Multi-level Multimodal Common Semantic Space for Image-Phrase Grounding

Hassan Akbari, Svebor Karaman, Surabhi Bhargava, Brian Chen,
Carl Vondrick, and Shih-Fu Chang

Columbia University, New York, NY, USA

{ha2436,sk4089,sb4019,bc2754,cv2428,sc250}@columbia.edu

Abstract

We address the problem of phrase grounding by learning a multi-level common semantic space shared by the textual and visual modalities. We exploit multiple levels of feature maps of a Deep Convolutional Neural Network, as well as contextualized word and sentence embeddings extracted from a character-based language model. Following dedicated non-linear mappings for visual features at each level, word, and sentence embeddings, we obtain multiple instantiations of our common semantic space in which comparisons between any target text and the visual content is performed with cosine similarity. We guide the model by a multi-level multimodal attention mechanism which outputs attended visual features at each level. The best level is chosen to be compared with text content for maximizing the pertinence scores of image-sentence pairs of the ground truth. Experiments conducted on three publicly available datasets show significant performance gains (20%-60% relative) over the state-of-the-art in phrase localization and set a new performance record on those datasets. We provide a detailed ablation study to show the contribution of each element of our approach and release our code on GitHub[1]*.*

1. Introduction

Phrase grounding [39, 32] is the task of localizing within an image a given natural language input phrase, as illustrated in Figure 1. This ability to link text and image content is a key component of many visual semantic tasks such as image captioning [10, 21, 18], visual question answering [2, 30, 48, 52, 11], text-based image retrieval [12, 40], and robotic navigation [44]. It is especially challenging as it requires a good representation of both the visual and textual domain and an effective way of linking them.

On the visual side, most of the works exploit Deep Convolutional Neural Networks but often rely on bounding box proposals [39, 42, 15] or use a global feature of the image [10], limiting the localization ability and freedom of the method. On the textual side, methods rely on a closed vocabulary or try to train their own language model using small image-caption pairs datasets [17, 59, 53, 9]. Finally, the mapping between the two modalities is often performed with a weak linear strategy [39, 51]. We argue that approaches in the literature have not fully leveraged the potential of the more powerful visual and textual model developed recently, and there is room for developing more sophisticated representations and mapping approaches.

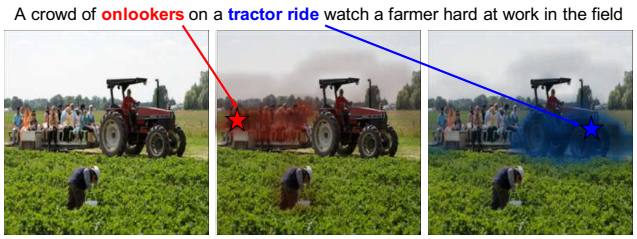

Figure 1. The phrase grounding task in the pointing game setting. Given the sentence on top and the image on the left, the goal is to point (illustrated by the stars here) to the correct location of each natural language query (colored text). Actual example of our method results on Flickr30k.

In this work, we propose to explicitly learn a non-linear mapping of the visual and textual modalities into a common space, and do so at different granularity for each domain. Indeed, different layers of a deep network encode each region of the image with gradually increasing levels of discriminativeness and context awareness, similarly single words and whole sentences contain increasing levels of semantic meaning and context. This common space mapping is trained with weak supervision and exploited at test-time with a multi-level multimodal attention mechanism, where a natural formalism for computing attention heatmaps at each level, attended features and pertinence scoring, enables us to solve the phrase grounding task elegantly and effectively. We evaluate our model on three commonly used datasets in the literature of textual grounding and show that it sets a new state-of-the-art performance by a large margin.

[1] https://github.com/hassanhub/MultiGrounding

Our contributions in this paper are as follows:

- We learn, with weak-supervision, a non-linear mapping of visual and textual features to a common region-word-sentence semantic space, where comparison between any two semantic representations can be performed with a simple cosine similarity;
- We propose a multi-level multimodal attention mechanism, producing either word-level or sentence-level attention maps at different semantic levels, enabling us to choose the most representative attended visual feature among different semantic levels;
- We set new state-of-the-art performance on three commonly used datasets, and give detailed ablation results showing how each part of our method contributes to the final performance.

2. Related works

In this section, we give an overview of related works in the literature and discuss how our method differs from them.

2.1. Grounding natural language in images

The earliest works on solving textual grounding [39, 42, 15] tried to tackle the problem by finding the right bounding box out of a set of proposals, usually obtained from pre-specified models [62, 45]. The ranking of these proposals, for each text query, can be performed using scores estimated from a reconstruction [42] or sentence generation [15] procedure, or using distances in a common space [39]. However, relying on a fixed set of pre-defined concepts and proposals may not be optimal and the quality of the bounding boxes defines an upper bound [15, 46] of the performance that can be achieved. Therefore, several methods [6, 61] have integrated the proposal step in their framework to improve the bounding box quality. These works often operate in a fully supervised setting [5, 53, 57, 11, 6], where the mapping between sentences and bounding boxes has to be provided at training time which is not always available and is costly to gather. Furthermore, methods based on bounding boxes often extract features separately for each bounding box [15, 4, 46], inducing a high computational cost.

Therefore, some works [41, 17, 59, 47, 54] choose not to rely on bounding boxes and propose to formalize the localization problem as finding a spatial heatmap for the referring expression. This setting is mostly weakly-supervised, where at training time only the image and the text (describing either the whole image or some parts of it) are provided but not the corresponding bounding box or segmentation mask for each description. This is the more general setting we are addressing in this paper. The top-down approaches [41, 59] and the attention-based approach [17] learn to produce a heatmap for each word of a vocabulary. At test time, all these methods produce the final heatmap by averaging the heatmaps of all the words in the query that exist in the vocabulary. Several grounding works have also explored the use of additional knowledge, such as image [46] and linguistic [47, 38] structures, phrase context [5] and exploiting pre-trained visual models predictions [4, 54].

In contrast to many works in the literature, we don't use a pre-defined set of image concepts or words in our method. We instead rely on visual feature maps and a character-based language model with contextualized embeddings which could handle any unseen word considering the context in the sentence.

2.2. Mapping to common space

It is a common approach to extract visual and language features independently and fuse them before the prediction [9, 4, 6]. Current works usually apply a multi-layer perceptron (MLP) [6, 4], element-wise multiplication [14], or cosine similarity [9] to combine representations from different modalities. Other methods have used the Canonical Correlation Analysis (CCA) [38, 39], which finds linear projections that maximize the correlation between projected vectors from the two views of heterogeneous data. [11] introduced the Multimodal Compact Bilinear (MCB) pooling method that uses a compressed feature from the outer product of two vectors of visual and language features to fuse them. Attention methods [51, 34] can also measure the matching of an image-sentence feature pair.

We use non-linear mappings of both visual features (in multiple semantic levels) and textual embeddings (both contextualized word and sentence embeddings) separately and use multi-level attention with multimodal loss to learn those mapping weights.

2.3. Attention mechanisms

Attention has proved its effectiveness in many visual and language tasks [23, 1, 7, 52, 50], it is designed to capture a better representation of image-sentence pairs based on their interactions. The Accumulated Attention method [8] propose to estimate attention on sentences, objects and visual feature maps in an iterative fashion, where at each iteration the attention of the other two modalities is exploited as guidance. A dense co-attention mechanism is explored in [34] to solve the Visual Question Answering task by using a fully symmetric architecture between visual and language representations. In their attention mechanism, they add a dummy location in attention map when no region or word the model should attend along with a softmax. In AttnGAN [51], a deep attention multimodal similarity model is proposed to compute a fine-grained image-text matching loss.

In contrast to these works, we remove the softmax on top of the attention maps to let the model decide which word-region could be related to each other by the guide of

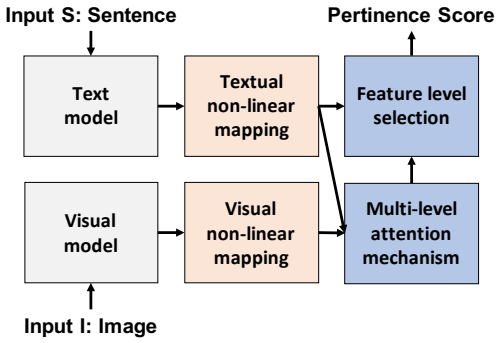

Figure 2. Overview of our method: the textual input is processed with a pre-trained text model followed by a non-linear mapping to the common semantic space. Similarly for the image input, we use a pre-trained visual model to extract visual features maps at multiple levels and learn a non-linear mapping for each of them to the common semantic space. A multi-level attention mechanism followed by a feature level selection produces the pertinence score between the image and the sentence. We train our model using only the weak supervision of image-sentence pairs.

the multimodal loss. Since we map the visual features to a multi-level visual representation, we give the model the freedom to choose any location at any level for either sentence or word. In other words, each word or sentence can choose which level of representation (and which region in that representation) to attend to. We directly calculate the attention map by cosine similarity in our common semantic space. We show that this approach significantly outperforms all the state of the art approaches on three commonly used datasets and set a new state of the art performance.

3. Method

In this section, we describe our method (illustrated in Figure 2) for addressing the textual grounding task and elaborate on each part with details. In Section 3.1, we explain how we extract multi-level visual features from an image and word/sentence embeddings from the text, and then describe how we map them to a common space. In Section 3.2 we describe how we calculate multi-level multimodal attention map and attended visual feature for each word/sentence. Then, in Section 3.3 we describe how we choose the most representative visual feature level for the given text. Finally, in Section 3.4 we define a multimodal loss to train the whole model with weak supervision.

3.1. Feature Extraction and Common Space

Visual Feature Extraction: In contrast to many vision tasks where the last layer of a pre-trained CNN is being used as visual representation of an image, we use feature maps from different layers and map them separately to a common space to obtain a multi-level set of feature maps to be compared with text. Intuitively, using different levels of visual representations would be necessary for covering a wide range of visual concepts and patterns [26, 55, 58]. Thus, we extract $L = 4$ sets of feature maps from L different levels of a visual network, upsample them by a bi-linear interpolation[2] to a fixed resolution $M \times M$ for all the L levels, and then apply 3 layers of 1x1 convolution (with LeakyRelu [31]) with D filters to map them into equal-sized feature maps. Finally, we stack these feature maps and space-flatten them to have an overall image representation tensor $V \in \mathbb{R}^{N \times L \times D}$, with $N = M \times M$. This tensor is finally normalized by the l_2-norm of its last dimension. An overview of the feature extraction and common space mapping for image can be seen in the left part of Figure 3.

In this work, we use VGG [43] as a baseline for fair comparison with other works [10, 47, 17], and the state of the art CNN, PNASNet-5 [29], to study the ability of our model to exploit this more powerful visual model.

Textual Feature Extraction: State-of-the-art works in grounding use a variety of approaches for textual feature extraction. Some use pre-trained LSTM or BiLSTMs on big datasets (e.g. Google 1 Billion [3]) based on either word2vec [33] or GloVe [36] representations. Some train BiLSTM solely on image-caption datasets (mostly MSCOCO) and argue it is necessary to train them from scratch to distinguish between visual concepts which may not be distinguishable in language (e.g. red and green are different in vision but similar in language as they are both colors) [34, 51, 17, 47, 9, 14, 61, 39, 57, 8]. These works either use the recurrent network outputs at each state as word-level representations or their last output (on each direction for BiLSTM) as sentence-level or a combination of both.

In this paper, however we use ELMo [37], a 3-layer network pre-trained on 5.5B tokens which calculates word representations on the fly (based on CNN on characters, similar to [19, 60]) and then feed them to 2 layers of BiLSTMs which produce contextualized representations. Thus, for a given sentence the model outputs three representations for each token (splitted by white space). We take a linear combination of the three representations and feed them to 2 fully connected layers (with shared weights among words), each with D nodes with LeakyRelu as non-linearity between each layer, to obtain each word representation $\mathbf{s}_t$ (green pathway in the right part of Figure 3). The resulting word-based text representation for an entire sentence would be a tensor $\mathbf{S} \in \mathbb{R}^{T \times D}$ built from the stacking of each word representation $\mathbf{s}_t$. The sentence-level text representation is calculated by concatenation of last output of the BiLSTMs at each direction. Similarly, we apply a linear combination on the two sentence-level representations and map it to the common space by feeding it to 2 fully connected layers of

[2] as transposed convolution produces checkerboard artifacts [35]

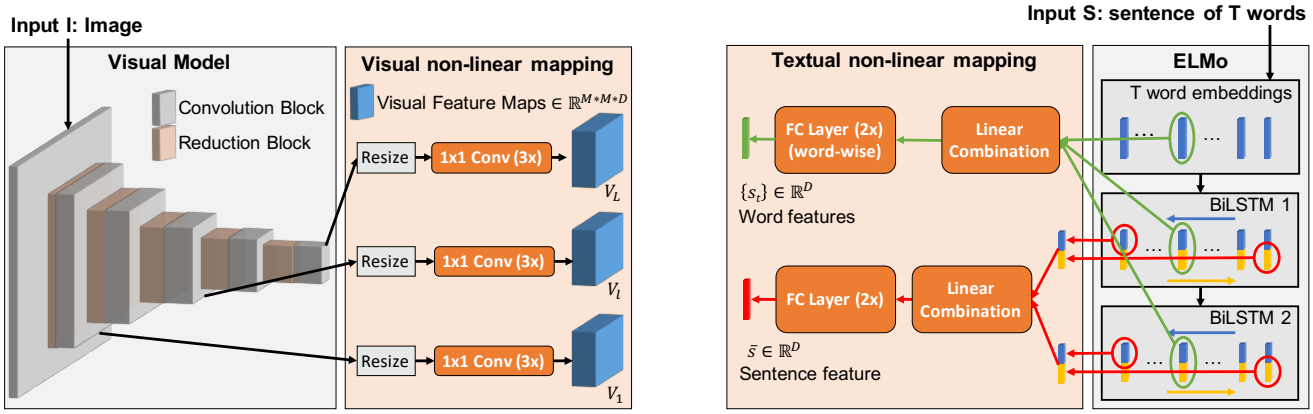

Figure 3. Left: we choose feature maps of different convolutional blocks of a CNN model, resize them to the same spatial dimensions using bi-linear interpolation, and map them to feature maps of the same size. Right: word and sentences embedding to the common space from the pre-trained ELMo [37] model. The green pathway is for word embedding, the red pathway for sentence embedding. All the orange boxes (1×1 convolutional layers of the visual mapping, linear combination and the two sets of fully connected layers of the textual mapping) are the trainable parameters of our projection to the common space.

D nodes, producing the sentence representation $\bar{\mathbf{s}}$ (red pathway in the right part of Figure 3). The word tensor and the sentence vector are normalized by their last dimension l_2-norm before being fed to the multimodal attention block.

3.2. Multi-Level Multimodal Attention Mechanism

Given the image and sentence, our task is to estimate the correspondences between spatial regions (n) in the image at different levels (l), and words in the sentence at different positions (t). We seek to estimate a correspondence measure, $H_{n,t,l}$, between each word and each region at each level. We define this correspondence by the cosine similarity between word and image region representations at different levels in common space:

$$H_{n,t,l} = \max(0, \langle \mathbf{s}_t, \mathbf{v}_{n,l} \rangle). \tag{1}$$

Thus, $\mathbf{H} \in \mathbb{R}^{N \times T \times L}$ represents a multi-level multi-modal attention map which could be simply used for calculating either visual or textual attended representation. We apply ReLU to the attention map to zero-out dissimilar word-visual region pairs, and simply avoid applying softmax on any dimension of the heatmap tensor. Note that this choice is very different in spirit from the commonly used approach of applying softmax to attention maps [50, 49, 8, 34, 17, 51, 41]. Indeed for irrelevant image-sentence pairs, the attention maps would be almost all zeros while the softmax process would always force attention to be a distribution over the image/words summing to 1. Furthermore, a group of words shaping a phrase could have the same attention area which is again hard to achieve considering the competition among regions/words in the case of applying softmax on the heatmap. We will analyze the influence of this choice experimentally in our ablation study.

Given the heatmap tensor, we calculate the attended visual feature for the l-th level and t-th word as

$$\mathbf{a}_{t,l} = \frac{\sum_{n=1}^{N} H_{t,n,l} \mathbf{v}_{n,l}}{\left\| \sum_{n=1}^{N} H_{t,n,l} \mathbf{v}_{n,l} \right\|_2}, \tag{2}$$

which is basically a weighted average over the visual representations of the l-th level with the attention heatmap values as weights. In other words, $\mathbf{a}_{t,l}$ is a vector in the hyperplane spanned by a subset of visual representations in the common space, this subset being selected based on the heatmap tensor. An overview of our multi-level multimodal attention mechanism for calculating attended visual feature can be seen in Figure 4. In the sequel, we describe how we use this attended feature to choose the most representative hyperplane, and calculate a multimodal loss to be minimized by weak supervision of image-sentence relevance labels.

3.3. Feature Level Selection

Once we find the attended visual feature, we calculate the word-image pertinence score at level l using cosine similarity for each word and the attended visual feature as

$$R_{t,l} = \langle \mathbf{a}_{t,l}, \mathbf{s}_t \rangle. \tag{3}$$

Intuitively, each visual feature map level could carry different semantic information, thus for each word we propose to apply a hard level-attention to get the score from the level contributing the most as

$$R_t = \max_l R_{t,l}. \tag{4}$$

This procedure can be seen as finding projection of the textual embeddings on hyperplanes spanned by visual features

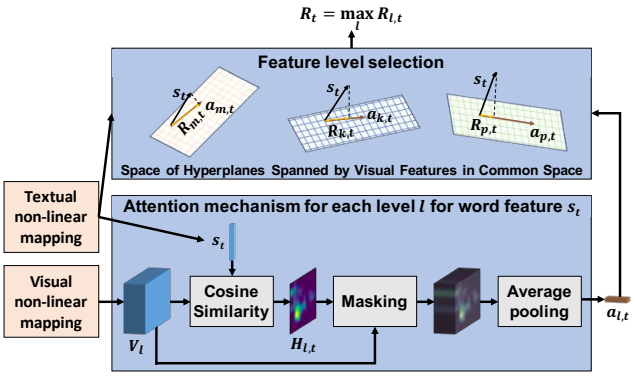

Figure 4. For each word feature s_t, we compute an attention map $H_{l,t}$ and an attended visual feature $a_{l,t}$ at each level l. We choose the level that maximizes similarity between the attended visual feature and the textual feature in the common space to produce the pertinence score R_t. This is equivalent to finding the hyperplane (spanned by each level visual feature vectors in the common space) that best matches the textual feature.

from different levels and choosing the one that maximizes this projection. Intuitively, that chosen hyperplane can be a better representation for visual feature space attended by word t. This can be seen in the top central part of Figure 2, where selecting the maximum pertinence score over levels is equivalent to selecting the hyperplane with the smallest angle with the t-th word representation (or the highest similarity between attended visual feature and textual feature). Thus, selecting the most representative hyperplane (or visual feature level).

Once we find the best word-image pertinence score, similar to [51] and inspired by the minimum classification error [20], we find the overall (word-based) sentence-image pertinence score as follows:

$$R_w(S,I) = \log\left(\left(\sum_{t=0}^{T-1}\exp\left(\gamma_1 R_t\right)\right)^{\frac{1}{\gamma_1}}\right). \tag{5}$$

Similarly, for the sentence we can repeat the same procedure (except that we no more need Eq. (5)) for finding the attention map, attended visual feature and sentence-image pertinence score as follows, respectively:

$$H^s_{n,l} = \max(0, \langle \bar{\mathbf{s}}, \mathbf{v}_{n,l}\rangle) \tag{6a}$$

$$\mathbf{a}^s_l = \sum_{n=1}^{N} H^s_{n,l}\mathbf{v}_{n,l} \tag{6b}$$

$$R_{s,l} = \langle \mathbf{a}^s_l, \bar{\mathbf{s}}\rangle \tag{6c}$$

$$R_s(S,I) = \max_l R_{s,l} \tag{6d}$$

3.4. Multimodal Loss

In this paper, we only use a weak supervision in the form of binary image-caption relevance. Thus, similar to [10, 16, 51] we train the network on a batch of image-caption pairs, $\{(S_b, I_b)\}_{b=1}^B$ and force it to have high sentence-image pertinence score for related pairs and low score for unrelated pairs. Thus, considering a pertinence score R_x (either R_w or R_s), we calculate the posterior probability of the sentence S_b being matched with image I_b by applying competition among all sentences in the batch using:

$$P_x(S_b|I_b) = \frac{\exp(\gamma_2 R_x(S_b, I_b))}{\sum_{b'}^{B}\exp(\gamma_2 R_x(S_{b'}, I_b))} \tag{7}$$

Similarly, the posterior probability of I_b being matched with S_b could be calculated using:

$$P_x(I_b|S_b) = \frac{\exp(\gamma_2 R_x(S_b, I_b))}{\sum_{b'}^{B}\exp(\gamma_2 R_x(S_b, I_{b'}))} \tag{8}$$

Then, similarly to [10, 51], we can define the loss using the negative log posterior probability over relevant image-sentence pairs as follows:

$$L^x = -\sum_{b=1}^{B}\Big(\log P_x(S_b|I_b) + \log P_x(I_b|S_b)\Big) \tag{9}$$

As we want to train a common semantic space for both words and sentences, we combine the loss L^w (that can be computed based on the word relevance R_w) and the sentence loss L^s (obtained using R_s) to define our final loss L as

$$L = L^w + L^s. \tag{10}$$

This loss is minimized over a batch of B images along with their related sentences. We found in preliminary experiments on held-out validation data, that the values $\gamma_1 = 5$, $\gamma_2 = 10$ work well and we keep them fixed for our experiments. In the next section, we will evaluate our proposed model on different datasets and will have an ablation study to show the reason for our choices in our model.

4. Experiments

In this section, we first present the datasets used in our experimental setup. We then evaluate our approach comparing with the state-of-the-art, and further present ablation studies showing the influence of each step of our method.

4.1. Datasets

MSCOCO 2014 [27] consists of 82,783 training images and 40,504 validation images. Each image is associated with five captions describing the image. We use the train split of this dataset for training our model.

Method	Settings	Training	Test Accuracy VG	Flickr30k	ReferIt
Baseline	Random	-	11.15	27.24	24.30
Baseline	Center	-	20.55	49.20	30.40
TD [59]	Inception-2	VG	19.31	42.40	31.97
SSS [17]	VGG	VG	30.03	49.10	39.98
Ours	BiLSTM+VGG	VG	50.18	57.91	**62.76**
Ours	ELMo+VGG	VG	48.76	60.08	60.01
Ours	ELMo+PNASNet	VG	**55.16**	**67.60**	61.89
CGVS [41]	Inception-3	MSR-VTT	-	50.10	-
FCVC [10]	VGG	MSCOCO	14.03	29.03	33.52
VGLS [47]	VGG	MSCOCO	24.40	-	-
Ours	BiLSTM+VGG	MSCOCO	46.99	53.29	47.89
Ours	ELMo+VGG	MSCOCO	47.94	61.66	47.52
Ours	ELMo+PNASNet	MSCOCO	**52.33**	**69.19**	**48.42**

Table 1. Phrase localization accuracy (pointing game) on Flickr30k, ReferIt and VisualGenome (VG) compared to state of the art methods.

Class	pointing game accuracy [41] Inc.3	Ours VGG	Ours PNAS	attention correctness [41] Inc.3	Ours VGG	Ours PNAS
bodyparts	0.194	0.408	**0.449**	0.155	0.299	**0.373**
animals	0.690	0.867	**0.876**	0.657	0.701	**0.826**
people	0.601	0.673	**0.756**	0.570	0.562	**0.724**
instrument	0.458	0.286	**0.575**	0.502	0.297	**0.555**
vehicles	0.645	0.781	**0.838**	0.615	0.554	**0.738**
scene	0.667	**0.685**	0.682	0.582	0.596	**0.639**
other	0.427	0.502	**0.598**	0.348	0.424	**0.535**
clothing	0.360	0.472	**0.583**	0.345	0.330	**0.473**
average	0.501	0.617	**0.692**	0.473	0.508	**0.639**

Table 2. Category-wise pointing game accuracy and attention correctness on Flickr30k Entities.

Flickr30k Entities [39] contains 224k phrases describing localized bounding boxes in ∼31k images each described by 5 captions. Images and captions come from Flickr30k [56]. We use 1k images from the test split of this dataset for evaluation.

VisualGenome [25] contains 77,398 images in the training set, and a validation and test set of 5000 images each. Each image consists of multiple bounding box annotations and a region description associated with each bounding box. We use the train split of this dataset to train our models and use its test split for evaluation.

ReferIt consists of 20,000 images from the IAPR TC-12 dataset [13] along with 99,535 segmented image regions from the SAIAPR-12 dataset [6]. Images are associated with descriptions for the entire image as well as localized image regions collected in a two-player game [22] providing approximately 130k isolated entity descriptions. In our work, we only use the unique descriptions associated with each region. We use a split similar to [15] which contains 9k training, 1k validation, and 10k test images. We use the test split of this dataset to evaluate our models.

4.2. Experimental Setup

We use a batch size of $B = 32$, where for a batch of image-caption pairs each image (caption) is only related to one caption (image). Image-caption pairs are sampled randomly with a uniform distribution. We train the network for 20 epochs with the Adam optimizer [24] with $lr = 0.001$ where the learning rate is divided by 2 once at the 10-th epoch and again at the 15-th epoch. We use $D = 1024$ for common space mapping dimension and $\alpha = 0.25$ for LeakyReLU in the non-linear mappings. We regularize weights of the mappings with l_2 regularization with reg_value $= 0.0005$. For VGG, we take outputs from {conv4_1, conv4_3, conv5_1, conv5_3} and map to semantic feature maps with dimension $18 \times 18 \times 1024$, and for PNASNet we take outputs from {Cell 5, Cell 7, Cell 9, Cell 11} and map to features with dimension $19 \times 19 \times 1024$. Both visual and textual networks weights are fixed during training and only common space mapping weights are trainable. In the ablation study, we use 10 epochs without dividing learning rate, while the rest of settings remain the same. We follow the same procedure as in [17, 18, 39, 47] for cleaning and pre-processing the datasets and use the same train/test splits for fair comparison in our evaluations.

4.3. Phrase Localization Evaluation

As stated in Section 4.1, we train our model on the train split of MSCOCO and Visual Genome (VG), and evaluate it on the test splits of Flickr30k, ReferIt, and VG. In test time, for Flickr30k we feed a complete sentence to the model and take weighted average of attention heatmaps of words for each query with the word-image pertinence scores from Eq. (4) as weights. For ReferIt and Visual Genome, we treat each query as a single sentence and take its sentence-level attention heatmap as the final query pointing heatmap. Once the pointing heatmaps are calculated, we find the max location (as pointing location for the given query) and evaluate the model by the pointing game accuracy: $\frac{\#hit}{\#hit+\#miss}$.

Pointing game accuracy results can be found in Table 1 for Flickr30k, ReferIt and Visual Genome datasets. The results show that our method significantly outperforms all state-of-the-art methods in all conditions and all datasets. For fair comparison with [17, 10, 47], we used a VGG16 visual model and replaced the pre-trained BiLSTM layers of ELMo with a single trainable BiLSTM. This model (BiLSTM+VGG) still gives a pointing game accuracy absolute improvement of 20.15% for VisualGenome, 7.81% for Flickr30k, and 23.28% for ReferIt, while giving relative improvement of 67.09%, 15.59%, and 56.98%, respectively. Results with the more recent PNASNet model are even better, especially for Flickr30k and VisualGenome.

To get a deeper understanding of our model, we first report in Table 2 category-wise pointing game accuracy and attention correctness [28] (percentage of the heatmap falling into the ground truth bounding box) and compare with the

Figure 5. Image-sentence pair from Flickr30k with four queries (colored text) and corresponding heatmaps and selected max value (stars).

Level / PNASNet Layers	Selection Rate (%)									
	bodyparts	animals	people	instrument	vehicles	scene	other	clothing	average	sentence
1 / Cell 5	2.6	10.4	7.5	0.9	2.0	5.4	5.4	5.3	6.3	0.7
2 / Cell 7	0.1	2.0	4.2	0.0	1.7	2.5	0.9	0.3	2.5	0.05
3 / Cell 9	85.9	48.4	64.6	88.6	68.3	49.5	70.9	86.1	66.5	86.51
4 / Cell 11	11.4	39.2	23.7	10.5	27.9	42.6	22.8	8.3	24.7	12.7

Table 3. Level selection rate for different layers of PNASNet on different categories in Flickr30k

state-of-the-art method [41] on Flickr30k. We observe that our method obtains a higher performance on almost all categories even when VGG16 is used as the visual backbone. The model based on PNASNet consistently outperforms the state-of-the-art on all categories on both metrics. We further perform a test on level selection rate for different types of queries and report them in Table 3. It shows that the 3rd level dominates the selection while the 4th level is also important for several categories such as scene and animals. The 1st level is exploited mostly for the animals and people categories. The full sentence selection relies mostly on the 3rd level as well, while for some sentences the 4th model has been selected. This demonstrates the power of the proposed method in selecting the right level of representation.

4.4. Ablation Study

In this section, we trained on MSCOCO multiple configurations of our approach, with a PNASNet visual model, to better understand which aspects of our method affects positively or negatively the performance. We report evaluation results on Flickr30k in Table 4. Results are sorted by performance to show the most successful combinations.

We first evaluated the efficacy of using multi-level feature maps (ML) with level selection compared to a fixed choice of visual layer (M: middle layer, L: last layer) for comparison to word and sentence embeddings (WL and SL). Specifically, we used *Cell 7* as middle layer, and *Cell 11* as last layer, to be compared with word and sentence embedding in Eq. (1) and Eq. (6a), respectively. The results in rows $1, 2$ show that using level-attention mechanism based on multi-level feature maps significantly improves the performance over single visual-textual feature comparison.

	SA	ELMo	NLT	NLV	WL	SL	Acc.
1		✓	✓	✓	ML	ML	67.73
2		✓	✓	✓	M	L	62.67
3			✓	✓	ML	ML	61.13
4		✓		✓	M	L	58.40
5		✓	✓		M	L	56.92
6			✓	✓	M	L	56.42
7		✓			M	L	54.75
8	✓	✓	✓	✓	M	L	47.20
9	✓				M	L	44.83

Table 4. Ablation study results on Flickr30k using PNASNet. SA: Softmax Attention; NLT: Non-Linear Text mapping; NLV: Non-Linear Visual mapping; WL: Word-Layer; SL: Sentence-Layer; Acc.: pointing game accuracy.

We then study the affect of non-linear mapping into the common space for the text and visual features (NLT and NLV). By comparing rows $2, 4, 5, 7$, we see that non-linear mapping in our model is really important, and replacing any mapping with a linear one significantly degrades the performance. We can also see that non-linear mapping seems more important on the visual side, but best results are obtained with both text and visual non-linear mappings.

We further study the use of ELMo for text embedding or the commonly used approach of training a Bi-LSTM. Specifically, we simply replaced the pre-trained BiLSTMs of ELMo model with a trainable BiLSTM (on top of word embeddings of ELMo), and directly feed the BiLSTM outputs to the attention model. The results in rows $1, 3$ and $2, 6$ show the importance of using a strong contextualized text embedding as the performance drops significantly.

We also study the use of softmax on the heatmaps, comparing rows $2, 8$, we see that applying softmax leads to a very negative effect on the performance. This makes sense, as elaborated in Section 3.2, since this commonly used approach forces unnecessarily the heatmap to have a distribution on either words or regions. Results in row 9 corresponds to a simple baseline on par with the state-of-the-art, showing how much improvement can be gained by not using softmax, the use of our multi-level non-linear common space representation and attention mechanism, and a powerful contextualized textual embedding.

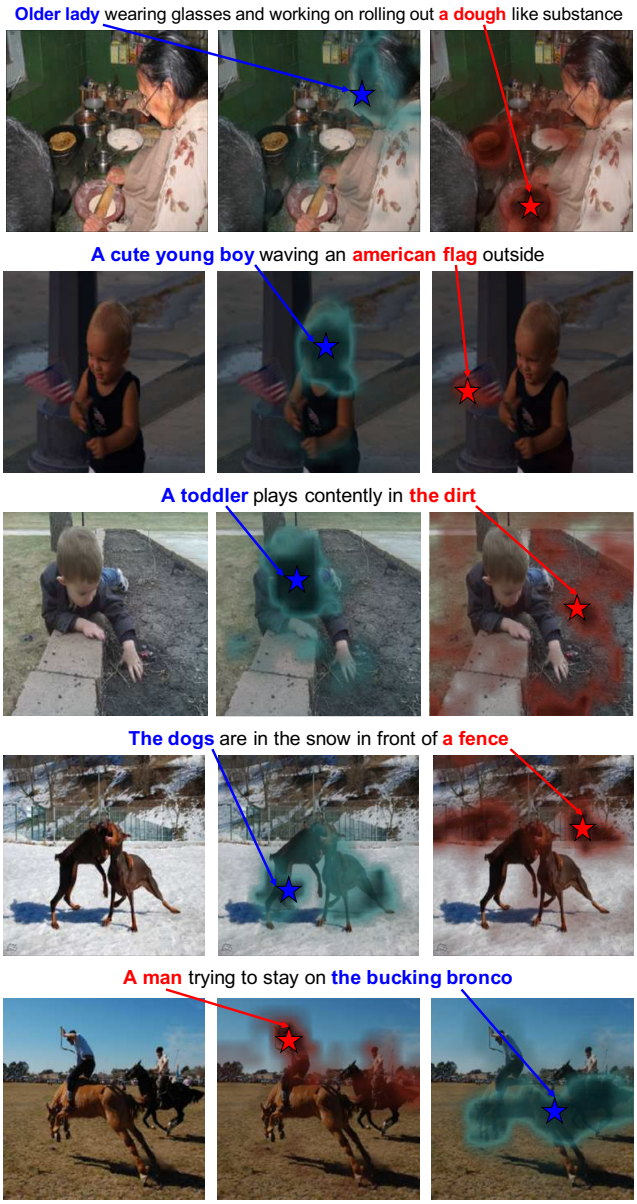

Figure 6. Some image-sentence pairs from Flickr30k, with two queries (colored text) and corresponding heatmaps and selected max value (stars).

4.5. Qualitative results

We give in Figure 5, 6, and 7 some examples of heat maps generated for some queries of the Flickr30k dataset. Specifically, we upsample the heatmaps from their original size of 18×18 (as we use the VGG backbone for these visualizations) by bilinear interpolation to the original image size. We can observe that the max (pointing) location in heatmaps point to correct location in the image and the heatmaps often capture relevant part of the image for each query. It can deal with persons, context and objects even if they are described with some very specific

Figure 7. Some failure cases of our model. The model makes some semantically reasonable mistakes in pointing to regions.

words (e.g. "bronco"), which shows the power of using a character-based contextualized text embedding. Finally, Figure 7 shows some localization failures involving concepts that are semantically close, and in challenging capture conditions. For example, the frames are mistakenly pointed for the query "window" which is overexposed.

5. Conclusion

In this paper, we present a weakly supervised method for phrase localization which relies on multi-level attention mechanism on top of multi-level visual semantic features and contextualized text embeddings. We non-linearly map both contextualized text embeddings and multi-level visual semantic features to a common space and calculate a multi-level attention map for choosing the best representative visual semantic level for the text and each word in it. We show that such combination sets a new state of the art performance and provide quantitative numbers to show the importance of 1. using correct common space mapping, 2. strong contextualized text embeddings, 3. freedom of each word to choose correct visual semantic level. Future works lies in studying other applications such as Visual Question Answering, Image Captioning, etc.

Acknowledgment

This work was supported by the U.S. DARPA AIDA Program No. FA8750-18-2-0014. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, volume 3, page 6, 2018. 2

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015. 1

[3] Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. One billion word benchmark for measuring progress in statistical language modeling. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014. 3

[4] Kan Chen, Jiyang Gao, and Ram Nevatia. Knowledge aided consistency for weakly supervised phrase grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[5] Kan Chen, Rama Kovvuri, Jiyang Gao, and Ram Nevatia. Msrc: Multimodal spatial regression with semantic context for phrase grounding. *International Journal of Multimedia Information Retrieval*, 7(1):17–28, 2018. 2

[6] Kan Chen, Rama Kovvuri, and Ram Nevatia. Query-guided regression network with context policy for phrase grounding. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 2, 6

[7] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6298–6306. IEEE, 2017. 2

[8] Chaorui Deng, Qi Wu, Qingyao Wu, Fuyuan Hu, Fan Lyu, and Mingkui Tan. Visual grounding via accumulated attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7746–7755, 2018. 2, 3, 4

[9] Martin Engilberge, Louis Chevallier, Patrick Pérez, and Matthieu Cord. Finding beans in burgers: Deep semantic-visual embedding with localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3984–3993, 2018. 1, 2, 3

[10] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482, 2015. 1, 3, 5, 6

[11] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2016. 1, 2

[12] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *European Conference on Computer Vision*, pages 241–257. Springer, 2016. 1

[13] Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *Int. Workshop OntoImage*, volume 5, 2006. 6

[14] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. Grounding visual explanations. In *European Conference on Computer Vision*. Springer, 2018. 2, 3

[15] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4555–4564, 2016. 1, 2, 6

[16] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 2333–2338. ACM, 2013. 5

[17] Syed Ashar Javed, Shreyas Saxena, and Vineet Gandhi. Learning unsupervised visual grounding through semantic self-supervision. *arXiv preprint arXiv:1803.06506*, 2018. 1, 2, 3, 4, 6

[18] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4565–4574, 2016. 1, 6

[19] Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*, 2016. 3

[20] Biing-Hwang Juang, Wu Hou, and Chin-Hui Lee. Minimum classification error rate methods for speech recognition. *IEEE Transactions on Speech and Audio processing*, 5(3):257–265, 1997. 5

[21] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 1

[22] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 6

[23] Mahmoud Khademi and Oliver Schulte. Image caption generation with hierarchical contextual visual spatial attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1943–1951, 2018. 2

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[25] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense

image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 6

[26] Tsung-Yi Lin, Piotr Dollár, Ross B Girshick, Kaiming He, Bharath Hariharan, and Serge J Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. 3

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[28] Chenxi Liu, Junhua Mao, Fei Sha, and Alan L Yuille. Attention correctness in neural image captioning. In *AAAI*, pages 4176–4182, 2017. 6

[29] Chenxi Liu, Barret Zoph, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan Yuille, Jonathan Huang, and Kevin Murphy. Progressive neural architecture search. *arXiv preprint arXiv:1712.00559*, 2017. 3

[30] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297, 2016. 1

[31] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *in ICML Workshop on Deep Learning for Audio, Speech and Language Processing*. Citeseer, 2013. 3

[32] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 1

[33] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, 2013. 3

[34] Duy-Kien Nguyen and Takayuki Okatani. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 3, 4

[35] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 2016. 3

[36] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 3

[37] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 2227–2237, 2018. 3, 4

[38] Bryan A Plummer, Arun Mallya, Christopher M Cervantes, Julia Hockenmaier, and Svetlana Lazebnik. Phrase localization and visual relationship detection with comprehensive image-language cues. In *Proc. ICCV*, 2017. 2

[39] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 1, 2, 3, 6

[40] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In *European conference on computer vision*, pages 3–20. Springer, 2016. 1

[41] Vasili Ramanishka, Abir Das, Jianming Zhang, and Kate Saenko. Top-down visual saliency guided by captions. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2017. 2, 4, 6, 7

[42] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pages 817–834. Springer, 2016. 1, 2

[43] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 3

[44] Jesse Thomason, Jivko Sinapov, and Raymond Mooney. Guiding interaction behaviors for multi-modal grounded language learning. In *Proceedings of the First Workshop on Language Grounding for Robotics*, pages 20–24, 2017. 1

[45] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013. 2

[46] Mingzhe Wang, Mahmoud Azab, Noriyuki Kojima, Rada Mihalcea, and Jia Deng. Structured matching for phrase localization. In *European Conference on Computer Vision*, pages 696–711. Springer, 2016. 2

[47] Fanyi Xiao, Leonid Sigal, and Yong Jae Lee. Weakly-supervised visual grounding of phrases with linguistic structures. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2017. 2, 3, 6

[48] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. In *International conference on machine learning*, pages 2397–2406, 2016. 1

[49] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision*, pages 451–466. Springer, 2016. 4

[50] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015. 2, 4

[51] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. *arXiv preprint*, 2018. 1, 2, 3, 4, 5

[52] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21–29, 2016. 1, 2

[53] Raymond Yeh, Jinjun Xiong, Wen-Mei Hwu, Minh Do, and Alexander Schwing. Interpretable and globally optimal prediction for textual grounding using image concepts. In *Advances in Neural Information Processing Systems*, pages 1912–1922, 2017. 1, 2

[54] Raymond A Yeh, Minh N Do, and Alexander G Schwing. Unsupervised textual grounding: Linking words to image concepts. In *Proc. CVPR*, volume 8, 2018. 2

[55] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015. 3

[56] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 6

[57] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 3

[58] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 3

[59] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018. 1, 2, 6

[60] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657, 2015. 3

[61] Fang Zhao, Jianshu Li, Jian Zhao, and Jiashi Feng. Weakly supervised phrase localization with multi-scale anchored transformer network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2, 3

[62] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *European conference on computer vision*, pages 391–405. Springer, 2014. 2