End-to-end Convolutional Semantic Embeddings

Quanzeng You
Microsoft
One Microsoft Way, Redmond, WA
quyou@microsoft.com

Zhengyou Zhang
Tencent
Shenzhen, China
zhengyou@tencent.com

Jiebo Luo
University of Rochester
Rochester, NY
jluo@cs.rochester.edu

Abstract

Semantic embeddings for images and sentences have been widely studied recently. The ability of deep neural networks on learning rich and robust visual and textual representations offers the opportunity to develop effective semantic embedding models. Currently, the state-of-the-art approaches in semantic learning first employ deep neural networks to encode images and sentences into a common semantic space. Then, the learning objective is to ensure a larger similarity between matching image and sentence pairs than randomly sampled pairs. Usually, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are employed for learning image and sentence representations, respectively. On one hand, CNNs are known to produce robust visual features at different levels and RNNs are known for capturing dependencies in sequential data. Therefore, this simple framework can be sufficiently effective in learning visual and textual semantics. On the other hand, different from CNNs, RNNs cannot produce middle-level (e.g. *phrase-level in text) representations. As a result, only global representations are available for semantic learning. This could potentially limit the performance of the model due to the hierarchical structures in images and sentences. In this work, we apply Convolutional Neural Networks to process both images and sentences. Consequently, we can employ mid-level representations to assist global semantic learning by introducing a new learning objective on the convolutional layers. The experimental results show that our proposed textual CNN models with the new learning objective lead to better performance than the state-of-the-art approaches.*

1. Introduction

Convolutional Neural Networks (CNNs) have achieved significant success on a wide range of computer vision tasks [21]. Meanwhile, Recurrent Neural Networks (RNNs) have also been extensively deployed to model sequential data, such as machine translation [1, 32] and speech recognition [8]. Until recently, CNNs are applied to natural language processing tasks and show competitive performance on several tasks, including sentence classification [16] and machine translation [7]. Part of the success can be attributed to their ability in modeling complex hierarchical structures of languages. In semantic embedding learning, we are also interested in the structures presented in images and sentences [28] for understanding of their semantic relations at different levels. However, the state-of-the-art approaches usually use RNNs for language modeling due to their success on processing sequential data, including sentences. However, RNNs have their limitations. On one hand, structural representations cannot be obtained from RNNs. On the other hand, they are known for their incapability of capturing long range dependencies in sequential data. To solve this issue, different variants, such as Long Short-Term Memory (LSTM) [11] and Gated Recurrent Units [3], are carefully designed to keep the information propagate smoothly and robustly. However, there still exist the possibility of unstable gradients.

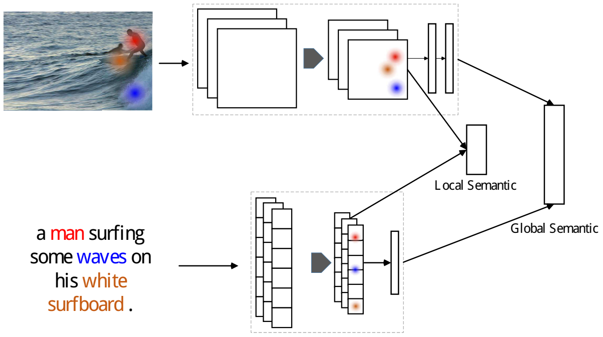


Figure 1. Motivation of using CNNs for semantic embeddings. CNNs can produce hierarchical feature representations, which can be exploited for semantic learning.

Different from RNNs, CNNs can represent visual content in multiple levels, from low-level filters to middle-level parts and the whole objects [35]. This kind of multiple level representations offers opportunity for visual and textual semantic embeddings. In particular, we first design a Convolutional Neural Network architecture for text to obtain multiple level (*e.g*. word-level, phrase-level and sentence-level) semantic representations of sentences. Next, we deploy the semantic learning objective at different levels of visual and textual representations. In such a way, we expect the model to learn more robust semantic representations.

For given image and sentence pairs, both the visual and textual CNNs can be trained in an end-to-end fashion, which we call Convolutional Semantic Embeddings (CSE). Figure 1 shows our motivation for convolutional semantic embedding. Convolutional Neural Networks can provide hierarchical feature maps for both visual and textual modalities. Different colors of image regions and phrases are represented at different convolutional feature maps. The local semantic mechanism, which considers the intermediate convolutional feature maps, will assist the learning of global semantic embeddings.

However, until recently, CNNs are rarely applied to NLP tasks. Notably, CNNs are first applied to sentence classification [16]. Since then, many other works applied CNN in NLP, including [36, 9, 2, 4, 7] for instance. These results, again, suggest the effectiveness of CNNs on processing language related tasks, which also provides supports for our motivation of using CNNs for cross-modality semantic learning.

Our contributions are summarized as follows:

- We design and apply Convolutional Neural Networks (CNNs) for visual and textual semantic embeddings. For given sentences and images, the network can be trained end-to-end.

- We employ the intermediate convolutional features as well as the global semantics features for local context feature learning. Then, we introduce the intermediate objective to assist the global semantic learning using local context features.

- The experimental results on both Flickr30k [27] and MS-COCO [22] datasets demonstrate the effectiveness of the proposed model.

2. Related Work

Our work is closely related to visual and textual semantic embeddings and Convolutional Neural Networks for text analysis. We discuss the main related publications on these two research topics as follows.

2.1. Semantic Embeddings

Semantic embedding understanding of visual and textual content has been considered as the fundamental task for retrieval applications [12]. The state-of-the-art approaches mainly employ deep learning to encode both text and images. Next, the retrieval results are obtained by computing their similarities using the learned semantic representations.

Deep visual semantic embedding model [6] firstly employed deep learning models to learn the embeddings of images and their labels. The learning objective is to make sure that matching image and label pairs have larger similarities than mismatching pairs. This task is extended to images and sentences retrieval in [19], where the encoding model for text becomes the Recurrent Neural Networks, instead of the Word2Vec model [25]. Following work has tried to develop novel learning objectives for this task [19, 30, 20].

More recently, the retrieval performance of using deep semantic embeddings has been significantly improved by novel architectures. Inspired by metric learning, neighborhood structure preserving was considered as a constraint in the learning objective in [31]. Selective multimodal Long Short-Term Memory network (sm-LSTM) [14] employed attention mechanism to select salient pairs from images and sentences. Then, they were passed to a multimodal LSTM network for local similarity measurement and aggregation. Euclidean loss [5] was utilized to approximate the correlation-based loss in a tied two-way network model and showed promising retrieval results. Meanwhile, recurrent residual fusion (RRF) model [23], which consists of several recurrent units with residual blocks and a fusion model, showed improved performance on both Flickr30K and MS-COCO benchmarking datasets.

2.2. CNNs for Text Analysis

Recently, Convolutional Neural Networks have also been employed on language processing tasks. The authors in [17] applied one dimensional Convolutional Neural Networks for sentence classification. Later, character-level CNNs [36, 18] were successfully applied to several different tasks including language modeling, sentiment analysis and classification. The work from [13] showed that CNNs can generate rich matching patterns at different levels for semantic sentence matching. Their architecture of convolutional network consists of several one-dimensional convolution layers and pooling layers to generate a fixed length global vector for each input sentence. Their results suggest that CNNs are capable of encoding sentence for a variety of tasks. He *et al*. [9] applied CNNs for sentence similarity modeling. Due to the hierarchical structure representations of CNNs, they developed a structured similarity measure to consider multiple levels of similarity granularities. Furthermore, attention model was taken into consideration for modeling sentence pairs in [34]. More recently, Gehring *et al*. [7] exploited CNNs for machine translation. Their model achieved the state-of-the-art results on several benchmarking datasets.

The above results suggest that with well-designed architecture, CNNs can be successful in natural language processing tasks. When designing our CNN architecture for semantic learning, we carefully include several one-dimensional convolutional layers for encoding phrases with different length. Next, we apply several convolutional highway network layers to further process these encoded phrases. Each additional highway network layer enables the network to cover a larger range of context. Furthermore, the carry gate of highway network [29] offers the mechanism of including semantics at different levels.

It should be noted that the authors in [24] also employed convolutional neural networks to encode sentence for the two retrieval tasks. However, their architecture of convolutional neural network is directly derived from the standard CNNs for images, which consists of several convolutional layers and pooling layers. This plainly transferred architecture, without much careful tuning for text, may not produce high-quality textual semantic representations. In addition, their learning objective does not consider the intermediate loss from convolutional layers, as proposed in this work.

3. Convolutional Semantic Embeddings

We employ Convolutional Neural Networks to encode both visual and textual content. We call this model Convolutional Semantic Embeddings (CSE).

3.1. Overview of the Framework

Our model is inspired by semantic learning from [19]. The main idea is to encode images and their descriptive sentences into a common semantic embedding space. This is achieved by first producing equal dimensional features for both visual and textual content. Next, the learning objective is designed to pull together matching image and sentence pairs, and separate mismatching pairs at the same time. This drives both visual and textual models to learn feature representations within the same semantic space.

Due to the sequential nature of sentences, Recurrent Neural Networks are widely employed to produce textual features. In our proposed framework, as shown in Figure 2, the image descriptions are encoded using a Convolutional Neural Network. Similar to the approach in [19], we map images and sentences into a common semantic space, but employing CNN for sentence encoding instead of RNN. Again, the objective is to make matching image and sentence pairs more similar than mismatching pairs in the global semantic space, as shown in the top of Figure 2.

Meanwhile, the hierarchical representations of CNNs make intermediate semantic learning in convolutional layers possible. Specifically, instead of only forcing a consistency in the semantic space of global features, we can also add the consistency constraints on the intermediate *regional* features. This additional constraint encourages the model to consider the *regional* semantics into consideration. Eventually, this design is expected to produce more robust and better global semantics.

3.2. Learning Objectives

As shown in Figure 2, our model consists of two learning objectives. The first learning objective, also defined as global objective, is to learn the semantic embeddings using the feature representations of the whole images as well as the whole sentences. Currently, margin-based ranking loss, also known as hinge ranking loss, has been widely deployed to guide the learning of visual and textual semantics [6, 19, 15]. This objective maintains the semantic state, which attempts to pull together the matching pairs and separate the mismatching pairs. To achieve this goal, it tries to assure that the similarity between each matching pair (v_i, s_i) is larger than the similarities between any mismatching pairs (v_i, s_j) or (v_k, s_i) $(i \neq j, i \neq k)$ by a margin α. Otherwise, the model will be penalized by a loss computed as follows

$$\begin{aligned}&\sum_i \sum_j (\alpha - f(v_i, s_i) + f(v_i, s_j)) \\ &+ \sum_i \sum_j (\alpha - f(v_i, s_i) + f(v_j, s_i)),\end{aligned} \tag{1}$$

where α is the margin, v_i and s_i, in this studied problem, are the global representations for the i-th image and i-th sentence respectively, and $f(\cdot, \cdot)$ is the similarity function between two vectors. Here we use the same settings as [19], where $f(\cdot, \cdot)$ is the cosine similarity function.

Following this definition, we introduce the second objective function, intermediate objective. Indeed, direct application of the loss in Eqn.(1) to convolutional features could be problematic. The first difficulty comes from the dimensionality of *intermediate* convolutional features. Usually, they have high dimensions, *e.g.* 2D convolutional features has three dimensions ($f \times h \times w$, f is number of filters, h and w are the height and width respectively). A possible solution is to flatten the convolutional features into a vector. However, this may lead to very high dimensionality of the feature space, which makes the local semantics difficult to learn. Second, the main motivation of this local objective function is to help the learning of global semantics. This implies that the global semantic embeddings are supposed to participate in the design of this local loss.

From the above discussions, we design the following computing mechanism, which intends to remedy the above discussed issues. The global visual features $v \in \mathbb{R}^d$ and textual features $s \in \mathbb{R}^d$ will serve as *reference* features for intermediate loss computation[1]. Let $l_v \in \mathbb{R}^{f_v \times h_v \times w_v}$ and

[1] v_i and s_i are features for the i-th image and sentence pair. However, to keep it concise, we drop the subindex i if no confusion arises.

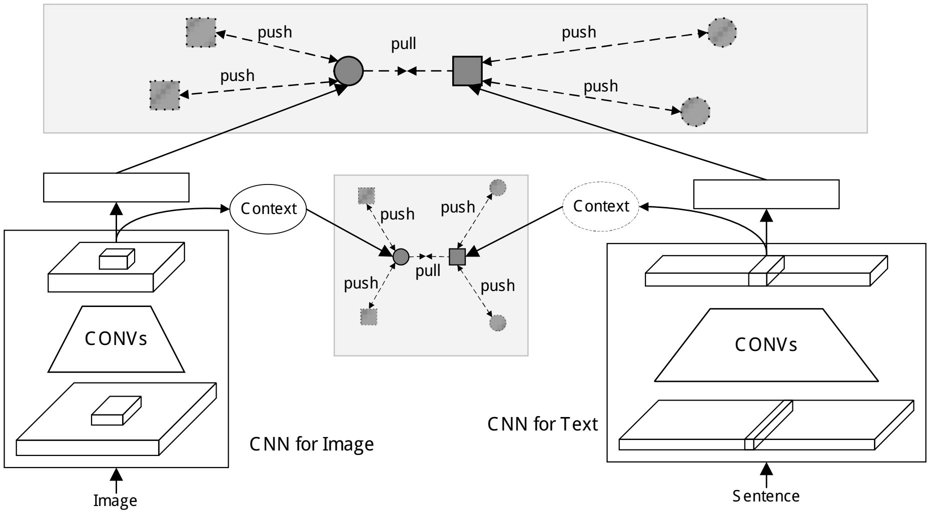


Figure 2. Framework of the proposed convolutional semantic embedding model. There are two objectives (in light shade area), the global and the intermediate objective. In the top, we learn the global semantic embedding space for visual and textual modalities. Both modalities are processed by Convolutional Neural Networks to obtain the global visual and textual representations. Meanwhile, the intermediate objective is designed to assist the convergence of the global semantics.

$l_s \in \mathbb{R}^{f_s \times h_s \times w_s}$ denote the intermediate convolutional features for image and text respectively. We employ a linear model to map the local features l_v into a space with the same dimensionality as the global features.

$$l'_v = T(l_v)W_v + b_v, \tag{2}$$

where $T(l_v) \in \mathbb{R}^{(h_v \times w_v) \times f_v}$ is a reshape operation, $W_v \in \mathbb{R}^{f_v \times d}$ and $b_v \in \mathbb{R}^d$ are the learnable parameters. Now, $l'_v \in \mathbb{R}^{(h_v \times w_v) \times d}$ becomes the new local features on a total of $h_v \times w_v$ local regions [33]. Instantly, we are able to compute the relevance between these local features and the global textual feature s as follows

$$\alpha_v = \texttt{softmax}(l'_v s). \tag{3}$$

Then, we manage to compute the contextual visual features as a weighted sum of the mapped local visual features

$$c_v = \sum_i \alpha_{vi} l'_{v(i,\cdot)}. \tag{4}$$

The textual context c_s can be computed in a similar approach.

The above computational scheme, to some extent, solves the issues of using local convolutional features to guide the semantic learning. In addition, since the role of this intermediate loss is to assist the global semantic learning, the local loss of the i-th image and sentence pair is only activated when its global loss defined in Eqn.(1) is positive. Given the context vector c_v and c_s, we define the local loss in a similar way to the global loss in Eqn.(1) as follows

$$\begin{aligned}&\sum_i I(i) \sum_j (\gamma - f(c_{vi}, s_i) + f(c_{vi}, s_j))\\ &\sum_i I(i) \sum_j (\gamma - f(v_i, c_{si}) + f(v_j, c_{si})).\end{aligned} \tag{5}$$

Again, γ is the margin and $f(\cdot,\cdot)$ is the cosine similarity function between two vectors and $I(i)$ is an indicator function as follows

$$I(i) = \begin{cases} 1 & \text{if global loss for the i-th pair is positive,} \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

3.3. Convolutional Neural Networks for Semantics

In this work, we employ the recently proposed ResNet [10] as image representation while designing a novel Convolutional Neural Network for text semantic learning.

As shown in Figure 3, the inputs are the words of sentences in sequential order, where each word is embedded into a feature space with dimension D. Following the design in [16], the word embeddings are initialized using pre-trained word2vec model on Google News corpus [25]. Next, we include four one dimensional convolutional layers with different kernel sizes to capture semantics of different n-grams. The concatenation of their outputs produces the inputs for the following convolutional highway layers [29], which have also been employed for language modeling [18]. In each convolutional highway network, the

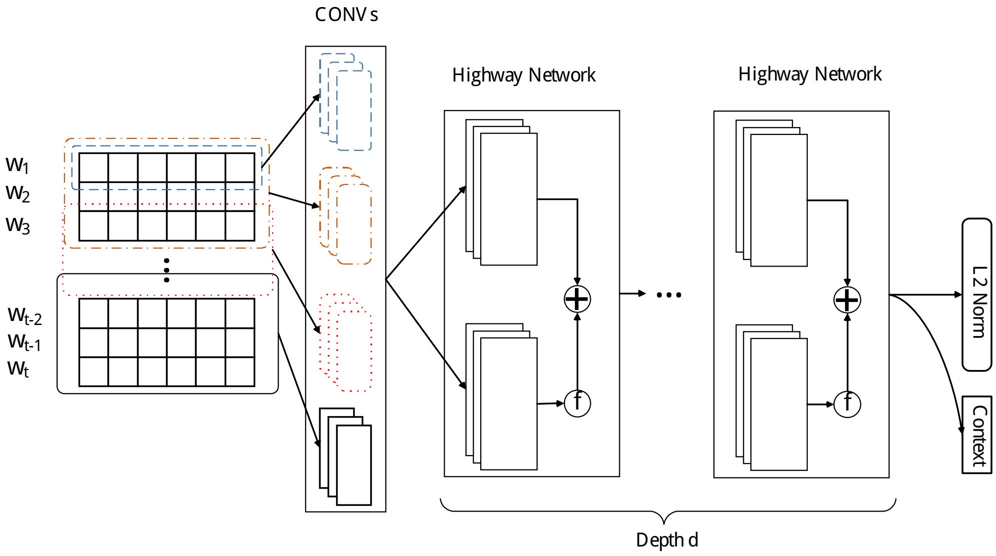


Figure 3. The proposed Convolutional Neural Network architecture for text representation learning. It consists of several convolutional layers with different kernel sizes and several highway network layers for hierarchical text representation learning.

inputs are firstly convolved with a one dimensional convolutional neural network. Then, given input x, the output z of the highway network is computed as

$$t = \sigma(\texttt{Conv1d}(x, k)) \tag{7}$$

$$z = t * \texttt{Conv1d}(x, k) + (1 - t) * x \tag{8}$$

where $\sigma(\cdot)$ is the sigmoid function and Conv1d is a one dimensional convolutional operator with kernel size k, left zero padding of size $k-1$, and stride of size 1. This design guarantees that the convolved outputs have the same dimensionality as the inputs.

Since the kernel size of the convolutional dimensionality is k, the model is expected to cover an additional $(k-1)$ input words in each added highway network layer. Moreover, with the mechanism of carry gate in Eqn.(8), the model can automatically decide how much to keep from previous context x and how much to add from current context $\texttt{Conv1d}(x, k)$. We append a Max-Pooling layer, which operates on the last dimension of the outputs from the last highway network, to generate the global semantic features for the input sentence $\{w_1, w_2, \ldots, w_T\}$.

4. Experiments

We evaluate the proposed model on two benchmark datasets on semantic embedding learning, Flickr30K and MS-COCO. Both datasets contain images collected from Flickr (https://www.flickr.com). Each image has about 5 sentences written by different Amazon Mechanical Turkers. The evaluation has two tasks: image retrieval and sentence retrieval. Public splits of training, testing and validating for both datasets are available online (https://github.com/karpathy/neuraltalk). We use these splits to train our model to make fair comparisons with other state-of-the-art models.

4.1. Implementation Details

We implement the proposed model using PyTorch (http://pytorch.org). As discussed in Section 3.3, we use the pre-trained ResNet [10] on ImageNet for visual feature extraction. For the text, we use the proposed text CNN and initialize the word embeddings using pre-trained Word2Vec [25] model on GoogleNews corpus. In both visual and textual CNN, we use the outputs of the second to the last convolutional layer to represent the local regional features. On one hand, this convolutional layer has a larger coverage context in both images and sentences. On the other hand, the last convolutional layer provides the freedom for global semantic learning.

We set the global semantic embedding size to 1024. The margin α in global loss (Eqn.(1)) is set to 0.5. Meanwhile, for the local objective, we only intend to make sure that the computed context of matching pairs should have a larger score than unmatched pairs. Thus we set γ (Eqn.(5)) to zero.

For the pre-trained ResNet, only the last two convolutional layers for local and global feature extraction are fine-tuned together with the text CNN. The rest layers of the ResNet are freezed without fine-tuning. For the proposed text CNN architecture, we have four one dimensional convolutional layers with kernel sizes one, three, five and seven. We have three convolutional highway network layers, with the same kernel size of three. The whole network is trained

Model	Sentence Retrieval				Image Retrieval				Sum
	R@1	R@5	R10	Median r	R@1	R@5	R@10	Median r	
1K testing split									
m-CNN (Ens) [24]	42.8	73.1	84.1	3	32.6	68.6	82.8	3	384
Order-embeddings [30]	46.7	-	88.9	2	37.9	-	85.9	2	-
DSPE [31]	50.1	79.7	89.2	-	39.6	75.2	86.9	-	420.7
SM-LSTM [14]	52.4	81.7	90.8	1	38.6	73.4	84.6	2	421.5
SM-LSTM (Ens) [14]	53.2	83.1	91.5	1	40.7	75.8	87.4	2	431.8
2WayNet [5]	55.8	75.2	-	-	39.7	63.3	-	-	
RRF-Net [23]	**56.4**	**85.3**	91.5	-	43.9	78.1	88.6	-	443.8
CSE (Ours)	56.3	84.4	**92.2**	**1**	**45.7**	**81.2**	**90.6**	**2**	**450.4**
5K testing split									
Order-embeddings [30]	23.3	-	65.0	5.0	18.0	-	57.6	7.0	-
CSE (ours)	**27.9**	**57.1**	**70.4**	**4**	**22.2**	**50.2**	**64.4**	**5**	**292.2**

Table 1. Evaluation results on MS-COCO testing split. CSE is the performance of the proposed model with the intermediate loss.

using Adam optimization algorithm with a learning rate of 0.001. The mini-batch size is 128 and both local and global loss are computed within each mini-batch. During the testing stage, we compute the Top-K ($K = 1, 5, 10$) recall for both image and sentence retrieval tasks.

4.2. Performance on MS-COCO Dataset

MS-COCO has a total of $82,783$ training images and $40,504$ validating images. We train our model on the training split. We do not include any validation images to augment the training dataset. The online public validating split is used to select the model and the performance on the testing split is reported for comparisons with other approaches.

The performance on MS-COCO dataset is shown in Table 1. For the 1K testing split, our model shows comparable performance with the recent proposed model RRF-Net [23] on image retrieval task. Both of RRF-Net and the proposed Convolutional Semantic Embeddings (CSE) demonstrate significant performance improvements over other state-of-the-art results. In addition, on the sentence retrieval task, CSE outperforms all other state-of-the-art models. This suggests the effectiveness of the proposed text Convolutional Neural Network for textual semantic learning. Overall, our model has the best result in terms of sums over all the recalls on both retrieval tasks. The bottom of Table 1 shows the results of our model on the 5K testing dataset of MS-COCO. Again, our model shows improved results on both tasks compared with baseline results.

4.3. Performance on Flickr30K Dataset

Flickr30K has a total of $29,783$ training images, 1000 validation images and 1000 testing images. Compared with MS-COCO, this dataset is quite smaller. However, it is also widely used to test the performance of different models with relatively insufficient training data. When training on this smaller dataset, we use the same settings on MS-COCO. The performance comparisons of our model with other state-of-the-art models are summarized in Table 2.

Overall, DAN (ResNet) [26] shows the best retrieval performance than other models. This could be due to the application of large image size (448×448 compared to 256×256 used by other models). Meanwhile, both the recent proposed RRF-Net [30] and CSE outperform other state-of-the-art baselines on the two sentence retrieval and image retrieval tasks. Our model does not show better performance than RRF-Net on sentence retrieval. However, the proposed CSE model still shows slightly better performance on the task of image retrieval. We argue that the performance gap between ours and RRF-Net may be attributed to the text part. Different from the visual CNN, which is pre-trained on the ImageNet dataset, the textual CNN is relatively trained from scratch. Without sufficient data, the textual CNN can only learn to encode the sentence to be semantically close to its matching image. However, it may be difficult to distinguish between different sentences without seeing sufficiently diverse examples.

On the other hand, the RRF-Net employs the Hybrid Gaussian-Laplacian Mixture Model (HGLMM) [20]. From this well-crafted discriminative textual feature extractor, RRF-Net can produce even better results. More importantly, our model can be trained end-to-end. Differently, RRF-Net firstly used HGLMM to produce a 18000-dimensional vector for each sentence. Then, they apply principal component analysis (PCA) to reduce the feature vector to 6000. It cannot be trained from end-to-end. In addition, this 6000 dimensional features also introduce more parameters (compared to our 1024 dimensional vector).

4.4. Analysis of Intermediate Objective

It is also interesting to investigate the impact of the intermediate loss defined in Eqn.(5). In particular, we train the network on both datasets with the same settings. However,

Model	Sentence Retrieval				Image Retrieval				Sum
	R@1	R@5	R10	Median r	R@1	R@5	R@10	Median r	
m-CNN (Ens) [24]	33.6	64.1	74.9	3	26.2	56.3	69.6	4	324.7
DSPE [31]	40.3	68.9	79.9	-	29.7	60.1	72.1	-	351
SM-LSTM [14]	42.4	67.5	79.9	2	28.2	57.0	68.4	4	343.4
SM-LSTM (Ens) [14]	42.5	71.9	81.5	2	30.2	60.4	72.3	3	358.7
2WayNet [5]	49.8	67.5	-	-	36.0	55.6	-	-	
RRF-Net [23]	47.6	77.4	87.1	-	35.4	68.3	**79.9**	-	395.7
DAN (ResNet) [26]	**55.0**	**81.8**	**89.0**	**1**	39.4	**69.2**	79.1	**2**	**413.5**
CSE (Ours)	44.6	74.3	83.8	2	**36.9**	69.1	79.6	**2**	388.4

Table 2. Evaluation results on Flickr30K testing split. CSE is the performance of the proposed model with the intermediate loss. Note that DAN [26] uses image size of 448 × 448 instead of 256 × 256 by other models.

this time we turn off the intermediate loss and only optimize the global loss. The results are summarized in Table 3. In both datasets, the performance of the model becomes worse when the intermediate loss is turned off. The differences in R@1 metric are the largest. This may partially prove that the intermediate loss can drive the model to converge to a better local optimum with better retrieval performance.

Dataset	Image to Text		Text to Image	
	R@1	R@5	R@1	R@5
Flickr30K (CSE)	44.6	74.3	36.9	69.1
Flickr30K (w/o inter.)	43.2	73.6	35.7	68.2
MS-COCO (CSE)	56.3	84.4	45.7	81.2
MS-COCO (w/o inter.)	51.6	83.3	43.4	79.0

Table 3. Performance comparisons between CSE and CSE without the intermediate loss.

Meanwhile, since we can compute the context embeddings in Eqn.(4), it is possible to use this context vector to represent each image or sentence for the retrieval task. In particular, we test the following cases. We use sentence retrieval as an example to explain the details of different approaches. Image retrieval is similar and can be easily derived for each case.

- **Context** In this setting, we use the image context vector c_v to retrieve sentences represented by global semantics s.
- **Early fusion** For sentence retrieval, we use $(c_v + v)/2$ to represent images and use global s to represent sentences.
- **Late fusion** Assuming f^i and $f^i_{c_v}$ are the similarities of sentence s_i with global semantics of v, and c_v respectively, we choose $F^i = (f^i + f^i_{c_v})/2$ as the final similarities and then compute the ranks.
- **Best** Let r^i and $r^i_{c_v}$ be the rank of i-th sentence using v and c_v respectively. We choose $R^i = min(r^i, r^i_{c_v})$. This is kind of the upper bound of fusing the two ranking results due to the usage of oracle labels. The purpose of this result is to show the overlaps of top ranks between the context and the global semantics.

Dataset	Image to Text		Text to Image	
	R@1	R@5	R@1	R@5
Flickr30K (CSE)	44.6	74.3	36.9	69.1
Flickr30K (Context)	36.9	70.7	26.6	57.5
Flickr30K (Early)	44.7	74.4	37.1	69.2
Flickr30K (Late)	44.7	74.4	37.1	69.2
Flickr30K (Best)	54.1	81.1	43.5	73.6
MS-COCO (CSE)	56.3	84.4	45.7	81.2
MS-COCO (Context)	41.5	78.7	35.5	71.7
MS-COCO (Early)	56.3	84.4	45.7	81.3
MS-COCO (Late)	56.3	84.4	45.7	81.3
MS-COCO (Best)	64.4	90.2	55.5	86.8

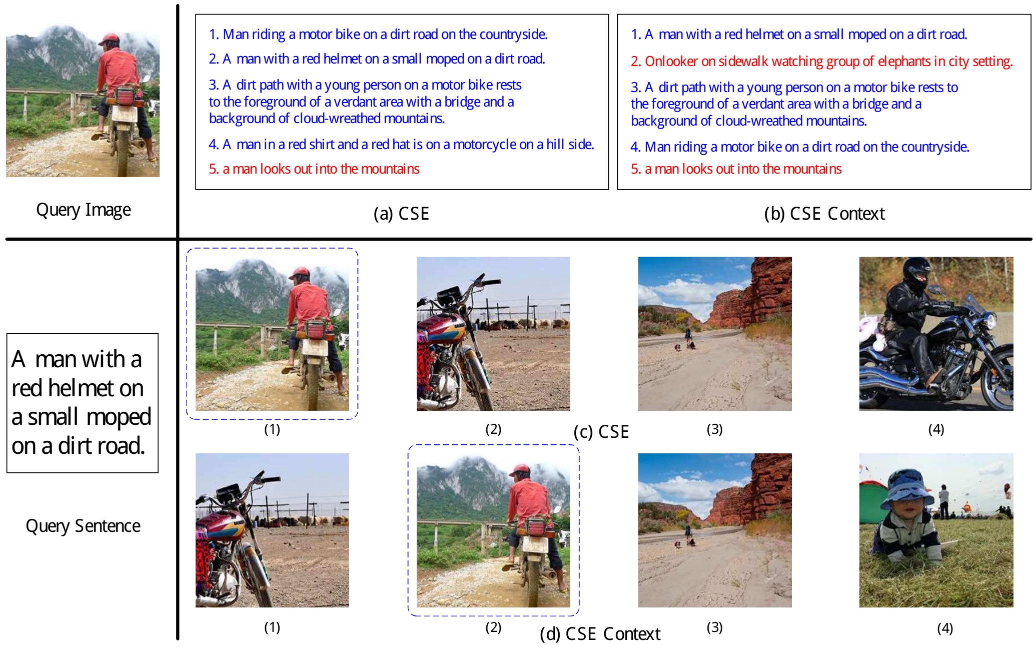
Table 4. Performance of using local context embeddings for retrieval. See the body text for details of each different case.

The fusing results are summarized in Table 4. It is interesting that even though the local objective is only designed to assist the global semantic learning, we can still obtain promising retrieval results using the context vector itself for the task. Meanwhile, both late fusion and early fusion have almost the same performance with the global semantics. This is expected as the goal of the intermediate loss is to make predictions consistent with the global semantics. However, from the "Best" column, in terms of the ranking positions, the two loss terms are indeed quite different.

Figure 4 also show some examples of the retrieval results for CSE and CSE Context. For this given example, we select one image and one of its descriptions for image retrieval and sentence retrieval respectively. As expected, CSE gives better results than CSE Context. However, using the context vector, the top ranked retrieval results are quite similar to CSE, but with different ranks. More analysis and inspection between the global and intermediate semantics

Cross-dataset	Method	Sentence Retrieval			Image Retrieval		
		R@1	R@5	R10	R@1	R@5	R@10
Train: Flickr30K, Test: MS-COCO	RRF-Net [23]	**24.8**	**53.0**	**64.8**	18.8	44.1	58.5
	CSE	24.6	49.2	62.5	**19.1**	**44.4**	**58.6**
Train: MS-COCO, Test: Flickr30K	RRF-Net [23]	28.8	53.8	66.4	21.3	42.7	53.7
	CSE	**30.6**	**59.3**	**71.0**	**26.0**	**52.1**	**64.3**

Table 5. Performance on cross-dataset evaluation.



1. Man riding a motor bike on a dirt road on the countryside.
2. A man with a red helmet on a small moped on a dirt road.
3. A dirt path with a young person on a motor bike rests to the foreground of a verdant area with a bridge and a background of cloud-wreathed mountains.
4. A man in a red shirt and a red hat is on a motorcycle on a hill side.
5. a man looks out into the mountains

1. A man with a red helmet on a small moped on a dirt road.
2. Onlooker on sidewalk watching group of elephants in city setting.
3. A dirt path with a young person on a motor bike rests to the foreground of a verdant area with a bridge and a background of cloud-wreathed mountains.
4. Man riding a motor bike on a dirt road on the countryside.
5. a man looks out into the mountains

A man with a red helmet on a small moped on a dirt road.

Figure 4. Examples of image retrieval and sentence retrieval for CSE and CSE Context. For sentence retrieval, blue sentences are the ground-truth in the top-5 results and red sentences are incorrect. For image retrieval, images with blue dashed box are the ground-truth.

could be an interesting future research topic.

4.5. Cross-dataset Evaluation

Following RFF-Net [23], we also evaluate the performance of our models in terms of cross-dataset generalization. As shown in Table 5, we employ the model trained on Flickr30K or MS-COCO to evaluate the testing split of the other dataset. The performance of the generation is similar to and positively correlated with the performance in Table 1 and Table 2. However, the performance gap between RRF-Net and CSE becomes smaller for models trained on Flickr30K. On the contrary, the generability of CSE are much larger than RRF-Net on the models trained on MS-COCO, in particular on the image retrieval task. These results suggest that with sufficient training data, our model shows better generability than RFF-Net.

5. Conclusions

In this work, we present an end-to-end convolutional neural network architecture for visual and textual semantic learning. Our work proposed model is inspired by the multi-level feature representations of Convolutional Neural Networks (CNNs). In particular, we design a CNN for text encoding and a simple yet effective intermediate objective function to assist the global semantic learning. The experimental results indicate that the proposed textual CNN improves the semantics for sentences and has better generalizability than the state-of-the-art. Meanwhile, the results on image retrieval suggest that it is effective on encoding sentences to its matching images. However, there is still room for improvement in sentence retrieval. More analysis on both intermediate representations and the correlation between visual CNN and textual CNN is necessary for designing better models.

Acknowledgment

We thank the support of New York State through the Goergen Institute for Data Science, the corporate research sponsors Adobe and Snap Inc., and NSF Award #1704309.

References

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. 1

[2] Y. Chen, L. Xu, K. Liu, D. Zeng, and J. Zhao. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 167–176, 2015. 2

[3] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 1

[4] A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun. Very deep convolutional networks for text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 1107–1116, 2017. 2

[5] A. Eisenschtat and L. Wolf. Linking image and text with 2-way nets. *arXiv preprint arXiv:1608.07973*, 2016. 2, 6, 7

[6] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013. 2, 3

[7] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin. Convolutional sequence to sequence learning. *arXiv preprint arXiv:1705.03122*, 2017. 1, 2

[8] A. Graves, A.-r. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE, 2013. 1

[9] H. He, K. Gimpel, and J. J. Lin. Multi-perspective sentence similarity modeling with convolutional neural networks. In *EMNLP*, pages 1576–1586, 2015. 2

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 5

[11] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 1

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

[13] B. Hu, Z. Lu, H. Li, and Q. Chen. Convolutional neural network architectures for matching natural language sentences. In *Advances in neural information processing systems*, pages 2042–2050, 2014. 2

[14] Y. Huang, W. Wang, and L. Wang. Instance-aware image and sentence matching with selective multimodal lstm. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2, 6, 7

[15] A. Karpathy and L. 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. 3

[16] Y. Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751, 2014. 1, 2, 4

[17] Y. Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014. 2

[18] Y. Kim, Y. Jernite, D. Sontag, and A. M. Rush. Character-aware neural language models. In *AAAI*, pages 2741–2749, 2016. 2, 4

[19] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014. 2, 3

[20] B. Klein, G. Lev, G. Sadeh, and L. Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4437–4446, 2015. 2, 6

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1

[22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2

[23] Y. Liu, Y. Guo, E. M. Bakker, and M. S. Lew. Learning a recurrent residual fusion network for multimodal matching. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2, 6, 7, 8

[24] L. Ma, Z. Lu, L. Shang, and H. Li. Multimodal convolutional neural networks for matching image and sentence. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2623–2631, 2015. 3, 6, 7

[25] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013. 2, 4, 5

[26] H. Nam, J.-W. Ha, and J. Kim. Dual attention networks for multimodal reasoning and matching. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 6, 7

[27] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. 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. 2

[28] R. Socher, C. C. Lin, C. Manning, and A. Y. Ng. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 129–136, 2011. 1

[29] R. K. Srivastava, K. Greff, and J. Schmidhuber. Highway networks. *arXiv preprint arXiv:1505.00387*, 2015. 3, 4

[30] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*, 2015. 2, 6

[31] L. Wang, Y. Li, and S. Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5005–5013, 2016. 2, 6, 7

[32] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016. 1

[33] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015. 4

[34] W. Yin, H. Schütze, B. Xiang, and B. Zhou. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. *arXiv preprint arXiv:1512.05193*, 2015. 2

[35] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 1

[36] X. Zhang and Y. LeCun. Text understanding from scratch. *arXiv preprint arXiv:1502.01710*, 2015. 2