Hierarchical Gumbel Attention Network for Text-based Person Search

Kecheng Zheng[1†], Wu Liu[2*], Jiawei Liu[1], Zheng-Jun Zha[1*], Tao Mei[2]
[1]University of Science and Technology of China, China, [2]AI Research of JD.com
zkcys001@mail.ustc.edu.cn,{liuwu1,tmei}@jd.com,{jwliu6,zhazj}@ustc.edu.cn,

ABSTRACT

Text-based person search aims to retrieve the pedestrian images that best match a given textual description from gallery images. Previous methods utilize the soft-attention mechanism to infer the semantic alignments between the regions of image and the corresponding words in sentence. However, these methods may fuse the irrelevant multi-modality features together which cause matching redundancy problem. In this work, we propose a novel hierarchical Gumbel attention network for text-based person search via Gumbel top-k re-parameterization algorithm. Specifically, it adaptively selects the strong semantically relevant image regions and words/phrases from images and texts for precise alignment and similarity calculation. This hard selection strategy is able to fuse the strong-relevant multi-modality features for alleviating the problem of matching redundancy. Meanwhile, a Gumbel top-k re-parameterization algorithm is designed as a low-variance, unbiased gradient estimator to handle the discreteness problem of hard attention mechanism by an end-to-end manner. Moreover, a hierarchical adaptive matching strategy is employed by the model from three different granularities, i.e., word-level, phrase-level, and sentence-level, towards fine-grained matching. Extensive experimental results demonstrate the state-of-the-art performance. Compared the existed best method, we achieve the 8.24% Rank-1 and 7.6% mAP relative improvements in the text-to-image retrieval task, and 5.58% Rank-1 and 6.3% mAP relative improvements in the image-to-text retrieval task on CUHK-PEDES dataset, respectively.

CCS CONCEPTS

• **Information systems** → **Image search**.

KEYWORDS

Gumbel attention, Text-based person search, Hierarchical adaptive matching

ACM Reference Format:
Kecheng Zheng[1†], Wu Liu[2*], Jiawei Liu[1], Zheng-Jun Zha[1*], Tao Mei[2]. 2020. Hierarchical Gumbel Attention Network for Text-based Person Search. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413864

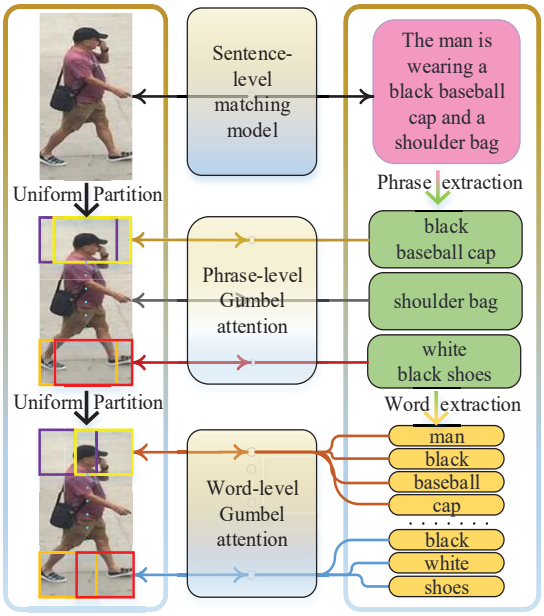

Figure 1: We factorize image-text matching into hierarchical levels including word-level, phrase-level and sentence-level to form a global to local structure. This enhances global matching with the help of detailed appearance description. The Gumbel attention module selects the strong semantically relevant image regions and words/phrases from images and texts for similarity calculation.

1 INTRODUCTION

Text-based person search aims to retrieve the target persons in an image database by a given natural language description. With the increasing scale of surveillance data [6, 16, 25, 28], it is unrealistic for manual person search in a large-scale dataset. Hence, automatic text-based person search algorithms are urgently demanded for handling this issue. This task possess various potential applications such as video surveillance and criminal investigation [18, 19, 25, 32]

Although general text-image retrieval methods [12, 20, 34, 36, 37, 39] recently have made progress on some popular benchmarks, e.g., MSCOCO [15], they often fail to generalize well on text-based person search. This is due to the fact that these general text-image

*Corresponding author.
†This work is done when Kecheng Zheng is an intern at JD AI Research.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413864

retrieval methods focus on encoding the global semantic information to exploit the relationship between objects, which neglect the subtle visual cues and text attribute descriptions. For fine-grained retrieval task, especially text-based person search, only a category of object (just a person, as shown in the Figure 1) exists in an image, and the text describes fine-grained visual appearance. Thus, we should consider fine-grained visual-textual matching under the premise of better global matching. In a word, the majority challenge of existing text-based person search methods is how to utilize the fine-grained and global information together to align the multi-granularity relevance.

To align the multi-granularity relevance, it is important to effectively extract corresponding visual contents to the human description between image and text. As shown in the Figure 1, "black baseball cap" corresponds to the visual content with the yellow boxes. Other phrases in this example are irrelevant to this visual content and should be filtered. Previous methods [12] adopt the soft attention mechanism to obtain weighted matching between image and text for alleviating the irrelevant matching. It inescapably fuses the irrelevant phrases or words into a region of visual feature towards unnecessary matching, such as the visual content with the yellow boxes to "shoulder bag" in the Figure 1. Therefore, it is necessary to design a module to select the strong semantically relevant image regions and words/phrases from images and texts for precise alignment and similarity calculation.

To address the above challenges, hard attention mechanism is able to select the strong alignments rather than the all soft selections of soft attention mechanism. Previous method [8] proposes a pose-guided hard attention mechanism to further select a subset of corresponding images regions and words from human pose information. Nevertheless, the posed hard attention module may have a high-variance, biased gradient estimator, which is difficult to learn the correct alignments between the image regions and words/phrases of a sentence. Besides, it also requires additional pose information which is time-consuming. Differently, the Gumbel distribution based attention mechanism is a low-variance, unbiased gradient estimator which can well solve this task. In probability theory and statistics, the Gumbel distribution is used to model the distribution of the maximum of several samples of various distributions. Gumbel-Max trick based on Gumbel distribution allows sampling from the categorical distribution, which is a natural choice for representing discrete structure in the world [7]. By adding independent noise from the Gumbel distribution, the log-probability of each category is perturbed to return the category with maximum perturbed log-probability. Therefore, this trick can be smoothly annealed into a kind of differentiable maximum selector. In this task, using top-k discrete variables is more commonly effective than maximum one. Thus, we design the Gumbel attention module based on Gumbel-Top-k trick, which is able to implement differentiable top-k hard attention mechanism.

In this work, we propose a novel hierarchical Gumbel attention network(HGAN) for text-based person search via Gumbel top-k reparameterization algorithm, towards discovering the strong multi-level alignments between visual content and textual descriptions. Specifically, the Gumbel attention module within the HGAN selects the strong semantically relevant image regions and words/phrases from images and texts for precise alignment and similarity calculation. This strong alignment selection strategy as a kind of hard attention mechanism is able to fuse the strong-relevant multi-modality features for alleviating the problem of matching redundancy. Meanwhile, a Gumbel top-k re-parameterization algorithm is designed as a low-variance, unbiased gradient estimator to handle the discreteness problem of hard attention mechanism by an end-to-end manner. Moreover, a hierarchical adaptive matching strategy is employed by the model, which exploits the multi-level semantic relevance between the textual descriptions of persons and corresponding visual content from three different granularities, i.e., word-level, phrase-level, and sentence-level. The sentence-level matching network aims to learn discriminative cross-modal representations and calculate similarity from global views. To further capture the meaningful fine-grained relations, we adopt the word-level and phrase-level matching model to compute the local similarity between image regions and textual descriptions with the guidance of the Gumbel attention module. Extensive experimental results have demonstrated the effectiveness of the proposed method on three challenging datasets, i.e., CUHK-PEDES, CUB and Flowers.

The main contributions are summarized as follow:

- We propose a novel Gumbel attention module to alleviate matching redundancy problem by selecting the strong semantically relevant regions for all the regions of images, and the corresponding words/phrase.
- We propose a hierarchical adaptive matching model to learn subtle feature representations from three different granularities, i.e., word-level, phrase-level and sentence-level, for fine-grained image-text retrieval tasks.
- Extensive experiments on three challenging datasets, i.e., CUHK-PEDES, CUB and Flowers, demonstrate the effectiveness of our proposed method.

2 RELATED WORK

Text-based Person Search Text-based person search retrieves the pedestrian through natural language description. In order to caption the affinity between the words of a sentence and corresponding image feature, a recurrent neural network with gated neural attention [14] is proposed to take a description sentence and a person image both as the input of LSTM. A dual-path model [40] is exploited for visual-textual embedding learning, which uses the instance loss to investigate the fine-grained difference in intra-modality. A graph attention based adversarial learning [17, 26] is adopted to exploit directed semantic scene graph for learning modality-invariant feature representations. There are also some attempts to address the fine-grained retrieval task by learning the semantic relevance between the corresponding local regions of images and the phrases in the textual description. Some soft-attention based cross-modal retrieve methods [12, 20, 39] have achieved promising results on several general image-text retrieve tasks by directly inferring the semantic alignment between the regions of images and the corresponding words in sentences, which are not ineffective and easily to produce matching redundancy. Ya et al. [8] attempt to build similarity models with the attention mechanism to compute the matching score of image-text pair under the guidance of human pose [24]. However, these methods try to infer the semantic alignments by attending all

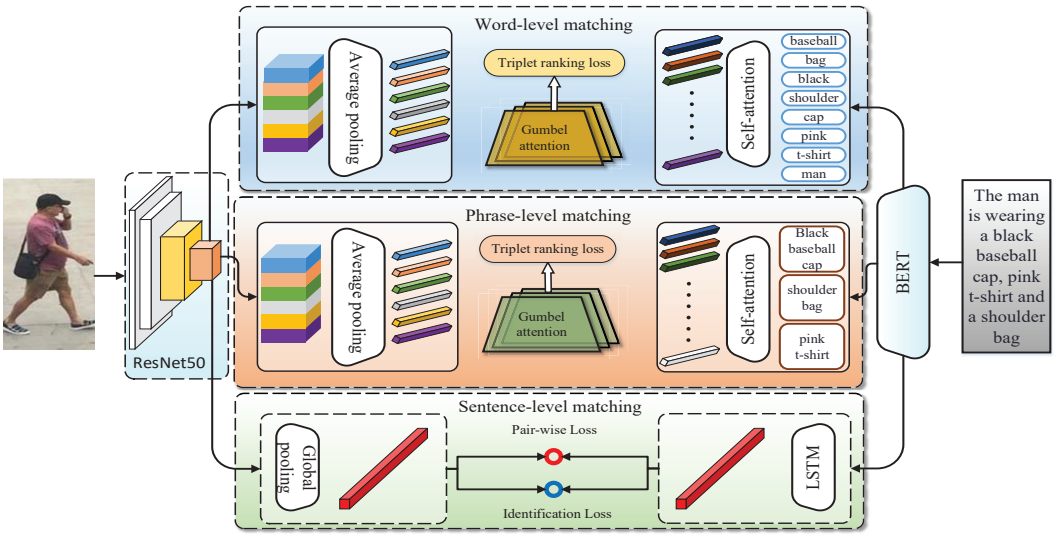

Figure 2: The overall architecture of the proposed hierarchical Gumbel attention network. It consists of a ResNet for extracting global and local visual features, a multi-level textual network based BERT for learning hierarchical textual features as well as a Gumbel attention module for discovering the strong latent alignments using both image regions and words in a sentence towards learning fine-grained textual and visual representations.

the regions of images and the corresponding words in sentences, which are not effective and easy to cause matching redundancy.

Attention mechanism Soft-attention mechanism [8, 17, 29] has gained great attentions in recent years, which is used for natural language tasks, image captioning and visual question answering. However, soft attention still carry some redundant information, even if the value is small. In this case, hard attention should been used for filtering the redundant keys.For the hard attention mechanis, its optimization is not differentiable. For solving this issue, some methods adopt REINFORCE [31] algorithm as the gradient estimator, which is unbiased, high variance in general. Xu et al. [33] use REINFORCE algorithm to guide soft attention to verge on hard attention. Zaremba et al. [35] adopt curriculum learning to make soft attention gradually become discrete. In order to achieve an unbiased gradient estimator, there are also some tricks to implement hard attention mechanism. It is mentioned that the widely used re-parameterization trick [7] named Gumbel re-parameterization is adopted to create a low-variance, unbiased gradient estimator. This gradient estimator [7] enables the variational auto-encoding model to generate more disentangled representations. In the translation task [11], this Gumbel trick can guide th modal to obtain a variety of high-quality translations. We design a Gumbel top-k re-parameterization algorithm to implement a hard attention mechanism for fine-grained image-text retrieval tasks.

3 HIERARCHICAL GUMBEL ATTENTION NETWORK

Figure 2 illustrates the overview of the proposed hierarchical Gumbel attention network (HGAN) model which consists of word-level matching, phrase-level matching and sentence-level matching. In this section, we firstly present the overall architecture of the proposed approach and then introduce each component in the following subsections.

3.1 Overall Architecture

Let $\boldsymbol{\mathcal{X}} = \{\boldsymbol{x}_i, \boldsymbol{t}_i\}_{i=1}^{N}$ be a training set of N image-text pairs which contain N image samples $\boldsymbol{x_i}$ and the corresponding textual descriptions $\boldsymbol{t_i}$ captured from K pedestrians whose IDs are $Y = \{y_i\}_{i=1}^{N}$. Given a query of textual description $\boldsymbol{t}$, the goal is to identify the most relevant pedestrian's images from the image gallery. In this work, we propose a novel hierarchical Gumbel attention matching model for text-based person search via Gumbel top-k re-parameterization algorithm, which explores hierarchically precise alignments between images and texts of pedestrians towards learning discriminating fine-grained textual-visual representations. As shown in Figure 2, our hierarchical Gumbel matching model consists of three different angularities adaptive matching model, i.e., word-level, phrase-level, and sentence-level, via the guidance of Gumbel attention mechanism. Specifically, to capture the subtle visual variations, we adopt the ResNet [5] with a partitioning strategy to extract the visual features of an image which are divided into three parts: one global context feature for sentence matching and two partial features for phrase matching and word matching, respectively. As for the textual representation extraction, BERT [4] is used to extract the sequences of word embedding. Then, we feed the sequences of word embedding into a bidirectional long short-term memory network (bi-LSTM) [38] to effectively summarize the content of input textual description as the sentence-level representation. To exploit the fine-grained textual descriptions we use the Natural Language ToolKit (NLTK) [21] to obtain several noun phrases. We feed the word embeddings and phrase embeddings into self-attention module [29] for encoding the textual representations. After obtaining

hierarchical textual embedding, the sentence-level module is used to match the global context of images and corresponding sentence description. Then, a top-k Gumbel attention module is designed to adaptively select a subset of image regions or words/phrases for similarity calculation towards alleviating matching redundancy problem and exploit the multi-level semantic relevance. Due to the discrete nature of the Gumbel attention module, we design the top-k Gumbel re-parameterization algorithm as a low-variance, unbiased gradient estimator to iteratively train this module.

3.2 Learning Hierarchical Representation

Learning Visual Representation Images also contain multiple aspects such as shoes, clothes and pants. However, it is time-consuming to parse image into semantic structures which require spatial segmentation or object detection. These extra pre-processing technologies are too time-consuming to apply into the real-time system. We thus adopt the feature partitioning strategy [27, 28] to focus on different levels of aspects in the image. This kind of partitioning strategy do not involve extra computing cost and is able to encode the local semantic information of images. As shown in Figure 2, we first adopt the convolutional neural networks, i.e., ResNet-50 model, to extract the visual features $\boldsymbol{V} \in \mathbb{R}^{H\times W\times C}$ before the last pooling layer. Then, we employ different convolutioal layers and mean pooling on respective image parts as multi-scale partitioning methods to obtain multi-level partial visual representations. It reduces the dimension of visual features from $H \times W$ into $N_v = H' \times W'$. A following 1×1 kernel-sized convolutional layer reduces the dimension of C into C'. It means that we adopt different weight to encode the visual feature into two level of embeddings. Finally, we obtain the two level of embeddings with the dimension of C':

$$\begin{aligned} \boldsymbol{V}_w &= [\boldsymbol{v}_1^w, \ldots, \boldsymbol{v}_{N_v}^w] = GAP_1(\boldsymbol{V}) \in \mathbb{R}^{N_v\times C'}, \\ \boldsymbol{V}_p &= [\boldsymbol{v}_1^p, \ldots, \boldsymbol{v}_{N_v}^p] = GAP_2(\boldsymbol{V}) \in \mathbb{R}^{N_v\times C'}, \end{aligned} \tag{1}$$

where GAP denotes global average pooling, $\boldsymbol{v} \in \mathbb{R}^{C'}$, there is $N_v = H'W'$ embeddings of $\boldsymbol{V}_w$ and $\boldsymbol{V}_p$ respectively. We feed the word embeddings and phrase embeddings into one transform layer [29] for encoding the textual representations.

As the global visual representation is defined as follows: $\boldsymbol{V}_g = \text{avgpool}(\boldsymbol{V}) \in \mathbb{R}^{1\times C}$, where avgpool denotes global average pooling along the first and second dimension. This partition method that extracts local features brings almost no extra computational cost, which is better than the pose-based or region-based approaches. In addition, visual partitioning strategy can be done along with the end-to-end training procedure and enables our method to adaptively obtain more disentangled visual features.

Learning Textual Representation Given a textual description T', we feed the whole words into a pretrained BERT model to extract the sequence of textual embeddings. Then we use the self-attention module [29] to extract word-level textual embeddings $\boldsymbol{T}_w = \left[\boldsymbol{t}_1^w, \ldots, \boldsymbol{t}_R^w\right]$. In order to encode the dependencies between adjacent words, we adopt bi-LSTM to handle the word-level textual embedding vectors. The global textual representation $\boldsymbol{T}_g$ is defined as the concatenation of the last hidden states $\boldsymbol{T}_g = \text{concat}(\overrightarrow{\boldsymbol{h}_r}, \overleftarrow{\boldsymbol{h}_1}) \in \mathbb{R}^{1\times C}$. To exploit phrase-level semantic relevance between visual contents and the corresponding textual description, we utilize the NLTK to extract the noun phrase N from a sentence of human description. After acquiring M noun phrases $\boldsymbol{N} = [\boldsymbol{n}_1, \boldsymbol{n}_2, \ldots, \boldsymbol{n}_M]$, we can also use one layer of self-attention [29] to obtain the representations of all noun phrases: $\boldsymbol{T}_p = \left[\boldsymbol{t}_1^p, \ldots, \boldsymbol{t}_M^p\right]$.

3.3 Gumbel Attention Module

Given visual and textual features, our goal is to design a model, which hierarchically maps these features into a common embedding space to infer the similarity of image-text pair. Thus, a Gumbel attention module is proposed to discover the strong latent alignments via Gumbel top-k sampling towards calculating the similarity score between textual and visual representations

The word-level Gumbel attention module accepts two inputs: a set of partial visual features $\boldsymbol{V}_w = [\boldsymbol{v}_1^w, \boldsymbol{v}_2^w, \ldots, \boldsymbol{v}_{N_v}^w]$ and a set of word-level features $\boldsymbol{T}_w = \left[\boldsymbol{t}_1^w, \boldsymbol{t}_2^w, \ldots, \boldsymbol{t}_R^w\right]$. The output is two attention weights between these image-text pairs. Firstly, the whole word-level features $\boldsymbol{T}_w \in \mathbb{R}^{R\times C'}$ and visual feature $\boldsymbol{V}_w \in \mathbb{R}^{N_v\times C'}$ are fed into a multi-layer perceptron (MLP) to calculate the similarity map:

$$\boldsymbol{\gamma} = MLP(\boldsymbol{T}_w, \boldsymbol{V}_w), \tag{2}$$

where $\boldsymbol{\gamma} \in \mathbb{R}^{R\times N_v}$ represents the respective similarity score map between the N_v image regions and the R words, MLP denotes to the multilayer perceptron.

Stacked soft attention Previous soft attention methods attend all the words to the image or attend all the image regions to the text for computing the similarity:

$$\boldsymbol{V}_{wa} = softmax(\boldsymbol{\gamma}) \cdot \boldsymbol{V}_w, \ \boldsymbol{T}_{wa} = softmax(\boldsymbol{\gamma}^T) \cdot \boldsymbol{T}_w, \tag{3}$$

where the softmax operation is executed along with the second dimension, $\boldsymbol{V}_{wa} = \{\boldsymbol{v}_i^a\}_{i=1}^R \in \mathbb{R}^{R\times C'}$ and $\boldsymbol{T}_{wa} = \{\boldsymbol{t}_i^a\}_{i=1}^R \in \mathbb{R}^{H'W'\times C'}$ denote to the attention-reinforced visual and textual embeddings, respectively. Then we use the attention-reinforced embeddings to calculate the similarity:

$$s_{i2t} = \sum_{i=0}^{R}(MLP(\boldsymbol{v}_i^a + \boldsymbol{t}_i)), \ s_{t2i} = \sum_{i=0}^{N_v}(MLP(\boldsymbol{t}_i^a + \boldsymbol{v}_i)), \tag{4}$$

where s_{i2t} and s_{t2i} denote to the Image-Text similarity score and Text-Image similarity score, respectively. This soft attention may bring matching redundancy problem as some irrelevant words are fused into the image region or some irrelevant image regions are fused into the word. Thus, the Gumbel top-k sampling method within the model is designed, which adaptively selects a subset of words corresponding to the certain image region for similarity evaluation. It enables to alleviate matching redundancy problem. We also define two complimentary formulations of Gumbel top-k Attention below: Image-Text and Text-Image.

Gumbel Top-k Sampling The input of our Gumbel top-k attention module are a sequence of visual features $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n]$ and a sequence of textual features $\boldsymbol{T} = [\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_m]$. We also need the E as the temperature coefficient of Gumbel softmax, and k as the number of selection of gumbel top-k module. This module outputs image-to-text attention weights $\boldsymbol{\alpha} = \{\boldsymbol{\alpha}_i\}_{i=1}^m$, where $\boldsymbol{\alpha}_i = \{\alpha_{i,j}\}_{j=1}^n, k_1 = \sum_{j=1}^n \alpha_{i,j}, \alpha_{i,j} \in \{0, 1\}$, and text-to-image attention

Algorithm 1 Gumbel Top-k Attention

Input: a sequence of visual features $\boldsymbol{V} = [\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_n}]$ and a sequence of textual features $\boldsymbol{T} = [\boldsymbol{t_1}, \boldsymbol{t_2}, \ldots, \boldsymbol{t_m}]$.
E is the temperature coefficient of Gumbel softmax.
k_1 and k_2 is the k of gumbel top-k module.
Output: image-to-text attention weights $\boldsymbol{\alpha} = \{\boldsymbol{\alpha}_i\}_{i=1}^{m}$, where $\boldsymbol{\alpha}_i = \{\alpha_{i,j}\}_{j=1}^{n}$, $k_1 = \sum_{j=1}^{n} \alpha_{i,j}$, $\alpha_{i,j} \in \{0, 1\}$,
and text-to-image attention weights $\boldsymbol{\beta} = \{\boldsymbol{\beta}_i\}_{i=1}^{n}$, where $\boldsymbol{\beta}_i = \{\beta_{i,j}\}_{j=1}^{m}$, $k_2 = \sum_{j=1}^{m} \beta_{i,j}$, $\beta_{i,j} \in \{0, 1\}$

Calculating the similarity score map:
for $i \leftarrow 1$ *to* m **do**
  for $j \leftarrow 1$ *to* n **do**
    $\gamma_{i,j} = MLP(\boldsymbol{v_i}, \boldsymbol{t_j})$,
$\boldsymbol{\gamma} \in \mathbb{R}^{m \times n}$ denotes to the local similarity score map.
Calculating image-to-text attention weights:
for $i \leftarrow 1$ *to* m **do**
  $\boldsymbol{\gamma}_i^{'} \leftarrow l2norm(\boldsymbol{\gamma}_{i,:})$
  $\boldsymbol{u}_i \leftarrow Uniform(0, 1)$ # Sample from uniform distribution
  $\boldsymbol{r}_i \leftarrow -log(-log(\boldsymbol{u}_i)) + \boldsymbol{\gamma}_i^{'}$ # Add Gumbel noise to similarity
  $\boldsymbol{r}_i \leftarrow softmax(\boldsymbol{r}_i, E)$
  $\boldsymbol{\alpha}_i \leftarrow Top\_K(\boldsymbol{r}_i, k_1)$
Calculating text-to-image attention weights:
for $j \leftarrow 1$ *to* n **do**
  $\boldsymbol{\gamma}_j^{'} \leftarrow l2norm(\boldsymbol{\gamma}_{:,j})$
  $\boldsymbol{u}_j \leftarrow Uniform(0, 1)$ # Sample from uniform distribution
  $\boldsymbol{r}_j \leftarrow -log(-log(\boldsymbol{u}_j)) + \boldsymbol{\gamma}_j^{'}$ # Add Gumbel noise to similarity
  $\boldsymbol{r}_j \leftarrow softmax(\boldsymbol{r}_j, E)$
  $\boldsymbol{\beta}_j \leftarrow Top\_K(\boldsymbol{r}_j, k_2)$
return $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$

weights $\boldsymbol{\beta} = \{\boldsymbol{\beta}_i\}_{i=1}^{n}$, where $\boldsymbol{\beta}_i = \{\beta_{i,j}\}_{j=1}^{m}, k_2 = \sum_{j=1}^{m} \beta_{i,j}, \beta_{i,j} \in \{0, 1\}$. Algorithm 1 illustrates the detailed procedure of our Gumbel top-k attention mechanism.

Firstly, we calculate the local similarity score map $\boldsymbol{\gamma}$ between $\boldsymbol{V}$ and $\boldsymbol{T}$ as Algorithm 1 describes. It is difficult to select the top-k strong alignments from the similarity score map $\boldsymbol{\gamma}$. Thus, we introduce the Gumbel trick that finds the top-k element by perturbing the similarity $\boldsymbol{\gamma}$ with Gumbel noise towards implementing hard attention mechanism. Formally, let Gumbel noise $G_i \sim$ Gumbel $(0) = 0 - \log(-\log U_i), i \in \{0, 1, ..., n\}$ $i.i.d., U_i = Uniform(0, 1)$, and let $r_i = \gamma_i + G_i$ as a re-parameterizing sample with Gumbel noise. The idea is to re-parameterize the sample as a transformation of the parameters $\boldsymbol{\gamma}$ and some independent Gumbel noise G_i. Then by relaxing the transformation (from max to softmax), the Gumbel-softmax trick allows for training with backpropagation [7].

We should achieve k-hot attention weights from the re-parameterizing sampless r. For $\boldsymbol{r}_i$ in the image-to-text attention, $l_1^*, \ldots, l_k^* =$ Top_K $(r_i, k1)$. Then $l_1^*, \ldots, l_k^*$ are indices of ordered sample sequence sampled from the Gumbel random keys $\boldsymbol{r}_i$ in order of decreasing value. Let $\boldsymbol{e} = [0, \ldots, 0, 1, 0, \ldots, 0] \in \{0, 1\}^n$ be a 1-hot vector, i.e., a vector with only one nonzero element at index j, where $e_j = 1$. Thus, the attention weights of Gumbel top-k are:

$$\boldsymbol{\alpha} = \{\boldsymbol{\alpha}_i\}_{i=1}^{m}, \boldsymbol{\alpha}_i = \sum_{j \in [1, k_1]} e_{l_j^*}, \tag{5}$$

After the Gumbel top-k sampling, we use the set of k-hot vectors $\boldsymbol{\alpha}_i$ to calculate k textual embeddings with k largest scores:

$$\boldsymbol{T_i'} = \boldsymbol{\alpha}_i^T \times \boldsymbol{T} = [0, \boldsymbol{t_1'}, 0, \ldots, \boldsymbol{t_k'}, 0, 0] \in \mathbb{R}^{n \times C}, \tag{6}$$

$\boldsymbol{T_i'}$ represents the k textual embeddings with k largest scores. We also can obtain the top-k soft attention map: $softmax(\boldsymbol{\alpha} \odot \boldsymbol{\gamma})$.

To attend on these k words with respect to the corresponding image region, we exploit a weighted fusion of these words:

$$\boldsymbol{t_i^f} = softmax(\boldsymbol{\alpha} \odot \boldsymbol{\gamma}) \times \boldsymbol{T_i'}, \tag{7}$$

To determine the importance of each image region given the sentence context, we define a MLP layer to evaluate the relevance between the attended sentence vector $\boldsymbol{t_i^f}$ and each image region feature $\boldsymbol{v_i}$. Finally, the similarity between image $\boldsymbol{V}$ and sentence $\boldsymbol{T}$ is calculated by average pooling,

$$S(\boldsymbol{V}, \boldsymbol{T}) = \frac{\sum_{i=1}^{n} \sigma(MLP(\boldsymbol{v_i} + \boldsymbol{t_i^f}))}{n} \tag{8}$$

3.4 Alignment Objective

Triplet loss is a common ranking objective for image-text matching. We also employ a hinge-based triplet ranking loss with margin for word-level and phrase-level matching, i.e.

$$\begin{aligned} \mathcal{L}_w = &\sum_{\hat{T}} [\gamma - S(\boldsymbol{V_w}, \boldsymbol{T_w}) + S(\boldsymbol{V_w}, \hat{\boldsymbol{T}}_{\boldsymbol{w}})]_+ \\ &+ \sum_{\hat{V}} [\gamma - S(\boldsymbol{V_w}, \boldsymbol{T_w}) + S(\hat{\boldsymbol{V}}_{\boldsymbol{w}}, \boldsymbol{T_w})]_+ \end{aligned} \tag{9}$$

where $[x]_+ \equiv \max(x, 0)$ and S is a similarity score function. The first sum term represents that taking overall negative sentences $\boldsymbol{T_w}$ given an image $\boldsymbol{V_w}$; the second sum considers all negative images $\boldsymbol{V_w}$ given a sentence $\boldsymbol{T_w}$. $\mathcal{L}_p$ is similar to the $\mathcal{L}_w$.

As for the sentence-level matching, we adopt the identification loss and pair-wise loss. The identification loss enables the visual and textual representations to be discriminative in the common feature space with the guidance of person IDs. The pairwise loss pushes the same person ID together and projects the representations of languages and images into the shared feature space. The identification loss is formulated as follows:

$$\mathcal{L}_{ide} = -\frac{1}{N} \sum_{i=1}^{N} \log softmax(\boldsymbol{W}_{y_i}^T \boldsymbol{x}_i + \boldsymbol{b}) \tag{10}$$

where y_i represents the person ID of the i-th sample, N is batch size $\boldsymbol{x_i}$ refers to image embedding $\boldsymbol{v}$ or text embedding $\boldsymbol{t}$. The pair-wise loss is formulated as follows:

$$p_{ij} = \frac{e^{z_{ij}}}{\sum_{j=1}^{2} e^{z_{ij}}}, \mathcal{L}_{pair} = -\sum_{i=1}^{M} \sum_{j=1}^{M} p_{ij} \log \frac{p_{ij}}{q_{ij} + \epsilon} \tag{11}$$

where $\boldsymbol{z_{ij}}$ refers to the fused feature by combining the visual feature $\boldsymbol{v_i}$ and textual feature $\boldsymbol{t_j}$, ϵ is set to a small number for preventing division by zero. The KL divergence is used to make the true matching distribution q_i to match the predicted probability p_i. The total loss for the sentence-level matching model is $\mathcal{L}_{sentence} = \mathcal{L}_{pair} + \lambda \mathcal{L}_{ide}$.

4 EXPERIMENTS

In this section, we conduct extensive experiments to evaluate the performance of the proposed Gumbel hierarchical matching model (HGAN) on the text-based person search dataset (CUHK-PEDES) compared with state-of-the-art methods. To evaluate our method more generic and convincing, we also conduct some experiments on two other fine-grained cross-modality retrieval datasets, *i.e.*, CUB and Flowers. Moreover, we investigate the effectiveness of each component of HGAN, including the word-level matching model, phrase-level matching model, sentence-level matching model and the Gumbel attention module.

Datasets. We mainly evaluate our method on the text-based person search dataset named CUHK-PEDES [14]. This dataset contains totally 40,206 pedestrian images with 13,003 identities. We adopt the office split on this dataset which is split into training set, validation set, and test set without overlapping person IDs. Specifically, the training set contains 34,054 images labeled 11,003 pedestrian identities with 68,108 sentence descriptions. The validation set contains 3,078 images labeled 1000 identities and the test set contains 3,074 images labeled 1,000 identities, respectively. The Caltech-UCSD Birds (CUB) dataset contains 11,788 bird images which have 200 fine-grained categories. There are 10 visual descriptions within each bird image. This dataset is split into 100 training, 50 validation and 50 test categories. The Oxford-102 Flowers (Flowers) dataset contains 8,189 flower images that possess 102 fine-grained categories, where each image is labeled with 10 textual descriptions. The office split contains 62 categories for training, 20 categories for validation and 20 categories for testing.

Implementation Details. The implementation of the proposed method is based on the Tensorflow framework with four Tesla P40 GPUs. We adopt adam optimization algorithm with a learning rate lr of 0.0002 to train our model, in which the weight decay is set to $5e^{-4}$. All the images are resized to $384 \times 192 \times 3$ resolution and are normalised with $1.0/256$. The mini-batches is set to 32 for the training phase. We employ the ResNet-50 as the visual extractor in our experiments. The embedding size of BERT is set to 768, the embedding size of Bi-LSTM is set to 512, and the dimensions of the FC layers used in multi-level matching modal are set to 512.

Evaluation Metric. We adopt different performance metric for each dataset. For the CUHK-PEDES datasets, Rank-1, Rank-5, and Rank-10 accuracy are adopted as our evaluation metric. For the CUB and Flowers datasets, the AP@50 metric is adopted for text-to-image retrieval and rank-1 for image-to-text retrieval. Given a textual query, the model first calculates the percentage of the first 50 retrieved images that match the category of a text query. Then, the average percentage of matching for all the test categories is expressed as *AP*@50.

4.1 Comparsion to State-of-the-Art Methods

CUHK-PEDES:. Table 1 shows the performance comparison of the proposed HGAN against the state-of-the-art methods in terms of top-k accuracy in the text-to-imgae retrieval task. The compared methods include CNN-RNN [23], Neural Talk [30], GNA-RNN [14], IATVM [13], PWM-ATH [3], GLA [2], CAN [8], CMPM-CMPC [39], A-GANet [17], PMA [9] and TIMAM [26]. The proposed HGAN achieves 57.44%, 77.96%, 85.51% of rank-1, rank-5, rank-10 accuracy, respectively. We can see that our method surpasses existing methods, which demonstrates the effectiveness of the proposed method. HGAN adopt the partitioning strategy to capture more effective fine-grained visual-textual information, which outperforms the TIMAM [26] and CMPM-CMPC [39] that ignore these detailed information. Moreover, HGAN achieves significant performance improvement as compared to the network based attention mechanism methods [9, 17], which indicates that the proposed HGAN is able to select better strong alignments between relevant image regions and words/phrases. Some retrieval visualized results are illustrated in Figure 3.

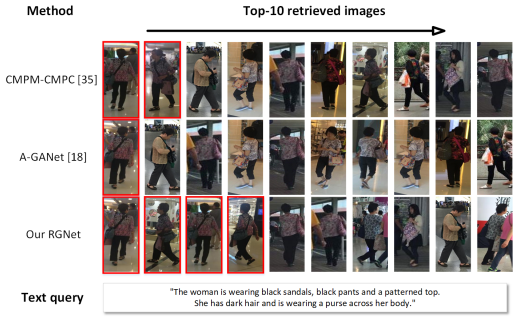

Figure 3: Examples of top-10 retrieved images from textual description by the proposed HGAN. Red boxes represent the corresponding images of the given textual description.

4.2 Ablation Studies

To demonstrate the effectiveness and contribution of each component of the HGAN, we conduct a series of ablation experiments on the CUHK-PEDES dataset. We evaluate the effectiveness of the BERT, the hierarchical matching module and Gumbel attention module, and compare their performance for text-based person search.

Impact of BERT:. Table 2 reports the performance comparison between the LA-Net models using different hidden layers within BERT as the word embedding. From the results, we can see that concatenating the last four hidden layers of BERT as the textual embedding achieves a better rank-1 accuracy above 54.2% in the final text-to-image matching results, rather than using last hidden layer or second last hidden layer. Compared with the other hidden layers, using the last four hidden layers in BERT is more universal for transferring to other tasks.

Impact of Proposed Components: Table 3 shows the effectiveness of each proposed component, i.e., sentence-level matching model, word-level matching model, and phrase-level matching model. We observe that the combination of losses ($\mathcal{L}_{ide}$ and $\mathcal{L}_{pair}$) in the sentence-level matching model obtains larger performance gain than using these losses individually. By sensibly adopting BERT, better textual descriptions enable to increase the accuracy to 54.2%. The word-level matching model can bring a clear performance gain that increases the rank-1 accuracy to 56.2%. Finally, the combination of hierarchical matching models obtain the 57.4% accuracy, which indicates that the hierarchical matching models with the guidance

Table 1: Performance comparison to the state-of-the-art methods on the CUHK-PEDES dataset.

Method	Publication	Text-to-Image				Image-To-Text			
		Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
CNN-RNN [23]	CVPR'2016	8.07	-	32.47	-	-	-	-	-
Neural Talk [30]	CVPR'2015	13.66	-	41.72	-	-	-	-	-
GNA-RNN [14]	CVPR'2017	19.05	-	53.64	-	-	-	-	-
IATVM [13]	ICCV'2017	25.94	-	60.48	-	-	-	-	-
PWM-ATH [3]	WACV'2018	27.14	49.45	61.02	-	-	-	-	-
GLA [2]	ECCV'2018	43.58	66.93	76.26	-	-	-	-	-
CAN [8]	arXiv'2018	45.52	67.12	76.98	-	-	-	-	-
CMPM-CMPC [39]	ECCV'2018	49.37	71.69	79.27	31.37	60.96	84.42	90.83	30.26
A-GANet [17]	ACMMM'2019	53.14	74.03	81.95	-	-	-	-	-
PMA [9]	AAAI'2020	54.12	75.45	82.97	-	-	-	-	-
TIMAM [26]	ICCV'2019	54.51	77.56	84.78	35.13	67.40	88.65	93.91	34.43
HGAN	-	**59.00**	**79.49**	**86.62**	**37.80**	**71.16**	**90.05**	**95.06**	**36.60**

Table 2: Evaluation of the effectiveness of using different hidden layers within BERT module as textual embedding on the CUHK-PEDES dataset.

Layers of BERT	Text-to-Image		
	Rank-1	Rank-5	Rank-10
Last Hidden	53.8	76.1	83.2
Second-to-last Hidden	54.3	77.1	84.3
Concat Last Four Hidden	**55.2**	**77.3**	**84.9**

Table 3: Ablation studies on the CUHK-PEDES dataset to evaluate the effectiveness of proposed components in terms of rank-1 and rank-10 accuracy.

$\mathcal{L}_{ide}$	$\mathcal{L}_{pair}$	BERT	word	phrase	R@1	R@10
✓					43.5	74.1
	✓				47.1	79.8
✓	✓				51.8	82.5
✓	✓	✓			55.2	84.9
✓	✓	✓	✓		58.3	85.7
✓	✓	✓	✓	✓	59.0	86.6

Table 4: Evaluation of the effectiveness of different visual partitioning strategy.

Pooling Layer		Number of Region (Size of Region)	Rank-1	Rank-5	Rank-10
kernel	stride				
2×2	2×2	32 (8×4)	54.29	76.15	83.67
4×4	2×2	21 (7×3)	57.41	78.07	85.09
4×4	3×2	15 (5×3)	**58.34**	79.08	85.82
4×4	4×4	8 (4×2)	57.24	78.74	85.56
4×8	2×2	7 (7×1)	58.03	78.74	**85.88**
4×8	3×2	5 (5×1)	58.14	**79.84**	**85.88**

of Gumbel module select the strong semantically relevant image regions and words/phrases from images and texts towards learning discriminative fine-grained textual-visual representations.

Impact of different visual partitioning strategy: We also exploit how the number of selected regions affects the performance of our HGAN. In experiments, we select different pooling layers for generating the different local visual features as the visual partitioning strategy. Table 4 shows the experimental results. We can see that the performance is best when the kernel is set to 4×4 and the stride is set to 3×2 in the pooling layer. When the number of region is 32, there's a lot of performance degradation. Because the overmuch partitionings in the visual feature lead to the too fine grained features, which is hard to include complete semantic information. And the performances of other settings vary little.

Visualization of the Gumbel top-k attention regions on several examples: To verify whether the proposed model can selectively attend to the corresponding regions and make our matching procedure more interpretable, we visualize the attention weights of our Gumbel top-k attention module. Fig 4 shows that this module selects different words according to each red visual region and gives these selected words attention weights. We can see that the Gumbel top-k attention module indeed selects the described words about the visual local regions and filters out the unrelated words, which illustrates our model can learn accurate aligned fine-grained matching. Specifically, for the foot part, the model mainly attends to the words which describe the "shoes". This fine-grained matching benefits the inference of image-text similarity.

Table 5: Evaluation of the effectiveness of each component within the Gumbel module on the CUHK-PEDES dataset. "HGAN w/o Gumbel" refers to that the gumbel attention module degrades to soft-attention module.

Method	Text-to-Image		
	Rank-1	Rank-5	Rank-10
HGAN w/o Gumbel	56.1	78.1	85.5
HGAN w Max Gumbel	57.6	78.5	86.0
HGAN	**59.0**	**79.5**	**86.6**

Table 5 compares each components of the Gumbel attention module. HGAN w/o Gumbel refers to HGAN without Gumbel top-k sampling which only exploits the hierarchical matching models

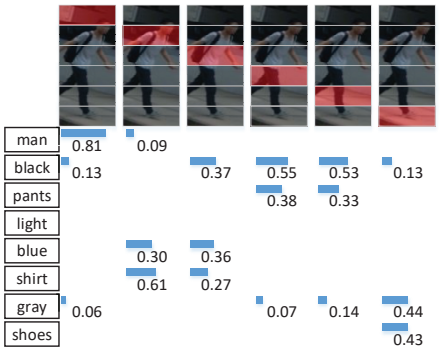

Figure 4: Visualization of weights in Gumbel attention module. This example shows that this module selects different words according to each red visual region and gives these selected words different attention weights.

with soft-attention. HGAN w/o reward refers to HGAN without reward function, and only use the Gumbel top-k sampling attention module. From Table 5, HGAN w/o Gumbel and HGAN w/o reward obtain 55.1% and 57.0% rank-1 accuracy, respectively. The performance improvement of HGAN over HGAN w/o Gumbel and HGAN w/o reward by 2.3% and 0.4% rank-1 accuracy respectively, indicates that the Gumbel attention mechanism can adaptively select a subset of image regions or words/phrases for similarity evaluation to alleviate the matching redundancy problem. We conduct experiment to evaluate the impact of k_1 and k_2 of the proposed Gumbel attention module in the word-level matching. The results are shown in Figure 5, we can see that when $k_1 = 4$ and $k_2 = 2$, our HGAN yields the best retrieval performance.

The k₁ of Gumbel attention module in the text

Rank-1 (k₂ \ k₁)	1	2	3	4	5
1	57.64	57.80	57.52	58.43	57.33
2	57.62	58.11	58.19	**58.97**	58.23
3	57.28	57.62	58.25	58.61	58.29
4	57.44	57.89	58.12	57.71	58.04
5	56.90	57.01	57.63	58.21	57.73

The k₂ of Gumbel attention module in the image

Figure 5: Evaluation of the proposed Gumbel attention module with different values of parameters k_1 and k_2.

4.3 Generalization

CUB and Flowers Datasets: We compare our proposed HGAN to the nine state-of-the-art methods on these two datasets and present our performance in table 6. The compared methods include Word2Vec [22], HGLMM [10], Word CNN [23], Word CNN-RNN [23], Attributes [13], IATV [13], CMPM-CMPC [39] and TIMAM [26]. Our method achieves consistent performance improvements results in terms of both image-to-text and text-to-image matching performance in both datasets as compared to the state-of-the-art methods. We observe that the performance increases 1.2% and 1.5% in terms of rank-1 accuracy as well as 1.8% and 1.4% in terms of *AP*@50.

Table 6: Performance comparison to the state-of-the-art methods on the CUB and Flowers datasets.

Method	CUB		Flowers	
	I2T	T2I	I2T	T2I
	Rank-1	AP@50	Rank-1	AP@50
Word2Vec[22]	38.6	33.5	54.2	52.1
HGLMM [10]	36.5	35.6	54.8	52.8
Word CNN [23]	51.0	43.3	60.7	56.3
Word CNN-RNN [23]	56.8	48.7	65.6	59.6
Attributes[1]	50.4	50.0	-	-
Triplet [13]	52.5	52.4	64.3	64.9
IATV [13]	61.5	57.6	68.9	69.7
CMPM-CMPC [39]	64.3	67.9	68.4	70.1
TIMAM [26]	67.7	70.3	70.6	73.7
HGAN	**68.9**	**71.7**	**72.4**	**75.1**

5 CONCLUSION

In this work, we propose a hierarchical adaptive matching model for text-based person search via Gumbel sampling, which discovers the strong latent alignments using both image regions and words in a sentence towards learning fine-grained textual and visual representations. Specifically, a Gumbel attention module within the model is designed, which adaptively selects a subset of image regions or words/phrases for similarity evaluation to alleviate matching redundancy problem. Due to the discrete nature of the Gumbel subset sampling module, we adopt the Gumbel top-k re-parameterization trick as a low-variance, unbiased gradient estimators to iteratively train the module. Moreover, a hierarchical adaptive matching strategy is employed by the model, which exploits the multi-level semantic relevance between the natural language descriptions of persons and corresponding visual content from three different granularities, i.e., word-level, phrase-level, and sentence-level. Extensive experimental results on three challenging benchmarks have demonstrated the effectiveness of the proposed method. In the future work, we will try to use the Gumbel attention in the more general image-text retrieval tasks.

ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China under Grant 2017YFB1300201, the National Natural Science Foundation of China (NSFC) under Grants U19B2038 and 61620106009.

REFERENCES

[1] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. 2015. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2927–2936.

[2] Dapeng Chen, Hongsheng Li, Xihui Liu, Yantao Shen, Jing Shao, Zejian Yuan, and Xiaogang Wang. 2018. Improving deep visual representation for person re-identification by global and local image-language association. In *Proceedings of the European Conference on Computer Vision*. 54–70.

[3] Tianlang Chen, Chenliang Xu, and Jiebo Luo. 2018. Improving Text-Based Person Search by Spatial Matching and Adaptive Threshold. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*. 1879–1887.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[6] Lingxiao He, Yinggang Wang, Wu Liu, He Zhao, Zhenan Sun, and Jiashi Feng. 2019. Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*. 8450–8459.

[7] Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. *ICLR* (2017).

[8] Ya Jing, Chenyang Si, Junbo Wang, Wei Wang, Liang Wang, and Tieniu Tan. 2018. Cascade Attention Network for Person Search: Both Image and Text-Image Similarity Selection. *arXiv preprint arXiv:1809.08440* (2018).

[9] Ya Jing, Chenyang Si, Junbo Wang, Wei Wang, Liang Wang, and Tieniu Tan. 2020. Pose-Guided Joint Global and Attentive Local Matching Network for Text-Based Person Search. In *AAAI*.

[10] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. 2015. Associating neural word embeddings with deep image representations using fisher vectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4437–4446.

[11] Wouter Kool, Herke Van Hoof, and Max Welling. 2019. Stochastic Beams and Where to Find Them: The Gumbel-Top-k Trick for Sampling Sequences Without Replacement. *ICML* (2019).

[12] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision*. 201–216.

[13] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang. 2017. Identity-aware textual-visual matching with latent co-attention. In *Proceedings of the IEEE International Conference on Computer Vision*. 1890–1899.

[14] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. 2017. Person search with natural language description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1970–1979.

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*. 740–755.

[16] Jiawei Liu, Zheng-Jun Zha, Xuejin Chen, Zilei Wang, and Yongdong Zhang. 2019. Dense 3D-convolutional neural network for person re-identification in videos. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 15, 1s (2019), 1–19.

[17] Jiawei Liu, Zheng-Jun Zha, Richang Hong, Meng Wang, and Yongdong Zhang. 2019. Deep Adversarial Graph Attention Convolution Network for Text-Based Person Search. In *Proceedings of the ACM International Conference on Multimedia*. 665–673.

[18] Jiawei Liu, Zheng-Jun Zha, QI Tian, Dong Liu, Ting Yao, Qiang Ling, and Tao Mei. 2016. Multi-scale triplet cnn for person re-identification. In *Proceedings of the 24th ACM international conference on Multimedia*. 192–196.

[19] Wu Liu, Tao Mei, Yongdong Zhang, Jintao Li, and Shipeng Li. 2013. Listen, look, and gotcha: instant video search with mobile phones by layered audio-video indexing. In *Proceedings of the ACM International Conference on Multimedia*. 887–896.

[20] Yu Liu, Yanming Guo, Erwin M Bakker, and Michael S Lew. 2017. Learning a recurrent residual fusion network for multimodal matching. In *Proceedings of the IEEE International Conference on Computer Vision*. 4107–4116.

[21] Edward Loper and Steven Bird. 2002. NLTK: the natural language toolkit. *arXiv preprint cs/0205028* (2002).

[22] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[23] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. 2016. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 49–58.

[24] Weijian Ruan, Wu Liu, Qian Bao, Jun Chen, Yuhao Cheng, and Tao Mei. 2019. Poinet: pose-guided ovonic insight network for multi-person pose tracking. In *Proceedings of the 27th ACM International Conference on Multimedia*. 284–292.

[25] M Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. 2018. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 420–429.

[26] Nikolaos Sarafianos, Xiang Xu, and Ioannis A Kakadiaris. 2019. Adversarial Representation Learning for Text-to-Image Matching. In *Proceedings of the IEEE International Conference on Computer Vision*. 5814–5824.

[27] Yifan Sun, Qin Xu, Yali Li, Chi Zhang, Yikang Li, Shengjin Wang, and Jian Sun. 2019. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 393–402.

[28] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision*. 480–496.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[30] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3156–3164.

[31] Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* (1992), 229–256.

[32] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. 2017. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5410–5419.

[33] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*. 2048–2057.

[34] Fei Yan and Krystian Mikolajczyk. 2015. Deep correlation for matching images and text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3441–3450.

[35] Wojciech Zaremba and Ilya Sutskever. 2015. Reinforcement learning neural turing machines-revised. *arXiv preprint arXiv:1505.00521* (2015).

[36] Zheng-Jun Zha, Jiawei Liu, Di Chen, and Feng Wu. 2020. Adversarial attribute-text embedding for person search with natural language query. *IEEE Transactions on Multimedia* (2020).

[37] Hanwang Zhang, Zheng-Jun Zha, Shuicheng Yan, Jingwen Bian, and Tat-Seng Chua. 2012. Attribute feedback. In *Proceedings of the 20th ACM international conference on Multimedia*. 79–88.

[38] Shu Zhang, Dequan Zheng, Xinchen Hu, and Ming Yang. 2015. Bidirectional long short-term memory networks for relation classification. In *Proceedings of the 29th Pacific Asia conference on language, information and computation*. 73–78.

[39] Ying Zhang and Huchuan Lu. 2018. Deep Cross-Modal Projection Learning for Image-Text Matching. In *Proceedings of the European Conference on Computer Vision*. 686–701.

[40] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, and Yi-Dong Shen. 2017. Dual-Path Convolutional Image-Text Embedding with Instance Loss. *arXiv preprint arXiv:1711.05535* (2017).