Multimodal Convolutional Neural Networks for Matching Image and Sentence

Lin Ma   Zhengdong Lu   Lifeng Shang   Hang Li
Noah's Ark Lab, Huawei Technologies
forest.linma@gmail.com, {Lu.Zhengdong, Shang.Lifeng, HangLi.HL}@huawei.com

Abstract

In this paper, we propose multimodal convolutional neural networks (m-CNNs) for matching image and sentence. Our m-CNN provides an end-to-end framework with convolutional architectures to exploit image representation, word composition, and the matching relations between the two modalities. More specifically, it consists of one image CNN encoding the image content, and one matching CNN learning the joint representation of image and sentence. The matching CNN composes words to different semantic fragments and learns the inter-modal relations between image and the composed fragments at different levels, thus fully exploit the matching relations between image and sentence. Experimental results on benchmark databases of bidirectional image and sentence retrieval demonstrate that the proposed m-CNNs can effectively capture the information necessary for image and sentence matching. Specifically, our proposed m-CNNs for bidirectional image and sentence retrieval on Flickr30K and Microsoft COCO databases achieve the state-of-the-art performances.

1. Introduction

Associating image with natural language sentence plays the essential role in many applications. Describing the image with natural sentences is useful for image annotation and caption [9, 23, 31], while retrieval image with query sentences is more convenient and helpful for the natural image search applications [14, 19]. The association between image and sentence can be formalized as a multimodal matching problem, where the semantically correlated image and sentence pairs should produce higher matching scores than uncorrelated ones.

The multimodal matching relations between image and sentence are complicated, which happen at different levels as shown in Figure 1. The words in the sentence, such as "grass", "dog", and "ball", denote the objects in the image. The phrases describing the objects and their attributes or activities, such as "black and brown dog", and "small black and
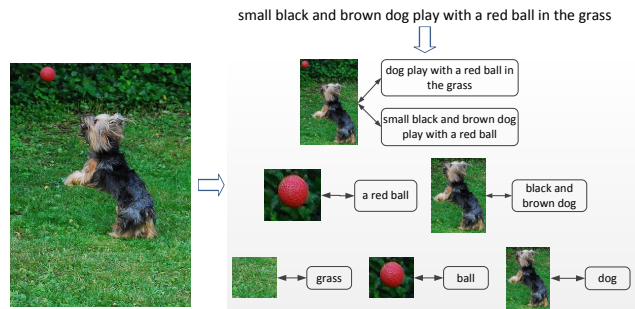

Figure 1. The multimodal matching relations between image and sentence. The words and phrases, such as "grass", "a red ball", and "small black and brown dog play with a red ball", correspond to the image areas of their grounding meanings. The global sentence "small black and brown dog play with a red ball in the grass" expresses the whole semantic meaning of the image content.

brown dog play with a red ball", correspond to the image areas of their grounding meanings. The whole sentence "small black and brown dog play with a red ball in the grass", expressing a complete semantic meaning, associates with the whole image content. These matching relations should be all taken into consideration for an accurate inter-modal matching between image and sentence. Recently, much research work focuses on modeling the image and sentence matching relation at the specific level, namely the word level [38, 39, 7], phrase level [46, 34], and sentence level [14, 19, 37]. However, to the best of our knowledge, there are no models to fully exploit the matching relations between image and sentence by considering the word, phrase, and sentence level inter-modal correspondences together.

The multimodal matching between image and sentence requires good representations of the image and sentence. Recently, deep neural networks have been employed to learn better image and sentence representations. Specifically, convolutional neural networks (CNNs) have shown their powerful abilities on image representation [11, 36, 41, 12] and sentence representation [17, 20]. However, the abil-

ity of CNN on multimodal matching, specifically the image and sentence matching problem, has not been studied.

In this paper, we propose a novel multimodal convolutional neural network (m-CNN) framework for the image and sentence matching problem. By training on a set of image and sentence pairs, the proposed m-CNNs are able to retrieve and rank the images given a natural sentence query, and vice versa. Our core contributions are:

1. CNN is firstly studied for the image and sentence matching problem. We employ convolutional architectures to summarize the image, compose words of the sentence into different semantic fragments, and learn the matching relations and interactions between image and the composed fragments.

2. The complicated matching relations between image and sentence are fully studied in our proposed m-CNN by letting image and the composed fragments of the sentence meet and interact at different levels. We validate the effectiveness of m-CNNs on bidirectional image and sentence retrieval experiments, in which we achieve performances superior to the state-of-the-art approaches.

2. Related Work

2.1. Association between Image and Text

There is a long thread of work on the association between image and text. Early work usually focuses on modeling the correlation between image and the annotating words [7, 38, 39, 10, 43] or phrases [34, 46]. These models cannot well capture the complicated matching relations between image and the natural sentence. Recently, the association between image and sentence has been studied for bidirectional image and sentence retrieval [14, 19, 37, 44, 25, 24, 33] and automatic image captioning [3, 6, 18, 21, 22, 29, 28, 42].

For bidirectional image and sentence retrieval, Hodosh *et al.* [14] proposed KCCA to discover the shared feature space between image and sentence. However, the highly non-linear inter-modal relations cannot be well exploited based on the shallow representations of image and sentence. Recent papers seek better representations of image and sentence from deep architectures. Socher *et al.* [37] proposed to employ the semantic dependency-tree recursive neural network (SDT-RNN) to map the sentence into the same semantic space as the image representation, and the association is then measured as the distance in that space. Yan *et al.* [44] stacked fully connected layers together to represent the sentence and used deep canonical correlation analysis (DCCA) for matching images and text. Klein *et al.* [25] used the Fisher vector (FV) for the sentence representation. Kiros *et. al* [24] proposed skip-thought vector (STV) to encode the sentence for matching the image. As such, the

global level matching relations between image and sentence are studied by representing the sentence as a global vector. However, they neglect the local fragments of the sentence and their correspondences to the image content. Compared with [37], Karpathy *et al.* [19] work on a finer level by aligning the fragments of sentence and regions of image. Plummer *et.al* [33] used the entities to collect region-to-phrase (RTP) correspondences for richer image-to-sentence models. The local inter-modal correspondences between image and sentence fragments are thus studied, where the global matching relations are not considered. As illustrated in Figure 1, the image content corresponds to different fragments of sentence from local words to the global sentence. To fully exploit the inter-modal matching relations, we propose m-CNNs to compose words of sentence to different fragments, let the fragments meet image at different levels, and learn their matching relations.

For automatic image captioning, the authors use recurrent visual representation (RVP) [3], multimodal recurrent neural network (m-RNN) [29, 28], multimodal neural language model (MNLM) [21, 22], neural image caption (NIC) [42], deep visual-semantic alignments (DVSA) [18], and long-term recurrent convolution networks (LRCN) [6] to learn the relation between image and sentence and generate the caption for a given image. Please note that those models naturally produce scores for image-sentence association (e.g., the likelihood of a sentence as the caption for a given image). It can thus be readily used for bidirectional retrieval.

2.2. Image and Sentence Representation

For image, CNNs have demonstrated their powerful abilities to learn the image representation from image pixels, which achieved the state-of-the-art performances on image classification [11, 36, 41, 12] and object detection [32, 8]. For sentence, there is a thread of neural networks for the sentence representation, such as CNN [17, 20], time-delay neural network [4], recursive neural network [15], and recurrent neural network [16, 37, 29, 40]. The obtained sentence representation can be used for the sentence classification [20], image and sentence retrieval [37, 29], language modeling [4], text generation [18, 40], and so on.

3. m-CNNs for Matching Image and Sentence

As illustrated in Figure 2, m-CNN takes the image and sentence as the inputs and generates the matching score between them. More specifically, m-CNN consists of three components.

- **Image CNN:** The image CNN is used to generate the image representation for matching the fragments composed from words, which is computed as follows:

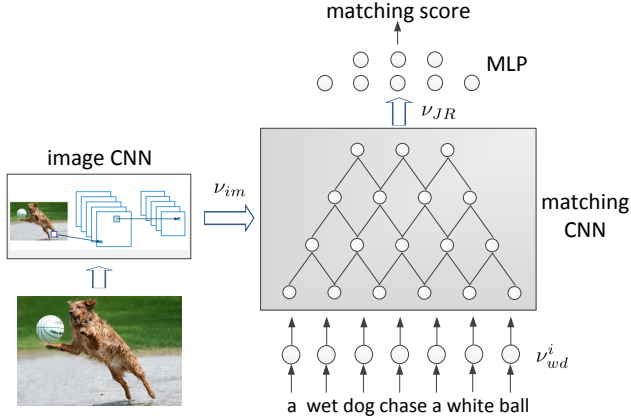$$\nu_{im} = \sigma(\mathbf{w}_{im}(CNN_{im}(I)) + b_{im}), \qquad (1)$$

Figure 2. The m-CNN architecture for matching image and sentence. Image representation is generated by the image CNN. Matching CNN composes words to different fragments of the sentence and learns the joint representation of image and sentence fragments. MLP summarizes the joint representation and outputs the matching score.

where $\sigma(\cdot)$ is the activation function (e.g., Sigmoid or ReLU [5]). CNN_{im} is an image CNN which takes the image as the input and generates a fixed length image representation. The successful image CNNs for image recognition, such as [35, 36], can be used to initialize the image CNN, which returns the 4096-dimensional activations of the fully connected layer immediately before the last ReLU layer. The matrix $\mathbf{w}_{im}$ is of the dimension $d \times 4096$, where d is set as 256 in our experiments. Each image is thus represented as one d-dimension vector ν_{im}.

- **Matching CNN** The matching CNN takes the encoded image representation ν_{im} and word representations ν_{wd}^i as the input and produces the joint representation ν_{JR}. As illustrated in Figure 1, the image content may correspond to sentence fragments with varying scales, which will be adequately considered in the learnt joint representation of image and sentence. Targeting at fully exploiting the inter-modal matching relation, our proposed matching CNNs firstly compose words to different semantic fragments and then let the image meet these fragments to learn their inter-modal structures and interactions. More specifically, different matching CNNs are designed to make the image interact with the composed fragments at different levels to generate the joint representation, from the word and phrase level to the sentence level. Detailed information of the matching CNNs at different levels will be introduced in the following subsections.

- **MLP** Multilayer perceptron (MLP) takes the joint representation ν_{JR} as the input and produces the final

matching score between image and sentence, which is calculated as follows.

$$\mathbf{s}_{match} = \mathbf{w}_s\big(\sigma(\mathbf{w}_h(\nu_{JR}) + b_h)\big) + b_s. \quad (2)$$

where $\sigma(\cdot)$ is the nonlinear activation function. $\mathbf{w}_h$ and b_h are used to map ν_{JR} to the representation in the hidden layer. $\mathbf{w}_s$ and b_s are used to compute the matching score between image and sentence.

The three components of our proposed m-CNN are fully coupled in the end-to-end image and sentence matching framework, with all the parameters (e.g., those for image CNN, matching CNN, MLP, $\mathbf{w}_{im}$ and b_{im} in Eq. (1), and word representations) can be jointly learned under the supervision from matching instances. Threefold benefits are provided. Firstly, the image CNN can be tuned to generate a better image representation for matching. Secondly, word representations can be tuned for further composition and matching processes. Thirdly, the matching CNN (as detailed in the following) composes word representations to different fragments and lets the image representation meet these fragments at different levels, which can fully exploit the inter-modal matching correspondences between image and sentence. With the nonlinear projection in Eq. (1), the image representations ν_{im} for different matching CNNs are expected to encode the image content for matching the composed semantic fragments of the sentence.

3.1. Different Variants of Matching CNN

To fully exploit the matching relations of image and sentence, we let the image representation meet and interact with different composed fragments of the sentence (roughly the word, phrase, and sentence) to generate the joint representation.

3.1.1 Word-level Matching CNN

In order to find the word-level matching relation, we let the image meet with the word-level fragments of sentence and learn their interactions and relations. Moreover, as most convolutional models [1, 26], we consider the convolution units with a local "receptive field" and shared weights to adequately model the rich structures for word composition and inter-modal interaction. The word-level matching CNN, denoted as MatchCNN$_{wd}$, is designed as in Figure 3 (a). After sequential layers of convolution and pooling, the joint representation of image and sentence is generated as the input of MLP for calculating the matching score.

Convolution Generally, with a sequential input ν, the convolution unit for feature map of type-f (among F_ℓ of them) on the ℓth layer is

$$\nu_{(\ell,f)}^i \overset{\text{def}}{=} \sigma(\mathbf{w}_{(\ell,f)}\vec{\nu}_{(\ell-1)}^i + b_{(\ell,f)}), \quad (3)$$
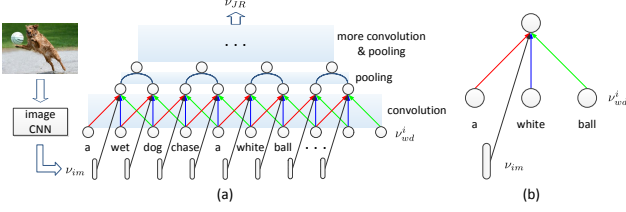
Figure 3. The word-level matching CNN. (a) The word-level matching CNN architecture. (b) The convolution units of multi-modal convolution layer of MatchCNN$_{wd}$. The dashed ones indicate the zero padded word and image representations, which are gated out after convolution process.

where $\mathbf{w}_{(\ell,f)}$ are the parameters for the f feature map on ℓ^{th} layer, $\sigma(\cdot)$ is the activation function, and $\vec{\nu}_{(\ell-1)}^{i}$ denotes the segment of $(\ell-1)^{th}$ layer for the convolution at location i , which is defined as follows

$$\vec{\nu}_{(\ell-1)}^{i} \overset{\text{def}}{=} \nu_{(\ell-1)}^{i} \parallel \nu_{(\ell-1)}^{i+1} \parallel \cdots \parallel \nu_{(\ell-1)}^{i+k_{rp}-1}. \qquad (4)$$

k_{rp} defines the size of local "receptive field" for convolution. "$\parallel$" concatenates the neighboring k_{rp} word vectors into a long vector. In this paper, k_{rp} is chosen as 3 for the convolution process.

As MatchCNN$_{wd}$ targets at exploring word-level matching relation, the multimodal convolution layer is introduced by letting the image meet the word-level fragments of sentence. The convolution unit of the multimodal convolution layer is illustrated in Figure 3 (b). The input of the multimodal convolution unit is denoted as:

$$\vec{\nu}_{(0)}^{i} \overset{\text{def}}{=} \nu_{wd}^{i} \parallel \nu_{wd}^{i+1} \parallel \cdots \parallel \nu_{wd}^{i+k_{rp}-1} \parallel \nu_{im}, \qquad (5)$$

where ν_{wd}^{i} is the vector representation of word i of the sentence, and ν_{im} is the encoded image feature for matching word-level fragments of sentence. It is not hard to see that this input will lead the "interaction" between words and image representation at the first convolution layer, which provides the local matching signal at word level. From the sentence perspective, the multimodal convolution on $\vec{\nu}_{(0)}^{i}$ composes the words $\nu_{wd}^{i}, \cdots, \nu_{wd}^{i+k_{rp}-1}$ in local "receptive field" to a higher semantic representation, such as the phrase "a white ball". From the matching perspective, the multimodal convolution on $\vec{\nu}_{(0)}^{i}$ captures and learns the inter-modal correspondence between image representation and the word-level fragments of sentence. The meanings of the word "ball" and the composed phrase "a white ball" are grounded in the image to make the inter-modal matching relations.

Moreover, in order to handle natural sentences of variable lengthes, the maximum length of sentence is fixed for MatchCNN$_{wd}$. Zero vectors are padded for the image and word representation, as the dashed ones in Figure 3 (a). The

output of the convolution process on zero vectors is gated to be zero. The convolution process in Eq. (3) is further formulated as:

$$\nu_{(\ell,f)}^{i} = g(\vec{\nu}_{(\ell-1)}^{i}) \cdot \sigma(\mathbf{w}_{(\ell,f)}\vec{\nu}_{(\ell-1)}^{i} + b_{(\ell,f)})$$

$$\text{where,} \quad g(x) = \begin{cases} 0, & x == \mathbf{0} \\ 1, & \text{otherwise} \end{cases} \qquad (6)$$

The gating function can eliminate the unexpected matching noise composed from the convolution process.

Max-pooling After each convolution layer, a max-pooling layer is followed. Taking a two-unit window max-pooling as an example, the pooled feature is obtained by:

$$\nu_{(\ell+1,f)}^{i} = \max(\nu_{(\ell,f)}^{2i}, \nu_{(\ell,f)}^{2i+1}) \qquad (7)$$

The effects of max-pooling are two-fold. 1) Together with the stride as two, the max-pooling process lowers the dimensionality of the representation by half, thus quickly making the final joint representation of the image and sentence. 2) It helps filter out the undesired interaction and relation between image and fragments of sentence. Take the sentence in Figure 3 (a) as an example, the composed phrase "dog chase a" matches more closely to the image than "chase a white". Therefore, we can imagine that a well-trained multimodal convolution unit will generate better matching representation of "dog chase a" and image. The max-pooling process will pool the matching representation out for further convolution and pooling processes.

The convolution and pooling processes explore and summarize the local matching signals explored at the word level. More layers of convolution and pooling can be further employed to form matching decisions at larger scales and finally reach a global joint representation. Specifically, in this paper another two more convolution and max-pooling layers alternate to summarize the local matching decisions and finally produce the global joint representation of matching, which reflects the inter-modal correspondence between image and word-level fragments of the sentence.

3.1.2 Phrase-level Matching CNN

Different from matching CNN at word-level, we let CNN work solely on words to certain levels before interacting with the image. Without seeing the image feature, the convolution process will compose the words in the "receptive field" into a higher semantic representation, while the max-pooling process will filter out the undesired compositions. These composed representations are named as phrase from the language perspective. We let image meet the composed phrases to reason their inter-modal matching relations.

As illustrated in Figure 4 (a), after one layer of convolution and max-pooling process, short phrases (denoted as
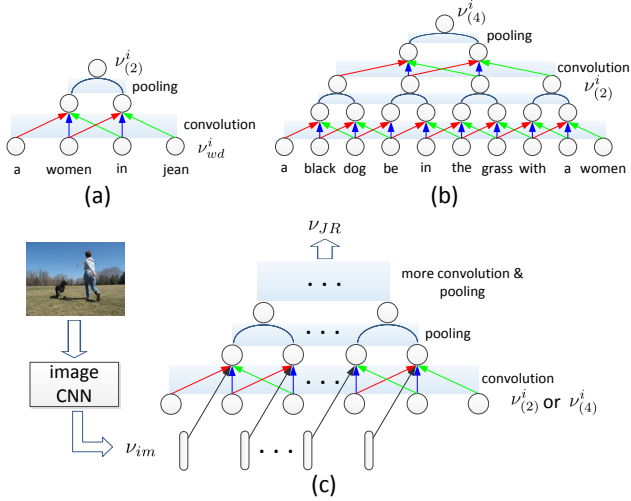
Figure 4. The phrase-level matching CNN and composed phrases. (a): The short phrase is composed by one layer convolution and pooling. (b): The long phrase is composed by two sequential layers of convolution and pooling. (c): The phase-level matching CNN architecture.

$\nu^i_{(2)}$) are composed from four words, such as "a woman in jean". These composed short phrases present richer and specific descriptions about the objects and their relationships compared with single words, such as "woman" and "jean". With an additional layer of convolution and max-pooling process on short phrases, long phrases (denoted as $\nu^i_{(4)}$) are composed from four short phrases (also from ten words), such as "a black dog be in the grass with a woman" in Figure 4 (b). Compared with the composed short phrases and single words, the long phrases present even richer and higher semantic meanings about the specific description of the objects, their activities, and their relative positions.

In order to reason the inter-modal relations between image and the composed phrases, a multimodal convolution layer is introduced by performing convolution on the image and phrase representations. The input of the multimodal convolution unit is:

$$\vec{\nu}^i_{ph} \stackrel{\text{def}}{=} \nu^i_{ph} \parallel \nu^{i+1}_{ph} \parallel \cdots \parallel \nu^{i+k_{rp}-1}_{ph} \parallel \nu_{im}. \qquad (8)$$

where ν^i_{ph} is the composed phrase representation, which can be either short phrases $\nu^i_{(2)}$ or long phrases $\nu^i_{(4)}$. The multimodal convolution process produces the phrase-level matching decisions. Then the layers after that (namely the max-pooling layer or convolution layer) can be viewed as further fusion of these local phrase-level matching decisions to a joint representation, which captures the local matching relations between image and composed phrase fragments. Specifically, for short phrases, two sequential layers of convolution and pooling are followed to generate the joint rep-
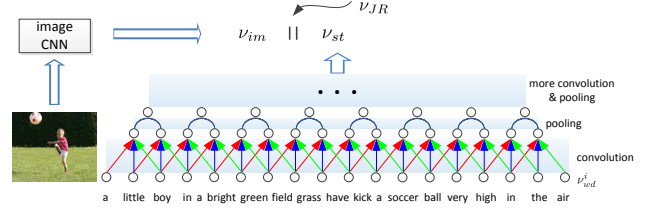
resentation. We name the matching CNN for short phrases and image as MatchCNN$_{phs}$. For long phrases, only one sequential layer of convolution and pooling is used to summarize the local matching to the joint representation. The matching CNN for long phrases and image is named as MatchCNN$_{phl}$.

3.1.3 Sentence-level Matching CNN

The sentence-level convolutional matching CNN, denoted as MatchCNN$_{st}$, goes one step further in the composition and defers the matching until the sentence is fully represented, as illustrated in Figure 5. More specifically, one image CNN encodes the image into a feature vector. One sentence CNN, consisting of three sequential layers of convolution and pooling, represents the whole sentence as a feature vector. The multimodal layer concatenates the image and sentence representation together as their joint representation:

$$\nu_{JR} = \nu_{im} \parallel \nu_{st}, \qquad (9)$$

where ν_{st} denotes the sentence representation by vectorizing the features in the last layer of the sentence CNN.

For the sentence "a little boy in a bright green field have kick a soccer ball very high in the air" illustrated in Figure 5, although word-level and phrase-level fragments, such as "boy", "kick a soccer ball", correspond to the objects as well as their activities in the image, the whole sentence needs to be fully represented to make a reliable association with the image. The sentence CNN with layers of convolution and pooling is used to encode the whole sentence as a feature vector representing its semantic meaning. Concatenating the image and sentence representation together, MatchCNN$_{st}$ does no non-trivial matching, but transfer the representations of the two modalities to the later MLP for fusing and matching.

3.2. m-CNNs with Different Matching CNNs

We can get different m-CNNs with different variants of Matching CNNs, namely m-CNN$_{wd}$, m-CNN$_{phs}$, m-

Figure 5. The sentence-level matching CNN. The joint representation is obtained by concatenating the image and sentence representations together.

MatchCNN$_{wd}$	MatchCNN$_{phs}$	MatchCNN$_{phl}$	MatchCNN$_{st}$
+ ν_{im}			
multi-conv-200	conv-200	conv-200	conv-200
max-2	max-2	max-2	max-2
	+ ν_{im}		
conv-300	multi-conv-300	conv-300	conv-300
max-2	max-2	max-2	max-2
		+ ν_{im}	
conv-300	conv-300	multi-conv-300	conv-300
max-2	max-2	max-2	max-2
			+ ν_{im}

Table 1. Configurations of MatchCNN$_{wd}$, MatchCNN$_{phs}$, MatchCNN$_{phl}$, and MatchCNN$_{st}$ in columns. (conv denotes convolution layer; multi-conv denotes the multimodal convolution layer; max denotes max pooling layer.)

CNN$_{phl}$, and m-CNN$_{st}$. To fully exploit the inter-modal matching relations between image and sentence at different levels, we use an ensemble m-CNN$_{ENS}$ of the four variants by summing the matching scores generated from these m-CNNs together.

4. Implementation details

In this section, we describe the detailed configurations of our proposed m-CNN models and how we train the proposed networks.

4.1. Configurations

We use two different image CNNs, OverFeat [35] (the "fast" network) and VGG [36] (with 19 weight layers), with which we take not only the architecture but also the original parameters (learnt on ImageNet dataset) for initialization. By chopping the top softmax layer and the last ReLU layer, the output of the last fully-connected layer is deemed as image representation, denoted as $CNN_{im}(I)$ in Eq. (1).

The configurations of MatchCNN$_{wd}$, MatchCNN$_{phs}$, MatchCNN$_{phl}$, and MatchCNN$_{st}$ are outlined in Table 1. We use three convolution layers, three max pooling layers, and an MLP with two fully connected layers for all these four networks. The first convolution layer of MatchCNN$_{wd}$, second convolution layer of MatchCNN$_{phs}$, and third convolution layer of MatchCNN$_{phl}$ are the multimodal convolution layers, which blend the image representation and fragments of the sentence together to compose a higher level semantic representation. The MatchCNN$_{st}$ concatenates the image and sentence representation together and leave the interaction to the final MLP. The matching CNNs are designed on fixed architectures, which need to be set to accommodate the maximum length of the input sentences. During our evaluations, the maximum length is set as 30. The word representations are initialized by the skip-gram model [30] with dimension 50. The joint representation obtained from the matching CNNs is fed into MLP with one hidden layer with size 400.

4.2. Learning

The m-CNN models can be trained with contrastive sampling using a ranking loss function. More specifically, for the score function $s_{match}(\cdot)$ as in Eq. (2), the objective function is defined as:

$$e_\theta(x_n, y_n, y_m) = \\ \max\left(0, \mu - s_{match}(x_n, y_n) + s_{match}(x_n, y_m)\right) \quad (10)$$

where θ denotes the parameters, (x_n, y_n) denotes the correlated image-sentence pair, and (x_n, y_m) is the randomly sampled uncorrelated image-sentence pair ($n \neq m$). The notational meaning of x and y varies with the matching task: for image retrieval from query sentence, x denotes the natural sentence and y denotes the image; for sentence retrieval from query image, it is just the opposite. The object is to force the matching score of the correlated pair (x_n, y_n) to be greater than the uncorrelated pair (x_n, y_m) by a margin μ, which is simply set as 0.5 for our training process.

We use stochastic gradient descent (SGD) with mini-batches of 100~150 for optimization. In order to avoid overfitting, early-stopping [2] and dropout (with probability 0.1) [13] are used. ReLU is used as the activation function throughout m-CNNs.

5. Experiments

In this section, we evaluate the effectiveness of our m-CNNs on bidirectional image and sentence retrieval. We begin by describing the datasets used for evaluation, followed by a brief description of competitor models. As our m-CNNs are bidirectional, we evaluate the performances on both image retrieval and sentence retrieval.

5.1. Datasets

We test our matching models on the public image-sentence datasets, with varying sizes and characteristics.

Flickr8K [14] This dataset consists of 8,000 images collected from Flickr. Each image is accompanied with 5 sentences describing the image content. This database provides the standard training, validation, and testing split.

Flickr30K [45] This dataset consists of 31,783 images collected from Flickr. Each image is also accompanied with 5 sentences describing the content of the image. Most of the images depict varying human activities. We used the public split as in [29] for training, validation, and testing.

Microsoft COCO [27] This dataset consists of 82,783 training and 40,504 validation images with 80 categories labeled for a total of 886,284 instances. Each image is also

	Sentence Retrieval				Image Retrieval			
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
Random Ranking	0.1	0.6	1.1	631	0.1	0.5	1.0	500
DeViSE [7]	4.8	16.5	27.3	28.0	5.9	20.1	29.6	29
SDT-RNN [37]	6.0	22.7	34.0	23.0	6.6	21.6	31.7	25
MNLM [22]	13.5	36.2	45.7	13	10.4	31.0	43.7	14
MNLM-vgg [22]	18.0	40.9	55.0	8	12.5	37.0	51.5	10
m-RNN [29]	14.5	37.2	48.5	11	11.5	31.0	42.4	15
Deep Fragment [19]	12.6	32.9	44.0	14	9.7	29.6	42.5	15
RVP (T) [3]	11.6	33.8	47.3	11.5	11.4	31.8	45.8	12.5
RVP (T+I) [3]	11.7	34.8	48.6	11.2	11.4	32.0	46.2	11
DVSA (DepTree) [18]	14.8	37.9	50.0	9.4	11.6	31.4	43.8	13.2
DVSA (BRNN) [18]	16.5	40.6	54.2	7.6	11.8	32.1	44.7	12.4
DCCA [44]	17.9	40.3	51.9	9	12.7	31.2	44.1	13
NIC [42]	20.0	*	61.0	6	19.0	*	64.0	**5**
FV (Mean Vec) [25]	22.6	48.8	61.2	6	19.1	45.3	60.4	7
FV (GMM) [25]	28.4	57.7	70.1	4	20.6	48.5	64.1	6
FV (LMM) [25]	27.7	56.6	69.0	4	19.8	47.6	62.7	6
FV (HGLMM) [25]	28.5	58.4	71.7	4	20.6	49.4	64	6
FV (GMM+HGLMM) [25]	**31.0**	**59.3**	**73.7**	**4**	**21.3**	**50.0**	**64.8**	**5**
OverFeat [35]:								
m-CNN$_{wd}$	8.6	26.8	38.8	18.5	8.1	24.7	36.1	20
m-CNN$_{phs}$	10.5	29.4	41.7	15	9.3	27.9	39.6	17
m-CNN$_{phl}$	10.7	26.5	38.7	18	8.1	26.6	37.8	18
m-CNN$_{st}$	10.6	32.5	43.6	14	8.5	27.0	39.1	18
m-CNN$_{ENS}$	14.9	35.9	49.0	11.0	11.8	34.5	48.0	11.0
VGG [36]:								
m-CNN$_{wd}$	15.6	40.1	55.7	8	14.5	38.2	52.6	9
m-CNN$_{phs}$	18.0	43.5	57.2	8	14.6	39.5	53.8	9
m-CNN$_{phl}$	16.7	43.0	56.7	7	14.4	38.6	52.2	9
m-CNN$_{st}$	18.1	44.1	57.9	7	14.6	38.5	53.5	9
m-CNN$_{ENS}$	24.8	53.7	67.1	5	20.3	47.6	61.7	**5**

Table 2. Bidirectional image and sentence retrieval results on Flickr8K.

associated with 5 sentences describing the content of the image. We used the public split as in [28] for training, validation, and testing.

5.2. Competitor Models

We compared our models with recently developed models on the performances of the bidirectional image and sentence retrieval, specifically DeViSE [7], SDT-RNN [37], DCCA [44], FV [25], STV [24], RTP [33], Deep Fragment [19], m-RNN [28, 29], MNLM [22], RVP [3], DVSA [18], NIC [42], and LRCN [6]. DeViSE and Deep Fragment are regarded as working on word-level and phrase-level, respectively. SDT-RNN, DCCA, and FV are all regarded as working on the sentence-level, which embed the image and sentence into the same semantic space. The other models, namely MNLM, m-RNN, RVP, DVSA, NIC, and LRCN, which are originally proposed for automatic image captioning, can also be used for retrieval in both directions.

5.3. Experimental Results and Analysis

5.3.1 Bidirectional Image and Sentence Retrieval

We adopt the evaluation metrics [19] for a fair comparison. More specifically, for bidirectional retrieval, we report the median rank (Med r) of the closest ground truth result in the list, as well as the R@K (with $K = 1, 5, 10$) which computes the fraction of times the correct result was found among the top K items. The performances of the proposed m-CNNs on bidirectional image and sentence retrieval of Flickr8K, Flickr30K, Microsoft COCO are illustrated in Table 2, 3, and 4. We highlight the best performance of each evaluation metric.

On Flickr8K, FV performs the best, suggesting the strong and beneficial bias of Fisher vector on modeling sentences, which is most obvious when the training data are relatively scarce. Our proposed m-CNN performs inferiorly to FV, but still superior to other methods. The reason, as suggested by the results of larger datasets (Flickr30K

	Sentence Retrieval				Image Retrieval			
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
Random Ranking	0.1	0.6	1.1	631	0.1	0.5	1.0	500
DeViSE [7]	4.5	18.1	29.2	26	6.7	21.9	32.7	25
SDT-RNN [37]	9.6	29.8	41.1	16	8.9	29.8	41.1	16
MNLM [22]	14.8	39.2	50.9	10	11.8	34.0	46.3	13
MNLM-vgg [22]	23.0	50.7	62.9	5	16.8	42.0	56.5	8
m-RNN [29]	18.4	40.2	50.9	10	12.6	31.2	41.5	16
m-RNN-vgg [28]	35.4	63.8	73.7	**3**	22.8	50.7	63.1	5
Deep Fragment [19]	14.2	37.7	51.3	10	10.2	30.8	44.2	14
RVP (T) [3]	11.9	25.0	47.7	12	12.8	32.9	44.5	13
RVP (T+I) [3]	12.1	27.8	47.8	11	12.7	33.1	44.9	12.5
DVSA (DepTree) [18]	20.0	46.6	59.4	5.4	15.0	36.5	48.2	10.4
DVSA (BRNN) [18]	22.2	48.2	61.4	4.8	15.2	37.7	50.5	9.2
DCCA [44]	16.7	39.3	52.9	8	12.6	31.0	43.0	15
NIC [42]	17.0	*	56.0	7	17.0	*	57.0	7
LRCN [6]	*	*	*	*	17.5	40.3	50.8	9
RTP (joint training) [33]	31.0	58.6	67.9	*	22.0	50.7	62.0	*
RTP (SAE) [33]	36.7	61.9	73.6	*	25.4	55.2	68.6	*
RTP (weighted distance) [33]	**37.4**	63.1	74.3	*	26.0	56.0	69.3	*
FV (Mean Vec) [25]	24.8	52.5	64.3	5	20.5	46.3	59.3	6.8
FV (GMM) [25]	33.0	60.7	71.9	**3**	23.9	51.6	64.9	5
FV (LMM) [25]	32.5	59.9	71.5	3.2	23.6	51.2	64.4	5
FV (HGLMM) [25]	34.4	61.0	72.3	**3**	24.4	52.1	65.6	5
FV (GMM+HGLMM) [25]	35.0	62.0	73.8	**3**	25.0	52.7	66.0	5
OverFeat [35]:								
m-CNN$_{wd}$	12.7	30.2	44.5	14	11.6	32.1	44.2	14
m-CNN$_{phs}$	14.4	38.6	49.6	11	12.4	33.3	44.7	14
m-CNN$_{phl}$	13.8	38.1	48.5	11.5	11.6	32.7	44.1	14
m-CNN$_{st}$	14.8	37.9	49.8	11	12.5	32.8	44.2	14
m-CNN$_{ENS}$	20.1	44.2	56.3	8	15.9	40.3	51.9	9.5
VGG [36]:								
m-CNN$_{wd}$	21.3	53.2	66.1	5	18.2	47.2	60.9	6
m-CNN$_{phs}$	25.0	54.8	66.8	4.5	19.7	48.2	62.2	6
m-CNN$_{phl}$	23.9	54.2	66.0	5	19.4	49.3	62.4	6
m-CNN$_{st}$	27.0	56.4	70.1	4	19.7	48.4	62.3	6
m-CNN$_{ENS}$	33.6	**64.1**	**74.9**	**3**	**26.2**	**56.3**	**69.6**	**4**

Table 3. Bidirectional image and sentence retrieval results on Flickr30K.

and Microsoft COCO), is mainly the insufficient training samples. Flickr8K consists of only 8,000 images, which are insufficient for adequately tuning the parameters of the convolutional architectures in m-CNNs. On Flickr30K and Microsoft COCO datasets, with more training samples, m-CNN$_{ENS}$ (with VGG) outperforms all the competitor models in terms of most metrics, as illustrated in Table. 3 and 4. Moreover, except FV, only NIC slightly outperforms m-CNN$_{ENS}$ (with VGG) on image retrieval task measured by R@10. Except the lack of training samples, another possible reason is that NIC uses a better image CNN [41], compared with VGG. As discussed in Section 5.3.3, the performance of image CNN greatly affects the performance of the bidirectional image and sentence retrieval.

On Flickr30K, with more training instances (30,000 im-

ages), the best performing competitor model becomes the RTP on both tasks. Only m-RNN-vgg, FV, and RTP outperform m-CNN$_{ENS}$ (with VGG) on sentence retrieval task measured by R@1. When it comes to image retrieval, m-CNN$_{ENS}$ (with VGG) is consistently better than all competitor models. One possible reason may be that m-RNN-vgg is designed for caption generation and is particularly good at finding the suitable sentence for any given image. One possible reason for RTP may be that the Flickr30K entities are specifically presented, where the bounding boxes corresponding to each entity are manually labeled. As such, much more information are available for image retrieval.

On Microsoft COCO, with more training instances (over 110,000 images), the performances of our proposed m-CNN in terms of all the evaluation metrics have been sig-

	Sentence Retrieval				Image Retrieval			
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
Random Ranking	0.1	0.6	1.1	631	0.1	0.5	1.0	500
m-RNN-vgg [28]	41.0	73.0	83.5	2	29.0	42.2	77.0	3
DVSA [18]	38.4	69.9	80.5	**1**	27.4	60.2	74.8	3
STV (uni-skip) [24]	30.6	64.5	79.8	3	22.7	56.4	71.7	4
STV (bi-skip) [24]	32.7	67.3	79.6	3	24.2	57.1	73.2	4
STV (combine-skip) [24]	33.8	67.7	82.1	3	25.9	60.0	74.6	4
FV (Mean Vec) [25]	33.2	61.8	75.1	3	24.2	56.4	72.4	4
FV (GMM) [25]	39.0	67.0	80.3	3	24.2	59.2	76.0	4
FV (LMM) [25]	38.6	67.8	79.8	3	25.0	59.5	76.1	4
FV (HGLMM) [25]	37.7	66.6	79.1	3	24.9	58.8	76.5	4
FV (GMM+HGLMM) [25]	39.4	67.9	80.9	2	25.1	59.8	76.6	4
VGG [36]:								
m-CNN$_{wd}$	34.1	66.9	79.7	3	27.9	64.7	80.4	3
m-CNN$_{phs}$	34.6	67.5	81.4	3	27.6	64.4	79.5	3
m-CNN$_{phl}$	35.1	67.3	81.6	2	27.1	62.8	79.3	3
m-CNN$_{st}$	38.3	69.6	81.0	2	27.4	63.4	79.5	3
m-CNN$_{ENS}$	**42.8**	**73.1**	**84.1**	2	**32.6**	**68.6**	**82.8**	**3**

Table 4. Bidirectional image and sentence retrieval results on Microsoft COCO.

image	sentence	m-CNN$_{wd}$	m-CNN$_{phs}$	m-CNN$_{phl}$	m-CNN$_{st}$
	three person sit at an outdoor table in front of a building paint like the union jack .	**-0.87**	**1.91**	**-1.84**	**2.93**
	like union at in sit three jack the person a paint building table outdoor of front an .	-1.49	1.66	-3.00	2.37
	sit union a jack three like in of paint the person table outdoor building front at an .	-2.44	1.55	-3.90	2.53
	table sit three paint at a building of like the an person front outdoor jack union in .	-1.93	1.64	-3.81	2.52

Table 5. The matching scores of the image and sentence. The natural sentence (in bold) is the true caption of the image, while the other three sentences are generated by random reshuffle of words.

nificantly improved, compared with those on Flickr8k and Flickr30K. Firstly, it demonstrates that with sufficient training samples, the parameters of the convolutional architecture in m-CNN can be more adequately tuned. Secondly, only DVSA outperforms the proposed m-CNN$_{ENS}$ (with VGG) on sentence retrieval in terms of Med r. On image retrieval, m-CNN$_{ENS}$ significantly and consistently outperforms all the competitor models.

5.3.2 Performances of Different m-CNNs

The proposed m-CNN$_{wd}$ and DeViSE [7] both target at exploiting word-level inter-modal correspondences between image and sentence. However, DeViSE treats each word equally and average their word vectors as the representation of the sentence, while our m-CNN$_{wd}$ let image interact with each word and compose them to higher semantic representations, which significantly outperforms DeViSE. On the other end, both SDT-RNN [37] and the proposed m-CNN$_{st}$ exploit the matching between image and sentence at the sentence level. However, SDT-RNN encodes each sentence recursively into a feature vector based on a pre-given dependency tree, while m-CNN$_{st}$ works on a more flexible

manner with sliding window on the sentence to finally generate the sentence representation. Therefore, a better performance is obtained by m-CNN$_{st}$.

Deep Fragment [19] and the proposed m-CNN$_{phs}$ and m-CNN$_{phl}$ match the image and sentence fragments at phrase levels. However, Deep Fragment uses edges of dependency tree to model the sentence fragments, making it unable to describe more complex relations in sentence. For example, Deep Fragment parses a relative complex phrase "black and brown dog" to two relations "(CONJ, black, brown)" and "(AMOD, brown, dog)", while m-CNN$_{phs}$ handles the same phrase as a whole to compose them to a higher semantic representation. Moreover, m-CNN$_{phl}$ can readily handle longer phrases and reason their grounding meanings in the image. Consequently, better performances of m-CNN$_{phs}$ and m-CNN$_{phl}$ (with VGG) are obtained compared with Deep Fragment.

Moreover, it can be observed that m-CNN$_{st}$ consistently outperform other m-CNNs. The sentence CNN can well summarize the natural sentence and make a better sentence-level association with image in m-CNN$_{st}$. Other m-CNNs captures the matching relations at word and phrase levels. The matching relations should be considered together to

fully depict the inter-modal correspondences between image and sentence. Thus m-CNN$_{ENS}$ achieves the best performances, which indicates that m-CNNs at different levels are complementary with each other to capture the complicated image and sentence matching relations.

5.3.3 Influence of Image CNN

We use OverFeat and VGG to initialize the image CNN in m-CNN for the retrieval tasks. It can be observed that m-CNNs with VGG significantly outperform that with Over-Feat by a large margin, which is consistent with their performance on classification on ImageNet (14% and 7% top-5 classification errors for OverFeat and VGG, respectively). Clearly the retrieval performance depends heavily on the efficacy of the image CNN, which might explain the good performance of NIC on Flickr8K. Moreover, region with CNN features [8] are used for encoding image regions to feature vectors, which are used as the image fragments in Deep Fragment and DVSA. In the future, we will consider to incorporate these image CNNs into our m-CNNs to make more accurate inter-modal matching.

5.3.4 Composition Abilities of m-CNNs

m-CNNs can compose words to different semantic fragments of the sentence for the inter-modal matching at different levels, and therefore posses the ability of word composition. More specifically, we want to check whether the m-CNNs can compose words of random orders into semantic fragments for matching the image content. As demonstrated in Table 5, the matching scores between an image and its accompanied sentence (from different m-CNNs) greatly decrease after the random reshuffle of words. It is a fairly strong evidence that m-CNNs will compose words in natural sequential order into high semantic representations and thus make the inter-modal matching relations between image and sentence.

6. Conclusion

We proposed multimodal convolutional neural networks (m-CNNs) for matching image and sentence. The proposed m-CNNs rely on convolution architectures to compose different semantic fragments of the sentence and learn the interaction between image and the composed fragments at different levels, therefore fully exploit the inter-modal matching relations. Experimental results on bidirectional image and sentence retrieval demonstrate the consistent state-of-the-art performances of our proposed models.

References

[1] O. Abdel Hamid, A. R. Mohamed, H. Jiang, and G. Penn. Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition. In *ICASSP*, 2012. 3

[2] R. Caruana, S. Lawrence, and C. L. Giles. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. *NIPS*, 2000. 6

[3] X. Chen and C. L. Zitnick. Learning a recurrent visual representation for image caption generation. *arXiv:1411.5654*, 2014. 2, 7, 8

[4] R. Colllobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *JMLR*, 2011. 2

[5] G. E. Dahl, T. N. Sainath, and G. E. Hinton. Improving deep neural networks for lvcsr using rectified linear units and dropout. In *ICASSP*, 2013. 3

[6] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darreell. Long-term recurrent convolutional networks for visual recognition and description. *arXiv:1411.4389*, 2014. 2, 7, 8

[7] A. Frame, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. A. Ranzato, and T. Mikolov. Devise: A deep visual-semantic embedding model. *NIPS*, 2013. 1, 2, 7, 8, 9

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierachies for accurate object detection and semantic segmentation. *CVPR*, 2014. 2, 10

[9] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. *ECCV*, 2014. 1

[10] D. Grangier and S. Bengio. A neural network to retrieve images from text queries. *ICANN*, 2006. 2

[11] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *ECCV*, 2014. 1, 2

[12] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. *arXiv:1502.01852*, 2015. 1, 2

[13] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Improving neural networks by proventing co-adaptation of feature detectors. *arXiv:1207.0580*, 2012. 6

[14] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013. 1, 2, 6

[15] O. Irsoy and C. Cardie. Deep recursive neural networks for compositionality in language. *NIPS*, 2014. 2

[16] N. Kalchbrenner and P. Blunsom. Recurrent convolutional neural networks for discourse compositionality. *arXiv:1306.3584*, 2013. 2

[17] N. Kalchbrenner, E. Grefenstette, and P. Blunsom. A convolutional neural network for modelling sentences. *ACL*, 2014. 1, 2

[18] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *arXiv:1412.2306*, 2014. 2, 7, 8, 9

[19] A. Karpathy, A. Joulin, and L. Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. *NIPS*, 2014. 1, 2, 7, 8, 9

[20] Y. Kim. Convolutional neural network for sentence classification. *EMNLP*, 2014. 1, 2

[21] R. Kiros, R. Salakhutdinov, and R. Zemel. Multimodal neural language model. *ICML*, 2014. 2

[22] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv:1411.2539*, 2014. 2, 7, 8

[23] R. Kiros and C. Szepesvári. Deep representations and codes for image auto-annotation. *NIPS*, 2012. 1

[24] R. Kiros, Y. Zhu, R. Salakhutdinov, R. S. Zemel, A. Torralba, R. Urtasun, and S. Fidler. Skip-thought vectors. *arXiv:1506.06726*, 2015. 2, 7, 9

[25] B. Klein, G. Lev, G. Sadeh, and L. Wolf. Associating neural word embeddings with deep image representations using fisher vectors. *CVPR*, 2015. 2, 7, 8, 9

[26] Y. LeCun and Y. Bengio. Convolutional networks for images, speech and time series. *The Handbook of Brain Theory and Neural Networks*, 3361, 1995. 3

[27] T. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, D. Ramanan, C. L. Zitinick, and P. Dollar. Microsoft coco: Common objects in context. *arXiv:1405.0312*, 2014. 6

[28] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille. Dep captioning with multimodal recurrent neural networks (m-rnn). *arXiv:1412.6632*, 2014. 2, 7, 8, 9

[29] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille. Explain images with multimodal recurrent neural networks. *arXiv:1410.1090*, 2014. 2, 6, 7, 8

[30] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv:1301.3781*, 2013. 6

[31] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2txt: Describing images using 1 million captioned photogrphs. *NIPS*, 2011. 1

[32] W. Ouyang, P. Luo, X. Zeng, S. Qiu, Y. Tian, H. Li, S. Yang, Y. Xiong, C. Qian, Z. Zhu, R. Wang, C. C. Loy, X. Wang, and X. Tang. Deepid-net: Multi-stage and deformable deep convolutional neural networks for object detection. *arXiv:1409.3505*, 2014. 2

[33] B. Plummer, L. Wang, C. Cervantes, J. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-snetence models. *arXiv:1505.04870*, 2015. 2, 7, 8

[34] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. *CVPR*, 2011. 1, 2

[35] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. Lecun. Overfeat: Intergrated recognition, localization and detection using convolutional networks. *arXiv:1312.6229*, 2014. 3, 6, 7, 8

[36] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014. 1, 2, 3, 6, 7, 8, 9

[37] R. Socher, Q. V. L. A. Karpathy, C. D. Manning, and A. Y. Ng. Grounded compositional semantics for finding and describing images with sentences. *TACL*, 2:207–218, 2014. 1, 2, 7, 8, 9

[38] N. Srivastava and R. Salakhutdinov. Learning representations for multimodal data with deep belief nets. *ICML Representation Learning Workshop*, 2012. 1, 2

[39] N. Srivastava and R. Salakhutdinov. Multimodal learning with deep boltzmann machines. *NIPS*, 2012. 1, 2

[40] I. Sutskever, J. Martens, and G. Hinto. Generating text with recurrent neural networks. *ICML*, 2011. 2

[41] C. Szegedy, W. Liu, Y. Jia, P. Sermannet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *arXiv:1409.4842*, 2014. 1, 2, 8

[42] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: a neural image caption generator. *arXiv:1411.4556*, 2014. 2, 7, 8

[43] J. Weston, S. Bengio, and N. Usunier. Wsabie: Scaling up to large vocabulary image annotation. *IJCAI*, 2011. 2

[44] F. Yan and K. mikolajczyk. Deep correlation for matching images and text. *CVPR*, 2015. 2, 7, 8

[45] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. *TACL*, 2014. 6

[46] C. L. Zitnick, D. Parikh, and L. Vanderwende. Learning the visual interpretation of sentences. *ICCV*, 2013. 1, 2