Weakly Supervised Phrase Localization with Multi-Scale Anchored Transformer Network

Fang Zhao Jianshu Li Jian Zhao Jiashi Feng
National University of Singapore
elezhf@nus.edu.sg {jianshu,zhaojian90}@u.nus.edu elefjia@nus.edu.sg

Abstract

In this paper, we propose a novel weakly supervised model, Multi-scale Anchored Transformer Network (MATN), to accurately localize free-form textual phrases with only image-level supervision. The proposed MATN takes region proposals as localization anchors, and learns a multi-scale correspondence network to continuously search for phrase regions referring to the anchors. In this way, MATN can exploit useful cues from these anchors to reliably reason about locations of the regions described by the phrases given only image-level supervision. Through differentiable sampling on image spatial feature maps, MATN introduces a novel training objective to simultaneously minimize a contrastive reconstruction loss between different phrases from a single image and a set of triplet losses among multiple images with similar phrases. Superior to existing region proposal based methods, MATN searches for the optimal bounding box over the entire feature map instead of selecting a sub-optimal one from discrete region proposals. We evaluate MATN on the Flickr30K Entities and ReferItGame datasets. The experimental results show that MATN significantly outperforms the state-of-the-art methods.

1. Introduction

Textual phrase localization in images is a very challenging problem and has attracted extensive attention in recent years [21, 27, 12, 28, 22, 13, 4]. It plays an important role in text based image retrieval, human-robot interaction and visual question answering. Unlike object detection over several semantic classes, phrase localization aims to find the image region corresponding to a specified free-form textual phrase about the image content.

Most of existing phrase localization methods solve this problem through selecting regions from finite pre-computed region proposals (*e.g.*, Selective Search [26] and Edge Boxes [32]). The performance of those methods thus heavily depends on the quality of the region proposals. However, the region proposals generated heuristically are usually not accurate, which severely limits the performance. As shown in Fig. 1, none of the candidate bounding boxes from region proposals corresponds to the ground truth and thus existing methods cannot accurately localize the phrase. Although some methods rely on region proposal network (RPN) to produce more accurate proposal bounding boxes, such as [13, 4], they are *fully supervised* and require bounding box annotations of phrases to learn RPN. Most of image-sentence datasets, such as MSCOCO [18] and Flickr30K [30], also have no such location information about phrases mentioned in the sentence. Manually annotating location of each phrase is very time-consuming and labor intensive, which hampers those fully-supervised methods from being scalable to larger datasets.

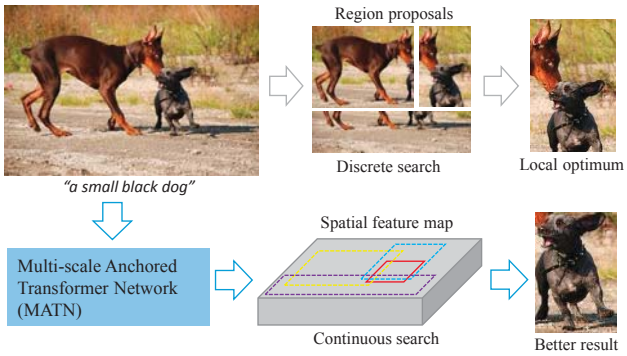

Figure 1. Main idea of the proposed Multi-scale Anchored Transformer Network (MATN). Using only image-level supervision, MATN performs continuous search over the entire spatial feature map by taking region proposals as anchors to offer more accurate phrase localization. This is different from existing weakly supervised methods that only perform selection over these finite discrete region proposals and also different from region proposal network (RPN) based methods that need bounding box annotations of phrases as supervised information.

To address these critical problems in phrase localization, we propose a novel weakly supervised model, namely Multi-scale Anchored Transformer Network (MATN)

(Fig. 1), which can search over the entire spatial feature map continuously to more accurately localize phrases using only image-level supervision. Inspired by the spatial transformer module [14], MATN predicts affine transformation parameters of region for a phrase by establishing multi-scale spatial correspondence between the phrase and image feature maps. Furthermore, MATN takes region proposals, such as Edge Boxes [32] that can be obtained very cheaply, as spatial position regularization for anchoring the prediction such that it can effectively alleviate the contamination of complex image background for localization. To optimize MATN, we propose a novel training strategy that encourages MATN to minimize a contrastive reconstruction loss between different phrases from a single image to produce more discriminative regions, and also minimize a set of triplet losses among multiple images with the similar phrases to explicitly leverage the shared knowledge across images.

Our main contributions include the following four aspects. 1) A novel Multi-scale Anchored Transformer Network is proposed to localize phrases in images without requiring any region-level strong supervision. The model can search for fine-grained bounding boxes continuously over the spatial feature maps instead of selecting only from bounding box candidates, thus offering appealing robustness to errors in region proposals. 2) A new training strategy is introduced that enables the model to learn to exploit the discrimination of different phrases and shared knowledge from similar images. 3) An anchored transformation is developed that exploits region proposals as a spatial position constraint to facilitate searching new regions. 4) Our proposed model boosts the benchmark of weakly supervised phrase localization, achieving new state-of-the-art performance on the Flickr30K Entities and ReferitGame datasets.

2. Related Work

Phrase Localization/Grounding. Several works have studied the textual phrase localization/grounding. [23] introduced visual phrases and a multiple detection decoding algorithm that considers properties of interacting objects in different levels of abstraction, which is the earliest work about phrase localization. Recently, [21] presented a region-to-phrase dataset, namely Flickr30K Entities, and gave a baseline by using a CCA model to learn a shared semantic space that associates phrases to image regions. [27] proposed a two-branch neural network to learn joint embeddings of image regions and phrases. The network is optimized using a large-margin objective that preserves both within-view and cross-view feature space structures. [31] formulated the top-down attention of a CNN classifier as a probabilistic Winner-Take-ALL process and utilized an excitation backprop scheme to pass along top-down signals downwards in the network hierarchy. [22], whose work is most related with ours, proposed to ground a phrase by using a soft attention model to weight feature vectors of region proposals for phrase reconstruction. In contrast to [22], our method is based on a fine-grained searching instead of discretely selecting.

Weakly Supervised Object Localization. This task aims to use only image level labels to detect objects in images without object bounding box annotations. [5] followed a multiple-instance learning approach that iteratively trains the detector and infers the object locations in the positive training images. [3] modified a pretrained deep convolutional neural network to operate at the level of image regions, which performs region selection and classification simultaneously. [15] addressed this problem by introducing two types of context-aware guidance models that leverage their surrounding context regions to improve localization. Although these works do not use ground truth bounding boxes for training either, they only consider a limited object class set, such as dog, cat and person. By contrast, our method can handle any form phrases in the training and test process.

Image Captioning. Image captioning focuses on the whole image and produces its textual description. [6] developed a recurrent convolutional model for large-scale visual learning which is end-to-end trainable and successfully applied to image captioning. [16] used a deep multi-modal embedding model for bidirectional retrieval of images and sentences and learnt a common space for fragments of images and sentences. Like [12], phrase localization can be implemented through applying the image captioning methods to image regions and computing scores on phrases.

Visual Attention. Our model utilizes a differentiable attention mechanism which is extended from spatial transformer [14]. In [14], a spatial differentiable transformation is applied to a feature map during the forward pass of the convolutional neural network to allow the spatial manipulation of data within the network. There exist different attention mechanisms which are proposed for computer vision tasks. [29] introduced a spatial attention based model including both soft and hard attentions, which automatically learns to attend to salient objects while predicting the corresponding words in the caption. [9] proposed to generate images using a sequential variational Auto-encoder model, which allows for the iterative construction of complex images through imitating the foveation of the human eye.

3. The Proposed Model

Given an image and a textual phrase, our goal is to search for the region corresponding to the specified phrase over the spatial feature map of the image. Fig. 2 illustrates the framework of our proposed MATN. A base convolutional neural network (CNN) is used to obtain the spatial feature map. Then a multi-scale correspondence network (MCN) is introduced to estimate affine transformation parameters

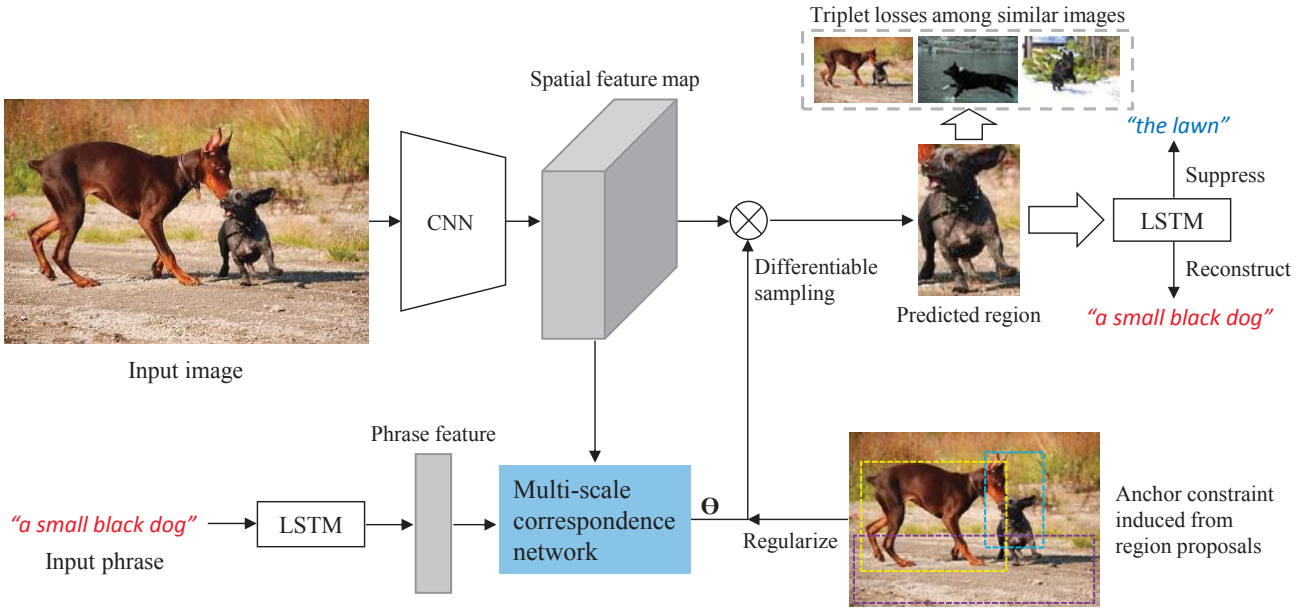

Figure 2. Framework of the proposed Multi-scale Anchored Transformer Network. It consists of a multi-scale correspondence network and an anchor constraint induced from a set of region proposals. A contrastive reconstruction loss of different phrases and a set of triplet losses among similar images are devised to train the model by differentiable feature map sampling. More details on each component are given in the texts and illustrated in following figures.

of region under an anchor constraint induced from region proposals. By differentiable sampling [14] over the spatial feature map, a contrastive reconstruction loss of different phrases associated with an image and a set of triplet losses computed w.r.t. multiple images with similar phrases are used to train the proposed MATN. We now proceed to explain each component of MATN in details.

3.1. Multi-scale Correspondence Network

The purpose of multi-scale correspondence network is to establish correspondence between phrases and image regions, laying foundations for the following exact phrase localization. As mentioned in the Related Work section, spatial transformer [14] is a differentiable module that can be inserted into convolutional architectures to learn spatial transformation over feature maps. However, the original spatial transformer only produces transformation parameters for region sampling, and thus cannot be applied for phrase localization which needs to predict the region associated with the phrase.

As shown in Fig. 3, we introduce a new multi-scale correspondence network (MCN) to learn regional transformation parameters through computing the correlation scores between the phrase and the spatial feature map of the image. To build such a correspondence network, given an input image of size $W \times H$, we first use the base CNN to obtain its feature map of size $W' \times H' \times C$. Such a feature map encodes appearance of the image and preserves valuable spatial information. We also add several extra convolutional layers on the input feature map to account for multiple scales to capture wider context information (see the Implementation section for the specific layer configuration). Given an input phrase with T words, we represent each word in the phrase as a one-hot vector and embed it into a lower dimensional feature vector by a fully connected layer. Then we use a Long-Short Term Memory (LSTM) network [10] to encode the embedded word sequence and use the hidden state h_T at the time step T as a feature representation of the phrase.

To obtain the correspondence map between the phrase and the spatial feature map, we tile and separately concatenate the phrase representation h_T to the feature vector $h_{i,j,s}$ at each scale s and spatial location (i, j) of the feature map giving a local descriptor for this location. Then taking the concatenated feature as the input, we compute the correspondence map $\mathbf{z}_s$ containing correlation scores $\{z_{i,j,s}\}$ through a two-layer fully connected network

$$z_{i,j,s} = \sigma(W_2\sigma(W_1[h_T; h_{i,j,s}])), \quad (1)$$

where σ is the ReLU function. This is implemented as two 1×1 convolutional layers with stride 1 in practice. The correspondence maps $\mathbf{z}_s$ of multiple scales then respectively go through a liner regression layer and are added up to produce the affine transformation parameters $\theta = (s_x, s_y, t_x, t_y)$,

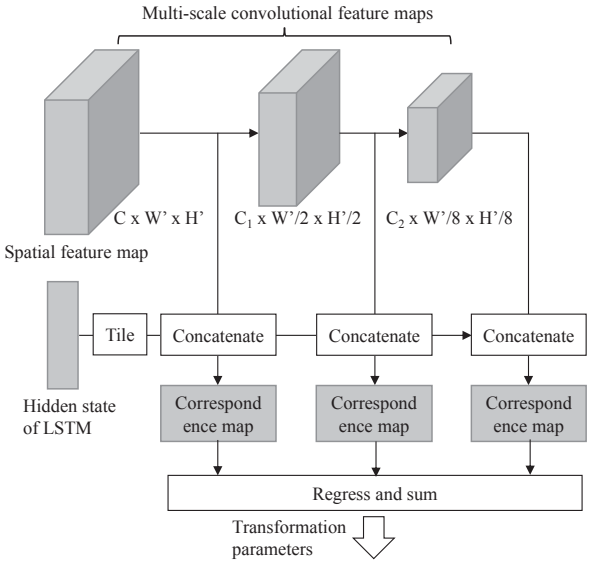

Figure 3. Architecture of multi-scale correspondence network (MCN). The spatial correspondence maps of multiple scales capture the correlation between the phrase and objects of different sizes.

where (s_x, s_y) and (t_x, t_y) represent respectively scale and translation along x-axis and y-axis. Finally, through differentiable sampling, the transformation parameters θ are applied on the spatial feature map to obtain the feature map of the corresponding region. The form of the affine transformation is given by

$$\begin{pmatrix} x_{i,j} \\ y_{i,j} \end{pmatrix} = \begin{pmatrix} s_x & 0 & t_x \\ 0 & s_y & t_y \end{pmatrix} \begin{pmatrix} x^r_{i,j} \\ y^r_{i,j} \\ 1 \end{pmatrix}, \qquad (2)$$

where $(x^r_{i,j}, y^r_{i,j})$ is the coordinate of spatial location (i, j) of the region feature map, and similarly $(x_{i,j}, y_{i,j})$ is the coordinate in the input feature map which defines the sample point. All coordinates are normalized and belong to $[-1, 1]$ when within the spatial bounds of feature maps.

3.2. Anchor Constraint with Region Proposals

One straightforward approach for localizing phrases is to employ the attention mechanism to learn the MCN described above. However, it is hard to guarantee that the prediction of the network converges to the desired region due to distracting factors presented in natural images like multiple objects and complex scenes. It is worth noting that region proposals, such as Edge Boxes [32], provide useful cues for the network to localize regions with high objectness. Thus we consider exploiting region proposals to gain additional spatial position guidance to alleviate the difficulties caused by scarce supervision information.

Concretely, given a set of region proposals $\{r_n\}_{n=1}^N$ for the target image, we introduce an anchor constraint induced from the spatial position of $\{r_n\}_{n=1}^N$ to regularize the regression for the transformation parameter θ. Specifically, we take $\{r_n\}_{n=1}^N$ as N anchors and enforce the predicted bounding box to be close to its nearest anchor. Here we only allow one of the anchors to affect the parameters regression because considering all of them at the same time would result in a meaningless average position over region proposals. Therefore, we define the anchor-based regularization term as

$$R_{\text{anchor}} = \left\| \theta - \underset{\mathbf{p} \in \{\mathbf{p}(r_n)\}_{n=1}^N}{\arg\min} \|\mathbf{p} - \theta\|_2^2 \right\|_2^2, \qquad (3)$$

where $\mathbf{p}(r_n)$ is the transformation parameter converted from the position of the region proposal r_n with center (x_n, y_n), width w_n, and height h_n by

$$\mathbf{p}(r_n) = \left[\frac{w_n}{W}, \frac{h_n}{H}, \frac{2x_n}{W} - 1 + \frac{w_n}{W}, \frac{2y_n}{H} - 1 + \frac{h_n}{H} \right]. \qquad (4)$$

Eqn. (3) can be seen as a soft constraint which enables MCN to focus on several possible regions containing the object described by the phrase. That is, the predicted bounding box is not necessarily one of the region proposals but can be located at a better position around them. Here the region proposals behave like anchors to keep the predicted position not far away from them.

Compared with selecting from discrete bounding box candidates, our model can explore more regions because it performs localization over the entire spatial feature map and meanwhile incorporates the prior position information from the bounding box candidates, *i.e.*, region proposals. In the case that the ground truth is partially or even entirely uncovered by these bounding box candidates, it is impossible to predict correctly for the region-selection based methods. In contrast, MATN is able to address the problem through refining the positions of these candidates based on differentiable transformation and anchor constraint.

3.3. Learning by Discrimination and Similarity

To train the proposed MATN, we first introduce a contrastive reconstruction loss consisting of two parts. One is to encourage the feature map $\mathbf{h}$ sampled by the predicted transformation parameters θ to be able to reconstruct the input phrase, and the other is to suppress the reconstruction of a different phrase from the same image for $\mathbf{h}$. In this way, the model can learn to predict a discriminative region for the specified phrase. Similar to the phrase encoding, we use LSTM to model the distribution of the reconstructed phrase. Then the loss is defined as the difference between

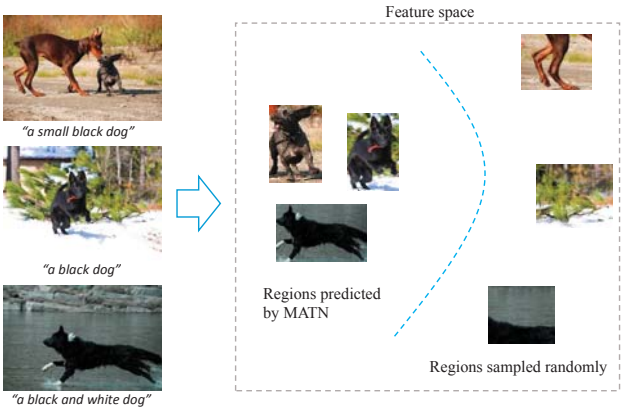

Figure 4. A set of triplet losses computed w.r.t. images with similar phrases. It is used to enforce regions predicted by MATN to be closer to each other than regions sampled randomly in a feature space.

the negative log likelihoods of the pair of phrases:

$$L_{\text{recon}} = -\sum_{t=1}^{T} \log P_t(\mathbf{w}_t|\mathbf{h}) + \lambda \sum_{t=1}^{T} \log P_t(\mathbf{w}_t^{\text{diff}}|\mathbf{h}), \tag{5}$$

where $\mathbf{w}_t$ and $\mathbf{w}_t^{\text{diff}}$ are the tth word in the input phrase and a different phrase respectively, and λ controls the effect of the contrastive term. To measure the semantic similarity between phrases, we compute their cosine distance based on the word2vec model [20]. Here we simply use the bag-of-words to encode words in a phrase, *i.e.*, pooling all the word vectors via sum to a phrase vector. Then a threshold of cosine distance is set to determine whether two phrase are similar or not.

Although the phrase reconstruction provides useful supervision information in the absence of strong region-level supervision, only using the semantic information of phrases from a single image to train MATN to learn the corresponding spatial transformation is insufficient due to complex content of natural images. Inspired by the object co-localization [25, 2], we consider leveraging the shared knowledge across multiple images with similar phrases to jointly optimize MATN. That is, these images should contain the common or similar object. As illustrated in Fig. 4, given a set of images containing objects described by similar phrases, *e.g.*, "*a small black dog*", "*a black dog*" and "*a black and white dog*", we propose to train MATN to make sure regions of those images predicted by MATN are closer to each other than regions sampled randomly in a feature space. To this end, we first use the transformation parameters θ obtained by MCN and a random θ^{rand} to sample regions from the spatial feature map and feed the sampled regional feature maps into a two-layer fully connected network $\mathbf{f}(.)$ to obtain feature vectors of the image regions. Then a set of triplet losses is computed by

$$L_{\text{triplet}} = \frac{1}{N(N-1)} \sum_{n=1}^{N} \sum_{m \neq n} \Big[||\mathbf{f}(\mathbf{h}_n) - \mathbf{f}(\mathbf{h}_m)||_2^2 - ||\mathbf{f}(\mathbf{h}_n) - \mathbf{f}(\mathbf{h}_m^{\text{rand}})||_2^2 + \rho \Big]_+, \tag{6}$$

where N is the image number of the set, $[.]_+ = \max(0, .)$, $\mathbf{h}$ and $\mathbf{h}^{\text{rand}}$ are the feature maps of predicted and random regions respectively, and ρ is a margin parameter (simply set to 1 here). The similarity between phrases is also measured by the cosine distance of their word2vec vectors. Here a subset of region proposals of an image is selected further as the anchors to speed up the convergence of the model, which only contains region proposals which are visually similar with those of other images in the set because dissimilar proposals obviously are not the ground truth.

Finally, the objective function of MATN is given by

$$L = L_{\text{triplet}} + \alpha L_{\text{recon}} + \beta R_{\text{anchor}}, \tag{7}$$

where α and β are weighting parameters. It can be optimized by the standard back-propagation algorithm in an end-to-end way as all the terms are differentiable.

Applying MATN for inference is straightforward. Firstly MCN takes features of both the testing image and phrase as inputs to predict affine transformation parameters θ. Then the bounding box of the phrase can be obtained by applying Eqn. (2) to the image, where the anchor constraint can be removed because the prior position information has been learned by MCN. Thus our model is more efficient for inference compared with previous methods, such as [22] and [31], because it gets rid of generating region proposals.

4. Experiments

We test the proposed MATN on two challenging image-sentence datasets for phrase localization, *i.e.*, Flickr30K Entities [21] and ReferItGame [17]. We present quantitative evaluations in terms of accuracy against different IoU (Intersection over Union) thresholds, and compare our model with the state-of-the-art weakly supervised phrase localization methods, *i.e.*, GroundeR [22] and c-MWP [31]. We also compare with other recent methods including the image captioning methods, *i.e.*, Deep Fragments [16] and LRCN [6], and the object classification method, *i.e.*, CAFFE-7K [11], which can be applied to phrase localization and also do not need the bounding box annotations of the phrases for training.

4.1. Datasets

The Flickr30k Entities dataset [21] is an extension of the Flickr30K dataset [30]. It associates captions of 31K images with 276K manually annotated bounding boxes and

Table 1. Accuracy (IoU > 0.5) of phrase localization on the Flickr30k Entities dataset.

Methods	Accuracy (IoU > 0.5)
Deep Fragments [16]	21.78
c-MWP (MCG) [31]	26.20
c-MWP (EB) [31]	27.00
GroundeR (VGG-CLS) [22]	24.66
GroundeR (VGG-DET) [22]	28.94
Ours: MATN-SC-RE	31.15
Ours: MATN-RE	32.61
Ours: MATN	**33.10**

thus makes the evaluation of phrase localization available. Similar to [22], we divide the dataset into three subsets, 1,000 images for validation, 1,000 for testing and the remaining images for training. Following [21], if multiple ground truth bounding boxes correspond to a single phrase (*e.g.*, a group of people), we use the union of the boxes to represent the phrase. The ReferItGame dataset [17] contains 20K images and 120K annotated descriptions collected in a two player game for image regions obtained from the segmentation regions in the SAIAPR-12 dataset [7]. We use the same dataset split with [22], *i.e.*, 10K images for testing, 1,000 for validation and the rest for training.

4.2. Implementation Details

In our experiments, we adopt the VGG-16 network [24] as our base network. MCN has the same architecture as conv5_3, fc7 and conv8_2 in the SSD network [19]. All input images are resized to 480×480. Similar to [22], the CNN parts are pretrained for the task of object detection on PASCAL [8] and fixed in the training process. We generate 100 region proposals for each image using Edge Boxes [32] to obtain the anchor constraint. To measure the visual similarity between region proposals, we compute the normalized L2 distances of their feature vectors, and select region proposals with top-K smallest distances as anchors. K is 20 for Flickr30K Entities and 50 for ReferItGame. For those images with unique phrases, we only use the contrastive reconstruction loss.

For the language model, the dimension of the embedding vector is 512 and the size of the LSTM memory is 512 for both phrase encoding and decoding. The word2vec model is trained on part of Google News dataset (about 100 billion words) and contains 300-dimensional vectors for 3 million words and phrases. The threshold of the cosine distance is set to 0.9 and 0.8 respectively for the judgment of similar and different phrases.

Stochastic gradient descent with RMSProp is used to optimize the network parameters. The learning rate is 0.0001, the RMS decay is 0.99 and the weight decay is 0.0005. At each iteration, we choose 4 images to construct a set of triplet losses. We set $\lambda = 0.5$ for the contrastive term. The weighting parameters α and β are set to 0.1 and 0.5, respectively. All the hyperparameters are obtained according to the evaluation on the validation sets.

4.3. Results

We report the phrase localization results in terms of accuracy, *i.e.*, the percentage of phrases correctly matched with regions. Here the predicted region for each phrase is deemed correct if the region overlaps with the ground truth bounding box with an IoU larger than a threshold.

4.3.1 Ablation Study

We first evaluate the contributions of some key components in our MATN model in Table 1 and 3 by examining several variants including (1) only using single scale and the general reconstruction loss (MATN-SC-RE) and (2) using multi-scale but without the contrastive reconstruction loss (MATN-RE). As one can see that MATN-RE performs better than MATN-SC-RE because multi-scale correspondence maps can capture the correlation of the phrase and objects of different sizes. Through encouraging the predicted region to be less relevant to other phrases, MATN outperforms MATN-RE because the predicted bounding boxes tend to be more discriminative for the input phrase.

Note that the objective function (7) contains three components: contrastive reconstruction loss of different phrases, triplet loss sampled from similar images and anchor constraint. For the two losses, we have tried to train the model using only one of them, but it did not converge because natural images contain multiple objects and complex scenes, and any one of the losses cannot individually provide sufficient guidance for the transformer network. Similarly, as described in Sect. 3.2, anchor constraint is also a key component. It cannot be removed, otherwise training of the model would not converge.

4.3.2 Flickr30k Entities

On the Flickr30k Entities dataset, we compare the proposed MATN with GroundeR [22], c-MWP [31] and Deep Fragments [16]. GroundeR and c-MWP are the state-of-the-art weakly supervised phrase localization methods. GroundeR (VGG-CLS) is pretrained for the image classification on ImageNet and GroundeR (VGG-DET) is pretrained for the object detection on PascalVOC. c-MWP (MCG) and c-MWP (EB) use MCG [1] and Edge Boxes respectively to generate region proposals. Deep Fragments is a recent image captioning method, which is trained on Flickr30k and evaluated with the ground truth phrases and bounding boxes of Flickr30k Entities. Here we use its result reported in [22].

Table 2. Accuracy (IoU > 0.5) of phrase localization for different phrase types on the Flickr30k Entities dataset.

Methods	People	Clothing	Body parts	Animals	Vehicles	Instruments	Scene	Other
GroundeR (VGG-CLS) [22]	36.01	9.54	0.76	24.13	32.50	15.43	**37.00**	13.43
GroundeR (VGG-DET) [22]	44.32	9.02	0.96	46.91	**46.00**	19.14	28.23	16.98
MATN	**54.71**	**13.38**	**2.87**	**58.21**	45.04	**19.48**	21.97	**17.02**

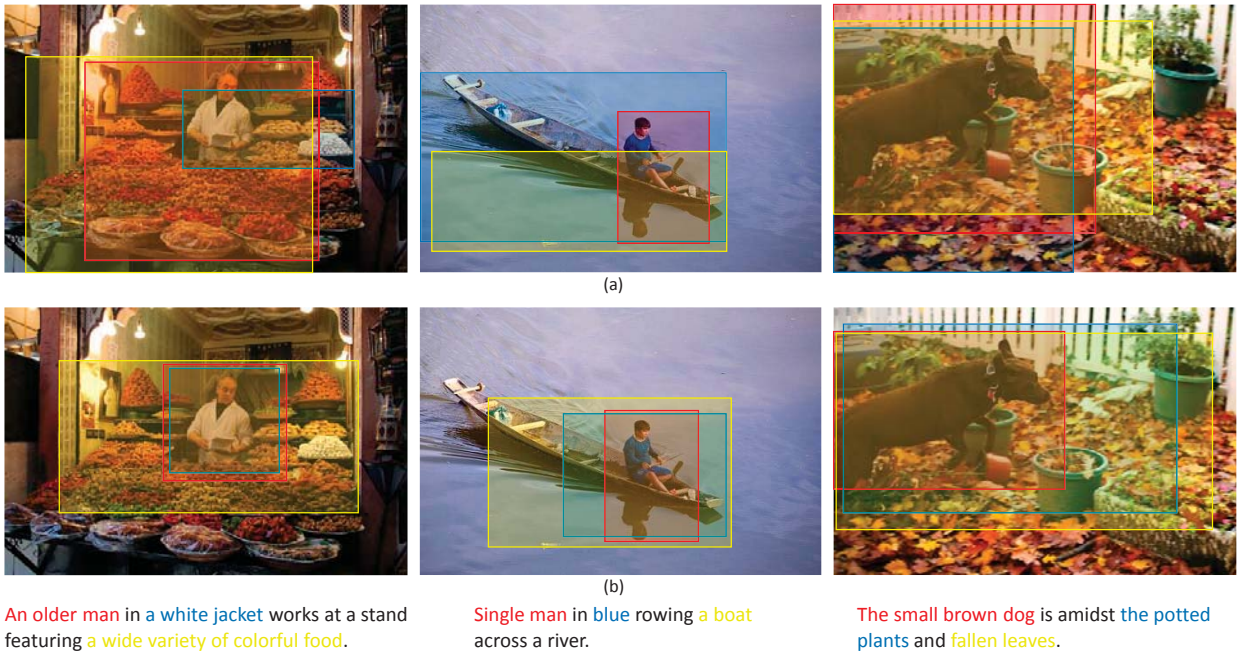

Figure 5. Qualitative results of (a) GroundeR (VGG-DET) and (b) MATN on the Flickr30K Entities dataset. Each phrase of a sentence is localized in an image using different color boxes (best viewed in color).

Table 3. Accuracy (IoU > 0.5) of phrase localization on the ReferItGame dataset.

Methods	Accuracy (IoU > 0.5)
LRCN [6]	8.59
CAFFE-7K [11]	10.38
GroundeR (VGG) [22]	10.69
GroundeR (VGG-SPAT) [22]	10.70
Ours: MATN-SC-RE	12.13
Ours: MATN-RE	13.30
Ours: MATN	**13.61**

Table 1 reports the accuracy of phrase localization for these methods under the condition of IoU > 0.5. We can see that our proposed MATN achieves the best performance and surpasses all the state-of-the-arts (*i.e.*, GroundR and c-MWP) with a large margin ($> 4.0\%$). It demonstrates effectiveness of searching over the entire spatial feature map by referring to the anchors from region proposals, which can find fine-grained bounding boxes compared with these region-selection based methods. Table 2 shows the accuracy of phrase localization for different types of phrases. It is worth noting that our results are better than other methods for most phrase types, especially for "People" and "Animals". As for the phrase type "Scene", it usually contains entire images, thus GroundeR (VGG-CLS), which uses the VGG classification network trained on entire images, is more suitable to handle this phrase type. Fig. 5 presents some qualitative results compared with GroundeR (VGG-DET) on the Flickr30K Entities dataset. We visualize each phrase of a sentence in an image using different color boxes. It can be observed that the bounding boxes predicted by MATN are more precise than GroundeR (VGG-DET). The relative large objects like"the small brown dog" can be localized better than the small objects like "a white jacket ", which is consistent with the quantitative results in Table 2, because there is no strong location constraint such as bounding box annotations in the weakly supervised scenario.

4.3.3 ReferItGame

On the ReferItGame dataset, we compare the proposed MATN with GroundeR, LRCN [6] and CAFFE-7K [11].

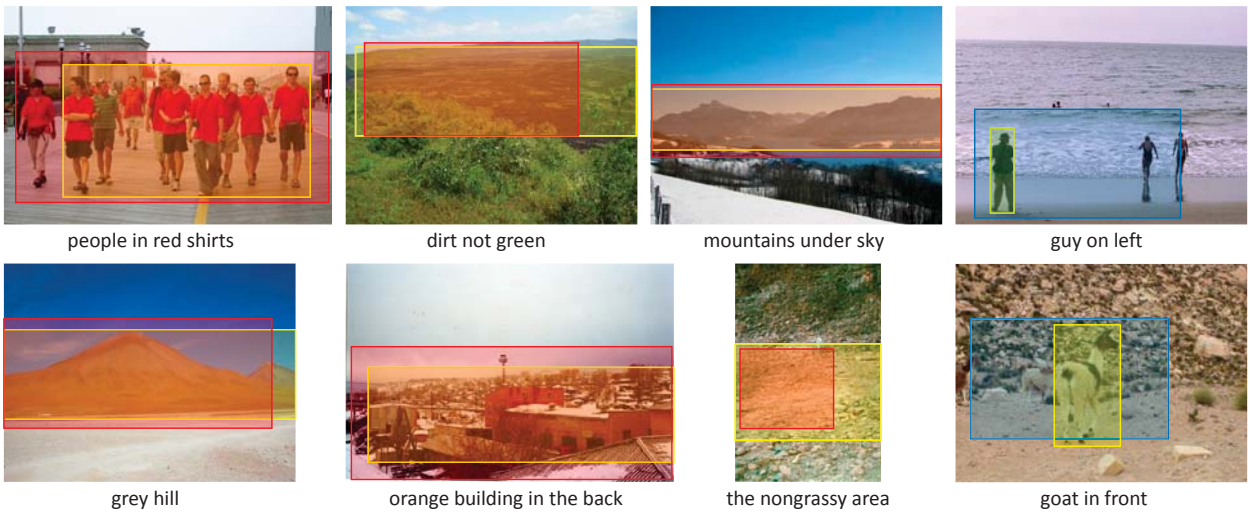

Figure 6. Qualitative results including correct examples (IoU $>$ 0.5) and failure examples (IoU $\leq$ 0.5) on the ReferItGame dataset. Yellow boxes indicate ground truths, red ones indicate correct results and blue ones indicate incorrect results (best viewed in color).

Table 4. Accuracy (IoU $>$ 0.75) of phrase localization on the Flickr30k Entities and ReferItGame datasets.

Methods	Accuracy (IoU $>$ 0.75)	
	Flickr30k Entities	ReferItGame
GroundeR [22]	7.42	1.95
MATN	**11.04**	**3.93**

GroundeR (VGG) directly crops the regions on original image pixels according to region proposals and extracts their features using the VGG classification network. GroundeR (VGG+SPAT) also uses additional spatial features. LRCN is an image captioning model which is trained on MSCOCO and used to score how likely the phrase is to be generated for the proposal box. CAFFE-7K is a large scale object classifier trained on ImageNet. It is used to predict a class for each region proposal and construct a word bag to match with the query phrase in a joint vector space. Both the methods are unsupervised with respect to the bounding box annotations of the phrases and we use the results reported in [22].

Table 3 reports the accuracy of phrase localization for these methods with IoU $>$ 0.5. Our method still significantly outperforms other methods although this dataset is more challenging than Flickr30k Entities due to fewer training samples and more complicated text descriptions. Besides, our model performs localization on top of a convolutional feature map shared by image regions. This makes our model more efficient than other methods which need to pass each image region through the deep model to obtain its own feature map.

Fig. 6 shows qualitative results on the ReferItGame dataset including success and failure cases. For the texts containing relative position statements, such as "on left", our model cannot localize it accurately because it is extremely hard to learn spatial relationships in the weakly supervised learning setting.

Furthermore, we report the accuracy with IoU $>$ 0.75 on both the datasets in Table 4 to validate the effectiveness of our fine-grained search. We compare with the sate-of-the-art GroundeR (VGG-DET) on Flickr30k Entities and GroundeR (VGG+SPAT) on ReferItGame. We can see that MATN still gives better results because MATN can refine the position of bounding box candidates over the spatial feature map thus can obtain good results even though under the stricter evaluation condition.

5. Conclusion

This paper proposes a Multi-scale Anchored Transformer Network to localize free-form textual phrases in images without the bounding box supervised information. According to the correlation scores between LSTM feature vectors of phrases and spatial feature maps of images, the multi-scale correspondence network predicts affine transformation parameters of phrase region under an anchor constraint induced from region proposals. The model is trained by simultaneously minimizing a contrastive reconstruction error between different phrases from a single image and a set of triplet losses among multiple images with similar phrases. Extensive experiments demonstrate that the proposed method outperforms state-of-the-art weakly supervised phrase localization methods by a significant margin.

Acknowledgments

Jiashi Feng was partially supported by NUS startup R-263-000-C08-133, MOE Tier-I R-263-000-C21-112, NUS IDS R-263-000-C67-646 and ECRA R-263-000-C87-133.

References

[1] P. Arbelaez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. In *Proc. IEEE Conf. Comp. Vis. Pattern Recogn. (CVPR)*, 2014. 6

[2] S. Bao, Y. Xiang, and S. Savarese. Object co-detection. In *Proc. Eur. Conf. Comp. Vis. (ECCV)*, 2012. 5

[3] H. Bilen and A. Vedaldi. Weakly supervised deep detection networks. In *Proc. IEEE Conf. Comp. Vis. Pattern Recogn. (CVPR)*, 2016. 2

[4] K. Chen, R. Kovvuri, and R. Nevatia. Query-guided regression network with context policy for phrase grounding. In *Proc. Int. Conf. Comp. Vis. (ICCV)*, 2017. 1

[5] R. G. Cinbis, J. Verbeek, and C. Schmid. Multi-fold mil training for weakly supervised object localization. In *Proc. IEEE Conf. Comp. Vis. Pattern Recogn. (CVPR)*, 2014. 2

[6] J. Donahue, L. A. Hendricks, S. Guadarrama, and M. Rohrbach. Long-term recurrent convolutional networks for visual recognition and description. In *Proc. IEEE Conf. Comp. Vis. Pattern Recogn. (CVPR)*, 2015. 2, 5, 7

[7] H. J. Escalante, C. A. Hernandez, J. A. Gonzalez, and et al. The segmented and annotated iapr tc-12 benchmark. *Computer Vision and Image Understanding (CVIU)*, 2010. 6

[8] M. Everingham, L. V. Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision (IJCV)*, 2010. 6

[9] K. Gregor, I. Danihelka, A. Graves, and D. Wierstra. Draw: A recurrent neural network for image generation. In *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015. 2

[10] S. Guadarrama, E. Rodner, K. Saenko, N. Zhang, R. Farrell, J. Donahue, and T. Darrell. Long short-term memory. *Neural Computation*, 1997. 3

[11] S. Guadarrama, E. Rodner, K. Saenko, N. Zhang, R. Farrell, J. Donahue, and T. Darrell. Open-vocabulary object retrieval. *Robotics: Science and Systems*, 2014. 5, 7

[12] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell. Natural language object retrieval. In *Proc. IEEE Conf. Comp. Vis. Pattern Recogn. (CVPR)*, 2016. 1, 2

[13] X. L. F. Z. J. L. T. X. J. F. J. Li, Y. Wei. Deep attribute-preserving metric learning for natural language object retrieval. In *ACM MM*, 2017. 1

[14] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial transformer networks. In *Proc. Adv. Neural Info. Process. Syst. (NIPS)*, 2015. 2, 3

[15] V. Kantorov, M. Oquab, M. Cho, and I. Laptev. Context-aware deep network models for weakly supervised localization. In *Proc. Eur. Conf. Comp. Vis. (ECCV)*, 2016. 2

[16] A. Karpathy, A. Joulin, and F. Li. Deep fragment embeddings for bidirectional image sentence mapping. In *Proc. Adv. Neural Info. Process. Syst. (NIPS)*, 2014. 2, 5, 6

[17] S. Kazemzadeh, V. Ordonez, M. Matten, and T. L. Berg. Referit game: Referring to objects in photographs of natural scenes. In *Proc. Conf. Empir. Methods Natural Lang. Process. (EMNLP)*, 2014. 5, 6

[18] T. Lin, M. Maire, S. Belongie, and et al. Microsoft coco: Common objects in context. *arXiv preprint arXiv:1405.0312*, 2014. 1

[19] W. Liu, D. Anguelov, D. Erhan, and et al. Ssd: Single shot multibox detector. In *Proc. Eur. Conf. Comp. Vis. (ECCV)*, 2016. 6

[20] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proc. Adv. Neural Info. Process. Syst. (NIPS)*, 2013. 5

[21] B. A. Plummer1, L. Wang, C. M. Cervantes, and J. C. Caicedo. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proc. Int. Conf. Comp. Vis. (ICCV)*, 2015. 1, 2, 5, 6

[22] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele. Grounding of textual phrases in images by reconstruction. In *Proc. Eur. Conf. Comp. Vis. (ECCV)*, 2016. 1, 2, 5, 6, 7, 8

[23] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. In *Proc. IEEE Conf. Comp. Vis. Pattern Recogn. (CVPR)*, 2011. 2

[24] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015. 6

[25] K. Tang, A. Joulin, L. Li, and F. Li. Co-localization in real-world images. In *Proc. IEEE Conf. Comp. Vis. Pattern Recogn. (CVPR)*, 2014. 5

[26] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International Journal of Computer Vision (IJCV)*, 2013. 1

[27] L. Wang, Y. Li, and S. Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proc. IEEE Conf. Comp. Vis. Pattern Recogn. (CVPR)*, 2016. 1, 2

[28] M. Wang, M. Azab, N. Kojima, R. Mihalcea, and J. Deng. Structured matching for phrase localization. In *Proc. Eur. Conf. Comp. Vis. (ECCV)*, 2016. 1

[29] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, and et al. Show, attend and tell: Neural image caption generation with visual attention. In *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015. 2

[30] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguist. (TACL)*, 2014. 1, 5

[31] J. Zhang, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff. Top-down neural attention by excitation backprop. In *Proc. Eur. Conf. Comp. Vis. (ECCV)*, 2016. 2, 5, 6

[32] C. L. Zitnick and P. Dollar. Edge boxes: Locating object proposals from edges. In *Proc. Eur. Conf. Comp. Vis. (ECCV)*, 2014. 1, 2, 4, 6