Person Search with Natural Language Description

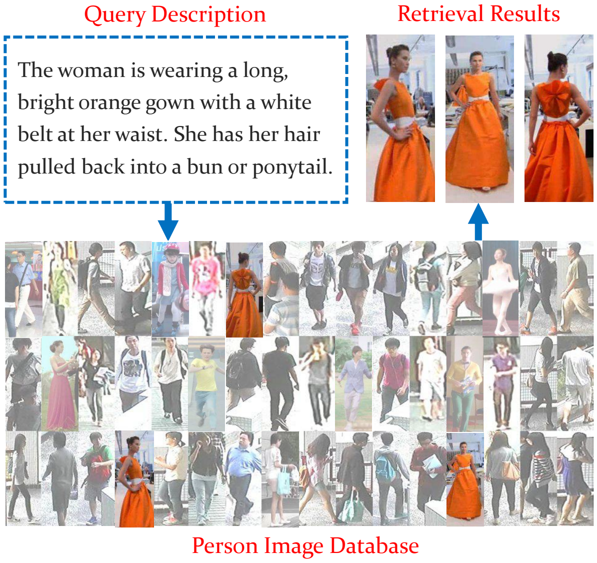
Shuang Li[1] Tong Xiao[1] Hongsheng Li[1]* Bolei Zhou[2] Dayu Yue[3] Xiaogang Wang[1]*
[1]The Chinese University of Hong Kong [2]Massachuate Institute of Technology [3]SenseTime Group Limited
{sli,xiaotong,hsli,xgwang}@ee.cuhk.edu.hk, bolei@mit.edu, yuedayu@sensetime.com

Abstract

Searching persons in large-scale image databases with the query of natural language description has important applications in video surveillance. Existing methods mainly focused on searching persons with image-based or attribute-based queries, which have major limitations for a practical usage. In this paper, we study the problem of person search with natural language description. Given the textual description of a person, the algorithm of the person search is required to rank all the samples in the person database then retrieve the most relevant sample corresponding to the queried description. Since there is no person dataset or benchmark with textual description available, we collect a large-scale person description dataset with detailed natural language annotations and person samples from various sources, termed as CUHK Person Description Dataset (CUHK-PEDES). A wide range of possible models and baselines have been evaluated and compared on the person search benchmark. An Recurrent Neural Network with Gated Neural Attention mechanism (GNA-RNN) is proposed to establish the state-of-the art performance on person search.

1. Introduction

Searching person in a database with free-form natural language description is a challenging problem in computer vision. It has wide applications in video surveillance and activity analysis. Nowadays urban areas are usually equipped with thousands of surveillance cameras which generate gigabytes of video data every second. To search possible criminal suspects from such large-scale videos manually might take tens of days or even months to complete. Thus automatic person search is in urgent need. Based on modalities of the queries, existing person search methods can be mainly categorized into the ones with image-based queries and attribute-based queries. However, both modalities have major limitations and might not be suitable for practical usages. Facing such limitations, we propose to study the problem of searching persons with natural language descriptions. Figure 1 illustrates one example of the person search.



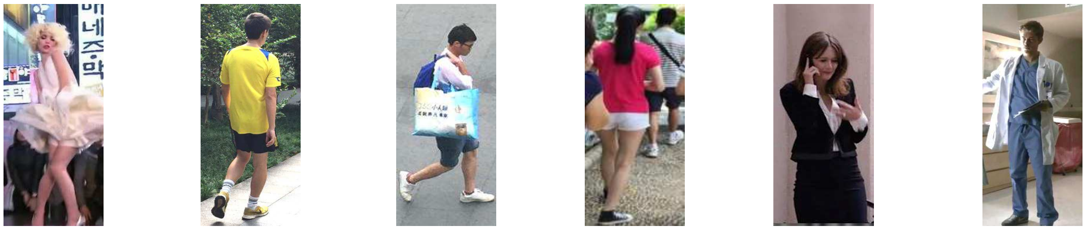
The woman is wearing a long, bright orange gown with a white belt at her waist. She has her hair pulled back into a bun or ponytail.

Figure 1. Given the natural language description of a person, our person search system searches through a large-scale person database then retrieve the most relevant person samples.

Person search with image-based queries is known as person re-identification in computer vision [44, 24, 39]. Given a query image, the algorithms obtain affinities between the query and those in the image database. The most similar persons can be retrieved from the database according to the affinity values. However, such a problem setting has major limitations in practice, as it requires at least one photo of the queried person being given. In many criminal cases, there might be only verbal description of the suspects' appearance available.

Person search could also be done through attribute-based queries. A set of pre-defined semantic attributes are used to describe persons' appearances. Classifiers are then trained on each of the attributes. Given a query, similar persons in the database can be retrieved as the ones with similar attributes [36, 35]. However, the attributes have many practical limitations as well. On the one hand, attributes have limited capability of describing persons' appearance. For instance, the PETA dataset [4] defined 61 binary and 4 multi-

*Corresponding authors



The woman is dressed up like Marilyn Monroe, with a white dress that is blowing upward in the wind, short curly blonde hair, and high heels.

The man is wearing yellow sneakers, white socks with blue stripes on the top of them, black athletic shorts and a yellow with blue t-shirt. He has short black hair.

The man has dark hair and is wearing glasses. He has on a pink shirt, blue shorts, and white tennis shoes. He has on a blue backpack and is carrying a re-useable tote.

The girl is wearing a pink shirt with white shorts, she is wearing black converse, with her hair in a pony tail.

The woman has long light brown hair, is wearing a black business suit with white low-cut blouse with large, white cuffs, a gold ring, and is talking on a cellphone.

The man is wearing blue scrubs with a white lab coat on top. He is holding paperwork in his hand and has a name badge on the left side of his coat.

Figure 2. Example sentence descriptions from our dataset that describe persons' appearances in detail.

class person attributes, while there are hundreds of words for describing a person's appearance. On the other hand, even with the exhausted set of attributes, labeling them for a large-scale person image dataset is expensive.

Facing the limitations of both modalities, we propose to use natural language description to search person. It does not require a person photo to be given as in those image-based query methods. Natural language also can precisely describe the details of person appearance, and does not require labelers to go through the whole list of attributes.

Since there is no existing dataset focusing on describing person appearances with natural language, we first build a large-scale language dataset, with 40,206 images of 13,003 persons from existing person re-identification datasets. Each person image is described with two sentences by two independent workers on Amazon Mechanical Turk (AMT). On the visual side, the person images pooled from various re-identification datasets are under different scenes, view points and camera specifications, which increases the image content diversity. On the language side, the dataset has 80,412 sentence descriptions, containing abundant vocabularies, phrases, and sentence patterns and structures. The labelers have no limitations on the languages for describing the persons. We perform a series of user studies on the dataset to show the rich expression of the language description. Examples from the dataset are shown in Figure 2.

We propose a novel Recurrent Neural Network with Gated Neural Attention (GNA-RNN) for person search. The GNA-RNN takes a description sentence and a person image as input and outputs the affinity between them. The sentence is input into a word-LSTM and processed word by word. At each word, the LSTM generates unit-level attentions for individual visual units, each of which determines whether certain person semantic attributes or visual patterns exist in the input image. The visual-unit attention mechanism weights the contributions of different units for different words. In addition, we also learn word-level gates that estimate the importance of different words for adaptive word-level weighting. The final affinity is obtained by averaging over all units' responses at all words. Both the unit-level attention and word-level sigmoid gates contribute to the good performance of our proposed GNA-RNN.

The contribution of this paper is three-fold. 1) We propose to study the problem of searching persons with natural language. This problem setting is more practical for real-world scenarios. To support this research direction, a large-scale person description dataset with rich language annotations is collected and the user study on the natural language description of person is given. 2) We investigate a wide range of plausible solutions based on different vision and language frameworks, including image captioning [19, 37], visual QA [45, 32], and visual-semantic embedding [31], and establish baselines on the person search benchmark. 3) We further propose a novel Recurrent Neural Network with Gated Neural Attention (GNA-RNN) for person search, with the state-of-the-art performance on the person search benchmark.

1.1. Related work

As there are no existing datasets and methods designed for the person search with natural language, we briefly survey the language datasets for various vision tasks, along with the deep language models for vision that can be used as possible solutions for this problem.

Language datasets for vision. Early language datasets for vision include Flickr8K [12] and Flickr30K [42]. Inspired by them, Chen *et al*. built a larger MS-COCO Caption [2] dataset. They selected 164,062 images from MS-COCO [25] and labeled each image with five sentences from independent labelers. Recently, Visual Genome [20] dataset was proposed by Krishna *et al*., which incorporates dense annotations of objects, attributes, and relationships within each image. However, although there are persons in the datasets, they are not the main subjects for descriptions and cannot be used to train person search algorithms with language descriptions. For fine-grained visual descriptions, Reed *et al*. added language annotations to Caltech-UCSD birds [38] and Oxford-102 flowers [29] datasets to describe contents of images for text-image joint embedding.

Deep language models for vision. Different from convolutional neural network which works well in image classification [21, 10] and object detection [18, 17, 16], recurrent neural network is more suitable in processing sequential data. A large number of deep models for vision tasks [40, 1, 13, 15, 8, 3, 5] have been proposed in recent years. For image captioning, Mao *et al.* [28] learned feature embedding for each word in a sentence, and connected it with the image CNN features by a multi-modal layer to generate image captions. Vinyal *et al.* [37] extracted high-level image features from CNN and fed it into LSTM for estimating the output sequence. The NeuralTalk [19] looked for the latent alignment between segments of sentences and image regions in a joint embedding space for sentence generation.

Visual QA methods were proposed to answer questions about given images [32, 30, 41, 34, 27, 7]. Yang *et al.* [41] presented a stacked attention network that refined the joint features by recursively attending question-related image regions, which leads to better QA accuracy. Noh *et al.* [30] learned a dynamic parameter layer with hashing techniques, which adaptively adjusts image features based on different questions for accurate answer classification.

Visual-semantic embedding methods [6, 19, 31, 26, 33] learned to embed both language and images into a common space for image classification and retrieval. Reed *et al.* [31] trained an end-to-end CNN-RNN model which jointly embeds the images and fine-grained visual descriptions into the same feature space for zero-shot learning. Text-to-image retrieval can be conducted by calculating the distances in the embedding space. Frome *et al.* [6] associated semantic knowledge of text with visual objects by constructing a deep visual-semantic model that re-trained the neural language model and visual object recognition model jointly.

2. Benchmark for person search with natural language description

Since there is no existing language dataset focusing on person appearance, we build a large-scale benchmark for person search with natural language, termed as CUHK Person Description Dataset (CUHK-PEDES). We collected 40,206 images of 13,003 persons from five existing person re-identification datasets, CUHK03 [23], Market-1501 [43], SSM [39], VIPER [9], and CUHK01 [22], as the subjects for language descriptions. Since persons in Market-1501 and CUHK03 have many similar samples, to balance the number of persons from different domains, we randomly selected four images for each person in the two datasets. All the image were labeled by crowd workers from Amazon Mechanical Turk (AMT), where each image was annotated with two sentence descriptions and a total of 80,412 sentences were collected. The dataset incorporates rich details about person appearances, actions, poses and interactions with other objects. The sentence descriptions are generally long (> 23 words in average), and has abundant vocabulary and little repetitive information. Examples of our proposed dataset are shown in Figure 2.

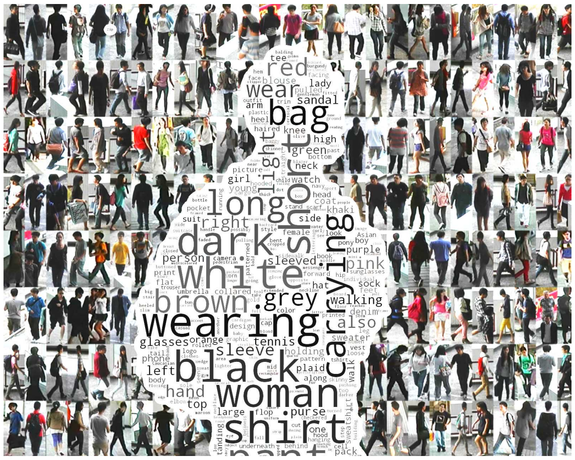


Figure 3. High-frequency words and person images in our dataset.

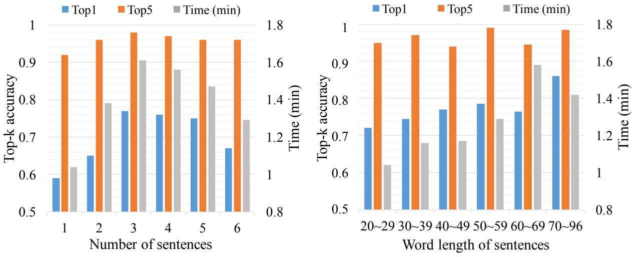


Figure 4. Top-1 accuracy, top-5 accuracy, and average used time of manual person search using language descriptions with different number of sentences and different sentence lengths.

2.1. Dataset statistics

The dataset consists of rich and accurate annotations with open word descriptions. There were 1,993 unique workers involved in the labeling task, and all of them have greater-than 95% approving rates. We asked the workers to describe all important characteristics in the given images using sentences with at least 15 words. The large number of workers means the dataset has diverse language descriptions and methods trained with it are unlikely to overfit to descriptions of just a few workers.

Vocabulary, phrase sizes, and sentence length are important indicators on the capacity our language dataset. There are a total of 1,893,118 words and 9,408 unique words in our dataset. The longest sentence has 96 words and the average word length is 23.5 which is significantly longer than the 5.18 words of MS-COCO Caption [25] and the 10.45 words of Visual Genome [20]. Most sentences have 20 to 40 words in length. Figure 3 illustrates some person examples and high-frequency words.

2.2. User study

Based on the language annotations we collect, we conduct the user studies to investigate 1) the expressive power of language descriptions compared with that of attributes, 2) the expressive power in terms of the number of sentences and sentence length, and 3) the expressive power of different word types. The studies provide us insights for understanding the new problem and guidance on designing our neural networks.

Language vs. attributes. Given a descriptive sentence or annotated attributes of a query person image, we ask crowd workers from AMT to select its corresponding image from a pool of 20 images. The 20 images consist of the ground truth image, 9 images with similar appearances to the ground truth, and 10 randomly selected images from the whole dataset. The 9 similar images are chosen from the whole dataset by the LOMO+XQDA [24] method, which is a state-of-the-art method for person reidentification. The other 10 distractor images are randomly selected and have no overlap with the 9 similar images. The person attribute annotations are obtained from the PETA [4] dataset, which have 1,264 same images with our dataset. A total of 500 images are manually searched by the workers, and the average top-1 and top-5 accuracies of the searches are evaluated. The searches with language descriptions have 58.7% top-1 and 92.0% top-5 accuracies, while the searches with attributes have top-1 and top-5 accuracies of 33.3% and 74.7% respectively. In terms of the average used time for each search, using language descriptions takes 62.18s, while using attributes takes 81.84s. The results show that, from human's perspective, language descriptions are much precise and effective in describing persons than attributes. They partially endorse our choice of using language descriptions for person search.

Sentence number and length. We design manual experiments to investigate the expressive power of language descriptions in terms of the number of sentences for each image and sentence length. The images in our dataset are categorized into different groups based on the number of sentences associated with each image and based on different sentence lengths. Given the sentences for each image, we ask crowd workers from AMT to manually retrieve the corresponding images from pools of 20 images. The average top-1 and top-5 accuracies, and used time for different image groups are shown in Figure 4, which show that 3 sentences for describing a person achieved the highest retrieval accuracy. The longer the sentences are, the easier for users to retrieve the correct images.

Word types. We also investigate the importance of different word types, including nouns, verbs, and adjectives by using manual experiments with the same 20-image pools. For this study, nouns, or verbs, or adjectives in the sentences are masked out before provided to the workers. For instance, “the girl has pink hair” is converted to “the **** has pink ****”, where the nouns are masked out. Results in Table 1 demonstrate that the nouns provide most information followed by the adjectives, while the verbs carry least information. This investigation provides us important insights that nouns and adjectives should be paid much attention to when we design neural networks or collecting new language data.

	orig. sent.	w/o nouns	w/o adjs	w/o verbs
top-1	0.59	0.38	0.44	0.57
top-5	0.92	0.81	0.85	0.92
time (min)	1.14	1.01	0.98	1.12

Table 1. Top-1 accuracy, top-5 accuracy, and average used time of manual person search results using the original sentences, and sentences with nouns, or adjectives, or verbs masked out.

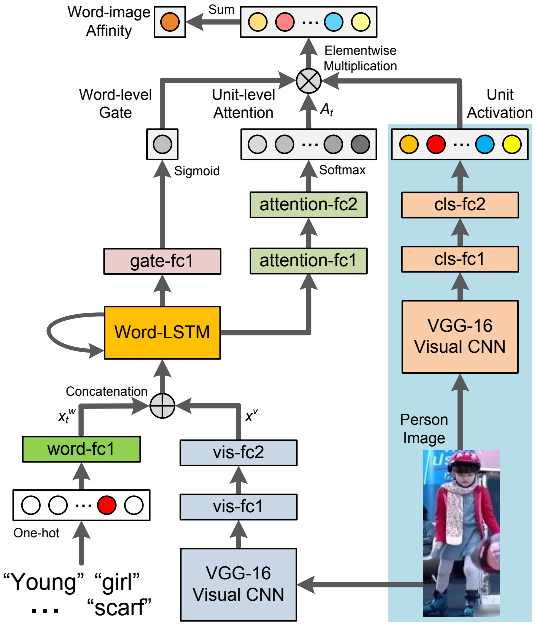


Figure 5. The network structure of the proposed GNA-RNN. It consists of a visual sub-network (right blue branch) and a language sub-network (left branch). The visual sub-network generates a series of visual units, each of which encodes if certain appearance patterns exist in the person image. Given each input word, The language sub-network outputs world-level gates and unit-level attentions for weighting visual units.

3. GNA-RNN model for pedestrian search

The key to address the person search with language description is to effectively build word-image relations. Given each word, it is desirable if the neural network would search related regions to determine whether the word with its context fit the image. For a sentence, all such word-image relations can be investigated, and confidences of all relations should be weighted and then aggregated to generate the final sentence-image affinity.

Based on this idea, we propose a novel deep neural network with Gated Neural Attention (GNA-RNN) to capture word-image relations and estimate the affinity between a sentence and a person image. The overall structure of the GNA-RNN is shown in Figure 5. The network model consists of a visual sub-network and a language sub-network. The visual sub-network generates a series of visual unit activations, each of which encodes if certain human attributes or appearance patterns (*e.g.*, white scarf) exist in the given person image. The language sub-network is a Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM) units, which takes words and images as input. At each word, it outputs unit-level attention and word-level gate to weight the visual units from the visual sub-network. The unit-level attention determines which visual units should be paid more attention to according to the input word. The word-level gate weight the importance of different words. All units' activations are weighted by both the unit-level attentions and word-level gates, and are then aggregated to generate the final affinity. By training such network in an end-to-end manner, the Gated Neural Attention mechanism is able to effectively capture the optimal word-image relations.

3.1. Visual units

The visual sub-network takes person images that are resized to 256×256 as inputs. It has the same bottom structure as VGG-16 network, and adds two 512-unit fully-connected layers at the "drop7" layer to generate 512 visual units, $\mathbf{v} = [v_1, ..., v_{512}]^T$. Our goal is to train the whole network jointly such that each visual unit determines whether certain human appearance pattern exist in the person image. The visual sub-network is first pre-trained on our dataset for person classification based on person IDs. During the joint training with language sub-network, only parameters of the two new fully-connected layers ("cls-fc1" and "cls-fc2" in Figure 5) are updated for more efficient training. Note that we do not manually constrain which units learn what concepts. The semantic meanings of the visual units automatically capture necessary semantic concepts via jointly training of the whole network.

3.2. Attention over visual units

To effectively capture the word-image relations, we propose a unit-level attention mechanism for visual units. At each word, the visual units having similar semantic meanings with the word should be assigned with more weights. Take Figure 5 as example, given the words "white scarf", the language sub-network would attend more the visual unit that corresponds to the concept of "white scarf". We train the language sub-network to to achieve this goal.

The language sub-network is a LSTM network [11], which is effective at capturing temporal relations of sequential data. Given an input sentence, the LSTM generates attentions for visual units word by word. The words are first encoded into length-K one-hot vectors, where K is the vocabulary size. Given a descriptive sentence, a learnable fully connected layer ("word-fc1" in Figure 5) converts the tth raw word to a word embedding feature x_w^t. Two 512-unit fully connected layers ("vis-fc1" and "vis-fc2" in Figure 5) following the "drop7" layer of VGG-16 are treated as visual features x_v for the LSTM. At each step, the LSTM takes $x_t = [x_t^w, x^v]^T$ as input, which is concatenation of tth word embedding x_t^w and image features x^v.

The LSTM consists of a memory cell c_t and three controlling gates, *i.e.* input gate i_t, forget gate f_t, and output gate o_t. The memory cell preserves the knowledge of previous step and current input while the gates control the update and flow direction of information. At each word, the LSTM updates the memory cell c_t and output a hidden state h_t in the following way,

$$\begin{aligned} i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i), \\ f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f), \\ o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o), \\ c_t &= f_t \odot c_{t-1} + i_t \odot h(W_{xc}x_t + W_{hc}h_{t-1} + b_c), \\ h_t &= o_t \odot h(c_t), \end{aligned} \tag{1}$$

where $\odot$ represents the element-wise multiplication, W and b are parameters to learn.

For generating the unit-level attentions at each word, the output hidden state h_t is fed into a fully-connected layer with ReLU non-linearity function and a fully-connected layer with softmax function to obtain the attention vector $A_t \in \mathbb{R}^{512}$, which has the same dimension as the visual units $\mathbf{v}$. The affinity between the sentence and the person image at the tth word can then be obtained by

$$a_t = \sum_{n=1}^{512} A_t(n) v_n, \quad \text{s.t.} \sum_{n=1}^{512} A_t(n) = 1, \tag{2}$$

where $A_t(n)$ denotes the attention value for the nth visual unit. Since each visual unit determines the existence of certain person appearance patterns in the image, the visual units alone cannot generate sentence-image affinity. The attention values A_t generated by the language sub-network decides which visual units' responses should be summed up to compute the affinity value. If the language sub-network generates high attention value at certain visual unit, only if the visual unit also has high response, which denotes existence of certain visual concepts, will the elementwise multiplication generates high affinity value at this word. The final sentence-image affinity is summation of affinity values at all words, $a = \sum_{t=1}^{T} a_t$, where T is the number of words in the given sentence.

3.3. Word-level gates for visual units

The unit-level attention is able to associate the most related units to each word. However, the attention mechanism requires different units' attentions competing with each other. In our case with the softmax non-linearity function, we have $\sum_{n=1}^{512} A_t(n) = 1$, and found that such constraints are important for learning effective attentions.

	NeuralTalk [37]	CNN-RNN [31]	EmbBoW	QAWord	QAWord-img	QABoW	GNA-RNN
top-1	13.66	8.07	8.38	11.62	10.21	8.00	**19.05**
top-10	41.72	32.47	30.76	42.42	44.53	30.56	**53.64**

Table 2. Quantitative results of the proposed GNA-RNN and compared methods on the proposed dataset.

However, according to our user study on different word types in Section 2.2, different words carry significantly different amount of information for obtaining language-image affinity. For instance, the word "white" should be more important than the word "this". At each word, the unit-level attentions always sum up to 1 and cannot reflect such differences. Therefore, we propose to learn world-level scalar gates at each word for learning to weight different words. The word-level scalar gate is obtained by mapping the hidden state h_t of the LSTM via a fully-connected layer with sigmoid non-linearity function $g_t = \sigma(W_g h_t + b_g)$, where σ denotes the sigmoid function, and W_g and b_g are the learnable parameters of the fully-connected layer.

Both the unit-level attention and world-level gate are used to weight the visual units at each word to obtain the per-word language-image affinity $\hat{a}_t$,

$$\hat{a}_t = g_t \sum_{n=1}^{512} A_t(n) v_n, \tag{3}$$

and the final affinity is the aggregation of affinities at all words $\hat{a} = \sum_{t=1}^{T} \hat{a}_t$.

3.4. Training scheme

The proposed GNA-RNN is trained end-to-end with batched Stochastic Gradient Descent, except for the VGG-16 part of the visual sub-network, which is pre-trained for person classification and fixed afterwards. The training samples are randomly chosen from the dataset with corresponding sentence-image pairs as positive samples and non-corresponding pairs as negative samples. The ratio between positive and negative samples is 1:3. Given the training samples, the training minimizes the cross-entropy loss,

$$E = -\frac{1}{N} \sum_{i=1}^{N} \left[y^i \log \hat{a}^i + (1 - y^i) \log(1 - \hat{a}^i) \right] \tag{4}$$

where $\hat{a}^i$ denotes the predicted affinity for the ith sample, and y^i denotes its ground truth label, with 1 representing corresponding sentence-image pairs and 0 representing non-corresponding ones. We use 128 sentence-image pairs for each training batch. All fully connected layers except for the one for word-level gates have 512 units.

4. Experiments

There is no existing method specifically designed for the problem. We investigate a wide range of possible solutions based on state-of-the-art language models for vision tasks, and compare those solutions with our proposed method. We also conduct component analysis of our proposed deep neural networks to show that our proposed Gated Neural Attention mechanism is able to capture complex word-image relations. Extensive experiments and comparisons with state-of-the-art methods demonstrate the effectiveness of our GNA-RNN for this problem.

	GNA-RNN	w/o pre-train	w/o gates	w/o attention
top-1	**19.05**	8.93	13.86	4.85
top-10	**53.64**	32.32	44.27	27.16

Table 3. Quantitative results of GNA-RNN on the proposed dataset without VGG-16 re-id pre-training, without world-level gates or without unit-level attentions.

# units	128	256	512	1024	2048
top-1	16.15	16.75	19.05	18.62	18.25
top-10	48.58	49.25	53.64	52.39	51.59

Table 4. Top-1 and top-10 accuracies of GNA-RNN with different number of visual units.

4.1. Dataset and evaluation metrics

The dataset is splitted into three subsets for training, validation, and test without having overlaps with same person IDs. The training set consists of 11,003 persons, 34,054 images and 68,108 sentence descriptions. The validation set and test set contain 3,078 and 3,074 images, respectively, and both of them have 1,000 persons. All experiments are performed based on this train-test split.

We adopt the top-k accuracy to evaluate the performance of person retrieval. Given a query sentence, all test images are ranked according to their affinities with the query. A successful search is achieved if any image of the corresponding person is among the top-k images. Top-1 and top-10 accuracies are reported for all our experiments.

4.2. Compared methods and baselines

We compare a wide range of possible solutions with deep neural networks, including methods for image captioning, visual QA, and visual-semantic embedding. Generally, each type of methods utilize different supervisions for training. Image captioning, visual QA, and visual-semantic embedding methods are trained with word classification losses, answer classification losses, and distance-based losses, respectively. We also propose several baselines to investigate the influences of detailed network structure design. To make fair comparisons, the image features for all compared methods are from our VGG-16 network pre-trained model.

Image captioning. Vinyals *et al.* [37] and Karpathy *et al.* [19] proposed to generate natural sentences describing an image using deep recurrent frameworks. We use the code provided by Karpathy *et al.* to train the image captioning model. We follow the testing strategy in [14] to use image captioning method for text-to-image retrieval. During the test phase, given a person image, instead of recursively using the predicted word as inputs of the next time step to predict the image caption, the LSTM takes the given sentence word by word as inputs. It calculates the per-word cross entropy losses between the given word and the predicted word from LSTM. Corresponding sentence-image pairs would have low average losses, while non-corresponding ones would have higher average losses.

Visual QA. Agrawal *et al.* [1] proposed the deeper LSTM Q + norm I method to answer questions about the given image. We replace the element-wise multiplication between the question and image features, with concatenation of question and image features, and replace the multi-class classifier with a binary classifier. Since the proposed GNA-RNN has only one layer for the LSTM, we change the LSTM in deeper LSTM Q + norm I to one layer as well for fair comparison. The norm I in [1] is also changed to contain two additional fully-connected layers to obtain image features instead of the original one layer following our model's structure. We call the modified model QAWord. Where to concatenate features of question and image modalities might also influence the classification performance. The QAWord model concatenates image features with sentence features output by the LSTM. We investigate concatenating the word embedding features and image features before inputting them into the LSTM. Such a modified network is called QAWord-img. We also replace the language model in QAWord with the simple language model in [45], which encodes sentences using the traditional Bag-of-Word (BoW) method, and call it QABoW.

Visual-semantic embedding. These methods try to map image and sentence features into a joint embedding space. Distances between image and sentence features in the joint space could then be interpreted as the affinities between them. Distances between corresponding sentence-image pairs should be small, and should be high between non-corresponding paris. Reed *et al.* [31] presented a CNN-RNN for zero-shot text-to-image retrieval. We utilize their code and compare it with our proposed framework. We also investigate replacing the language model in CNN-RNN with the simple BoW language model [45] for sentence encoding and denote it as EmbBoW.

4.3. Quantitative and qualitative results

Quantitative evaluation. Table 2 shows the results of our proposed framework and the compared methods. We use a single sentence as query to do the person search. Our approach achieves the best performance in terms of both top-1 and top-10 accuracies and outperforms other methods by a large margin. It demonstrates that our proposed network can better capture complex word-image relations than the compared ones.

For all the baselines, the image captioning method NeuralTalk outperforms the other baselines. It calculates the average loss at each word as the sentence-image affinity, and obtains better results than visual QA and visual embedding approaches, which encode the entire sentence into a feature vector. Such results show that the LSTM might have difficulty encoding complex person descriptive sentences into a single feature vector. Word-by-word processing and comparison might be more suitable for the person search problem. We also observe that QAWord-img and QAWord has similar performance. This demonstrates that, the modality fusion between image and word before or after LSTM has little impact on the person search performance. Both ways capture word-image relations to some extent. For the visual-semantic embedding method, the CNN-RNN does not perform well in terms of top-k accuracies with the provided code. The distance-based losses might not be suitable for learning good models for the person search problem. EmbBoW and QABoW use the traditional Bag-of-Word method to encode sentences and have worse performances than their counterparts with RNN language models, which show that the RNN framework is more suitable in processing natural language data.

Component analysis. We pre-train the visual VGG model for person re-id task first, and then fine-tune whole network for text-to-person search. Without the person re-id pre-training, top-1 and top-10 accuracies drop apparently as shown in Table 3. This means the initial training affects the final performance a lot. To investigate the effectiveness of the proposed unit-level attentions and word-level gates, we design two baselines for comparison. For the first baseline (denoted as "w/o gates"), we remove the word-level gates and only keep the unit-level attentions. In this case, different words are equally weighted in estimating the sentence-image affinity. For the second baseline (denoted as "w/o attention"), we try to keep the world-level gates, and replace the unit-level attentions with average pooling over units. We list top-1 and top-10 accuracies of the two baselines in Table 3. Both the unit-level attention and word-level gates are important for achieving good performance by our GNA-RNN.

Investigation on the impact of the number of visual units. Results of different number of visual units are listed in Table 4. Models with more visual units might over-fit the dataset. 512 units achieves the best result.

Qualitative evaluation. We conduct qualitative evaluation for our proposed GNA-RNN. Figure 6 shows 6 person search results with natural language descriptions by our proposed GNA-RNN. The four cases in the top 2 rows show successful cases where corresponding images are within the top-6 retrieval results. For the successful cases, we can observe that each top image has multiple regions that fit parts of the descriptions. Some non-corresponding images also

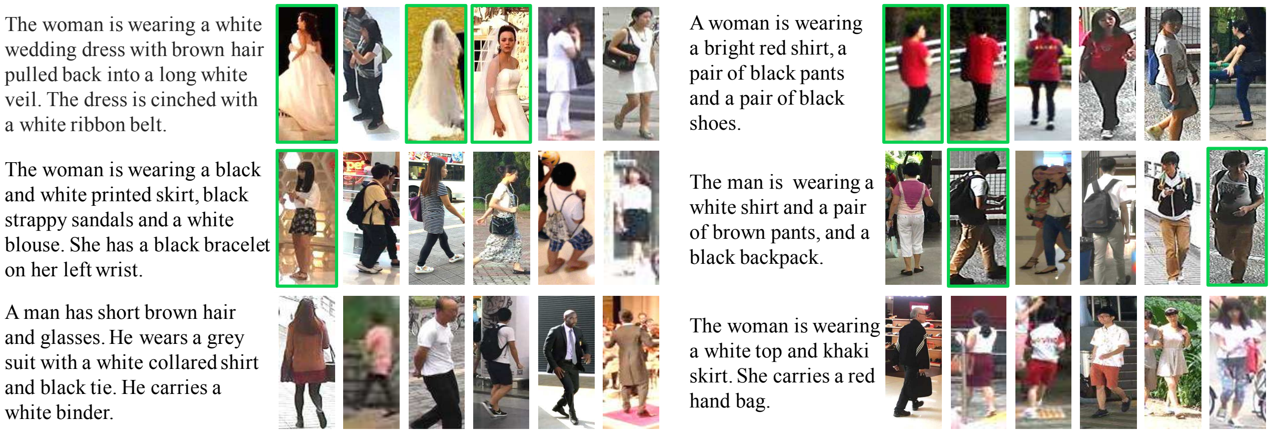

The woman is wearing a white wedding dress with brown hair pulled back into a long white veil. The dress is cinched with a white ribbon belt.
A woman is wearing a bright red shirt, a pair of black pants and a pair of black shoes.
The woman is wearing a black and white printed skirt, black strappy sandals and a white blouse. She has a black bracelet on her left wrist.
The man is wearing a white shirt and a pair of brown pants, and a black backpack.
A man has short brown hair and glasses. He wears a grey suit with a white collared shirt and black tie. He carries a white binder.
The woman is wearing a white top and khaki skirt. She carries a red hand bag.

Figure 6. Examples of top-6 person search results with natural language description by our proposed GNA-RNN. Corresponding images are marked by green rectangles. (Rows 1-2) Successful searches where corresponding persons are in the top-6 results. (Row 3) Failure cases where corresponding persons are not in the top-6 results.

show correlations to the query sentences. In terms of failure cases, there are two types of them. The first type of failure searches do retrieve images that are similar to the language descriptions, however, the exact corresponding images are not within the top retrieval results. For instance, the bottom right case in Figure 6 does include persons (top-2, top-3, and top-4) similar to the descriptions, who all wear white tops and red shorts/skirts. Other persons have some characteristics that partially fits the descriptions. The top-1 person has a "hand bag". The top-4 person wears "white top", and the top-6 person carries a "red bag". The second type of failure cases show that the GNA-RNN fails to understand the whole sentence but only captures separate words or phrases. Take the bottom left case in Figure 6 as an example, the phrase "brown hair" is not encoded correctly. Instead, only the word "brown" is captured, which leads to the "brown" suit for the top-1 and top-6 persons, and "brown" land in the top-2 image. We also found some rare words/concepts or detailed descriptions are difficult to learn and to locate, such as "ring", "bracelet", "cell phones", *etc.*, which might be learned if more data is provided in the future.

Visual unit visualization. We also inspect the learned visual units to see whether they implicitly capture common visual patterns in person images. We choose some frequent adjectives and nouns. For each frequent word, we collect its unit-level attention vectors for a large number of training images. Such unit-level attention vectors are averaged to identify its most attended visual units. For each of such units, we retrieve the training images that have the highest responses on the units. Some examples of the visual units obtained in this way are shown in Figure 7. Each of them captures some common image patterns.

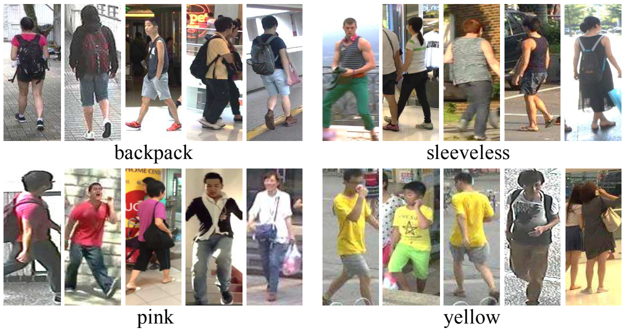


Figure 7. Images with the highest activations on 4 different visual units. The 4 units are identified as the one with the maximum average attention values in our GNA-RNN with the same word ("backpack", "sleeveless", "pink", "yellow") and a large number of images. Each unit determines the existence of some common visual patterns.

5. Conclusions

In this paper, we studied the problem of person search with natural languages. We collected a large-scale person dataset with 80,412 sentence descriptions of 13,003 persons. Various baselines are evaluated and compared on the benchmark. A GNA-RNN model was proposed to learn affinities between sentences and person images with the proposed gated neural attention mechanism, which established the state-of-the art performance on person search.

Acknowledgement This work is supported in part by SenseTime Group Limited, in part by the General Research Fund through the Research Grants Council of Hong Kong under Grants CUHK14207814, CUHK14213616, CUHK14206114, CUHK14205615, CUHK419 412, CUHK14203015 and CUHK14239816, in part by the Hong Kong Innovation and Technology Support Programme Grant ITS/121/15FX, in part by National Natural Science Foundation of China under Grant 61371192, in part by the Ph.D. Program Foundation of China under Grant 20130185120039, and in part by the China Postdoctoral Science Foundation under Grant 2014M552339.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *ICCV*, pages 2425–2433, 2015. 3, 7

[2] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 2

[3] X. Chen and C. L. Zitnick. Learning a recurrent visual representation for image caption generation. *arXiv preprint arXiv:1411.5654*, 2014. 3

[4] Y. Deng, P. Luo, C. C. Loy, and X. Tang. Pedestrian attribute recognition at far distance. In *ACM MM*, pages 789–792, 2014. 1, 4

[5] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, et al. From captions to visual concepts and back. In *CVPR*, pages 1473–1482, 2015. 3

[6] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NIPS*, pages 2121–2129, 2013. 3

[7] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016. 3

[8] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. Are you talking to a machine? dataset and methods for multilingual image question. In *NIPS*, pages 2296–2304, 2015. 3

[9] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, number 5, 2007. 3

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3

[11] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 5

[12] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013. 2

[13] R. Hu, M. Rohrbach, and T. Darrell. Segmentation from natural language expressions. *arXiv preprint arXiv:1603.06180*, 2016. 3

[14] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell. Natural language object retrieval. *CVPR*, 2016. 7

[15] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. *arXiv preprint arXiv:1511.07571*, 2015. 3

[16] K. Kang, H. Li, T. Xiao, W. Ouyang, J. Yan, X. Liu, and X. Wang. Object detection in videos with tubelet proposal networks. In *CVPR*, 2017. 3

[17] K. Kang, H. Li, J. Yan, X. Zeng, B. Yang, T. Xiao, C. Zhang, Z. Wang, R. Wang, X. Wang, et al. T-cnn: Tubelets with convolutional neural networks for object detection from videos. *arXiv preprint arXiv:1604.02532*, 2016. 3

[18] K. Kang, W. Ouyang, H. Li, and X. Wang. Object detection from video tubelets with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 817–825, 2016. 3

[19] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015. 2, 3, 7

[20] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *arXiv preprint arXiv:1602.07332*, 2016. 2, 3

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 3

[22] W. Li, R. Zhao, and X. Wang. Human reidentification with transferred metric learning. In *ACCV*, pages 31–44, 2012. 3

[23] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, pages 152–159, 2014. 3

[24] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, pages 2197–2206, 2015. 1, 4

[25] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 2, 3

[26] W. Liu, T. Mei, Y. Zhang, C. Che, and J. Luo. Multi-task deep visual-semantic embedding for video thumbnail selection. In *CVPR*, pages 3707–3715, 2015. 3

[27] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *ICCV*, pages 1–9, 2015. 3

[28] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*, 2014. 3

[29] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Proceedings of Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729, 2008. 3

[30] H. Noh, P. H. Seo, and B. Han. Image question answering using convolutional neural network with dynamic parameter prediction. *arXiv preprint arXiv:1511.05756*, 2015. 3

[31] S. Reed, Z. Akata, B. Schiele, and H. Lee. Learning deep representations of fine-grained visual descriptions. *arXiv preprint arXiv:1605.05395*, 2016. 2, 3, 6, 7

[32] M. Ren, R. Kiros, and R. Zemel. Exploring models and data for image question answering. In *NIPS*, pages 2953–2961, 2015. 2, 3

[33] M. Ren, R. Kiros, and R. Zemel. Image question answering: A visual semantic embedding model and a new dataset. *CoRR, abs/1505.02074*, 7, 2015. 3

[34] K. Saito, A. Shin, Y. Ushiku, and T. Harada. Dualnet: Domain-invariant network for visual question answering. *arXiv preprint arXiv:1606.06108*, 2016. 3

[35] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian. Deep attributes driven multi-camera person re-identification. *arXiv preprint arXiv:1605.03259*, 2016. 1

[36] D. A. Vaquero, R. S. Feris, D. Tran, L. Brown, A. Hampapur, and M. Turk. Attribute-based people search in surveillance environments. In *WACV*, pages 1–8, 2009. 1

[37] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164, 2015. 2, 3, 6, 7

[38] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-ucsd birds 200. 2010. 3

[39] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. End-to-end deep learning for person search. *arXiv preprint arXiv:1604.01850*, 2016. 1, 3

[40] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2015. 3

[41] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. *arXiv preprint arXiv:1511.02274*, 2015. 3

[42] P. Young, A. Lai, M. Hodosh, and J. 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

[43] L. Zheng, L. Shen, L. Tian, S. Wang, J. Bu, and Q. Tian. Person re-identification meets image search. *arXiv preprint arXiv:1502.02171*, 2015. 3

[44] W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR*, pages 649–656, 2011. 1

[45] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus. Simple baseline for visual question answering. *arXiv preprint arXiv:1512.02167*, 2015. 2, 7