet-101 network is used in this work). The feature map of the last residual block is projected to the dimensionality of the feature vector using global average pooling and a fully-connected layer. We opted for the original backbone architecture without any attention blocks \([6, 42]\) in order to keep the backbones simple and easy-to-reproduce in any framework and to avoid having to learn more parameters.

Learning discriminative representations from both modalities is of paramount importance for text-to-image matching. While for the image domain, most existing methods \([6, 21, 25, 66]\) rely on deep architectures that have demonstrated their capability of extracting discriminative features for a wide range of tasks, this is not the case for the text domain. Prior work usually relies on a single LSTM \([17]\) to model the textual input and learn the features that correspond to the input sentence. We argue that one of the main reasons that prevent existing computer vision methods from performing well on text-to-image matching problems is due to the fact that the textual features are not discriminative enough. To address this limitation, we borrow from the NLP community a recently proposed language representation model named BERT. The sequence of word embeddings extracted from BERT is then fed to a bidirectional LSTM \([17]\), which effectively summarizes the content of the input textual description. Finally, the textual representation denoted by \(\tau(T_i)\) is obtained by projecting the output of the LSTM to the dimensionality of the feature vector using a fully-connected layer. The reason an LSTM is employed on the output word embeddings is because it gives us the flexibility to initially “freeze” the weights of the language model and fine-tune only the LSTM along with the fully-connected layer and thus, significantly reducing the number of parameters. Once an adequate performance is observed, we “unfreeze” the weights of the language model and the whole network is trained end-to-end.

---

**Figure 2:** TIMAM consists of three modules: (i) the feature extraction module which extracts textual and visual features using their corresponding backbone architectures, (ii) the identification and cross-modal projection losses that match the feature distributions originating from the same identity, and (iii) an adversarial discriminator that pushes the model to learn modality-invariant representations for effective text-image matching.
3.2. Cross-Modal Matching

Given the visual and textual features, our aim is to introduce loss functions that will bring the features originating from the same identity/category close together and push away features originating from different identities. To accomplish this task, we introduce two loss functions for identification and cross-modal matching. The identification loss is a norm-softmax cross entropy loss which is commonly used in face identification applications [29, 52, 58] that introduces an $L_2$-normalization on the weights of the output layer. By doing so, it enforces the model to focus on the angle between the weights of the different samples instead of their magnitude. For the visual features, the norm-softmax cross entropy loss can be described as follows:

$$L_V^I = -\frac{1}{B} \sum_{i=1}^{B} \log \left( \frac{\exp(W_i^T \phi(V_i) + b_i)}{\sum_j \exp(W_j^T \phi(V_i) + b_j)} \right),$$

s.t. $||W_j|| = 1, \forall j \in [1, B],$

where $I$ stands for identification, $V$ corresponds to the visual modality, $B$ is the batch size, and $W_i, b_i$ are the weights and the bias of the classification layer for the visual feature representation $\phi(V_i).$ The loss for the textual features $L_T^V$ is computed in a similar manner and the final classification loss for identification $L_I = L_Y^I + L_T^V.$ It is worth noting that for datasets that do not have ID labels but only image-text pairs (e.g., the Flickr30K dataset [36]), we assign a unique ID to each image and use that ID as ground-truth for the identification loss. However, focusing solely on performing accurate identification is not sufficient for cross-modal matching since no association between the representations of the two modalities has been introduced thus far. To address this challenge, we use the cross-modal projection matching loss [65] which incorporates the cross-modal projection into the KL divergence measure to associate the representations across different modalities. The text representation is first normalized $\tau(T_j) = \frac{\tau(T_j)}{||\tau(T_j)||},$ and then the probability of matching $\phi(V_i)$ to $\tau(T_j)$ is given by:

$$p_{i,j} = \frac{\exp \left( \phi(V_i)^T \tau(T_j) \right)}{\sum_{k=1}^{B} \exp \left( \phi(V_i)^T \tau(T_k) \right)}.$$

The multiplication between the transposed image embedding and the normalized textual embedding reflects the scalar projection between $\phi(V_i)$ onto $\tau(T_j),$ while the probability $p_{i,j}$ represents the proportion of this scalar projection among all scalar projections between pairs in a batch. Thus, the more similar the image embedding is to the textual embedding, the larger the scalar projection is from the former to the latter. Since in each mini-batch there might be more than one positive matches (i.e., visual and textual features originating from the same identity) the true matching probability is normalized as follows: $q_{i,j} = \frac{p_{i,j}}{\sum_{k=1}^{B} (p_{i,k}).$ The cross-modal projection matching loss of associating $\phi(V_i)$ with the correctly matched text features is then defined as the KL divergence from the true matching distribution $q_i$ to the probability of matching $p_i$. For each batch this loss is defined as:

$$L_M^V = -\frac{1}{B} \sum_{i=1}^{B} \sum_{j=1}^{B} p_{i,j} \log \left( \frac{p_{i,j}}{q_{i,j} + \epsilon} \right),$$

where $M$ denotes matching, and $\epsilon$ is a very small number for preventing division by zero. The same procedure is followed to perform the opposite matching (i.e., from text to image) to compute loss $L_M^T$ during which the visual features are normalized instead of $\tau(T_i)$ to compute Eq. (2). Finally, the summation of the two individual losses constitutes the cross-modal projection matching loss $L_M = L_M^V + L_M^T.$

3.3. Adversarial Cross-Modal Learning

When training adversarial neural networks [5, 13, 48] a two-player minimax game is played between a discriminator $D$ and a feature generator $G.$ Both $G$ and $D$ are jointly trained so as $G$ tries to fool $D$ and $D$ tries to make accurate predictions. For the text-to-image matching problem, the two backbone architectures discussed in Section 3.1 serve as the feature generators $G^V$ and $G^T$ for the visual and textual modalities that produce feature representations $\phi(V_i)$ and $\tau(T_i),$ respectively. The key idea is to learn a good general representation for each input modality that maximizes the matching performance, yet obscures the modality information. By learning to fool the modality discriminator, better feature representations are learned, that are capable of performing text-to-image matching. The generated embeddings are fed to the modality discriminator, which classifies whether the input feature representation is drawn from the visual or the textual modality. The discriminator consists of two fully-connected layers that reduce the embedding size to a scalar value which is used to predict the input modality. The discriminator is optimized according to the following GAN [14] loss function:

$$L_D = -\mathbb{E}_{V_i \sim V} \log D(\phi(V_i)) - \mathbb{E}_{T_i \sim T} \log (1 - D(\tau(T_i))),$$

where $V$ and $T$ correspond to the image and text modalities respectively where samples are drawn and fed through the backbone architectures.

3.4. Training and Testing Details

The loss function that is used to train TIMAM is the summation of two identification losses ($L_I$), the two cross-modal matching losses ($L_M$) and the adversarial loss of the discriminator ($L_D$):

$$L = L_I + L_M + L_D$$
Figure 3: The three learning objectives. Top-left: We learn to classify each embedding based on the input ID. Top-right: We learn modality-invariant features using the discriminator. Bottom: We compute cross-modal projections between all samples in a batch in which samples from the same pair have a larger scalar projection, and learn to match using the predicted and the true matching probabilities.

An illustrative explanation of the three learning objectives is presented in Figure 3. We used stochastic gradient descent (SGD) with momentum equal to 0.9 to train the image and discriminator networks and the Adam optimizer [22] for the textual networks. The learning rate was set to $2 \times 10^{-4}$ and was divided by ten when the loss plateaued at the validation set until $2 \times 10^{-6}$. The batch-size was set to 64 and the weight decay to 4.

The distances are then sorted and rank-1 through rank-10 results are reported. For image-to-text matching the same process is followed by using the image features as probe and retrieving the most relevant textual descriptions.

### 4. Experiments

**Datasets:** To evaluate our method, four widely-used publicly available datasets were used and their evaluation protocols were strictly followed. We opted for these datasets in order to test TIMAM on a wide range of tasks ranging from pedestrians and flowers to objects and scenes. TIMAM was tested on (i) the CUHK-PEDES [26] that contains images of pedestrians accompanied by two textual descriptions, (ii) the Flickr30K dataset [36], which contains a wide variety of images (humans, animals, objects, scenes) with five descriptions for each image, (iii) the Caltech-UCSD Birds (CUB) [39] dataset that consists of images of birds with 10 descriptions for each image and finally, (iv) the Flowers [39] dataset that consists of images of flowers originating for 102 categories with 10 descriptions for each image.

**Evaluation Metrics:** The evaluation metrics used in each dataset are adopted. Thus, for the CUHK-PEDES and Flickr30K datasets rank-1, rank-5, and rank-10 results are presented for each method. For the CUB and Flowers datasets, the $AP@50$ metric is utilized for text-to-image retrieval and rank-1 for image-to-text matching. Given a query textual class, the algorithm first computes the percentage of top-50 retrieved images whose identity matches that of the textual query class. The average matching percentage of all test classes is denoted as $AP@50$. Finally, note that in each dataset TIMAM is evaluated against the eight best-performing methods. The complete results against all methods tested in each dataset as well as all details regarding the datasets (e.g., train/val/test splits, pre-processing) are included in the supplementary material.

### 4.1. Quantitative Results

**CUHK-PEDES Dataset:** We evaluate our approach against the eight best-performing methods that have been tested on the CUHK-PEDES dataset and present text-to-image matching results in Table 1. Some key methods that have
Table 2: Matching results on the Flickr30K dataset. The results are ordered based on their text-to-image rank-1 accuracy.

<table>
<thead>
<tr>
<th>Method</th>
<th>Image Backbone</th>
<th>Image-to-Text Rank-1</th>
<th>Rank-5</th>
<th>Rank-10</th>
<th>Text-to-Image Rank-1</th>
<th>Rank-5</th>
<th>Rank-10</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAN [33]</td>
<td>VGG-19</td>
<td>41.4</td>
<td>73.5</td>
<td>82.5</td>
<td>31.8</td>
<td>61.7</td>
<td>72.5</td>
</tr>
<tr>
<td>RRF-Net [30]</td>
<td>ResNet-152</td>
<td>47.6</td>
<td>77.4</td>
<td>87.1</td>
<td>35.4</td>
<td>68.3</td>
<td>79.9</td>
</tr>
<tr>
<td>CMPM + CMPC [65]</td>
<td>ResNet-152</td>
<td>49.6</td>
<td>76.8</td>
<td>86.1</td>
<td>37.3</td>
<td>65.7</td>
<td>75.5</td>
</tr>
<tr>
<td>DAN [33]</td>
<td>ResNet-152</td>
<td>55.0</td>
<td>81.8</td>
<td>89.0</td>
<td>39.4</td>
<td>69.2</td>
<td>79.1</td>
</tr>
<tr>
<td>NAR [28]</td>
<td>ResNet-152</td>
<td>55.1</td>
<td>80.3</td>
<td>89.6</td>
<td>39.4</td>
<td>68.8</td>
<td>79.9</td>
</tr>
<tr>
<td>VSE++ [12]</td>
<td>ResNet-152</td>
<td>52.9</td>
<td>80.5</td>
<td>87.2</td>
<td>39.6</td>
<td>70.1</td>
<td>79.5</td>
</tr>
<tr>
<td>SCO [20]</td>
<td>ResNet-152</td>
<td>55.5</td>
<td>82.0</td>
<td>89.3</td>
<td>41.1</td>
<td>70.5</td>
<td>80.1</td>
</tr>
<tr>
<td>GXN [15]</td>
<td>ResNet-152</td>
<td>56.8</td>
<td>89.6</td>
<td></td>
<td>41.5</td>
<td>-</td>
<td>80.1</td>
</tr>
<tr>
<td>TIMAM</td>
<td>ResNet-152</td>
<td>53.1</td>
<td>78.8</td>
<td>87.6</td>
<td>42.6</td>
<td>71.6</td>
<td>81.9</td>
</tr>
</tbody>
</table>

been evaluated on this dataset include (i) IATV [25] which learns discriminative features using two attention modules working on both modalities at different levels but it is not end-to-end; (ii) GLA [6] which identifies local textual phrases and aims to find the corresponding image regions using an attention mechanism; (iii) and CMPM [65] in which two projection losses are proposed to learn features for text-to-image matching. TIMAM outperforms all previous works by a large margin. We observe an absolute improvement of more than 5% in terms of rank-1 over the previous best performing method [65] which originates from learning better feature representations through the identification and cross-modal matching losses as well as the proposed adversarial learning framework.

CUB and Flowers Datasets: We test TIMAM against all eight methods evaluated on these datasets and present our matching results in Table 3. Our method achieves state-of-the-art results in both image-to-text and text-to-image matching in both datasets. We observe performance increases of 2.2% and 3.4% in terms of rank-1 accuracy as well as 3.6% and 2.4% in terms of AP@50.

Flickr30K Dataset: In Table 2, we report cross-modal retrieval results on the Flickr30K dataset against the top-8 best-performing methods. Similar to the best-performing methods, and only in this dataset, a ResNet-152 is employed to allow for a fair comparison. TIMAM surpasses all methods by a large margin in text-to-image matching but demonstrates inferior performance compared to GXN [15] in image-to-text matching. Most of the best-performing image-to-text matching methods employ multi-step attention blocks and thus, are able to learn “where to look” in an image which results in better image features. Unlike the rest of the datasets that contain a single primary object (i.e., flowers/birds/pedestrians only), Flickr30K contains a wide range of primary components. This image variance coupled with the relatively small number of training images make cross-modal matching a challenging task. While our approach achieves state-of-the-art results in the text-to-image matching task and is capable of learning correct associa-

Table 3: Cross-modal matching results on the CUB and Flowers datasets. The results are ordered based on the text-to-image AP@50 performance.

<table>
<thead>
<tr>
<th>Method</th>
<th>CUB Img2Txt Txt2Img</th>
<th>Flowers Img2Txt Txt2Img</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rank-1 AP@50</td>
<td>Rank-1 AP@50</td>
</tr>
<tr>
<td>Word2Vec [31]</td>
<td>38.6</td>
<td>33.5</td>
</tr>
<tr>
<td>GMM+HGLMM [23]</td>
<td>36.5</td>
<td>35.6</td>
</tr>
<tr>
<td>Word CNN [39]</td>
<td>51.0</td>
<td>43.3</td>
</tr>
<tr>
<td>Word CNN-RNN [39]</td>
<td>56.8</td>
<td>48.7</td>
</tr>
<tr>
<td>Attributes [1]</td>
<td>50.4</td>
<td>50.0</td>
</tr>
<tr>
<td>Triplet Loss [25]</td>
<td>52.5</td>
<td>52.4</td>
</tr>
<tr>
<td>IATV [25]</td>
<td>61.5</td>
<td>57.6</td>
</tr>
<tr>
<td>CMPM+CMPC [65]</td>
<td>64.3</td>
<td>67.9</td>
</tr>
<tr>
<td>TIMAM</td>
<td>67.7</td>
<td>70.3</td>
</tr>
</tbody>
</table>

4.2. Ablation Studies

Impact of Proposed Components: In our first ablation study (Table 4), we assess how each proposed component of TIMAM contributes to the final text-to-image matching performance on the CUHK-PEDES dataset. We observe that the identification ($L_I$) and cross-modal projection ($L_M$) losses result in a rank-1 accuracy of 49.85% when used together and considerably less when used individually. By introducing BERT, better word embeddings can be learned that increase the accuracy to 52.97%. Finally, when the proposed adversarial representation learning paradigm (ARL) is used, additional improvements are observed, regardless of whether BERT is used or not. We observe relative improvements of 2.9% and 3% with and without BERT respectively, which demonstrates that ARL helps the network learn modality-invariant representations that can successfully be used at deployment-time to perform cross-modal matching. Similar results are obtained in the Flickr30K
dataset in which ARL improved the rank-1 matching performance from 51.2% to 53.1% and from 41.0% to 42.6% in image-to-text and text-to-image respectively.

**Impact Backbone Depth:** In the second ablation study, we investigate to what extent the depth of the backbone networks affects the final performance. Similar to previous well-performing methods [6, 25, 65], a fully-connected layer was used to learn the word embeddings (denoted by FC-Embed.) and its impact was compared with the deep language model of BERT. For the image modality, two different ResNet backbones were employed while the rest of our proposed methodology remained the same. Rank-1 matching results in both directions are reported on the CUHK-PEDES dataset in which ARL improved the rank-1 matching performance from 51.2% to 53.1% and from 41.0% to 42.6% in image-to-text and text-to-image respectively.

**Qualitative Results:** Figure 4 depicts cross-modal retrieval results for all four datasets. We observe that TIMAM is capable of learning cloth and accessory-related correspondences as it can accurately retrieve images of people carrying bags with the correct set of clothing. The proposed approach retrieves consistent images given a textual query (e.g., group of people in snow) as well as similar textual descriptions, given an image query (e.g., all three descriptions describe dogs or soccer players in the first and second row on the right). Finally, some interesting observations can be made from the failure cases. While all retrieved images in Figure 5 have a different ID than the true label of the textual description, TIMAM can still retrieve images that match the textual input. For example, all images in the second row contain a female with a white t-shirt, black pants, carrying a bag.

**4.3. Discussion of Alternatives**

**Loss Functions:** The wide variety of loss functions available in the literature that could potentially be applicable to our problem, begs the question of why did we choose these specific identification and matching losses instead of other alternatives? Our first goal was to refrain from using losses that sample triplets or quadruplets within each batch [8, 11, 18, 43]. The reason is that such losses introduce a computational overhead during training [43, 47] and additional hard-mining schemes [11] would have to be incorporated to ensure that hard negatives are provided to such loss functions. Second, we relied on the experimental investigation of Zhang and Lu [65] that showed the KL-based loss described in Eq. (3) achieves superior matching performance among other alternatives [30, 45, 49] and is robust to small or large batch sizes. Finally, we demonstrate through the ablation studies described in Section 4.2, that the identification and matching losses we introduced result in substantial improvements in terms of rank-1 accuracy.

**Text Augmentation:** Aiming to introduce some noise to the textual input as a form of data-augmentation that could potentially improve the performance we experimented with the conditional augmentation technique of Zhang and Lu [65] that showed the KL-based loss described in Eq. (3) achieves superior matching performance among other alternatives [30, 45, 49]. Conditional augmentation resembles the reparameterization trick used in variational autoencoders, as it maps the text embedding to a mean and a variance feature vector which are then added together using some noise sampled from \( \mathcal{N}(0, I) \). In that way, more training pairs are generated given a small number of image-text pairs, and the method is robust to small perturbations in the conditioning manifold. However, when we experimented with this technique on the CUHK-PEDES dataset, we observed consistently worse results compared to the original approach. We thus believe that conditional augmentation is not suitable for text-to-image matching as not only two additional embeddings need to be learned (i.e., more parameters), but also the noise that is introduced ends up obfuscating the model instead of augmenting its learning and generalization capabilities.

**Text-to-Image Reconstruction:** Aiming to learn textual features capable of retrieving the most relevant images, we experimented with text-to-image reconstruction as an additional learning objective trained in an end-to-end setup. The textual embedding was fed to a decoder comprising upsampling and convolution layers that reconstructed the corresponding input image at different scales. While this method demonstrated good reconstruction results in the

---

Table 4: Ablation studies on the CUHK-PEDES dataset to investigate the additions in terms of rank-1 and rank-10 accuracy for the identification (\( L_I \)) and cross-modal projection (\( L_M \)) losses, the addition of BERT as a backbone architecture for language modeling, and the adversarial representation learning paradigm.

<table>
<thead>
<tr>
<th>( L_I )</th>
<th>( L_M )</th>
<th>BERT</th>
<th>ARL</th>
<th>Rank-1</th>
<th>Rank-10</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>40.1</td>
<td>70.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>44.9</td>
<td>77.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>49.8</td>
<td>81.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>51.3</td>
<td>82.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>52.9</td>
<td>83.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>54.5</td>
<td>84.8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Ablation studies on the Flickr30K dataset to assess the impact of the depth of different backbone architectures.

<table>
<thead>
<tr>
<th>Image Backbone</th>
<th>Text Backbone</th>
<th>Img2Txt</th>
<th>Txt2Img</th>
<th>Rank-1</th>
<th>Rank-10</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>52.1</td>
<td>71.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>55.2</td>
<td>72.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>57.3</td>
<td>73.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>59.4</td>
<td>74.6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Table 4 and Table 5 are placeholders and should be replaced with actual data from the document.]
Learning discriminative representations for cross-modal matching has significant challenges such as the large variance of the linguistic input and the difficulty of measuring the distance between the multi-modal features. To address these challenges, we introduced TIMAM: a text-image matching approach that employs an adversarial discriminator that aims to identify whether the input is originating from the visual or textual modality. When the discriminator is jointly trained with identification and cross-modal matching objectives, it results in discriminative modality-invariant embeddings. In addition, we observed that a deep language model can boost the cross-modal matching capabilities since better textual embeddings are learned. We demonstrated through extensive experiments, that (i) adversarial learning is well-suited for text-image matching, and (ii) a deep language model can successfully be utilized in cross-modal matching applications. State-of-the-art results were obtained in four publicly available datasets all of which are widely used in this domain. To facilitate further investigation by future research, we performed ablation studies, discussed alternative approaches that were explored, and presented qualitative results that provide an insight into the performance of our approach.

Acknowledgments

This work has been funded in part by the UH Hugh Roy and Lillie Cranz Cullen Endowment Fund. All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of the sponsors.
References


[61] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics, 2:67–78, 2014. 2


