

Structured Matching for Phrase Localization

Mingzhe Wang^(✉), Mahmoud Azab, Noriyuki Kojima,
Rada Mihalcea, and Jia Deng

Computer Science and Engineering, University of Michigan, Ann Arbor, USA
{mzwang,mazab,kojimano,mihalcea,jiadeng}@umich.edu

Abstract. In this paper we introduce a new approach to phrase localization: grounding phrases in sentences to image regions. We propose a structured matching of phrases and regions that encourages the semantic relations between phrases to agree with the visual relations between regions. We formulate structured matching as a discrete optimization problem and relax it to a linear program. We use neural networks to embed regions and phrases into vectors, which then define the similarities (matching weights) between regions and phrases. We integrate structured matching with neural networks to enable end-to-end training. Experiments on Flickr30K Entities demonstrate the empirical effectiveness of our approach.

Keywords: Vision · Language

1 Introduction

This paper addresses the problem of phrase localization: given an image and a textual description, locate the image regions that correspond to the noun phrases in the description. For example, an image may be described as “a man wearing a tan coat signs papers for another man wearing a blue coat”. We wish to localize, in terms of bounding boxes, the image regions for the phrases “a man”, “tan coat”, “papers”, “another man”, and “blue coat”. In other words, we wish to ground these noun phrases to image regions.

Phrase localization is an important task. Visual grounding of natural language is a critical cognitive capability necessary for communication, language learning, and the understanding of multimodal information. Specifically, understanding the correspondence between regions and phrases is important for natural language based image retrieval and visual question answering. Moreover, by aligning phrases and regions, phrase localization has the potential to improve weakly supervised learning of object recognition from massive amounts of paired images and texts.

Recent research has brought significant progress on the problem of phrase localization [1–3]. Plummer et al. introduced the Flickr30K Entities dataset, which includes images, captions, and ground-truth correspondences between regions and phrases [1]. To match regions and phrases, Plummer et al. embedded

regions and phrases into a common vector space through Canonical Correlation Analysis (CCA) and pick a region for each phrase based on the similarity of the embeddings. Subsequent works by Wang et al. [2] and Rohrbach et al. [3] have since achieved significant improvements by embedding regions and phrases using deep neural networks.

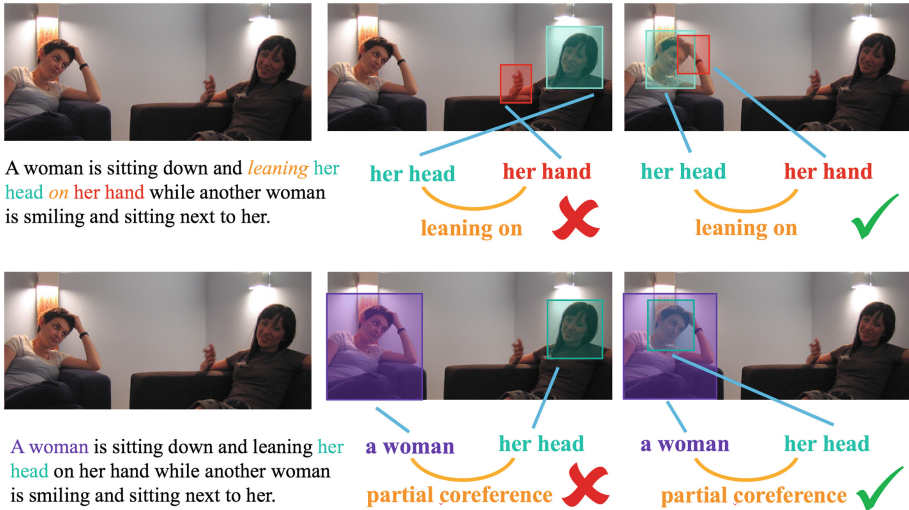


Fig. 1. Structured matching is needed for phrase localization: it is not enough to just match phrases and regions individually; the relations between phrases also need to agree with the relations between regions.

But existing works share a common limitation: they largely localize each phrase independently, ignoring the semantic relations between phrases. The only constraint used in previous research was that different phrases should describe different regions, i.e., that each region should be matched to no more than one phrase [3]. But phrases have more complex semantic relations between each other, and phrase localization is often impossible without a deep understanding of those semantic relations. For example, in Fig. 1, an image from Flickr30K Entities is captioned as “a woman is sitting down and leaning her head on her hand while another woman is smiling and sitting next to her.” Consider the localization of “her head” and “her hand” from “leaning her head on her hand”. There are two women, two heads, and two hands visible in the image, but only one head and one hand have a “leaning on” relation. So “her head” and “her hand” cannot be localized independently without verifying whether the head is actually leaning on the hand.

This brings forward the problem of *structured matching* of regions and phrases, that is, finding an optimal matching of regions and phrases such that not only does the visual content of each individual region agree with the meaning

of its corresponding phrase (e.g. the regions must individually depict “head” and “hand”), but the visual relation between each pair of regions also agrees with the semantic relation between the corresponding pair of phrases (e.g. the pair of regions together must depict “leaning her head on her hand”). The problem of structured matching is closely related to the standard (maximum weighted) bipartite matching, although significantly harder: the nodes on the same side have relations between them, and thus not only the nodes but also the relations need to be matched with the other side.

In this paper we introduce a new approach to phrase localization based on the idea of structured matching. We formulate structured matching as a discrete optimization problem and relax it to a linear program. We use neural networks to embed regions and phrases into vectors, which then define the similarities (matching weights) between regions and phrases. We integrate structured matching with neural networks to enable end-to-end training. Experiments on Flickr30K Entities demonstrate the empirical effectiveness of our approach.

2 Related Work

Grounding image descriptions. Many works have studied the alignment of textual descriptions and visual scenes [1–6]. Kong et al. [6] align mentions in textual descriptions of RGB-D scenes with object cuboids using a Markov random field. Karpathy et al. [4] generate image fragments from object detection and sentence fragments from dependency parsing, and match them in a neural embedding framework. Karpathy and Fei-Fei [5] use a similar framework but replace dependency parsing with bidirectional recurrent neural networks in order to generate image descriptions. It is worth noting that Karpathy et al. [4] and Karpathy and Fei-Fei [5] only evaluate the performance on proxy tasks (image retrieval and captioning) and do not evaluate the quality of matching.

Plummer et al. [1] introduced the Flickr30K Entities dataset, making it possible to directly evaluate image-sentence alignments. Using this dataset, Wang et al. [2] learn neural embeddings of phrases and regions under a large-margin objective and localize each phrase by retrieving the closest region in the embedding space. Instead of producing explicit embeddings, Rohrbach et al. [3] train a neural network to directly predict the compatibility of a region and a phrase; their framework also allows unsupervised training when ground truth correspondences are unavailable.

All these works differ from ours because when they align sentences and images, they do not explicitly consider how relations between parts of a sentence match the relations between parts of an image.

Object retrieval using natural language. The task of object retrieval using natural language is to locate visual objects based on natural language queries. Guadarrama et al. [7] generate textual descriptions for each region proposal and retrieval objects by matching a query with the descriptions. In contrast, Arandjelovic and Zisserman [8] convert a query into multiple query images using Google image search and retrieve objects by matching the query images with

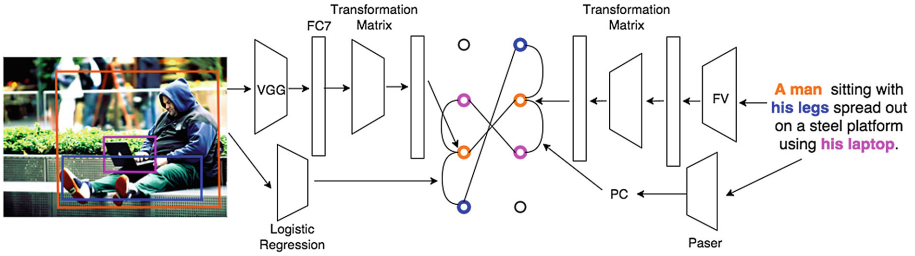


Fig. 2. We embed regions and phrases into a common vector space and perform structured matching that encourages not only the individual agreement of regions with phrases but also the agreement of phrase-phrase relations with region-region relations. In particular, we consider “partial match coreference” (PC) relations—the relation between phrases such as “a man” and “his legs”.

images in a database. Hu et al. [9] use a recurrent neural network to score each object proposal given an input text query and an input image, and retrieve the objects with the highest score. The difference between object retrieval using natural language and phrase localization is that the former aims to match an image region to a whole sentence whereas the latter aims to match an image region to only one part of a sentence.

Image captioning and retrieval. There has been a large body of prior work on image captioning or retrieval [2, 4, 5, 10–19]. Typical approaches include recurrent neural networks [4, 5, 14, 16, 17], Canonical Correlation Analysis [10], encoder-decoder architectures [13], and Discriminative Component Analysis [12]. These works differ from ours in that the emphasis of learning and evaluation is placed on matching images and sentences as a whole rather than matching their individual components such as regions and phrases.

3 Approach

Figure 2 illustrates our approach. Given an image and a description, our goal is to localize the image region that corresponds to each phrase in the description. Following prior work [1], we assume that short noun phrases (“a man”, “tan coat”) have already been extracted from the description. We also assume that pronouns and non-visual phrases have been removed. Also following prior work [1], we generate a set of region proposals in the form of bounding boxes.

Given these phrases and regions, the next step is to match them. To this end we adopt the same approach by Wang et al. [2]: we extract visual and phrasal features, embed them into a common vector space using neural networks, and measure region-phrase similarities in this common vector space. Using these similarities we then solve a structured matching problem: finding the optimal matching subject to two types of constraints: (1) a region can be matched to no more than one phrase, and (2) if two phrases have a certain semantic relation,

their corresponding regions should have a visual relation that is consistent with the semantic relation.

If we have only the first type of constraints, we arrive at a standard maximum weighted bipartite matching problem, which can be solved exactly by linear programming. The second type of constraints, however, pose significant new difficulties because it appears intractable to obtain exact solutions. As a result, we propose a relaxation to a linear program that gives approximate solutions.

We learn end to end with a structured prediction loss. That is, the learnable parameters of our framework are jointly optimized with the objective that for each image-sentence pair in the training set the ground truth matching should have a higher score than all other possible matchings. It is worth noting that although prior work on phrase localization has considered the constraint that a region should be matched to no more than one phrase [3], they have only used it as a post-processing heuristic, whereas we integrate this constraint into end-to-end training.

3.1 Representing Regions and Phrases

We generate regions proposals using Edgebox [20]. These regions serve as the candidates to be matched with phrases. To represent each region, we use Fast-RCNN [21] features, that is, features from a 16-layer VGG convolutional network that is pre-trained on the ImageNet [22] classification dataset and fine-tuned on the VOC2007 detection dataset [23]. In particular, we extract the fc7 layer activations to represent each region with a 4,096 dimensional feature vector.

To represent phrases, we use Fisher vectors [10]. Following [10], we extract Fisher Vectors of 18,000 dimensions by first applying ICA on the 300-dimensional word2vec [24] word vectors and then constructing a codebook with 30 centers from a Hybrid Gaussian-Laplacian mixture model (HGLMM). Similar to [2], to save time during training, we apply PCA to reduce the dimensionality of the Fisher vectors from 18,000 to 6,000.

Next, we apply a linear transformation to the fc7 activations of the regions and another linear transformation to the Fisher Vectors of the phrases in order to embed them in the same vector space. That is, given for a phrase p and a region r and their feature vectors x_p and x_r , we compute the embedded features \tilde{x}_p and \tilde{x}_r as

$$\begin{aligned}\tilde{x}_p &= M_1 x_p + b_1, \\ \tilde{x}_r &= M_2 x_r + b_2,\end{aligned}\tag{1}$$

where M_1, M_2, b_1, b_2 are learnable parameters. We define the similarity w_{ij} between a phrase p and a region r as

$$\cos(x_p, x_r) = \frac{\langle x_p, x_r \rangle}{\|x_p\| \|x_r\|},\tag{2}$$

i.e. the cosine similarity between the embedded vectors.

Given phrases $p_1, p_2 \dots p_n$ and regions $r_1, r_2 \dots r_m$, we obtain a similarity matrix $W_\theta = \{w_{ij}\}$, where

$$w_{ij} = \cos(x_{p_i}, x_{r_j}) \quad (3)$$

and θ represents the learnable parameters M_1, M_2, b_1, b_2 .

3.2 Bipartite Matching

Given the similarities between regions and phrases, we are ready to solve the matching problem. We first consider bipartite matching, matching with only the constraint that each region should be matched to no more than one phrase. We refer to this constraint as the exclusivity constraint.

It is worth noting that in the most general case, this constraint is not always valid. Two phrases can refer to the same region: for example, “a man” and “he”, “a man” and “the man”. In these cases we need to perform coreference resolution and group the coreferences before phrase localization. Here we assume that this step has been done. This is a valid assumption for the Flickr30K Entities dataset we use in our experiments—coreferences such as “he” and “she” have been removed.

Given phrases p_1, p_2, \dots, p_n , regions r_1, r_2, \dots, r_m , and their similarity matrix W_θ , we have a standard bipartite matching problem if we consider only the exclusivity constraints. We first formulate this problem as an integer program.

We define a binary variable $y_{ij} \in \{0, 1\}$ for each potential region-phrase pair $\{p_i, r_j\}$ to indicate whether r_j is selected as a match for p_i in a matching y . Let $S(W, y)$ be a score that measures the goodness of a matching y , computed as the sum of similarities of the matched pairs:

$$S(w, y) = \sum_{i=1}^n \sum_{j=1}^m w_{ij} y_{ij}. \quad (4)$$

The best matching maximizes this score and can be found by a linear program that relaxes y_{ij} to continuous variables in $[0, 1]$.

$$\begin{aligned} & \max_y S(W_\theta, y) \\ \text{s.t. } & \sum_{j=1}^m y_{ij} = 1, \quad i = 1, 2, \dots, n \\ & \sum_{i=1}^n y_{ij} \leq 1, \quad j = 1, 2, \dots, m \\ & 0 \leq y_{ij} \leq 1, \quad i = 1, \dots, n, j = 1, \dots, m. \end{aligned} \quad (5)$$

Here, the first constraint guarantees that each phrase is matched with exactly one region, and the second constraint guarantees that each region is matched with no more than one phrase. This linear program is guaranteed to have an

integer solution because all of its corner points are integers and this integer solution can be found by the simplex method.

To learn the embedding parameters, we optimize an objective that encourages the ground truth matching to have the best matching score. Let $y^{(l)}$ be the ground truth matching for the l^{th} image-sentence pair in a training set, and let $W_{\theta}^{(l)}$ be the region-phrase similarities. We define the training loss L as

$$L(\theta) = \sum_l \max(0, \max_{y'} S(W_{\theta}^{(l)}, y') - S(W_{\theta}^{(l)}, y^{(l)})), \quad (6)$$

where θ represents the learnable parameters. Note that although this loss involves a max operator that ranges over all possible matchings, computing the gradient of L with respect to θ only involves the best matching, which we can find by solving a linear program. It is also worth noting that this loss function is a simplified version of the structured SVM loss [25] with a margin of zero: although the ground truth matching needs to have the highest score among all possible matchings, the score does not need to be higher by a fixed positive margin than that of the best non-ground-truth matching. With no margin requirement, this loss is not as stringent as the original structured SVM formulation but is significantly easier to implement.

3.3 Partial Match Coreference

We now address the matching of relations. In this work, we consider “partial match coreference” (PC) relations, a specific type of semantic relations between phrases. Partial match coreference is composed of a noun phrase and another noun phrase including a possessive pronoun such as “his” or “her”. Such relations indicate that the second phrase refers to an entity that belongs to the first phrase. For instance, the following are examples of partial match coreference:

1. **A woman** is dressed in Asian garb with a basket of goods on **her hip**.
2. **An instructor** is teaching **his students** how to escape a hold in a self-defense class.

Partial match coreference points to a strong connection between the two noun phrases, which places constraints on the visual relation between their corresponding regions. In particular, partial match coreference relations can give strong cues about the appearance and spatial arrangement of the corresponding regions. We thus use PC relations in the task of phrase localization and study if it can bring any improvements.

We extract partial match coreference relations using the Stanford CoreNLP library [26]. Since a partial match coreference is indicated by possessive pronouns such as “his” or “her”, we first extract coreference relations between entity mentions in each image caption.

Among the extracted coreferences, some are “full matches”, such as “he” as a coreference of “a man” where the entire phrase “he” and the entire phrase “a man” are mutual coreferences. The rest are “partial matches” such as “her hip”

as a coreference of “a woman” where only a part of the phrase “her hip”, i.e. the possessive pronoun “her”, is a coreference of “a woman”. We discard all “full match” coreferences and keep only the “partial match” coreferences. Note that full match coreference is also useful because it indicates that two phrases should be matched to the same region. We discard them only because in Flickr30K Entities, the dataset we use for experiments, all pronouns that are full match coreferences are annotated as non-visual objects and excluded in evaluation.

3.4 Structured Matching with Relation Constraints

Given a certain type of semantic relations between phrases, we would like to enforce the constraint that the visual relations between regions should agree with these semantic relations. In the rest of this paper we use the partial coreference relation as an example.

Formally, consider two arbitrary phrases p_i, p_s . Let r_j, r_t be two arbitrary regions, which are potential matches for the two phrases. Given a matching y , let $z_{ijst} \in \{0, 1\}$ be a binary variable indicating whether phrases p_i and p_s are simultaneously matched to regions r_j and r_t . In other words,

$$z_{ijst} = y_{ij} \wedge y_{st}. \quad (7)$$

Let $g(r_j, r_t)$ be a non-negative function that measures whether two regions r_j, r_t have a visual relation that agrees with the partial match coreference (PC) relation, that is, whether the two regions have a “visual PC” relation.

We can now modify our matching objective to encourage the agreement between relations:

$$\max_y \sum_{i=1}^n \sum_{j=1}^m w_{ij} y_{ij} + \lambda \sum_{(i,s) \in Q} \sum_{j,t} z_{ijst} g(r_j, r_t), \quad (8)$$

where Q is the set of all pairs of phrases with PC relations. The term $z_{ijst} g(r_j, r_t)$ makes a matching y more desirable if whenever a pair of phrases have a PC relation the corresponding pair of regions have a visual PC relation.

This new objective poses additional challenges for finding the best matching. It is an integer program that appears difficult to solve directly, and it is not obvious how to relax it to a linear program with the Boolean term $z_{ijst} = y_{ij} \wedge y_{st}$.

We propose a linear program relaxation by introducing a probabilistic interpretation. We relax the binary variables y and z into real values in $[0, 1]$. We imagine that the matching is generated through a probabilistic procedure where each phrase p_i chooses, not necessarily independently, a region from all regions according to a multinomial distribution parametrized by the relaxed, continuous variables y_{ij} . That is, we interpret y_{ij} as $\Pr(R(p_i) = r_j)$, where $R(p_i)$ represents the region chosen by phrase p_i . This interpretation naturally requires that

$$\sum_j y_{ij} = \sum_j \Pr(R(p_i) = r_j) = 1, \quad (9)$$

which is the same constraint used earlier in bipartite matching that requires a phrase to match with exactly one region. We also add the exclusivity constraint that each region is matched to no more than one phrase:

$$\sum_i y_{ij} = \sum_i \Pr(R(p_i) = r_j) \leq 1. \tag{10}$$

We treat z_{ijst} as the joint probability $\Pr(R(p_i) = r_j, R(p_s) = r_t)$, that is, the probability that we match p_i to r_j and p_s to r_t simultaneously. It follows from the rule of marginalization that

$$\begin{aligned} \sum_{t=1}^m z_{ijst} &= \sum_t \Pr(R(p_i) = r_j, R(p_s) = r_t) = \Pr(R(p_i) = r_j) = y_{ij} \\ \sum_{j=1}^m z_{ijst} &= \sum_j \Pr(R(p_i) = r_j, R(p_s) = r_t) = \Pr(R(p_s) = r_t) = y_{st}. \end{aligned} \tag{11}$$

Putting all the constraints together we have the following linear program for structured matching:

$$\begin{aligned} \max_{y \in \mathcal{Y}} & \sum_{i=1}^n \sum_{j=1}^m w_{ij} y_{ij} + \lambda \sum_{(i,s) \in Q} \sum_{j,t} z_{ijst} g(r_j, r_t) \\ \text{s.t.} & \sum_{j=1}^m y_{ij} = 1, \text{ for } i = 1, 2, \dots, n \\ & \sum_{i=1}^n y_{ij} \leq 1, \text{ for } j = 1, 2, \dots, m \\ & \sum_{t=1}^m z_{ijst} = y_{ij} \text{ for any } i, j, s \\ & \sum_{j=1}^m z_{ijst} = y_{st} \text{ for any } i, s, t \\ & 0 \leq y_{ij} \leq 1, \text{ for all } i, j \\ & 0 \leq z_{ijst} \leq 1, \text{ for all } i, j, s, t. \end{aligned} \tag{12}$$

In this linear program, each pair of phrases with a partial match coreference relation p_i, p_s will lead to n^2 instances of z and g . This means that the linear program may have too many variables to be solved in a reasonable amount of time. To remedy this issue we adopt a heuristic that only applies the relation constraints to a subset of regions that are the most likely to be matched to phrases. Specifically, for a pair of phrases p_i and p_s , we only introduce z variables for the top 10 regions of p_i and the top 10 regions of p_s as measured by the cosine similarity. That is, the index j in z_{ijst} ranges over only the top 10 most similar regions of p_i and the index t ranges over only the top 10 most similar regions of p_s . This heuristic helps avoid a bloated linear program.

This linear program is easy to solve but is not guaranteed to produce an integer solution. A fractional solution indicates multiple possible regions for some

phrases. In such cases we run a depth first search to enumerate all feasible solutions contained in this fractional solution and find the best matching. Since we limit the number of z variables, the search space is usually small and our approach remains efficient.

To learn or fine-tune parameters for structured matching we use the same loss function as defined in Eq. 6, except that the matching score S is given by the solution value of this new linear program.

We implement the “visual PC” scoring function $g(r_j, r_t)$ as a logistic regressor. For a pair of regions r_j, r_t , we concatenate their fc7 feature vectors and pass the longer feature into the logistic regressor. The parameters of this logistic regressor can be learned jointly with all other parameters of our method.

4 Experiments

Setup. We evaluate our approach using the Flickr30k Entities dataset [1]. Flickr30k Entities is built on Flickr30K, which contains 31,783 images, each annotated with five sentences. In each sentence, the noun phrases are provided along with their corresponding bounding boxes in the image. These region-to-phrase correspondences enable the evaluation of phrase localization. There are more than 500k noun phrases (a total of 70k unique phrases) matched to 275k bounding boxes. Following [1], we divide these 31,783 image into three subsets, 1,000 images for validation, 1,000 for testing, and the rest for training. Also following [1], if a phrase is matched with multiple ground truth bounding boxes, we merge them into a new enclosing box. After this merging, every phrase has one and only one ground truth bounding box.

Following prior work [1–3], we generate 100 region proposals for each image using Edgebox [20] and localize each phrase by selecting from these regions. We select one region for each phrase and the selection is deemed correct if the region overlaps with the ground truth bounding box with an IoU (intersection over union) over 0.5. We evaluate the overall performance in terms accuracy, the percentage of phrases that are correctly matched to regions. Note that some prior works [1, 2] have reported performance in terms of recall@K: each phrase can select K regions; recall@K is 1 if one of them overlaps with the ground truth and 0 otherwise. Our definition of accuracy is the same as recall@1. We do not report recall@K with a K larger than 1 because it is unclear what it means to select more than one region for each phrase when we jointly localize phrases subject to the exclusivity constraints and relation constraints.

We use the same evaluation code released by [1]. It is also worth noting that since each phrase can only be localized to one of the region proposals, the quality of the region proposals establishes an upperbound of performance. Consistent with prior work, with EdgeBox the upperbound in our implementation is 76.91%.

Implementation. We implement the following approaches:

1. *CCA+Fast-RCNN*: We produce CCA embeddings using the code from Klein et al. [10] except we replace the VGG features pretrained on ImageNet with

the Fast-RCNN features (VGG features pretrained on ImageNet and fine-tuned on the VOC2007 detection dataset).

2. *Bipartite Matching*: using the same features as CCA+Fast-RCNN, we embed the features into a common vector space through (shallow) neural networks and perform bipartite matching with exclusivity constraints.
3. *Structure Matching*: Same as Bipartite Matching except we perform structured matching with relation constraints.

In training we modify the set of candidate regions: for each image we start with the 100 region proposals from EdgeBox; then we remove those with an IoU larger than 0.5 and add ground truth bounding boxes. The reason for this modification is that the EdgeBox region proposals may not contain the ground truth matches for all phrases. This modification ensures that all ground truth matches are included and each phrase has only one ground truth region.

For all training we use Stochastic Gradient Descent (SGD) with a learning rate of $1e-4$. We decrease the learning rate slightly after each epoch. We use a momentum of 0.9 and a weight decay of 0.0005. The hyperparameter λ in the matching loss (Eq. 6) is selected on the validation set. For bipartite matching, we initialize the embedding matrices M_1, M_2 with Canonical Correlation Analysis (CCA) [27] and fine-tune all parameters end to end for 3 epochs, optimizing the matching loss defined in Eq. 6. Since CCA provides a good initialization, the matching loss converges quickly.

For structured matching, we pre-train the “visual PC” logistic regressor using the 10,325 pairs of regions in the training set that have a ground truth “visual PC” relation and an equal number of negative pairs of regions. This pre-trained logistic regressor has an accuracy of 78% on the validation set. Then we initialize all other parameters using the pre-trained bipartite matching model and fine-tune all parameters (including those of the logistic regressor) for 2 epochs optimizing the structured matching loss with relation constraints.

Table 1. Accuracy (Recall@1) of our approach compared to other methods. Results in parentheses were released after the submission of this paper for peer review and are concurrent with our work.

Methods	Accuracy (Recall@1)
CCA [1]	25.30
NonlinearSP [2]	26.70 (43.89)
SCRC [9]	27.80
GroundR [3]	29.02 (47.70)
MCB [28]	(48.69)
CCA [29]	(50.89)
Ours: CCA+Fast-RCNN	39.44
Ours: Matching	41.78
Ours: Structured Matching	42.08

Table 2. Performance of bipartite matching and structured matching on only phrases with partial match coreference (PC) relations.

Methods	Accuracy (Recall@1) on PC phrases only
Bipartite Matching	47.8
Structured Matching	49.3

Results. Table 1 summarizes our results and compares them with related work. It is worth noting that some of the results from related work are concurrent with ours as they were released after the submission of this paper for peer review. Table 3 provides accuracy of phrase localization for different categories of phrases. Figures 3 and 4 show qualitative results including success and failure cases.

Our results show that Fast-RCNN features leads to a large boost of performance, as can be seen by comparing the CCA result from [1] with our CCA+Fast-RCNN result. Similar results have also been reported in [29] and the latest version of [2].

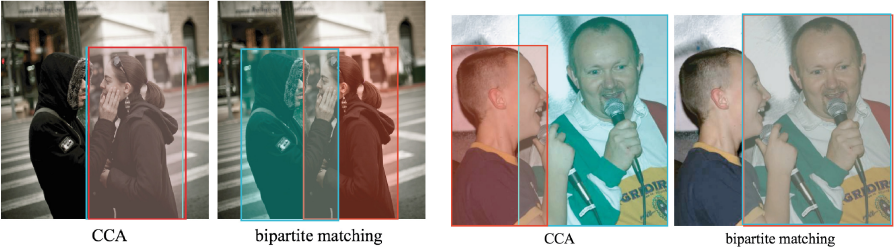
Also we see that Bipartite Matching further improves CCA+Fast-RCNN, which demonstrates the effectiveness of end-to-end training with our matching based loss function. Structured Matching with relation constraints provides a small additional improvement over Bipartite Matching. It is worth noting that the improvement from Structured Matching appears small partly because in the test set only 694 phrases out of a total of 17519 are involved in partial match coreference relations, limiting the maximum possible improvement when averaged over all phrases. If we consider only these 694 phrases and their accuracy as shown in Table 2, we see that Structured Matching achieves a more significant improvement over Bipartite Matching.

Table 3. Performance within categories. “Upperbound” is the maximum accuracy (recall@1) possible given the region proposals. Results in parentheses were released after the submission of this paper for peer review.

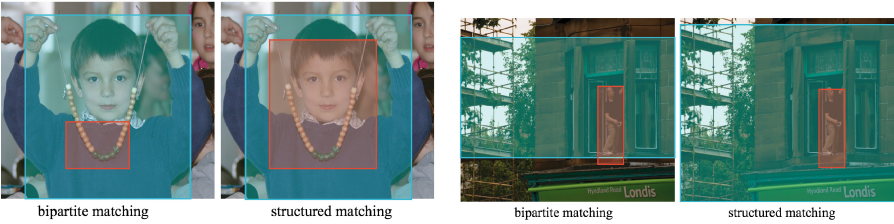
Methods	Person	Clothing	Body parts	Animals	Vehicles	Instruments	Scene	Other
CCA [1]	29.58	24.20	10.52	33.40	34.75	35.80	20.20	20.75
GroundR [3]	44.24	9.93	1.91	45.17	46.00	20.99	30.20	16.12
	(53.80)	(34.04)	(7.27)	(49.23)	(58.75)	(22.84)	(52.07)	(24.13)
CCA [29]	(64.73)	(46.88)	(17.21)	(65.83)	(68.75)	(37.65)	(51.39)	(31.77)
Ours: CCA+FRCN	55.39	32.78	16.25	53.86	48.50	19.14	28.97	23.56
Ours: Bipartite	57.94	34.43	16.44	56.56	51.50	27.16	33.42	26.23
Ours: Structured	57.89	34.61	15.87	55.98	52.25	23.46	34.22	26.23
Upperbound	89.36	66.48	39.39	84.56	91.00	69.75	75.05	67.40



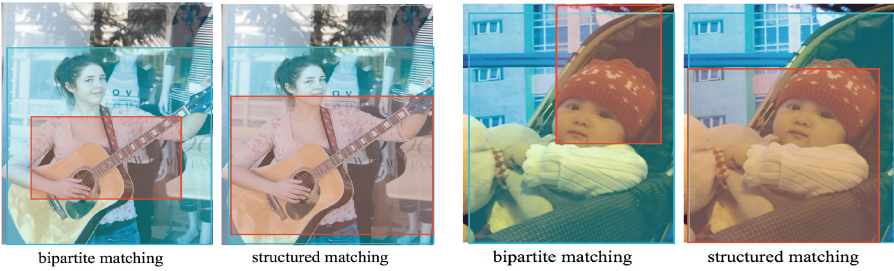
(a) A man with a helmet is using an ATM. (b) Two women and a man discuss notes in a classroom.



(c) A man wearing a black jacket with a woman wearing a black jacket are standing close to each other. (d) A man and a boy holding microphones.

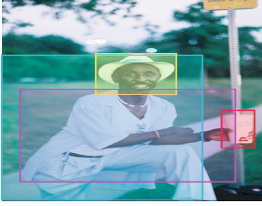


(e) A young boy shows his brown and green bead necklace. (f) A man is working on his house by repairing the windows.



(g) This lady is wearing a pink shirt and tuning her guitar. (h) A baby with red hat sit in his stroller.

Fig. 3. Qualitative results. The first two rows compare CCA with bipartite matching. The rest compare bipartite matching with structured matching. (Color figure online)



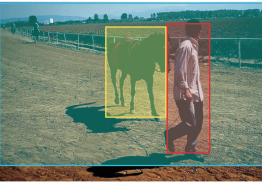
(a) A black man wearing a white suit and hat is holding a paper cup.



(b) A boy wearing an orange shirt is playing on a swing.



(c) A blond woman speaks at a podium labeled Holiday Stars Project in front of a blue wall.



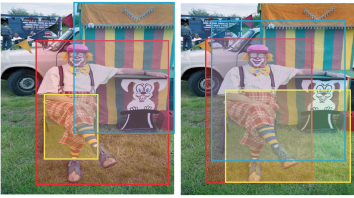
(d) A man is walking his horse on a racetrack.



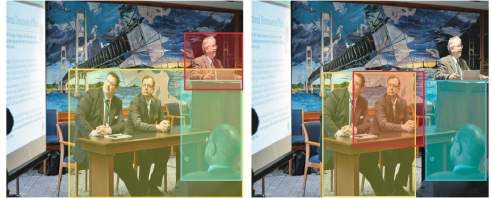
(e) A dark-haired bearded man wearing a turquoise shirt with a yellow peace sign on it.



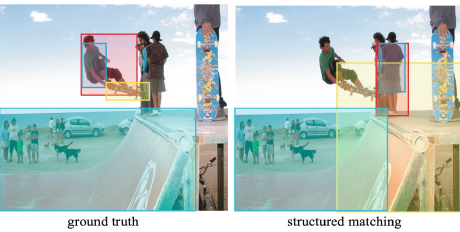
(f) A woman wearing a black helmet riding on a bike.



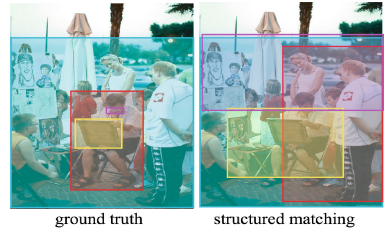
(g) A clown in red plaid overalls relaxes in the back of his tent.



(h) A man at a podium is speaking to a group of men at a conference.



(i) A man in a green shirt is jumping a ramp on his skateboard.



(j) Several people are standing on a street corner watching a cartoonist with glasses draw on his sketch pad.

Fig. 4. Qualitative results. The first two rows compare CCA with bipartite matching. The rest compare bipartite matching with structured matching. (Color figure online)

5 Conclusion

In this paper we have introduced a new approach to phrase localization. The key idea is a structured matching of phrases and regions that encourages the relations between phrases to agree with the relations between regions. We formulate structured matching as a discrete optimization problem and relax it to a linear program. We integrate structured matching with neural networks to enable end-to-end training. Experiments on Flickr30K Entities have demonstrated the empirical effectiveness of our approach.

References

1. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2641–2649 (2015)
2. Wang, L., Li, Y., Lazebnik, S.: Learning deep structure-preserving image-text embeddings. arXiv preprint [arXiv:1511.06078](https://arxiv.org/abs/1511.06078) (2015)
3. Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T., Schiele, B.: Grounding of textual phrases in images by reconstruction. arXiv preprint [arXiv:1511.03745](https://arxiv.org/abs/1511.03745) (2015)
4. Karpathy, A., Joulin, A., Li, F.F.F.: Deep fragment embeddings for bidirectional image sentence mapping. In: Advances in Neural Information Processing Systems, pp. 1889–1897 (2014)
5. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3128–3137 (2015)
6. Kong, C., Lin, D., Bansal, M., Urtasun, R., Fidler, S.: What are you talking about? text-to-image coreference. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3558–3565 (2014)
7. Guadarrama, S., Rodner, E., Saenko, K., Zhang, N., Farrell, R., Donahue, J., Darrell, T.: Open-vocabulary object retrieval. In: Robotics: Science and Systems (2014)
8. Arandjelovic, R., Zisserman, A.: Multiple queries for large scale specific object retrieval. In: BMVC, pp. 1–11 (2012)
9. Hu, R., Xu, H., Rohrbach, M., Feng, J., Saenko, K., Darrell, T.: Natural language object retrieval. arXiv preprint [arXiv:1511.04164](https://arxiv.org/abs/1511.04164) (2015)
10. Klein, B., Lev, G., Sadeh, G., Wolf, L.: Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation. arXiv preprint [arXiv:1411.7399](https://arxiv.org/abs/1411.7399) (2014)
11. Gong, Y., Wang, L., Hodosh, M., Hockenmaier, J., Lazebnik, S.: Improving image-sentence embeddings using large weakly annotated photo collections. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part IV. LNCS, vol. 8692, pp. 529–545. Springer, Heidelberg (2014)
12. Hoi, S.C., Liu, W., Lyu, M.R., Ma, W.Y.: Learning distance metrics with contextual constraints for image retrieval. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006, vol. 2. IEEE pp. 2072–2078 (2006)
13. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint [arXiv:1411.2539](https://arxiv.org/abs/1411.2539) (2014)

14. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. arXiv preprint [arXiv:1411.4555](https://arxiv.org/abs/1411.4555) (2014)
15. Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A.C., Berg, T.L.: Baby talk: understanding and generating image descriptions. In: Proceedings of the 24th CVPR, Citeseer (2011)
16. Mao, J., Xu, W., Yang, Y., Wang, J., Yuille, A.: Deep captioning with multimodal recurrent neural networks (m-RNN). arXiv preprint [arXiv:1412.6632](https://arxiv.org/abs/1412.6632) (2014)
17. Donahue, J., Hendricks, L.A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. arXiv preprint [arXiv:1411.4389](https://arxiv.org/abs/1411.4389) (2014)
18. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguist.* **2**, 67–78 (2014)
19. Socher, R., Karpathy, A., Le, Q.V., Manning, C.D., Ng, A.Y.: Grounded compositional semantics for finding and describing images with sentences. *Trans. Assoc. Comput. Linguist.* **2**, 207–218 (2014)
20. Zitnick, C.L., Dollár, P.: Edge boxes: locating object proposals from edges. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 391–405. Springer, Heidelberg (2014). doi:[10.1007/978-3-319-10602-1_26](https://doi.org/10.1007/978-3-319-10602-1_26)
21. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
22. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 248–255. IEEE (2009)
23. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2007 (VOC 2007) Results(2007). <http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html>
24. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
25. Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y.: Large margin methods for structured and interdependent output variables. *J. Mach. Learn. Res.* **6**(Sep), 1453–1484 (2005)
26. Clark, K., Manning, C.D.: Entity-centric coreference resolution with model stacking. In: Association for Computational Linguistics (ACL) (2015)
27. Hardoon, D.R., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis: an overview with application to learning methods. *Neural Comput.* **16**(12), 2639–2664 (2004)
28. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multi-modal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint [arXiv:1606.01847](https://arxiv.org/abs/1606.01847) (2016)
29. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. arXiv preprint [arXiv:1505.04870v3](https://arxiv.org/abs/1505.04870v3) (2015)